Applications of a Component Model for Structuring Housing Unit-Based Total Population Estimates into Categories of Age and Sex: A Pilot Test in New Mexico

DRAFT FOR PRESENTATION AT PAA, SPRING 2008: DO NOT CITE WITHOUT CORRESPONDING AUTHOR'S PERMISSION

Jack Baker*, Adelamar Alcantara, Xiaomin Ruan

Population Estimates Program

Bureau of Business and Economic Research

University of New Mexico

Brief Abstract: Housing unit-based estimates are popular, but lack a technique for estimating age/sex structure of a population. This paper evaluates the use of short-term component projections for allocating these estimates to categories of age/sex. The method is piloted for 2005 in Bernalillo County, New Mexico. The results suggest that the method will perform well where (1) population age/sex dynamics and construction trends may be assumed to be unrelated, (2) age-specific migration is well-measured, or (3) the magnitude of migration is very small. Here, the method produced estimates of population age/sex structure that matched closely with those produced by the Census Bureau using administrative records. This finding is promising because it is clear that Bernalillo County has recently experienced both significant housing unit growth as well as changes in age/sex structure. These results are discussed in light of their potential application to areas for which these conditions may not be met.

* Jack Baker, Ph.D.

(505) 277-2212

kali@unm.edu

Introduction

When accurately applied, housing unit-based population estimates may arguably better reflect local population trends than do traditional estimation techniques such as the -component model (Starcinyk and Zitter, 1968; Smith and Cody, 1994; Murdock and Ellis, 1991). The method, however, suffers from the potential drawback that it provides no way to simultaneously produce estimates of population characteristics such as age and sex. In cases where the housing unit-based total estimate is preferred (the ongoing debate over which method is more valid is outside the scope of this paper), Bryan (2004) has suggested that a short-term projection of age/sex structure might be used to allocate the total population estimate to these categories. While promising, this method relies on a fundamental assumption that either age/sex population dynamics are independent of residential construction trends or that the relationship between the two is well-captured in migration data. The first assumption is unlikely to be true in rapidly-growing areas since residential construction is marketed to particular demographic segments of the population and linked, therefore, to migration patterns with their own age/sex dynamics and trends. Residential construction is demand-driven and houses are strategically marketed to a target audience with a specific set of demographic characteristics. Such developments attract both local and non-local buyers and/or renters to sometimes previously undeveloped or redeveloping neighborhoods. Some developers will target young families while others might market to retirees. In either case, it may not be assumed that population age/sex structure and local trends in housing unit construction are unrelated. If historical age-specific migration trends are used, then short-term effects of increases in housing unit construction may not be adequately captured in a -component model and estimates of population characteristics will be biased in the direction of the error in the migration estimates. With these concerns in mind, any method of integrating housing unit-based estimates with a component model of population characteristics should be carefully evaluated prior to implementation. The model may be useful where no relationship exists between age/sex population dynamics and housing-unit growth (or they may be assumed to be unrelated—as when migration is very small), or where this relationship is adequately captured in the migration estimates.

Fortunately, a fairly straightforward method for evaluating the adequacy of this method—and its underlying assumptions—is possible. First, the resulting estimates may be compared to alternative component estimates, such as those produced by the United States Census Bureau at the County level using administrative records to evaluate their performance. Additional, indirect assessments are also possible based on analysis of the discrepancy between a component total population estimate and one based on housing units. In New Mexico, housing unit-based population estimates tend to be higher than those produced by the component method. If age/sex population dynamics are unrelated to housing unit growth, then we would expect that an equal distribution of the difference between the two total population estimates would make no difference in the age/sex structure of the population estimated by employing allocation factors to break the housing unit-based estimate according to proportions predicted by a component model. If these two are related, but this relationship is not captured in age-specific migration estimates, then an even distribution of this difference would distort the accuracy of the estimated age-structure. If migration is adequately captured, or age-structure and housing unit growth are not related, then this even distribution of the residual would not be observed to distort the age/sex structure. In this case, down-distribution of the total population estimate would produce an age-structure identical to that observed by increasing each age/sex category by a constant value capturing the overall difference between the two sets of population estimates. In a case where no significant difference between the two methods (rolling up versus distributing down) is found, then the assumptions of such a procedure would be met and the technique might be acceptable for use in applied demographic estimation.

This paper pilots the use of a short-term component projection for structuring housing unit-based total population estimates into categories of age and sex. It pilots the method in Bernalillo County, New Mexico for the July 1, 2005 population. Between 1990 and 2000, the County experienced a clear shift towards a more “rectangular” age-structure and the trend appears to have continued into the current decade. Furthermore, growth in housing units has been positive and strong during the same period suggesting that the County may represent an adequate place to test the adequacy of the proposed method. Housing unit-based total population estimates are provided by the University of New Mexico’s Bureau of Business and Economic Research for 2005. A component projection model is operationalized based on birth, death, and migration rates for the County and used to project the Census 2000 population age/sex structure forward to 2005 in single-year intervals. This estimate of the population age/sex structure is used to derive allocation factors against which the total housing unit-based estimate is multiplied as a scalar to arrive at numerical estimates for each age/sex category. Results are presented in five-year intervals for ease of interpretation and are compared to 2005 age/sex estimates produced by the United States Census Bureau based on administrative records. The study indicates that the model worked well for Bernalillo County in 2005, producing errors far less than one percent in any age/sex category. Further diagnostics on the adequacy of the method are conducted by comparing a “rolled-up” estimate to the “distributed down” version. In this case, errors were also extremely small, suggesting adequate measurement of migration for Bernalillo County. The implication of these results for population estimation, including a discussion of potential problems with its application to smaller populations, are briefly reviewed.

