Demographic Research a free, expedited, online journal
of peer-reviewed research and commentary
in the population sciences published by the
Max Planck Institute for Demographic Research
Doberaner Strasse 114 · D-18057 Rostock · GERMANY
www.demographic-research.org

DEMOGRAPHIC RESEARCH
VOLUME 1, ARTICLE 5
PUBLISHED 22 SEPTEMBER 1999
www.demographic-research.org/Volumes/Vol1/5

Estimating Parametric Fertility Models with Open Birth Interval Data

Carl Schmertmann

André Junqueira Caetano

Estimating Parametric Fertility Models with Open Birth Interval Data

by

Carl P. Schmertmann (schmertmann@fsu.edu)
André Junqueira Caetano (caetano@prc.utexas.edu)

Abstract:

In the past thirty years, more than 100 censuses gathered fertility data through questions on women's date of last birth. The standard "births last year" (BLY) approach for such data truncates timing information, using binary indicators for births in the prior year only. The first author recently proposed consistent, maximum-likelihood estimation approaches using untruncated date of last birth (DLB). In this paper we extend DLB techniques to parametric models. We construct estimators for Coale-Trussell M and m parameters from open interval lengths. We apply the new procedure to Brazilian census data, producing maps and spatial statistics for BLY and DLB m estimates in 723 municipalities in Minas Gerais. DLB estimators are less sensitive to sampling error than BLY estimators. This increased precision leads to clearer spatial patterns of fertility control, and to improved regression.

1 Introduction

Census data for calculating period age-specific fertility rates typically come in one of two forms. For each woman of childbearing age, census questionnaires usually record either the number of children born in the last year (*BLY*), or the date of the woman's last live birth (*DLB*). DLB and BLY questions are both common, and some censuses ask both. In recent surveys, the United Nations [12, 13] reported that among 262 national censuses taken between 1965 and 1994 in Africa, Asia, South America, and North America (excluding the USA and Canada), 63 asked DLB questions only, while another 50 asked both DLB and BLY questions (see [8] for more detail). In the most recent round of censuses nineteen countries, including Kenya, Indonesia, Sudan, Vietnam, Colombia, and Brazil, collected DLB data exclusively.

When estimating fertility rates from census data (still a common situation in many countries, particularly when estimates are for subnational areas), efficient use of DLB data is often an important concern. In principle, DLB data contain more information than BLY data, because a researcher can observe not only the fertility histories of the sampled women in the past year, but also many other births and periods of exposure that occurred more than one year earlier. In practice, demographers generally do not use all of the fertility information inherent in DLB data. Standard procedures for estimating age-specific fertility rates from DLB data merely convert to BLY form:

$$BLY \;=\; \begin{cases} 1 \;\; if \;\; DLB \le 1 \, year \\ 0 \;\; otherwise \end{cases} \qquad \{1\}$$

and then utilize this censored version of DLB in all subsequent calculations. Caution with DLB data stems in part from an early history of statistical problems with more ambitious uses (see [10] and [11] for proposed applications; [9] and [14] for critiques), and in part from the fact that DLB data do not provide researchers with histories of uniform lengths for all women.

In a recent paper [8], the first author proposed a new method for consistent estimation of period fertility from DLB information. The essential intuition is to change the unit of analysis from *women* to *woman-years*. A sample of N women who report the date of their last live birth will, in general, contain fertility information on many more than N woman-years. For example, a woman who is interviewed on her 32nd birthday and reports that her last live birth occurred 46 months earlier provides information on not one, but four, years of exposure to fertility risks: She had one birth in age interval (28,29], followed by no births in age intervals (29,30], (30,31], and (31,32].

The previous paper [8] derived maximum likelihood procedures for estimating fertility models from open-interval data. Like standard BLY calculations based on {1}, DLB estimators are consistent under the strong mathematical assumptions of many formal demographic models (unchanging fertility schedules and identical fertility rates for all women of a given age, regardless of parity). DLB estimators also have low bias under more realistic conditions. In contrast to BLY methods, estimators based on the multiple woman-years implicit in open-interval DLB data have much lower sampling variability. Thus, when basic fertility information comes from DLB data, it is possible to produce far more accurate fertility estimates from small samples or for small populations.

The earlier paper [8] derived the mathematical structure for estimating any fertility model from DLB data, but gave examples only for one simple type of fertility schedule (piecewise-constant, with five-year age groups, and no parametric restrictions on the shape of the age schedule). In this paper we demonstrate more fully how to estimate parametric fertility models

from DLB data. As a specific example we illustrate maximum likelihood estimation for the *M* and *m* parameters in a Coale-Trussell marital fertility model [4].

We also provide two examples of the type of analysis for which increasing the accuracy of fertility estimates is useful. We use a set of small-area data from the state of Minas Gerais, Brazil, to illustrate the analytical gain from using the full DLB data in place of the censored BLY version. We produce maps and spatial statistics from alternative 1991 estimates of Coale and Trussell's *m* parameter for the state's 723 municipalities, and show how the increased precision of DLB estimates leads to clearer spatial patterns of fertility control. In addition, we illustrate improvements in regression analysis of fertility when using DLB, rather than the usual BLY, fertility data.

2 Statistical Background

2.1 DLB Data

In order to fix ideas, consider the hypothetical sample in Table 1. Suppose that a survey is taken at time τ, and that each of six women reports her age (a), and the number of months since her last live birth. These data appear on the left-hand side of the table. In many data sets (including the public use samples of the Brazilian census that we use later in the paper), DLB data are available only in integer-truncated form; this version appears in the table in the "Years" column. From this point forward we will treat the number of years, rather than the number of months, as if it were the DLB data observed by the researcher.

Fertility information from too far in the past may be unrepresentative of current patterns. It is therefore desirable to restrict the analysis to the relatively recent past. The researcher can do this by considering only woman-years lived within T years of the survey, where T is a value selected by the researcher. The value of T should be chosen after weighing the benefits of increased sample sizes against the costs of possible biases (see [8]). In Table 1 we use T=5.

TABLE 1
Hypothetical Last-Birth Data for a Survey taken at time τ

		TIME SINCE LAST BIRTH			---------- *FIVE-YEAR HISTORY**--------					
i	Age at survey (a)	Months	Years	Years truncated at T=5 (u)	Birth (δ)	τ-5 to τ-4	τ-4 to τ-3	τ-3 to τ-2	τ-2 to τ-1	τ-1 to τ
1	32	46	3	3	1	–	**{29}**	30	31	32
2	28	33	2	2	1	–	–	**{26}**	27	28
3	21	no births	no births	5	0	17	18	19	20	21
4	40	121	10	5	0	36	37	38	39	40
5	25	16	1	1	1	–	–	–	**{24}**	25
6	31	5	0	0	1	–	–	–	–	**{31}**

**Histories consist of fertility information for the one-year periods (τ-5,τ-4], (τ-4,τ-3] and so on. For each woman, periods for which her fertility information is known are denoted by her (integer) age at the end of that period. Years in bold face and brackets are those in which a birth is reported.*

The column labeled (u) displays the number of complete years since last birth, truncated at the upper limit of T=5. This variable appears in many subsequent calculations. By construction $u \in \{0,1,...,T\}$, and the researcher observes MIN(T,u+1) woman-years from each individual sampled.

