

Research on regression-based county population estimates for Colorado
(work paper)

Eddie Hunsinger, Colorado Department of Local Affairs, September 2010

Readers and reviewers: I'll start with a brief description of the background and plan, then describe the method used, and data and results. At the end I'll give a link to the related, simple R script (with linked data), and the sources. Skip as you like, of course—some review or thoughts are better than no review or thoughts—you can't go wrong. It looks a little formal at first glance, but it's not meant to (forgive any poor or mixed grammar). If more information would be useful, don't hesitate to ask.

Especially useful thoughts would be on ideas to change or append the model (including variables and relationships) to improve the fit or logic, and any opinions on problems with the model. Please do point out mistakes if you see them.

1. Background and plan

The Colorado Department of Local Affairs, State Demography Office (SDO), used to make alternative, regression-based total population estimates for Colorado counties, and would like to start offering a form of them again for a couple of reasons: (1) They would offer something to compare the annual estimates from the current methodology to, and (2) If they fit very well, they could be used exclusively, and the SDO estimates would be less dependent on the US Census Bureau's administrative-record-based migration numbers.

The plan is to make a regression-based estimate for 2010, then compare that to the 2010 Census data when it comes out, and have some rough idea of what level of error should be expected for future, annual estimates.

The regression-based estimates presented here rely on the very simple and well-reviewed and -documented "ratio-correlation" techniques that are described similarly in several publications (including Shyrock and Segal, 1980, and Feeney, Hibbs and Gillaspay, 1995), and described below:

2. Ratio-correlation method

The ratio-correlation method is the most widely used regression-based method for county population estimates. It is based on the simple assumption that changes in the county shares of state record-counts of "symptomatic" data, such as birth certificates and voter registrations, will correlate with change in the county shares of state population (the total state population is estimated independently). It's been in use since the 1950's and is currently used in some form by many state demography offices (California, North Carolina, Texas, Virginia, among others) for official county population estimates. It may be described with the following formula:

$$PR_{i,t} / PR_{i,t-k} = b_1 * (XR_{i,t} / XR_{i,t-k}) + b_2 * (YR_{i,t} / YR_{i,t-k}) + a$$

Where " $PR_{i,t-k}$ " is the ratio of county population ("i") to the state population at a specified time (" $t-k$ "), " $XR_{i,t-k}$ " and " $YR_{i,t-k}$ " are county ratios of specified variables that are supposed to correlate with population, to the state total for that variable, at a specified time. " b_1 ", " b_2 " and " a " are

coefficients for a multiple linear regression model that are estimated with data from the last two census years. (Note: Some users of the model leave “a” in, while others drop it.) Independent variables that may be used for the ratio-correlation method include:

- Birth certificates
- Death certificates
- School enrollment
- Voter registrations
- Vehicle registrations
- Driving licenses
- Occupied housing units
- Employment data
- Tax records (income or sales)

To account for certain types of counties in the model, which may respond differently to the specified correlations, users can also use “dummy variables” (having a value of either 1 or 0) that indicate something about the county’s population (such as rural, or largely-imprisoned) to adjust the zero-intercept for these, and improve the overall fit (Pursell, 1970).

Additionally, it seems that users could set up an interaction for an area with a specified independent variable. For instance, if prisons are thought to be an important indicator of population change for certain counties, a prison population variable could be added to the model, and multiplied by a dummy variable that indicates whether the county was significantly affected by change in the prison population.

Stratification (multiple models) is an option if the number of counties (sample size) is large enough (Rosenberg, 1968). It should be noted that use of either stratification or dummy variables has not been shown to consistently improve ratio-correlation estimates (O’Hare, 1980).

Problems with ratio-correlation estimates include (1) Timing: Model coefficients based on censuses that are 10 years apart can’t clearly account for annual lags in the model correlations; also, data from the census years have an April 1 reference date, while data for the estimate years have a July 1 reference date, (2) Temporal instability: The modeled correlations will change to some degree over time, and this change will weaken the model, and (3) No clear interpretation and risk of multicollinearity: Rather than careful formulation and testing of a clear hypothesis, the independent variables are selected based only on some broadly-assumed relationships, and whether or not they improve the overall fit of the model (usually measured by the coefficient of determination, R^2). (O’Hare, 1980.)