Materials and Methods

Study Population

Bernalillo County is situated roughly in the center of the State of New Mexico (Map 1). Historically, it has contributed a majority share to the New Mexico population. The presence of the City of Albuquerque and recent strong growth suggests that this trend will continue into the foreseeable future. Between 1990 and 2000, the County grew at a rate of 1.47% per year. Between 2000 and 2005, this rate accelerated to 1.98% per year (Figure 1). This rapid growth increase was driven by unprecedented housing construction in the City of Albuquerque (Figure 2), which was filled by both natural increase and positive net-migration between 2000 and 2005. While natural increase between 1990 and 2000 amounted to approximately 45,000[1] new persons in Bernalillo County, a total increase of 76,101 suggests positive net-migration of 31,101 persons across the decade. Between 1990 and 2000, the Bernalillo County population experienced a shift toward a more rectangular age structure (Figures 3 and 4), primarily in response to negative net-migration in younger age intervals and positive net-migration in the 20-24 age interval between 1995 and 2000. The rapid growth in housing units and population, coupled with clear changes in age-structure, make Bernalillo County an ideal setting for testing the usefulness of a component model for structuring housing unit-based population estimates into categories of age and sex.

Database Development

Annually, the University of New Mexico’s Bureau of Business and Economic Research’s (BBER) Population Estimates Program updates its demographic estimation database annually. This process includes collecting data from various state and local governmental agencies as well as through direct phone survey of Group Quarters facilities throughout the State. Housing unit stock estimates are derived from the Census 2000 baseline, which is updated based on building permits issued by 26 self-permitting agencies throughout the State, as well as the New Mexico Construction Industries Division. Administrative data on vital events including births and deaths are provided by the State Department of

[1] This figure uses 1990 to 1999 births (www.unm.edu/~bber) and an average for the 1999 to 2000 period.

Health. Migration data arrives from several sources, including summaries of Internal Revenue Service data from the Census Bureau, driver license issuance data from the Motor Vehicle Division, and are used to supplement the Public Use Microdata Sample (PUMS) data on 1995 to 2000 age-specific residence from the 2000 Census. These data were used to produce housing unit-based total population estimates for Bernalillo County, as well as to construct age-specific rates of births and deaths, and to model age-specific inputs for in and out-migration. Rates were constructed using 1999-2001 data on births and deaths, centered upon the Census 2000 population. Age-specific migration vectors were developed by smoothing 1995 to 2000 migration data to one-year time increments ($Mig_{95\text{-}00}/5$) and using this value as a constant between 2000 and 2005. This procedure assumed that in-migrants take on the fertility and mortality characteristics of the resident 2000 population of the County, which may not be true and suffers from the drawback that rates are kept at a constant level across the projection horizon while they might, in reality, fluctuate from year to year. On the other hand, this approach enjoys the strengths of simplicity and directness, with no requirements for additional modeling of the vital rates of in and out-migrating populations that would likely be based on a myriad of assumptions for which a large amount of additional data and sub-analyses would be required.

Housing-Unit Based Estimation Procedures

Housing unit-based population estimation relies upon correct specification of the number of housing units (HU) within a given geography, in conjunction with adequate data on the average household size (or average persons-per-household--PPH),the occupancy rate (OR), and the group quarters (GQ) population (Murdoch and Ellis, 1991; Starcynik and Zitter, 1968; Smith and Cody, 2004). Once these quantities are estimated, a simple equation is employed to produce a population estimate:

$$\mathbf{Pop_t = (HU_t * PPH_t * OR_t) + GQ_t}$$

The accuracy of the method is entirely dependent upon the accuracy with which these quantities are estimated. For this study, housing units were estimated based on Census 2000 County “stock” and building permits issued between April 01, 2000 and December 31, 2004 (assuming a six-month lag between building permit issuance and actual occupancy of the unit). Average Household Size (or persons per household—PPH) and occupancy rates were estimated by employing constants from the Census 2000. Group Quarters totals were derived from the BBER’s annual survey of facilities within Bernalillo County.