The set of columns under the heading "Five-Year History" illustrates the available fertility histories from this sample. Each of the five right-hand columns corresponds to a one-year period. Cells for which histories are known contain the woman's age at the end of the year; other cells are blank. Woman-years that include births are emphasized with bold face and brackets. The column labeled (δ) contains a dummy indicator, equal to one if the woman had a birth within the five-year period, and equal to zero otherwise.

All women in the sample contribute one person-year of exposure to the rightmost column, corresponding to the period (τ-1,τ]. Each woman's age at the end of this period is simply a, her age on the survey date. All women with $u \geq 1$ also contribute information about fertility in the period (τ-2,τ-1], all women with $u \geq 2$ contribute information about (τ-3,τ-2], and so forth.

Standard methods for deriving age-specific fertility rates from a sample like that in Table 1 use only the rightmost column. For each age group, the researcher sums the births *in the past year only*, and divides by the number of women in that age group on the survey date. For Table 1 these calculations are trivial. Because there is only one birth recorded in the year before the survey (to woman #6) all estimated fertility rates would be zero, except $f_{30\text{-}34}=0.50$. This is an unrealistic age schedule for fertility, of course, the main cause of which is the very small sample size.

As Table 1 makes clear, however, there is considerably more fertility information embedded in the (a,u,δ) data than the last year alone reveals. The rightmost column corresponding to the year before the sample contains 6 woman-years and 1 birth. In contrast, the available information from the same women for the last T=5 years contains 20 woman-years and 4 births.

2.2 Estimating a Simple Fertility Schedule with DLB Data

The previous paper [8] showed that using the expanded sample of woman-years from a DLB data set can substantially reduce the sampling variance of estimators for small samples. Furthermore, if the researcher restricts the sampling period to five or fewer years before the survey (as in Table 1 above), potential biases caused by unobserved heterogeneity and time trends in fertility rates appear to be small.

For one simple fertility schedule – a piecewise-constant function with no restrictions on the pattern of fertility across different age groups – estimation with DLB data is extremely simple. Specifically, for the fertility schedule

$$f(a) = \begin{cases} \lambda_1 \;\; if \;\; a \in age\ group\ 1 \\ \ldots \\ \lambda_G \;\; if \;\; a \in age\ group\ G \end{cases} \qquad \{2\}$$

the previous paper [8] demonstrated that maximum likelihood estimates for parameters $\lambda_1 ... \lambda_G$ from DLB data are

$$\hat{\lambda_g} \;=\; B_g \,/\, Y_g \qquad g = 1 \ldots G \qquad \{3\}$$

where B_g and Y_g are the counts of births and woman-years, respectively, for age group g. As an example, in the five-year DLB information in Table 1, $B_{25\text{-}29}$=2 births (woman #1 and #2 each had a birth in this age group) and $Y_{25\text{-}29}$=5 (1 year each from women #1 and #5, and 3 years from woman #2). The estimated $\lambda_{25\text{-}29}$ is 0.40, compared to the estimate of zero from the last-year only data.

The estimators in {3} are simple and familiar. However, they are somewhat counterintuitive, because much of the measured exposure (Y) occurs after, rather than before, the measured events (B). Despite this inversion of the usual order of time, the maximum likelihood estimators are still familiar-looking event/exposure ratios. Allison [1] showed that similar counterintuitive results hold for backward recurrence times (times since last event) in many stochastic models.

2.3 Summary Indices for Piecewise-Constant Models

One very important aspect of {3} is that maximum likelihood estimation requires only a handful of summary indices ($B_1 \ldots B_G, Y_1 \ldots Y_G$), rather than the full set of individual-level DLB data. This section demonstrates that this is a property of any fertility model in which the age schedule is piecewise-constant across age groups, a fact that simplifies the estimation of many parametric models, including Coale-Trussell.

For any model with piecewise-constant rates, fertility at age a may be written as:

$$f(a) = \sum_g I_g(a)\ \lambda_g \qquad \{4\}$$

where λ_g is the model's fertility level for age group g, and $I_g(a)$ is an indicator function equal to 1 if age a belongs to group g, and equal to zero otherwise. The age-group rates $\lambda_1 \ldots \lambda_G$ may be unrestricted, as in {2}, or they may be required to conform to some parameterized schedule $\lambda_g = \lambda_g(\theta)$. The important point for the exposition is that fertility levels are identical at all ages within each group.

From [8], the log likelihood for an individual observation (a_i, u_i, δ_i) is

$$L_i = \delta_i \ln f(a_i - u_i) - \int_{a_i - u_i}^{a_i} f(z)\, dz \qquad \{5\}$$

The first term in this equation corresponds to the probability of a birth at age a_i-u_i (if δ_i=1 and a birth was reported), and the second term corresponds to the probability of surviving without a birth over the age interval $(a_i\text{-}u_i, a_i)$. When the model fertility schedule is piecewise-constant, this may be rewritten in terms of age groups as

$$L_i = \sum_g \left\{ \left[\delta_i\, I_g(a_i - u_i) \right] \left[\ln \lambda_g \right] \right\} - \sum_g \left\{ \left[\int_{a_i - u_i}^{a_i} I_g(z)\, dz \right] \left[\lambda_g \right] \right\} \qquad \{6\}$$

Summing over observations *i* yields the sample log likelihood

$$L = \sum_g \left\{ \left(\sum_i \left[\delta_i I_g(a_i - u_i) \right] \right) \left[\ln \lambda_g \right] \right\} - \sum_g \left\{ \left(\sum_i \left[\int_{a_i - u_i}^{a_i} I_g(z)\, dz \right] \right) \left[\lambda_g \right] \right\} \quad \{7\}$$

or, more intuitively,

$$L = \sum_g \left(B_g \ln \lambda_g - Y_g \lambda_g \right) \quad \{8\}$$

The derivation of {8}shows that the indices ($B_1 ... B_G, Y_1 ... Y_G$) contain all the information necessary for maximum likelihood estimation of any piecewise-constant fertility model from last-birth data. Event and exposure totals for each age group are sufficient summaries of the observable fertility histories.

2.4 Poisson Estimation

Equation {8} is closely related to the Poisson distribution. The natural logarithm of the probability that a Poisson process with rate λ generates B events in Y years is

$$B \ln \lambda - Y \lambda + C \quad \{9\}$$

where C = B ln Y - ln B!. Except for the C terms (which do not vary with λ), the log likelihood in {8} is a sum of the logs of G Poisson probabilities, one per age group. Thus, for any individual-level model with piecewise-constant fertility levels, one can calculate maximum likelihood estimators for parameters by pretending that the aggregate-level DLB data have distributions

$$B_g \sim Poisson(Y_g \lambda_g) \qquad g = 1 \ldots G \quad \{10\}$$

This estimation procedure, derived here for open-interval DLB data, is identical to that used by Broström [3] for standard fertility data.