Based on review of estimate errors through comparison to censuses, it seems that county-level ratio-correlation estimates for 10 years after the estimate base year (last census) have a Mean Absolute Percent Error (MAPE) of approximately 5. Below are examples of ratio-correlation estimate error analyses (each for 10 years after the estimate base year) that have been conducted by different states:

- Florida 1980 ratio-correlation population estimates error:
Variables: Birth certificates, school enrollment and occupied housing units
MAPE: 5.4
(Smith and Mandell, 1984)

-Texas 1990 ratio-correlation population estimates:

Variables: Birth certificates, death certificates, elementary school enrollment, vehicle registrations and voter registrations

MAPE: 4.8

(Hoque and Murdock, 1999)

-Arizona 2000 ratio-correlation population estimates:

Variables: School enrollment, federal tax returns and driving licenses

MAPE: 5.5

(Brown, 2003)

-Texas 2000 ratio-correlation population estimates:

Variables: Birth certificates, death certificates, elementary school enrollment, vehicle registrations and voter registrations

MAPE: 5.7

(Hoque, 2008)

3. Preparing 2010 ratio-correlation estimates for Colorado counties

The steps in making 2010 ratio-correlation estimates for Colorado counties are (1) Select and create the dependent and independent variables from 1990 and 2000 for the ratio-correlation model, (2) Estimate model coefficients using multiple-regression methods or statistical software, (3) Apply the model and coefficients to 2000 and 2010 data.

The official April 1, 1990 Census and April 1, 2000 Census household population counts for Colorado counties are used to create the dependent variables (population shares of state total) for a ratio-correlation model. These data don't include Broomfield County, which was created in 2001. Only the household (non-group quarters) population is modeled for estimation because the group quarters population can significantly affect the ratio-correlation model for certain counties, and much of the group quarters population (such as prisons and university dorms) can be tracked directly.

For independent variables in the ratio-correlation model, the following data sources are available for consideration (single variable 1990-2000 ratio-correlation model R^2 in parentheses):

-Birth Certificate counts for the fiscal year ending on July 1 of the estimate year (.68)

-Death Certificate counts for the fiscal year ending on July 1 of the estimate year (.29)

-Housing Units on April 1 of the census year, or July 1 of the estimate year (.71)

-QCEW Employment data for the first two quarters of the estimate year (.13)

-School Enrollment for fall of the estimate year (.79)

-Vehicle Registration counts for the previous calendar year (.90)

-Voter Registrations on November 1 of the estimate year (.74)

Some of the dummy variables that are considered are:

-Small (<50,000 people in 2000)

-Denver Metro (Region 3 counties)

-Tourism (population is significantly affected by tourism and resort communities)

-Prison (population is significantly affected by prisons)

After review of the model fit from various combinations of independent variables in the multiple regression formula, the following variables are selected:

- Birth certificates
- School enrollment
- Vehicle registrations
- Voter registrations

Neither Death Certificates, QCEW Employment data, nor any of the listed dummy variables are found to meaningfully improve the model fit beyond the use of those four selected variables. Housing Unit counts are not included, even though they do marginally improve the model fit, because instability in residential construction (across time and space) seems so great.

For a 1990-2000 ratio-correlation model (1990-2000 model) with respective census data on the household population, those selected variables give the following coefficients and error range (from the "lm" function in the R statistical software package):

	Estimate	Standard Error
Intercept	-0.06706	0.02835
Birth Certificates	0.15033	0.02729
School Enrollment	0.31686	0.04984
Vehicle Registrations	0.39634	0.05951
Voter Registrations	0.22403	0.03798

Residuals:

Minimum	First Quartile	Median	Third Quartile	Maximum
-0.1339356	-0.0323842	0.0006768	0.0307365	0.1049861

Multiple R-squared: 0.9665, Adjusted R-squared: 0.9642

Comparison of the 1990-2000 model predictions for 2000 to the 2000 Census data gives a MAPE of 3.90.