Component Model Procedures

Component population estimates are operationalized by using administrative data to "update" a base population in light rates of births, deaths, in-migration, and out-migration (Shyrock and Segal, 1973) in what amounts to an accounting-based procedure. The procedure is defined mathematically in the context of the "population balancing equation" (Murdock and Ellis, 1991; Smith, Tayman, and Swanson, 2001; Bryan, 2004).

$$\mathbf{Pop_t = Pop_{t\text{-}n} + (Births_{t\text{-}n,t} - Deaths_{t\text{-}n,t}) + (In\text{-}Migration_{t\text{-}n,\,t} - Out\text{-}migration_{t\text{-}n,\,t})}$$

Since the difference between births and deaths may be defined as the "natural increase" and that between in and out-migration as the "net-migration", this more cumbersome equation may be reduced to:

$$\mathbf{Pop_t = Pop_{t\text{-}n} + (Natural\ Increase_{t\text{-}n,\,t} + Net\text{-}Migration_{t\text{-}n,\,t})}$$ [2]

In a population without age-structure, this model would be sufficient to make total population estimates. In projecting population, as was necessary for this study, these data must be transformed into age-specific rates. We computed these rates, then operationalized the model using the well-known Leslie projection method (Leslie, 1945, 1948; Caswell, 2000) to update the 2000 base population to 2005. To operationalize the Leslie method, a matrix was created whose top row represented the products of survivorship and fertility rates for each age-interval (which yields the total births—age class 0 in the next year—when multiplied against an initial age-specific column vector) and contains 0s and the survivorship values for each age in the subsequent rows. In matrix notation, the estimating equation represented the initial age-distribution vector (N_t), the Leslie matrix of age-specific fertility and mortality rates (L) and the age-specific column vector of estimated net-migration counts (M):

---

[2] It is understood that either natural increase or net migration may be either positive or negative. In cases where one is negative, it is useful to remember that subtraction is simply addition of a negative number. This is reflected in the notation where net-migration is "added" but is often a negative quantity.

$$N_t = [\ N_0 * L\ ] + M$$

Model Fit Assessment

The true age/sex structured population of Bernalillo County in 2005 is unknown. The Census Bureau bases their estimates primarily upon Internal Revenue Service and Medicaid data, collected annually. Their annual population estimates reflect a model very similar to that employed here to estimate population characteristics (it is an administrative-records driven model) but as the final data for a given year is produced, the series of estimates is updated at a two-year lag (Bureau of the Census, 2005). This paper compared the 2005 Census Bureau estimates, updated in 2007 based on 2005 data, to those produced using a component model to structure the total population estimate. This method of evaluation assumes that the Census Bureau's estimates of population characteristics reasonably approximated those observed in Bernalillo County in 2005. The percent contribution of each age/sex category to the total population in the Census Bureau's estimates were compared to the percent contribution implied by the component model operationalized here. In addition, the absolute counts of each were computed for the model tested in this paper and compared to those of the Census Bureau. Absolute and percent residuals were computed for each age interval, along with overall average differences (both numerical and percentage-based) for both results.

Evaluation of Model Assumptions

An adequate application of the component projection for structuring total population estimates produced using the housing unit method would produce residuals that were evenly distributed across all age/sex categories. This would occur if one of two assumptions (or both) are met: (1) there is no relationship between age/sex structure dynamics and housing unit construction or (2) that migration data for Bernalillo County adequately captures this relationship. The third assumption of a potential model like the one tested here is that migration effects are small; obviously, this was not necessary to examine because it is clear that migration has been a strong force in population growth in Bernalillo County. Only a much more complicated analysis could discriminate between the first two assumptions. In essence, simple diagnostics may address both simultaneously because each would produce a similar effect. These assumptions were evaluated in this study by computing the total discrepancy between the component total population estimate and the housing unit-based total population estimate (the HU method is uniformly higher in New Mexico), allocating this discrepancy evenly across all age-sex categories and assessing discrepancies between this "null" age-structure and the one produced by distributing the total housing unit-based estimate down to age/sex categories based on the proportions produced by the component model. An adequate model would be one that showed little discrepancy between the two estimated age/sex population proportions, with only insignificant differences between a "down-distributed" population (based on the component model) and a "rolled up" model based on an even allocation of the residual between the two estimates across all age/sex categories.

Results

The overall results (Table 1) indicates a close fit between the estimated proportions for each age/sex category derived from the component model and those estimated by the Census Bureau using administrative records. In all age/sex intervals, the absolute error for each sex was computed as the difference between the Census Bureau estimate and the one produced in this study. Since the estimates were of percentages, these errors represent the fraction of a one-percent that occurred as a difference between the two estimates. A value of 0.001, for example, would indicate an error or 1/1000th of a percentage point. These very minor differences between the two sets of estimates likely stem from the fact that the component model implemented here made use of rates, not accounting of administrative data as is the case for the Census Bureau 2005 estimates. The average error across all male five-year age intervals was 2.0/1000ths of a percentage point, with a range between a low of 3.9/10,000ths of a percentage point and a high of 1.9/1,000ths of a percentage point. For females, the average error across all of the five-year age intervals was 2.2/1,000ths of a percentage point, ranging from a low of 1.9/10,000ths and a high of 5.8/1,000ths of a percentage point. These difference are, obviously, extremely low. In fact, they are much lower than expected using a method including rates that create a potential for rounding error. Overall, the results suggest a close congruence between the Census Bureau estimates and the component model estimates derived for this study.