2.5 Discussion

The distributional result in {10} leads us to one of this paper's main points. *A researcher using DLB data may use standard methods for estimating rates or fertility parameters. Estimation procedures for open-birth interval (DLB) or last-year (BLY) data differ only in the manner in which the data sets are assembled, not in the quantitative methods used. DLB data require no special statistical techniques, despite the unusual sampling scheme that generates the DLB versions of* Y_g, B_g, *or other data summaries.*

This conclusion held for the simple model presented in [8] (Equations (2) and (3)), and the exposition here shows that it is equally true for any fertility model in which rates are a function of age group. Furthermore, because age groups may be arbitrarily narrow, we expect (although we have

not formally proven it here) that the main result – i.e., that appropriate estimation methods are identical for BLY and DLB data – also applies to models in which f(a) is a continuous function of exact age.

3 Empirical Examples: Methods and Data

3.1 Poisson Regression for Coale-Trussell Parameters

We now apply these results to a well known fertility model, Coale-Trussell, using open-interval data from public use samples of Brazil's 1991 census. The simplest version of the Coale-Trussell model schedule for marital fertility [4] assumes that fertility levels for five-year age groups are related to one another by the parametric specification

$$\lambda_g(M,m) = M\, N_g^* \exp(m\, v_g^*) \qquad g = 1 \ldots G \qquad \{11\}$$

or, defining a new, mathematically more convenient parameter k=ln(M),

$$\lambda_g(M,m) = N_g^* \exp(k + m v_g^*) \qquad g = 1 \ldots G \qquad \{12\}$$

where G=6, the age groups are 20-24, 25-29, ..., 45-49, and the N_g* and v_g* values are known constants ([4], p. 188).

The results for piecewise-constant models in the previous section therefore imply that with aggregate DLB data generated by a Coale-Trussell fertility schedule with parameters (k,m), the researcher can estimate (k,m) by maximizing the sample likelihood under the assumption that:

$$B_g \sim Poisson\,[\,Y_g\, N_g^* \exp(k + m v_g^*)\,] \qquad \{13\}$$

Most modern statistical software packages can estimate (k,m) from $\{B_g, Y_g\}$ using the generalized linear modeling approach. Estimation is based on the relationship

$$\ln E(B_g) = \ln[\, Y_g\, N_g^*\,] + k + m v_g^* \qquad \{14\}$$

Broström [3] provided an example program for GLIM software [6]. Table 2 gives additional examples for the SAS and S-PLUS software systems.

Table 2
Code for Estimating Poisson Model

In both examples below, we assume that the researcher has prepared a data set called DLB. DLB must have 6 observations (one per age group 20-24...45-49), and must contain variables called N and V (the Coale-Trussell constants) and B and Y (the aggregate totals of births and years, respectively, from the last-birth data).

Both examples produce an estimated intercept, k=ln(M), and an estimated slope (m).

SAS

```
proc genmod data = DLB ;
off = log(Y*N);
model B = V / dist=poisson offset=off ;
```

S-PLUS

```
off <- log(Y*N)
glm(B~V, data=DLB, family=poisson, offset=off)
```

3.2 Brazilian Census Data

In all of our examples we use data from public use samples of Brazil's 1991 demographic census, which collected current fertility information exclusively in DLB form. Our focus is on subnational estimates. We analyze fertility in 723 small areas, called municipalities (*municípios* in Portuguese) from the state of Minas Gerais. Municipalities are roughly equivalent to U.S. counties, and these 723 administrative units cover the state completely, with no overlap. We selected the 1991 Minas Gerais data as a test case because of earlier work by colleagues [2], who applied a very different set of statistical methods – Bayesian spatial smoothing of the standard BLY data – to estimate municipal-level fertility control.

Table 3 presents information on the 1991 census sample for Minas Gerais. All data in this table refer to unweighted samples of women 20-49, regardless of marital status, on the census date. The overall sample is very large, with approximately 392,000 women. Municipal-level sample sizes vary widely, however. Column (1) provides information on the number of woman-years available from the year before the census; by construction, this equals the number of women surveyed. Many municipalities have extremely small sample sizes: there is information for fewer than 100 women in 49 of the 723 municipalities, and for fewer than 200 women in 206 (49+157) municipalities. The smallest municipal-level sample contains information for only 30 women aged 20-49, and the median size for the municipal-level samples is 311 women. Column (3) displays data on the cross-municipality distribution of births in the year prior to the census (i.e., BLY birth data). Last-year births are in single digits (0-9) for 62 of the municipalities, and the majority of municipalities (556 of 723) have fewer than 50 last-year births to interviewed women.

TABLE 3
Distribution of Unweighted Samples Sizes across 723 Municipalities in Minas Gerais, Brazil 1991 Public Use Census Samples

	# of municipalities in sample size range			# of municipalities in sample size range	
	(1)	*(2)*		(3)	*(4)*
WOMAN-YEARS	Past year (BLY)*	*Five years (DLB)*	BIRTHS	Past year (BLY)	*Five years (DLB)*
0-99	49	*0*	0-9	62	*0*
100-199	157	*3*	10-19	157	*12*
200-499	361	*102*	20-49	337	*89*
500-999	98	*198*	50-99	105	*195*
1000+	58	*420*	100+	62	*427*
Total	723	*723*	Total	723	*723*
Minimum Size	30	*125*	Minimum Size	0	*13*
Median	311	*1,162*	Median	30	*119*
Maximum	47,865	*192,869*	Maximum	3,285	*13,526*

* Woman-years over the past year equals the number of women interviewed

The small sample sizes for many municipalities clearly create severe challenges for estimating sensible local-level fertility indices, and for analyzing inter-municipality differences. With such small samples of women and last-year births, estimated fertility indicators may vary widely across municipalities merely because of coincidental sampling noise, not because of any real features of the fertility regime. Variability in small samples is likely to be a particularly bad problem for the Coale-Trussell *m* parameter, which typically has high standard errors and wide confidence intervals even in large samples ([3], Table 3).

As an extreme example of sampling variability, consider the municipality with the smallest number of women interviewed, Serra da Saudade, in central-western Minas Gerais. The 1991 census sample for Serra da Saudade includes only 30 women – four each in the 20-24, 30-34, and 40-44 age groups, eight each in the 25-29 and 35-39 groups, and two women 45-49. (Readers can view and manipulate the entire census sample for this municipality in the Addendum's spreadsheet, *Serra da Saudade.xls*.) Only two women, one 25-29 and one 35-39, reported births in the year before the census. A demographer who heroically (and naively) estimated the Coale-Trussell *m* parameter from these data would arrive at a value of -1.43. In contrast, estimated *m* values for the four (more populous) municipalities that border Serra da Saudade are 1.01, 2.00, 0.71, and 1.35. Serra da Saudade appears, then, to be an anomalous island in sea of fairly high fertility control. This is nonsense, of course. Differences in *m* between Serra da Saudade and its neighbors are caused almost entirely by the coincidental fact that half of the reported births for 1991 (1 of 2) were in the 35-39 age group, and because the sample weight for the older of the two mothers is higher. As one might expect, sampling noise, rather than real fertility differences, is the main cause of the local variation in *m*.