At the end of this document are graphs describing the 1990-2000 model error:

-Figure 1 is a histogram of the 1990-2000 model's residuals

-Figure 2 is a histogram of the residuals (percent) of population estimates for 2000, based on the 1990-2000 model

-Figure 3 is a histogram of the residuals of population estimates for 2000, based on the 1990-2000 model

-Figure 4 is a point-plot to compare the estimates for 2000 from the 1990-2000 model to the respective data from the 2000 Census

-Figure 5 is a point-plot to compare the estimates for 2000 from the 1990-2000 model, to the respective data from the 2000 Census, for areas with less than 50,000 people in 2000

The errors in the 2010 estimates won't be known until the 2010 Census data is released, of course, but should be as large as those for the above-described 2000 estimates, plus any effect of temporal instability in the model, and error in the state total population estimate.

In reviewing the selected model's errors for 2000, no clear biases by type of county (e.g. high tourism counties generally underestimated, Front Range generally overestimated, etc.) are discerned. Finding any of these, or recognizing any interactions of county-types with the independent variables, would be an ideal way to improve the model fit.

Using the 1990-2000 model, the next step will be to make population estimates for 2010. Because the independent variables for the 2010 estimates aren't yet available (should all be available by January of 2011), it's not possible them at this time, but it is possible to prepare and review 2009 estimates based on the 1990-2000 model, and compare them to the official Colorado State Demography Office county population estimates for 2009 (SDO estimates). The comparison can't give information on the accuracy of the selected model, but can provide description of the differences in shares of total population from the different estimate models.

The dependent variable for the 2009 estimates are the ratios of county populations to the state total population for 2009, divided by those for 2000, with any adjustments to 2000 for geography changes, such as the addition of Broomfield County. The independent variables are the ratios of county symptomatic data to the state total for 2009, divided by those for 2000, with any adjustments to 2000 for geography changes. The 1990-2000 model coefficients are not changed.

Comparison of the 1990-2000 model prediction for 2009 to the SDO estimates for 2009 gives a MAPE of 4.87. (Note: This MAPE includes an anomaly in 2007-2009 school district data for Sedgwick County. I'm trying to get in contact with Sedgwick County to make sense of it. With Sedgwick County removed, the MAPE is 3.84.)

At the end of this document are graphs describing the differences between the SDO estimates for 2009, and ratio-correlation 1990-2000 model estimates for 2009:

- Figure 6 is a histogram of the differences (percent) between the SDO estimates for 2009, and the 1990-2000 model estimates for 2009

- Figure 7 is a histogram of the differences between the SDO estimates for 2009, and the 1990-2000 model estimates for 2009

- Figure 8 is a point-plot to compare the SDO estimates for 2009 to the 1990-2000 model estimates for 2009

- Figure 9 is a point-plot to compare the SDO estimates for 2009 to the 1990-2000 model estimates for 2009, for areas with less than 50,000 people in 2000

4. R Code

The R script (with linked data) to make the described ratio-correlation estimates, is available at:

<http://www.demog.berkeley.edu/~eddieh/RatioCorrelationEstimates/RCScript.txt>

Just paste into R to run it. Unused variables are included in the script, and may be added to the model as well.

5. Sources

- W. Brown (2003). "Evaluation of July 1, 2000 County and Municipal Population Estimates by the Arizona Department of Economic Security." Unpublished report for the Arizona Department of Economic Security.
Available online at:
http://www.workforce.az.gov/admin/UploadedPublications/1834_WABEstEvalReport-050205.pdf

- D. Feeney, J. Hibbs, and T. Gillaspay (1995). "Ratio-Correlation Method."
In N. Rives, W. Serow, A. Lee, H. Goldsmith, and P. Voss (eds), *Basic Methods for Preparing Small-Area Population Estimates* (pp 118-136). University of Wisconsin-Madison/Extension.

- N. Hoque (2008). "An Evaluation of Population Estimates for Counties and Places in Texas for 2000"
In S. Murdock and D. Swanson (eds), *Applied Demography in the 21st Century* (pp 125-148). Springer Science and Business Media.

- N. Hoque and S. Murdock, (1999). "Evaluation of Texas population and estimates and projections programs population estimates for 1990." Presented at the Population Estimates Methods Conference, U.S. Census Bureau.