Further diagnostics comparing a "rolled-up" and "distributed down" set of estimates supports the conclusion that the necessary assumptions of the method were met. Very small discrepancies, again much less than 1% across each age category (with an average difference in the percentages of 0.00000) between the rolled-up and distributed-down estimates of the age/sex structure were observed (Table 2). For males, the observed differences ranged between 1/1000th of a percentage point to 2/1000th of a percentage point. For females, the discrepancies were even smaller, ranging between a high of 2.1/1000th of a percentage point to a low of as little as 4/100,000th of a percentage point. The Female estimates also showed an average difference of 0.00000. This result suggests that in the case of

Bernalillo County, the assumptions that either housing unit construction and age/sex population dynamics are unrelated, or that the relationship is well-captured in the migration data used for this study, appear to hold for both sexes.

Discussion

In 2005, Bernalillo was a large County and while significant shifts in age-structure and strong housing unit growth have co-occurred for the 1990 to 2005 period, it was to be expected that the method described and tested here would work well. Input data on births, deaths, and migration are collected with a relatively high level of accuracy and although the model used here assumed continuing trends in the components of growth based on 2000 values, over the short horizon for which these projections were used it was to be expected that in highly-populated region, short-term fluctuations would have negligible effects on the estimated percentages in this geographic context. The implication of the results of this study is that for any geography for which good input data exist, the component model will provide a viable alternative for estimating population age/sex structure using short-term projections. The results agree well with estimates based on administrative data (the Census Bureau's 2005 estimates). There are negligible differences between estimates produced by allocating the total housing unit-based population estimate downward to age/sex categories and one that evenly spreads the error between a component estimate and a housing unit-based estimate, suggesting that although in-migration is large to Bernalillo County, it is well-measured in the Census 2000 data. Moreover, when spread across 18 age categories, the effect of this residual is very small.

What is much less clear is how these results relate to smaller geographies with less room to absorb error in age-specific birth, death, and migration rates. In these settings, small errors in the estimate of migration rates, for example, could have profound effects upon estimates of population structure because their relative magnitude would be larger. While small errors in estimates of migration rates in a large County with over 600,000 persons is likely to wash out in a component projection driven estimate, in a smaller County with 1,000 or perhaps even fewer residents these effects are likely to have a much more pronounced effect. Many of these smaller areas are characterized by their own migration dynamics. In New Mexico they often display either a stability of the population structure due to permanence of a "native" population, or significant losses in younger age intervals that can radically re-shape age-structure over even short horizons. In Counties with smaller populations, estimation of rates themselves is more sensitive to measurement error and when compounded with difficulties in measuring migration in these areas it is possible that the assumptions that either housing unit construction and migration are unrelated or that migration is well-captured in the input data may not be met. In many of these areas, it is possible that migration effects may be small, but such evaluation would have to be made on a case-by-case basis rather than based upon a generalized procedural evaluation.

On this note, it is worth pointing out that the current study does not pertain directly to small-area population estimates either, which will suffer from many of the same challenges as those involved with smaller population Counties. In Bernalillo County, the method worked well and provided that the assumptions of using a component model for the purpose outlined here hold, it should work well for any larger population with adequate data sources for constructing a component model. Further research validating the model in other large-scale settings should be conducted prior to consideration of this method as a standard practice. Future research should also focus on the adaptation of this technique to smaller populations.

References

Bryan, Thomas (2004) *Population Estimates*: In: The Methods and Materials of Demography. Jacob Siegel and David Swanson, Edrs. New York: Elsevier.

Bureau of the Census (2005) Estimates and Projections Area Methodology County Population Estimates by Age, Sex, Race, and Hispanic Origin For July 01, 2005. Published online at: http://www.census.gov/popest/topics/methodology/State_AgeSex_v2005.pdf.

Caswell, H (2000) Matrix Population Models: Construction, Analysis, and Interpretation. 2nd edition. New York: Sinauer.

Leslie PH (1945) On the Use of Matrices in Certain Population Mathematics. *Biometrika.* 3(3):183-212.