Researchers can ameliorate the problem of small sample sizes when fertility data are collected in DLB form (as they are in the 1991 Brazilian census) by using information from woman-years that occurred more than one year before a survey. Columns (2) and (4) of Table 3 show how expansion of the Minas Gerais sample back to T=5 years before 1991 increases sample sizes. The numbers of observed births and woman-years in each municipality are approximately quadrupled by this procedure, and the distribution of municipal-level sample sizes shifts dramatically. With DLB data, the majority of municipalities have over 1000 woman-years and 100 births from which to estimate fertility. In contrast, only the very largest municipalities had equivalent sample sizes with the last-year-only data.

DLB sample sizes are still fairly small, but it is far more plausible that one can extract meaningful fertility information from the DLB than from the BLY samples. Roughly speaking, sample sizes quadruple, which should halve the standard errors of estimators. This represents a significant improvement in accuracy, and by reducing the level of noise in the data researchers can often "hear the signal" (i.e., identify systematic patterns of interest) much better.

3.3 Simulated Small-Sample Properties of BLY and DLB estimators

As demonstrated in [8], under the strong, idealized assumptions of many formal demographic models (constant age schedules and complete homogeneity within age groups), DLB data produce consistent parameter estimators that have lower variance than BLY estimators. Theoretical tests and empirical simulations in [8] also demonstrated that DLB estimators outperformed BLY under more realistic conditions, when age schedules change and fertility rates vary within age groups.

However, Schmertmann [8] compared DLB and BLY estimators only in models without parametric restrictions on the set of age-specific rates $\{\lambda_{15\text{-}19},\ldots,\lambda_{45\text{-}49}\}$. The Coale-Trussell model imposes parametric restrictions, and it is possible that the comparative performance of BLY and DLB estimators therefore differs. Most importantly, when fertility falls rapidly before the census date, as it did in Minas Gerais over the 1980s, DLB estimates of *m* for the census date may be biased downward, because the DLB data include earlier years in which fertility control was lower. Adding these woman-years to the DLB sample may therefore "contaminate" the estimate of current *m*.

The earlier simulations with changing rates in [8] suggest that any such bias is likely to be small. However, before calculating (*M*,*m*) estimates for hundreds of municipalities, it is instructive to compare small-sample properties of Coale-Trussell estimators based on actual BLY and DLB data from Minas Gerais.

We investigated these properties by drawing large numbers of subsamples of different sizes from the Minas Gerais 1991 public use sample. For each subsample we calculated Poisson regression estimates of (*M*,*m*) from both BLY and DLB versions of the data. We focus here on the second parameter *m*; results are nearly identical for *M* or *k*=ln *M*.

The distribution of *m* estimates over many subsamples allows us to assess (1) the magnitude of DLB biases caused by including woman-years from earlier periods of (presumably) lower fertility control, and (2) the reduction in sampling variability achieved by including these additional woman-years in the DLB sample.

Census public use files contain DLB data for approximately 265,000 married women 20-49 in Minas Gerais in 1991. Table 4 contains the weighted counts of these women by (a,u,δ) cell, using T=5 as the maximum sampling period. For simulation purposes we assume that a population of women is distributed across (a,u,δ) cells with exactly these proportions.

BLY estimates from the Minas Gerais sample, which use data from the last year only, are

$$\boldsymbol{k^*= -0.449 \qquad M^*=0.638 \qquad m^*=1.036 \qquad \text{[full sample BLY]}}.$$

We assume that these are the "true" population parameters to be estimated from small samples. DLB estimates from Table 4, which add potentially contaminating fertility information from 1-5 years before the 1991 census, are

$$\boldsymbol{k= -0.433 \qquad M=0.649 \qquad m=1.001 \qquad \text{[full sample DLB, T=5]}}.$$

The two estimation methods produce similar parameter estimates in the large sample in Table 4, but we wish to investigate their comparative performance in small samples. In particular, we wish to learn which method is more likely to produce estimates of *m* for 1991 that are close to the "true" value *m**=1.036, and to learn how performance of BLY and DLB estimators varies with sample size.

Table 4
Time Since Last Birth for Currently Married Women in Minas Gerais, 1991
Weighted Totals from Public Use Sample

	Years Since Last Live Birth						
AGE	0-1	1-2	2-3	3-4	4-5	5+/never	TOTAL
20	13,808	9,962	5,378	2,028	693	12,934	44,804
21	15,441	11,407	6,989	3,745	1,724	14,132	53,438
22	17,898	13,356	9,596	5,681	2,510	15,888	64,928
23	18,160	14,961	10,918	6,612	3,772	17,407	71,829
24	18,194	15,365	12,558	8,008	4,823	17,644	76,592
25	18,195	16,326	13,270	8,992	6,303	20,046	83,133
26	17,621	16,612	14,005	10,044	7,093	22,148	87,523
27	16,686	15,518	14,165	11,518	8,483	25,445	91,813
28	16,341	15,419	14,118	11,122	8,685	27,764	93,449
29	14,203	13,965	13,585	11,411	8,978	30,813	92,954
30	13,413	12,994	12,134	10,637	9,070	32,502	90,751
31	11,159	11,151	11,187	10,316	9,096	36,002	88,910
32	10,336	10,470	10,390	9,547	9,375	41,639	91,758
33	8,487	9,674	9,779	8,971	8,877	44,627	90,415
34	7,707	8,120	8,190	8,008	8,030	46,806	86,861
35	6,362	7,091	6,705	7,275	7,487	47,965	82,885
36	5,701	6,269	6,626	6,250	6,571	51,869	83,287
37	5,148	5,503	5,831	5,214	5,737	52,947	80,380
38	3,881	4,564	4,980	4,649	5,569	52,261	75,904
39	3,387	3,821	4,769	4,124	4,692	52,046	72,839
40	2,805	3,794	3,899	3,629	4,079	52,416	70,622
41	2,162	2,377	3,390	3,159	3,428	49,106	63,622
42	1,856	2,179	2,644	2,532	2,721	48,779	60,711
43	1,344	1,685	2,314	2,425	2,608	50,428	60,804
44	917	1,262	1,786	1,952	2,295	47,061	55,274
45	634	834	1,456	1,543	2,015	46,482	52,965
46	515	613	1,303	1,317	1,727	45,853	51,328
47	268	348	646	1,039	1,339	42,750	46,391
48	234	386	437	664	950	42,124	44,795
49	200	144	374	449	788	41,749	43,705
TOTAL	253,066	236,169	213,423	172,859	149,519	1,129,632	2,154,669

For each of several sample sizes $N \in \{100, 200, 500, 1000, 2000, 5000\}$ we conducted a Monte Carlo study by repeating the following procedure 200 times:

- draw a pseudo-random sample of *N* women from the distribution of (a,u,δ) in Table 4
- construct BLY and DLB values for B_g and Y_g, g=20-24,...,45-49
- estimate Coale-Trussell parameters *k* and *m* by Poisson regression and record their values

Table 5 displays summary results from these studies. [Figure 1] displays the distribution of *m* estimates for the *N*=200 case, representing a typical municipal-level sample size in our example data.