- W. O'Hare (1980). "A Note on the Use of Regression Estimates in Population Estimates."
Demography, 17 (pp 341-343). Johns-Hopkins University Press.

- D. Pursell (1970). Improving Population Estimates with the Use of Dummy Variables."
Demography, 7 (pp 87-91). Johns-Hopkins University Press.

- H. Rosenberg (1968). "Improving Current Population Estimates through Stratification." *Land Economics*, 44 (pp 331-338). University of Wisconsin Press.

- H. Shyrock and J. Segal (1980). *The Methods and Materials of Demography*, Volume 2. U.S. Department of Commerce.

- S. Smith and M. Mandell (1984). "A Comparison of Population Estimates Methods: Housing Unit Versus Component II, Ratio Correlation, and Administrative Records." *Journal of the American Statistical Association*, 79 (386) (pp 282-289). American Statistical Association.

Figure 1: A histogram of the 1990-2000 model's residuals

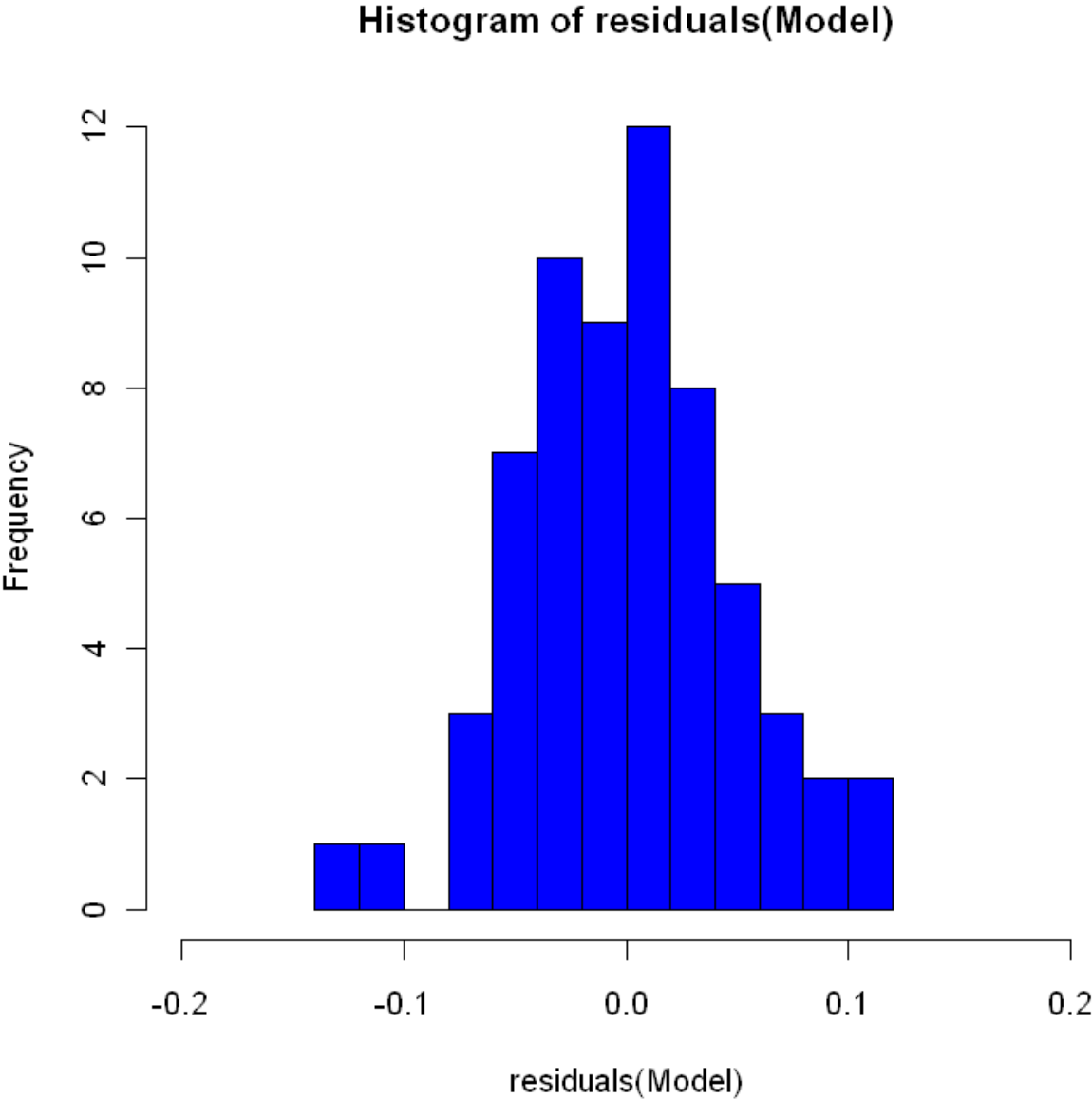


Figure 2: A histogram of the residuals (proportional to 2000 population) of population estimates for 2000, based on the 1990-2000 ratio-correlation model