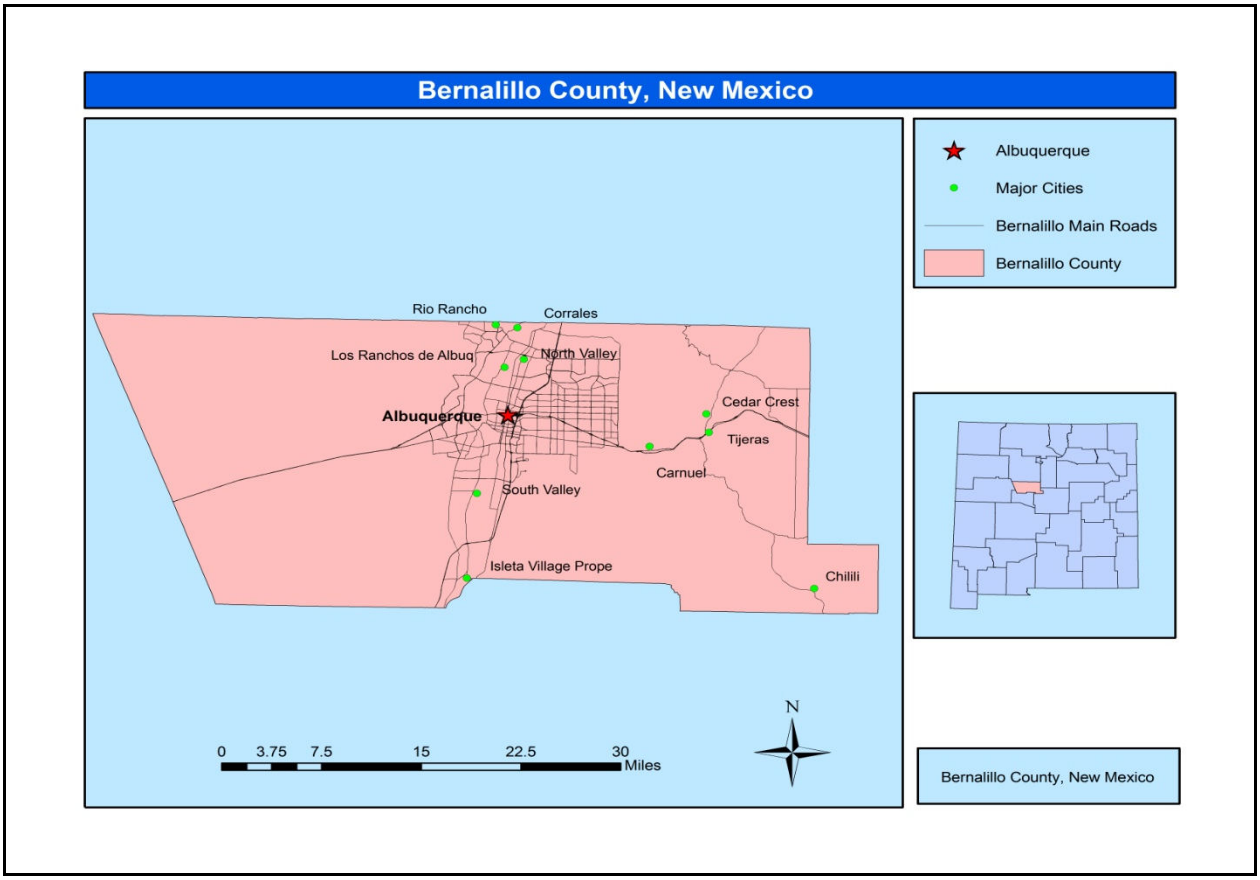
Leslie PH (1948) Some Further Notes on the Use of Matrices in Population Mathematics. *Biometrika*. 35(3/4):213-245.

Murdoch Steve and Ellis David (1986) Applied Demography: An Introduction to Basic Concepts, Methods, and Data. Boulder: Westview.

Smith Stanley, Tayman Jeff, and Swanson David (2001) State and Local Population Projections: Methodology and Analysis. New York: Kluwer.

Smith, Stanley and Cody, Scott (2004) An Evaluation of Population Estimates in Florida: April 1, 2000. *Population Research and Policy Review*. 23:1-24.

Starsinic, DE, Zitter M (1968) Accuracy of the Housing Unit Method in Preparing Population Estimates for Cities. *Demography*. 5:475-484.



Map 1. Bernalillo County, NM (2005)