Table 5
Summary measures for *m* estimates
over 200 Monte Carlo Samples at each Sample Size *N*

	BLY			*DLB*			% of samples in which DLB estimate
N	mean[a]	bias[b]	MAE[c]	*mean*[a]	*bias*[b]	*MAE*[c]	is closer to m*
100	1.17	0.13	0.69	*1.01*	*-0.03*	*0.27*	72
200	1.10	0.06	0.43	*1.02*	*-0.01*	*0.20*	72
500	1.04	0.00	0.26	*0.97*	*-0.06*	*0.13*	76
1,000	1.05	0.01	0.17	*1.01*	*-0.03*	*0.09*	74
2,000	1.03	-0.01	0.13	*1.00*	*-0.04*	*0.08*	73
5,000	1.04	0.01	0.08	*1.00*	*-0.04*	*0.05*	67
...							...
Population[d]	1.036	0	0	*1.001*	*-0.036*	*0.036*	0

[a] mean ≡ $[\Sigma_s m_s] / 200$, where s=1...200 indexes Monte Carlo samples
[b] bias ≡ mean - 1.036
[c] MAE ≡ $[\Sigma_s |m_s - 1.036|] / 200$
[d] Values on this row represent a single calculation from the full sample in Table 4, rather than Monte Carlo simulations. Under sampling without replacement, all possible samples of this size are identical.

The results in Table 5 illustrate that, in this particular case, DLB estimators produce markedly better results – indicated by lower mean absolute errors – at all sample sizes up to *N*=5000. As expected, falling fertility in Minas Gerais prior to 1991 leads to a tendency to underestimate fertility control *m* in DLB samples, as illustrated by the negative biases in the DLB column. This "contamination effect" is small, however. As one switches from BLY to DLB data, gains from

decreased sampling variance overwhelm disadvantages of bias from "contaminated" samples that include fertility information from earlier years. The net gain is especially large when N=100 or 200, because in this range of sample sizes there is evidence that BLY estimators have positive small-sample biases, as well as high variance. (An asymmetry in the data causes the right-skewed distribution in the estimates: small samples in which births are far below population averages are more likely than small samples in which births are far above.)

The most important column of Table 5 is the rightmost, which displays the percentage of Monte Carlo samples in which the DLB estimate of *m* was closer than the standard BLY estimate to *m**=1.036. The simulations show that despite small negative biases, the DLB estimate is approximately three times more likely to win this contest in any single sample of size $N \leq 2000$, and approximately twice as likely to be closer to *m** when $N=5000$.

In sum, simulation results with census data from Minas Gerais 1986-1991 illustrate that replacing the standard, truncated BLY form of open-interval data with DLB information produces superior estimators of Coale-Trussell parameters in small to moderately-sized samples such as those for the 1991 municipalities. In this particular case, as in the examples in the earlier paper [8], Monte Carlo evidence strongly suggests that DLB estimators yield better results. If the researcher's objective is to arrive at a sample estimate that is close to the population parameter, the benefits of decreased sampling variance with DLB data greatly exceed the small costs of increased bias. DLB is a far better bet to produce a good guess from a small sample.

4 Improvements in Small-Area Demographic Analysis with DLB Data

We now turn to two brief examples that illustrate how analysis can improve when the researcher uses all of the information in DLB data, as opposed to the usual censored form in {1}. We wish to demonstrate how improving fertility estimates improves demographic analysis, by means of some (simplified) examples of an increasingly common research task: analysis of demographic patterns over a large set of small geographic areas.

In order to illustrate the potential analytical gains from full use of DLB data, we estimated the Coale-Trussell *M* and *m* parameters twice for each of the 723 municipalities – first using the standard, BLY form of the data, and next using the full DLB sample back to a limit of five years before the census. In the absence of data on marital duration, we adopted the following simple procedure to convert from total to marital-only fertility rates within each municipality before estimating the Coale-Trussell parameters:

- tabulate B_g and Y_g values for all women (either BLY or DLB versions)
- calculate π_g, the proportion of women in each age group who were married in 1991
- approximate marital births (B_g') with total births: $B_g' = B_g$
- approximate marital exposure (Y_g') as $Y_g' = \pi_g Y_g$
- estimate $(k,m) = (\ln M, m)$ by Poisson regression of B_g' on v_g* with offset $\ln(N_g^* Y_g')$

Unlike the procedure for approximating marital fertility in the Monte Carlo simulations, this method uses only aggregate data on marital status. Differences between the two procedures are small. Our exposition again focuses on *m*, the estimated index of marital fertility control.

4.1 Example 1: Spatial Analysis

Table 6 displays summary information on the distribution of *m* estimates over municipalities. The BLY and DLB columns correspond to the two sources of data. Both data sources indicate that, overall, marital fertility control in Minas Gerais is high. The mean value of *m* (weighting all municipalities equally) is 0.94 from BLY data, and 0.84 from DLB data. BLY estimates of *m* are more variable across municipalities, however, with a higher standard deviation (0.68, versus 0.43 for DLB estimates). BLY estimates are also more prone to extreme, implausible values for *m*, at both the high and low ends of the distribution.

Table 6
Distribution of Coale-Trussell *m* estimates across 723 Municipalities in Minas Gerais, 1991 Census

	BLY data	*DLB data*
Mean	0.94	*0.84*
Std. Dev.	0.68	*0.43*
Minimum	-1.88	*-0.65*
5%ile	0.00	*0.15*
25%ile	0.51	*0.56*
Median	0.92	*0.82*
75%ile	1.27	*1.08*
95%ile	2.05	*1.65*
Maximum	4.96	*2.67*

Spatial patterns also emerge more clearly with the more stable DLB estimates. [Figure 2] displays two maps of the *m* estimates for Minas Gerais. Panels (a) and (b) map the BLY and DLB estimates, respectively. Neither map illustrates a clean, simple spatial structure of fertility control. (This might be too much to expect, since the spatial organization of other relevant variables may not be clean and simple.) Minas Gerais appears to have the highest levels of fertility control in the western “beak”, the lowest levels in the north. Broadly speaking, there is a northeast-to-southwest gradient of increasing fertility control. The DLB map in panel (b) shows this pattern more coherently

Visual inspection of the maps tells only part of the story, however, because the eye naturally focuses more on the larger municipalities (e.g., those in the northwest). Many important details in the smaller southern and eastern municipalities may be difficult to see from the full map. Table 7 contains (informal) measures of the maps’ global “smoothness”.

Table 7
Informal Measures of "Smoothness" for Minas Gerais maps of *m*

	BLY data	*DLB data*
Number of Municipalities	723	*723*
Number of Municipalities with $m < 0$	38	*7*
Fraction of Neighboring Municipalities in Identical *m* Categories*	.285	*.467*
Mean Absolute Difference $\|m_i - m_j\|$ between Neighbors	.641	*.324*
Fraction of Neighbors with $\|m_i - m_j\| \leq 0.25$	.274	*.486*
Fraction of Neighbors with $\|m_i - m_j\| \leq 0.50$	.505	*.783*
Fraction of Neighbors with $\|m_i - m_j\| \leq 0.75$	.679	*.920*

* "Neighboring" municipalities are those that share a boundary. Categories of *m* are defined in the map legends. All fractions in the table are calculated using row-standardized weights, so that municipalities with many neighbors and those with few neighbors receive equal treatment. The implicit question for all of the calculated fractions is "Select a municipality at random, then select one of its neighbors at random. What is the probability that the two selected cells satisfy the stated criterion?".