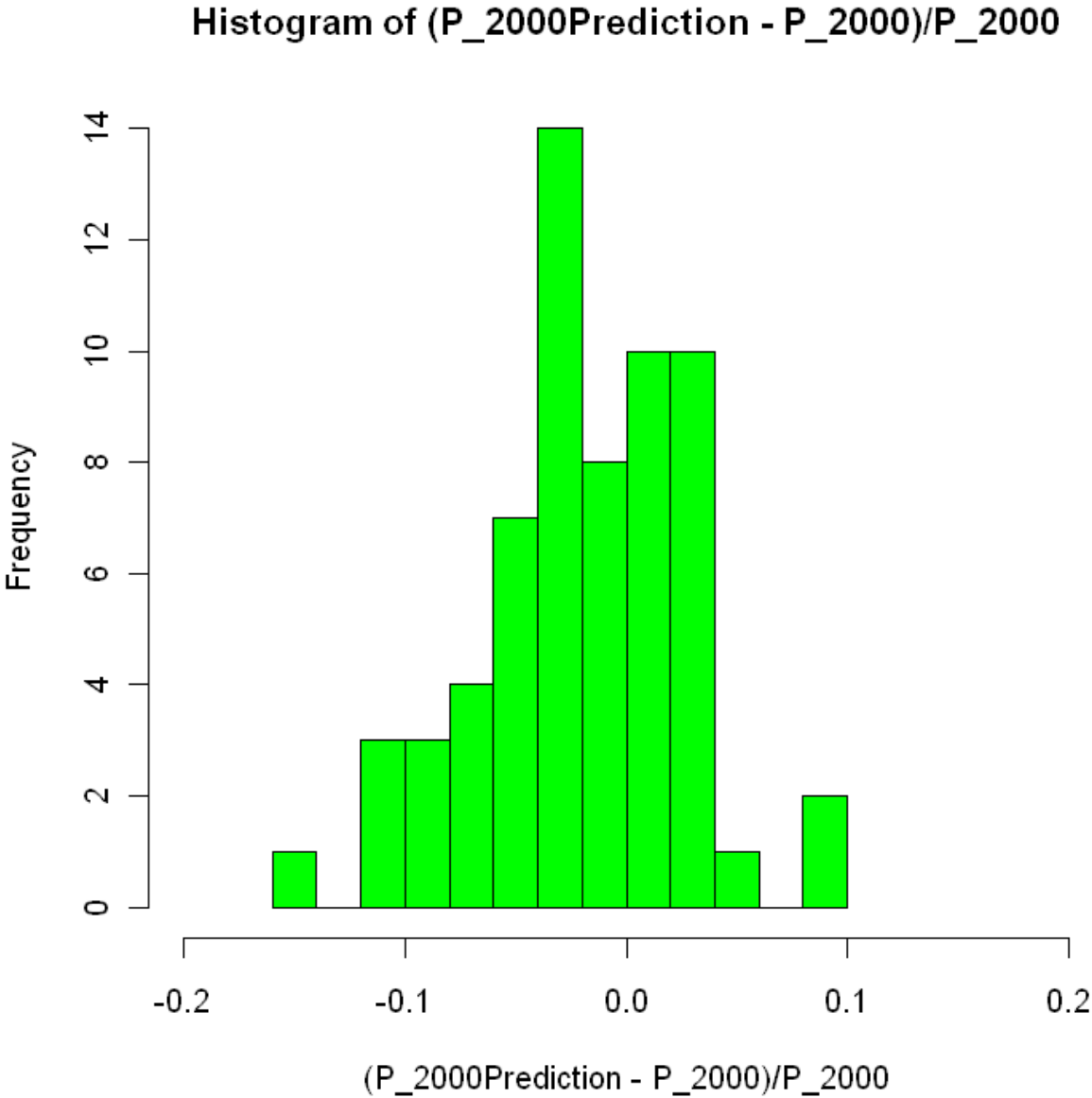


Figure 3: A histogram of the residuals of population estimates for 2000, based on the 1990-2000 model