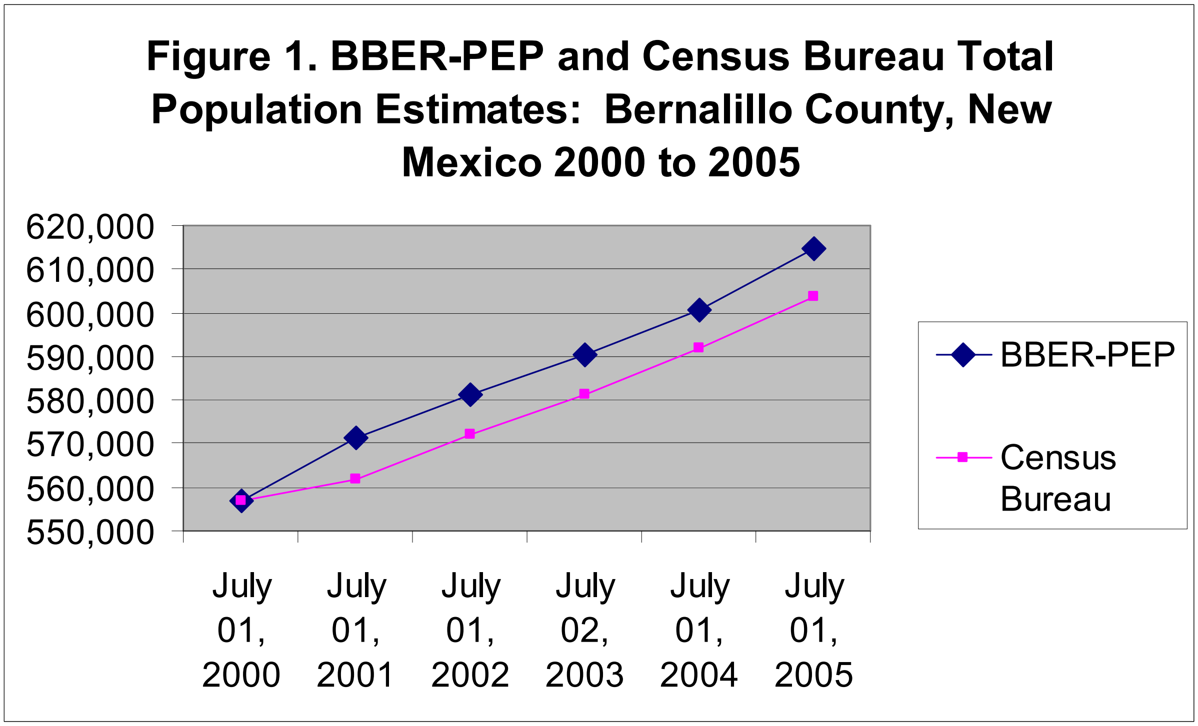


Figure 1. BBER-PEP and Census Bureau Total Population Estimates: Bernalillo County, New Mexico 2000 to 2005
620,000
610,000
600,000
590,000
580,000
570,000
560,000
550,000
July 01, 2000
July 01, 2001
July 01, 2002
July 02, 2003
July 01, 2004
July 01, 2005
BBER-PEP
Census Bureau

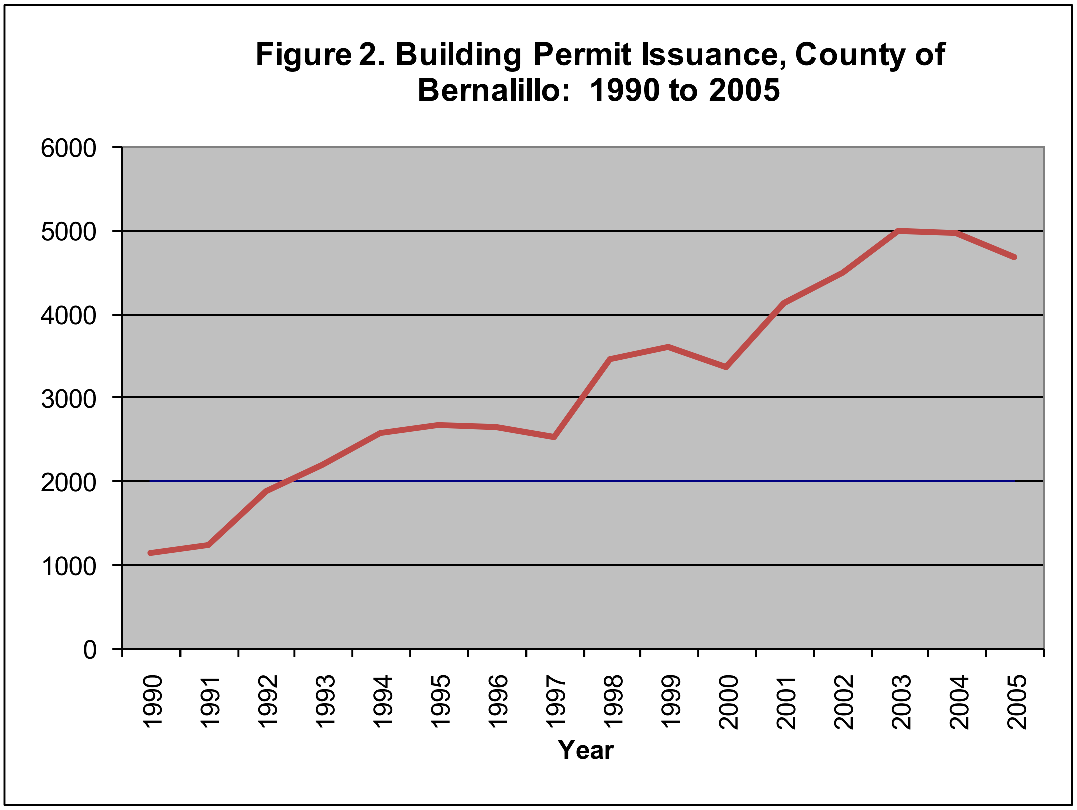


Figure 2. Building Permit Issuance, County of Bernalillo: 1990 to 2005
6000
5000
4000
3000
2000
1000
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
Year