Data in the table indicate that pairs of neighboring municipalities fall in the same range of *m* (and therefore have identical map colors) approximately 47% of the time in the DLB map, versus 29% in the BLY map. The mean absolute difference in *m* estimates between neighboring municipalities is only half as large in the DLB map (0.32) as in the BLY map (0.64). Neighboring municipalities have *m* values within ±0.25 of one another nearly half of the time (49%) on the DLB map, but only about one fourth of the time (27%) in the BLY map. In short, a series of measures all suggest that the DLB map in panel (b) provides a smoother, more coherent, less mottled-looking picture of fertility control in 1991 than the BLY map. Although the spatial structure of fertility control is still complicated, DLB data filter out enough sampling noise from the BLY data to make the overall picture more sensible and more intelligible.

DLB estimation also improves formal statistical analysis of spatial patterns. [Figure 3] illustrates spatial autocorrelation in the BLY and DLB estimates of *m*, as measured by a common spatial statistic, Moran's *I*. The horizontal axis in the figure represents a simple distance measure

(neighboring municipalities are at distance 1, neighbors of neighbors are at distance 2, and so forth; for reference, the radius of the main area of Minas Gerais, minus the western 'beak', is approximately 12 – that is, municipalities on the northern edge of Minas Gerais are about 12 steps from municipalities at the center of the map). The vertical axis represents Moran's I, which is essentially the average correlation in estimated fertility control *m* between randomly chosen pairs of municipalities at the specified distance [7]. Positive values of *I* indicate that municipalities at a given distance from one another tend to have similar levels of fertility control, with higher values indicating stronger associations.

Both the DLB and BLY estimates show positive and declining correlations up to about 10 spatial lags, and negative correlations between more distant municipalities. This autocorrelation pattern is consistent with an overall gradient in *m* over Minas Gerais, as opposed to a set of locally homogeneous but globally heterogeneous "patches" of low or high fertility control ([7], p. 67). It is also the statistical manifestation of the impression made by the maps (particularly the DLB map in panel b) in [Figure 2], which suggest a fairly clear northeast-to-southwest pattern of increasing fertility control.

Both BLY and DLB data exhibit the same pattern in [Figure 2]. However, this correlation pattern is stronger, clearer, and empirically more convincing when one uses the DLB data. Correlations among adjacent municipalities are +0.46 for the DLB estimates of *m*, compared to only +0.12 for the BLY estimates. At six spatial lags, correlations are +0.20 (DLB) and +0.07 (BLY). At 16 lags (close to corner-to-corner distances on the state map), correlations are -0.14 (DLB) and -0.05 (BLY). As before, expanding the sample size by using the DLB data clearly eliminates much of the sampling noise in the *m* estimates, and brings the spatial patterns into much sharper relief.

4.2 Example 2: Regression Analysis

As a second example, we use the Minas Gerais data in a manner more familiar to demographers: regression of municipal fertility control levels (*m*) on municipal characteristics. Like the spatial example above, the actual analysis that we present here is somewhat simplistic. Our objective is to provide a modest illustration of the value of DLB estimates, not to present a thorough, nuanced analysis of Brazilian fertility patterns.

In this spirit, consider a municipal-level analysis of the impact of growth by evangelical Protestant churches on marital fertility in Minas Gerais. Brazil is officially Catholic, but evangelicals are an increasing minority. In Minas Gerais in 1991, 87% of the population was Catholic, 8% evangelical. As with fertility, there are significant inter-municipal differences. Define EVANG as the fraction of women 20-49 in a municipality who report their religion as evangelical Protestant. Seven municipalities have EVANG=0, while at the other extreme eight have EVANG > 0.25, with a maximum (in the municipality of Itueta, on the eastern edge of the state) of 0.41.

It is unclear *a priori* whether a high percentage of evangelicals in a region should be correlated positively or negatively with fertility control. On one hand, many evangelical churches are family-oriented and emphasize traditional gender roles, which would suggest a pro-natalist influence and lower levels of control in regions with more evangelicals. On the other hand, areas that are "more evangelical" are by construction "less Catholic", which might mean that women tend to use more effective methods of birth control, making *m* higher in such regions.

A recent study comparing the family planning practices of evangelicals with Catholics in Rio de Janeiro, a state which borders southeastern Minas Gerais, Machado [5] found that evangelicals were more likely to use modern birth control methods, especially sterilization. Machado also notes that one prominent evangelical denomination (the Universal Church) has openly criticized the Catholic Church's position on birth control, has encouraged its members to pool resources in order to fund one another's sterilizations, and seems to have used female participation in debates on sexuality and family planning as a conscious strategy to recruit new members. These findings support the hypothesis of a positive correlation between a municipality's fraction evangelical and its estimated *m* value.

The left side of Table 8 reports results from a simple exploratory regression with the BLY data, in which the only independent variable other than EVANG is the fraction of the municipality's population residing in urban areas (URB):

$$E(m_i) = \beta_0 + \beta_1 \cdot URB_i + \beta_2 \cdot EVANG_i \qquad i = 1 \ldots 723 \qquad \{15\}$$

These BLY results suggest a positive correlation, but the EVANG coefficient of 0.68 is not significantly different from zero (p=.191), and the overall regression fit is poor (R^2=.064). In contrast, using the DLB estimates of *m* as the dependent variable yields a strongly significant positive correlation between *m* and EVANG (β=0.90; p=.004), and a much better overall fit for the regression (R^2=.197).

Table 8
OLS Regression Estimates, Dependent Variable = *m*

	BLY DATA			*DLB DATA*		
Coefficient	Estimate	t-stat	P-value	*Estimate*	*t-stat*	*P-value*
Intercept	0.43	5.63	0.0000	*0.27*	*6.01*	*0.0000*
URB	0.81	6.63	0.0000	*0.90*	*12.37*	*0.0000*
EVANG	0.68	1.31	0.1911	*0.90*	*2.92*	*0.0036*
	R^2= 0.064			*R^2= 0.197*		

Once again, switching to DLB data rids the data of sample noise that obscures important patterns. The DLB regression in Table 8 is only a beginning in the analysis of the effect of religion on Brazilian fertility, but (contrary to the results from the BLY version) the significant coefficient on EVANG is an important signal to the researcher to investigate further.

5 Conclusion

We have two main messages in this paper. The first is that parametric fertility models may be estimated from open-interval (DLB) birth data in straightforward fashion. Differences between open-interval estimation and standard methods lie mainly in the construction of the data sets, not in the application of statistical methods.

The second message is that use of DLB data can make a critical difference to the quality of statistical results. This is particularly true when analyzing populations at highly disaggregated levels. In fertility estimates from small samples of women, sampling noise can drown out signal. This is a familiar problem, of course, but using more of the information inherent in DLB data can greatly improve the precision of statistical estimators. More precise estimators make for better analysis and stronger conclusions. Our examples with Brazilian census data make this point in several ways: maps of demographic parameters are more coherent, spatial statistics have more power, and regressions provide clearer answers to questions about fertility's relations with other social and demographic variables.

DLB data are often available to researchers, but they are seldom used to their full potential. When fertility data are collected in last-birth or open-interval form, the methods elaborated in this paper can significantly improve the demographic analysis of small samples.