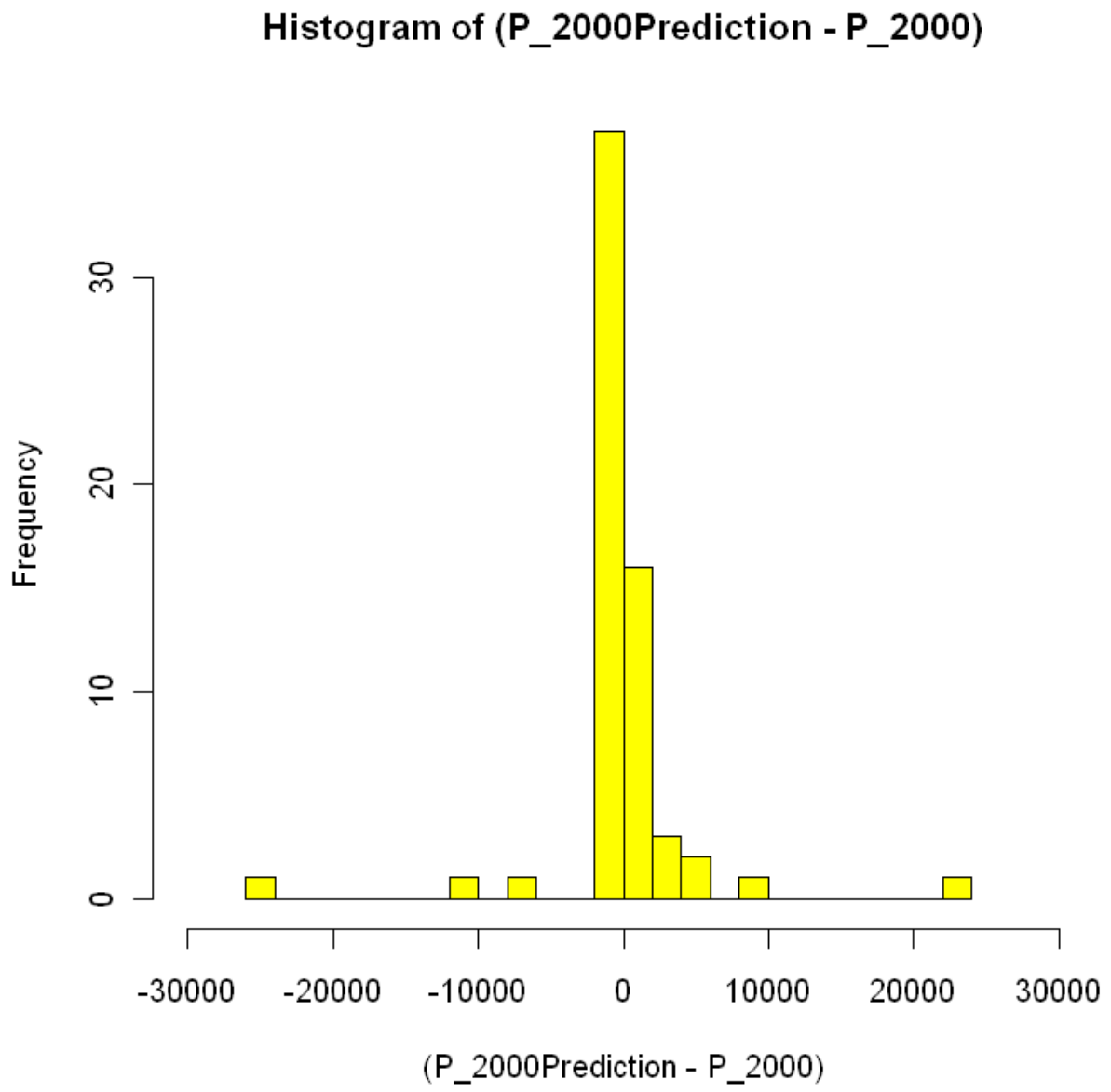


Figure 4: A point-plot to compare the estimates for 2000 from the 1990-2000 model to the respective data from the 2000 Census