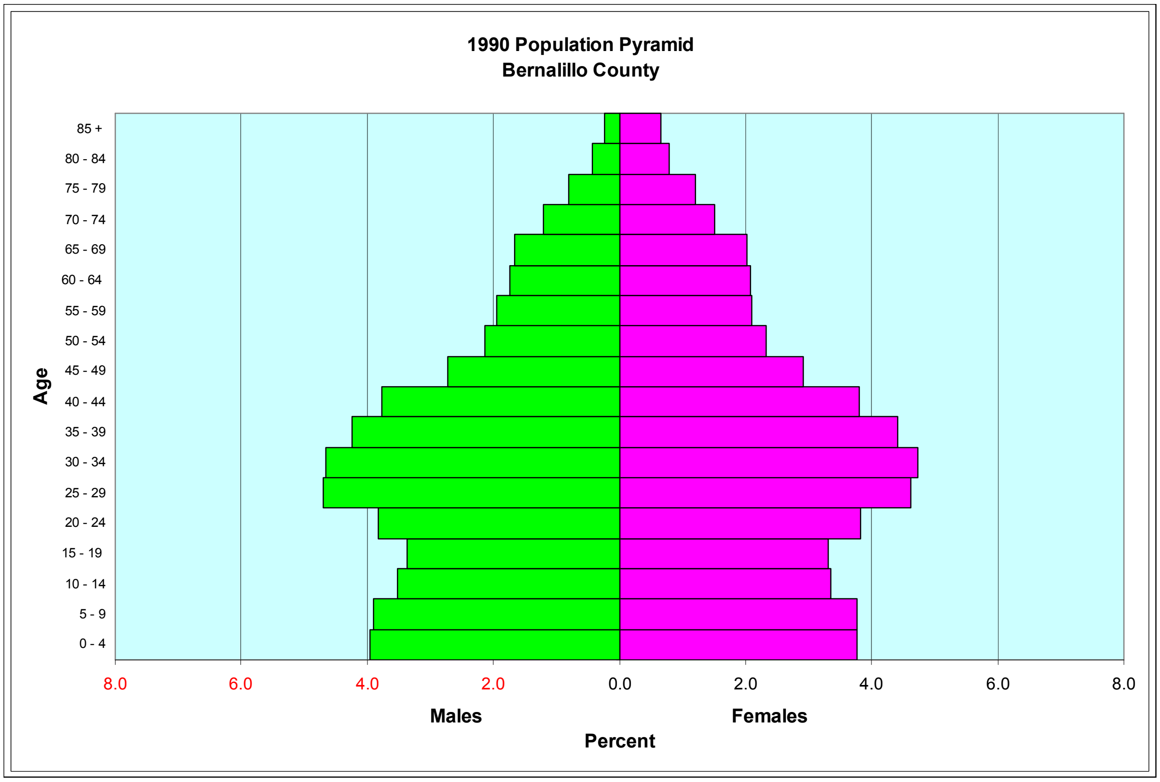


Figure 3. 1990 Bernalillo County Population Structure

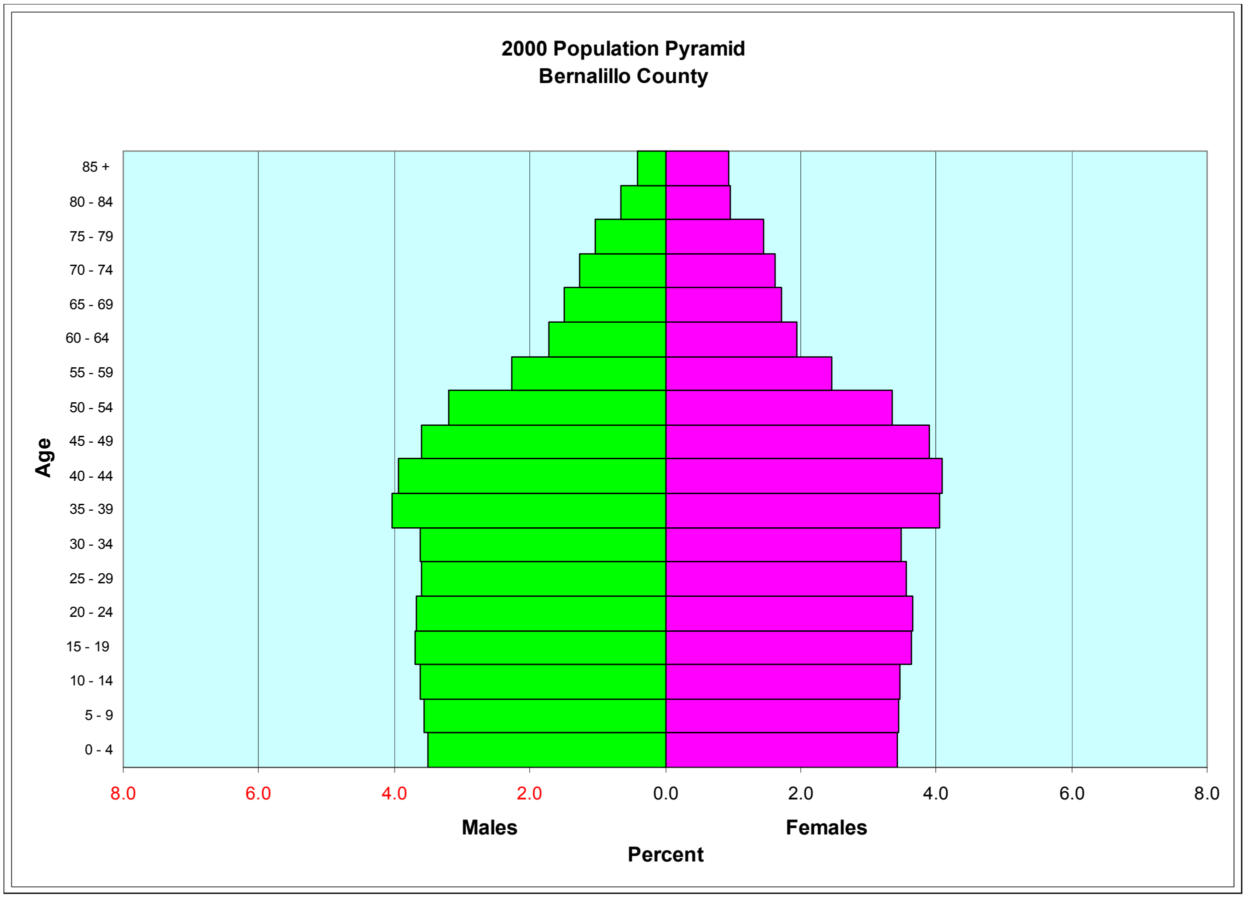


Figure 4. 2000 Bernalillo County Population Structure

Table 1. 2005 Component Based Age/Sex Specific Percent Estimates for Bernalillo County: Comparison with Census 2005 Estimates						
Age	**2005 Cohort Component Estimated Percent Male**	**2005 Cohort Component Estimated Percent Female**	**2005 Census Bureau Estimate Percent Male**	**2005 Census Bureau Estimate Percent Female**	**Absolute Difference in Percents Male**	**Absolute Difference in Percents Female**
0 to 4	7%	7%	7%	7%	0.00087	0.00140
5 to 9	7%	7%	7%	6%	0.00381	0.00336
10 to 14	7%	7%	7%	6%	0.00261	0.00363
15 to 19	7%	7%	7%	7%	0.00146	0.00125
20 to 24	7%	7%	8%	7%	0.00898	0.00888
25 to 29	7%	7%	7%	7%	0.00041	0.00127
30 to 34	7%	7%	7%	6%	0.00368	0.00588
35 to 39	7%	7%	7%	7%	0.00256	0.00184
40 to 44	8%	8%	8%	8%	0.00247	0.00107
45 to 49	8%	8%	8%	8%	0.00028	0.00019
50 to 54	7%	7%	7%	7%	0.00173	0.00284
55 to 59	6%	6%	6%	6%	0.00256	0.00230
60 to 64	4%	4%	4%	5%	0.00196	0.00271
65 to 69	3%	3%	3%	3%	0.00025	0.00070