6 Acknowledgments

This research was supported by the Mellon Foundation's Program in Brazilian Demography and Area Studies at the University of Texas Population Research Center. We thank Joe Potter, Tom Pullum, and Dan Powers for their helpful insights and comments.

REFERENCES

[1] P.D. Allison, 1985. “Survival Analysis of Backward Recurrence Times”. *Journal of the American Statistical Association* 80(390):315-322.

[2] R. Assunção, J.E. Potter, and S.M. Cavenaghi, 1998. “Estimating Fertility Schedules with Bayesian Spatial Models”. Paper presented at the annual meeting of the Population Association of America, Chicago.

[3] G. Broström, 1985. “Practical Aspects on the Estimation of the Parameters in Coale’s Model for Marital Fertility”. *Demography* 22(4):625-631.

[4] A.J. Coale and T.J. Trussell, 1974. “Model Fertility Schedules: Variations in the Age Structure of Childbearing in Human Populations”. *Population Index* 40:185-258.

[5] M. das D.C. Machado, 1996. “Sexual Values and Family Planning among Charismatic and Pentecostal Movements in Brazil”. *Reproductive Health Matters* 8:76-85.

[6] P. McCullagh and J.A. Nelder, 1983. *Generalized Linear Models.* Chapman and Hall: London.

[7] J. Odland, 1988. *Spatial Autocorrelation*. Scientific Geography Series, Vol. 9. Sage: Newbury Park, CA.

[8] C.P. Schmertmann, 1999. “Fertility Estimation from Open Birth Interval Data”. *Demography* 36(4):505-519.

[9] M.C. Sheps, J.A. Menken, J.C. Ridley, and J.W. Lingner, 1970. “Truncation Effect in Closed and Open Birth Interval Data”. *Journal of the American Statistical Association* 65(330):678-693.

[10] K. Srinivasan, 1968. “A Set of Analytical Models for the Study of Open Birth Intervals”. *Demography* 5(1):33-44.

[11] K. Srinivasan, 1970. “Findings and Implications of a Correlation Analysis of the Closed and the Open Birth Intervals”. *Demography* 5(1):33-44.

[12] United Nations, 1992. *Handbook of Population and Housing Censuses, Part II: Demographic and Social Characteristics*. Department of International Economic and Social Affairs, Statistical Office, Studies in Methods, Series F, Number 54. New York.

[13] United Nations, 1996. “Topics on Fertility and Mortality Collected in Population Censuses 1985-1994". Manuscript. UN Statistical Division and UN Population Division.

[14] K. Venkatacharya, 1972. “Some Problems in the Use of Open Birth Intervals as Indicators of Fertility Change”. *Population Studies* 26(3):495-505.

Figure 1. Distribution of BLY and DLB estimates of *m* over 200 Monte Carlo samples of size N=200, drawn from Minas Gerais 1991 public use samples

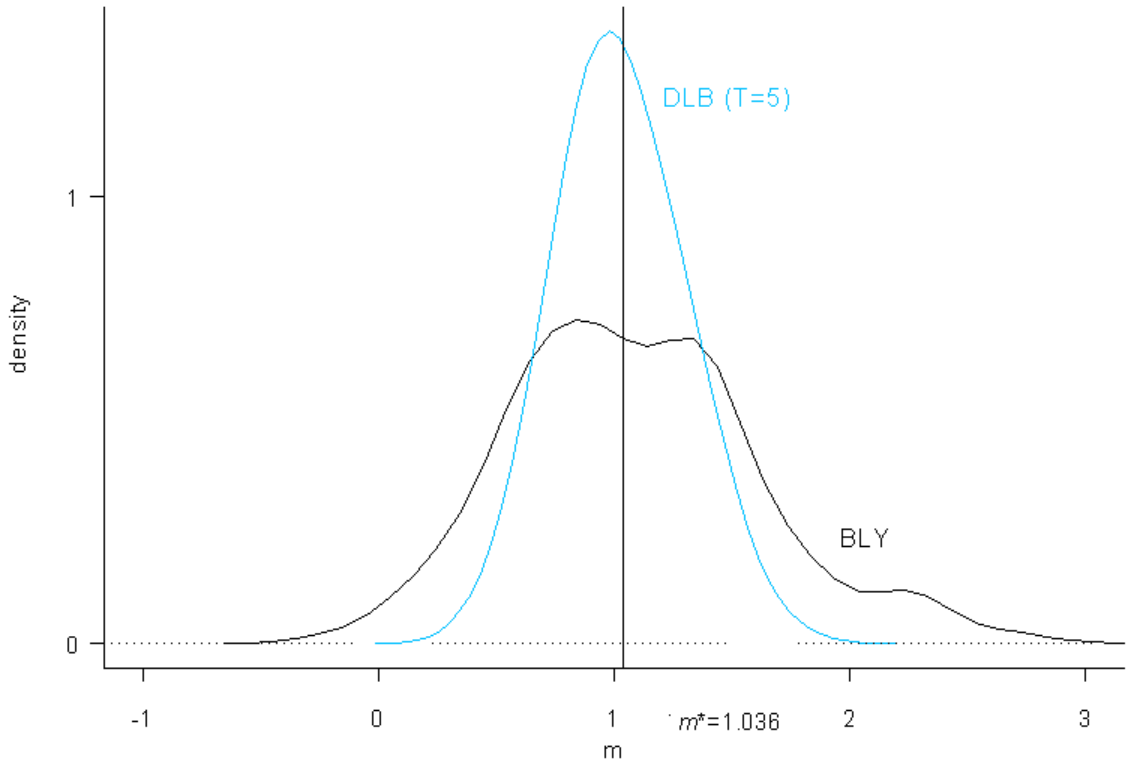
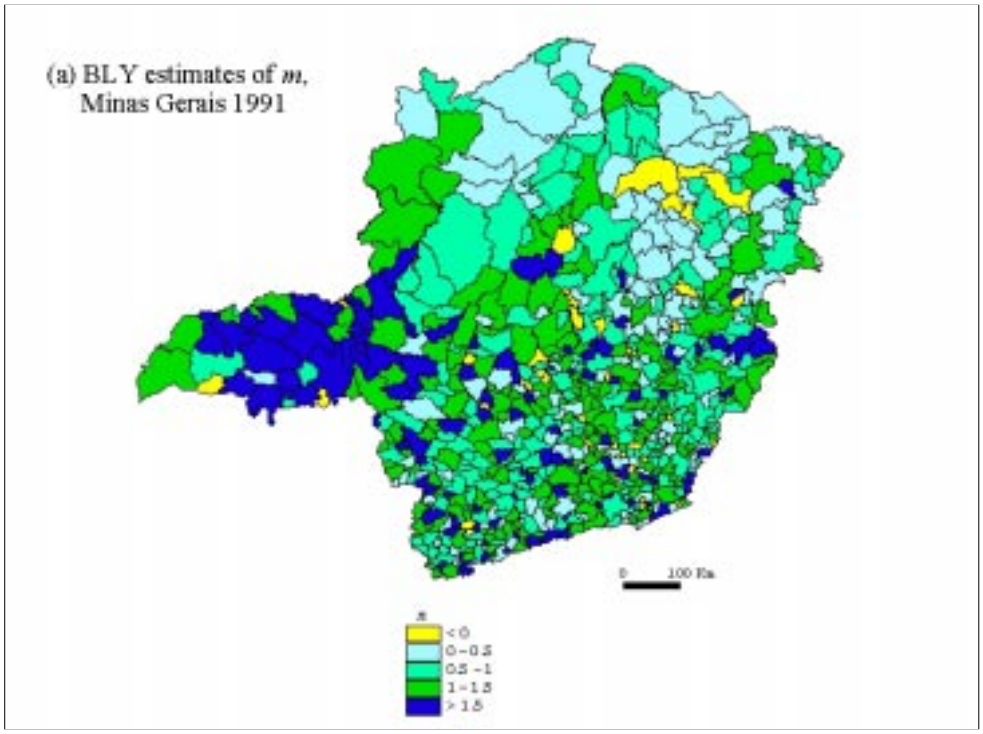
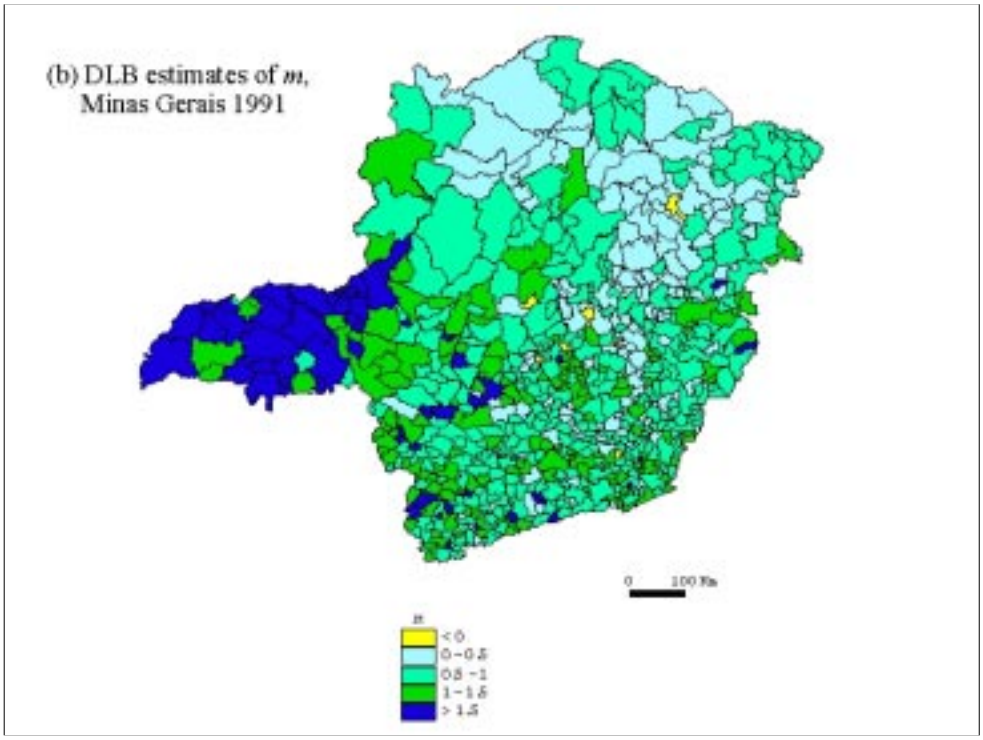