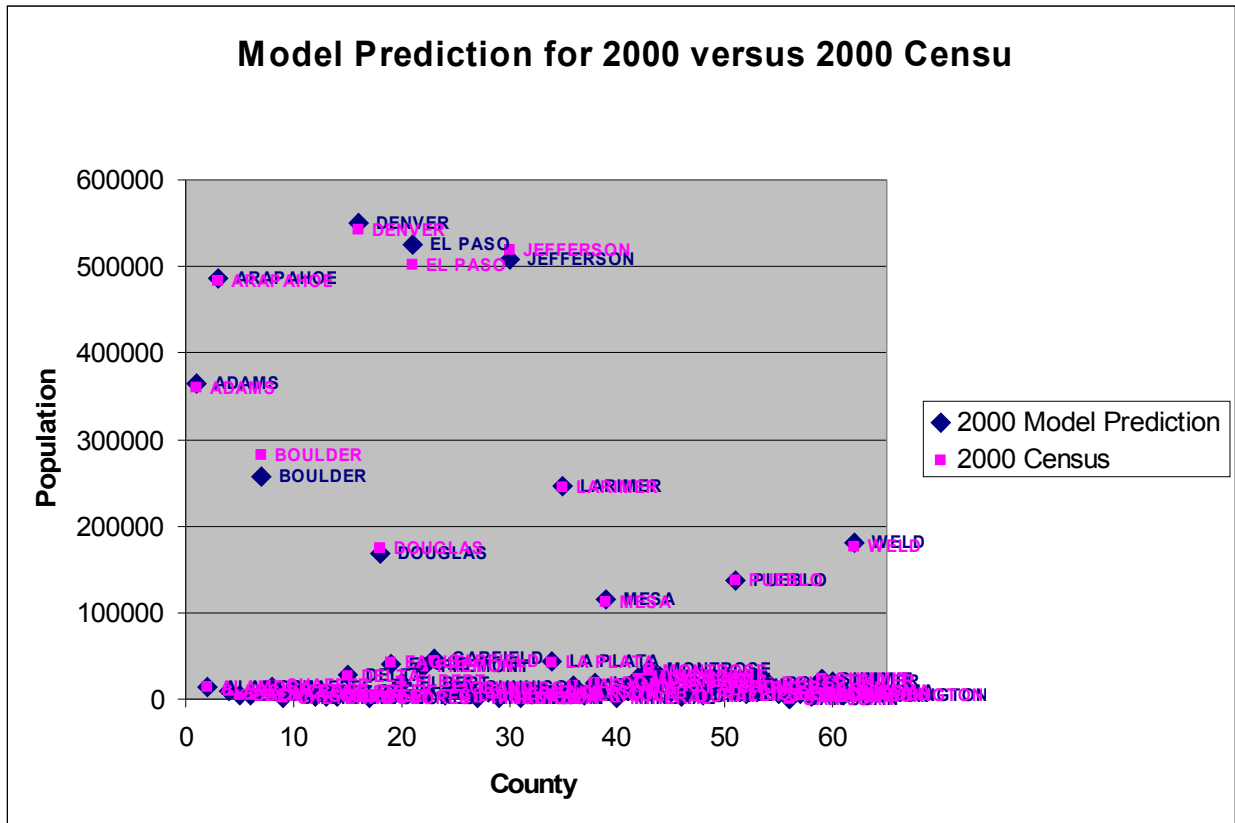


Figure 6: A histogram of the differences (proportional to 2009 population) between the SDO estimates for 2009, and the 1990-2000 model estimates for 2009