70 to 74	3%	3%	3%	3%	0.00074	0.00150
75 to 79	2%	3%	2%	3%	0.00079	0.00074
80 to 84	1%	2%	1%	2%	0.00071	0.00039
85+	1%	2%	1%	2%	0.00039	0.00015
Total	100%	100%	100%	100%		
				Average Difference	0.00201	0.00223

Table 2. Method Diagnostics: Roll-up vs Distribute Down						
Male			Female		Absolute Difference in Percents	
Age	2005 Distributed Down	2005 Rolled Up	2005 Distributed Down	2005 Rolled Up	Male	Female
0 to 4	7%	7%	7%	7%	0.00084	0.00314
5 to 9	7%	7%	7%	7%	0.00074	0.00392
10 to 14	7%	7%	7%	7%	0.00075	0.00343
15 to 19	7%	7%	7%	6%	0.00075	0.00496
20 to 24	7%	7%	7%	6%	0.00066	0.00296
25 to 29	7%	7%	7%	7%	0.00079	0.00143
30 to 34	7%	7%	7%	7%	0.00077	0.00138
35 to 39	7%	7%	7%	7%	0.00081	0.00439
40 to 44	8%	8%	8%	8%	0.00118	0.00196
45 to 49	8%	8%	8%	8%	0.00104	-0.00127
50 to 54	7%	7%	7%	7%	0.00068	-0.00287
55 to 59	6%	6%	6%	6%	0.00026	-0.00124
60 to 64	4%	4%	4%	5%	-0.00065	-0.00165
65 to 69	3%	3%	3%	4%	-0.00119	-0.00234
70 to 74	3%	3%	3%	3%	-0.00147	-0.00265
75 to 79	2%	2%	3%	3%	-0.00173	-0.00433
80 to 84	1%	2%	2%	2%	-0.00202	-0.00520
85+	1%	1%	2%	2%	-0.00222	-0.00603
Total	100%	100%	100%	100%		
				Average Difference	0.00000%	0.00000%