Figure 2. Municipal-level Estimates of *m* from 1991 Public Use Census Data.





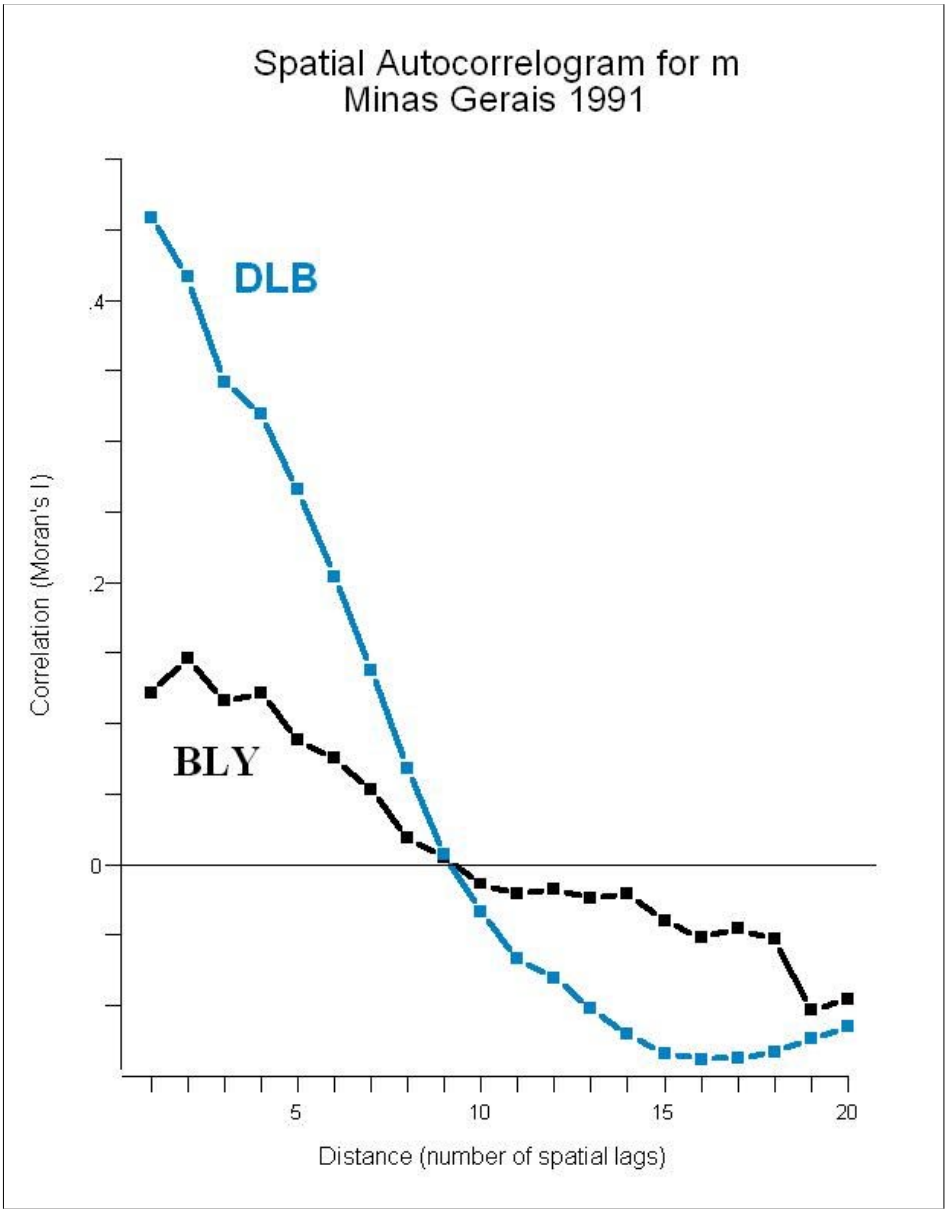


Figure 3. Moran's *I* for Various Spatial Lags, Minas Gerais estimates of *m*.