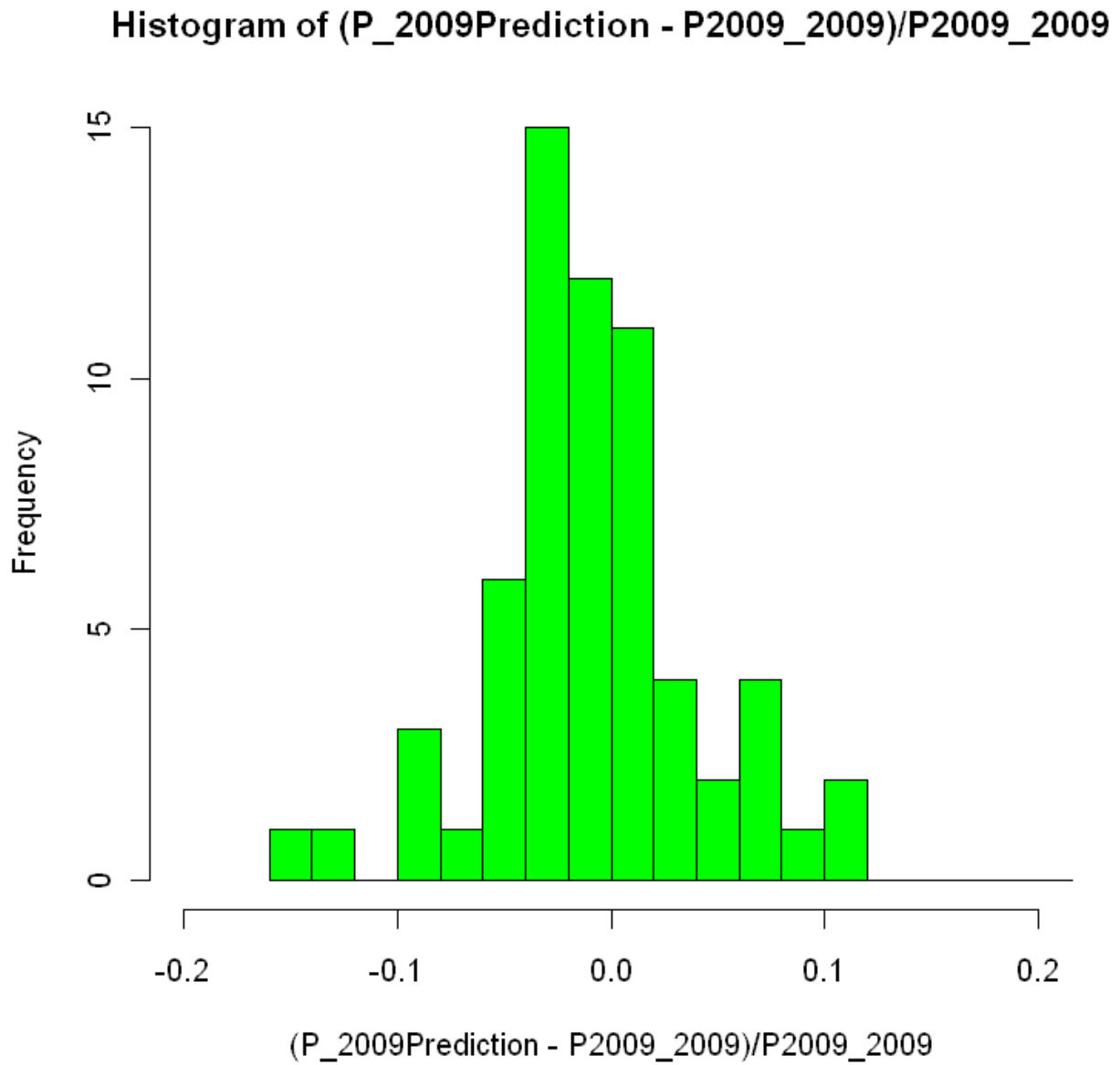
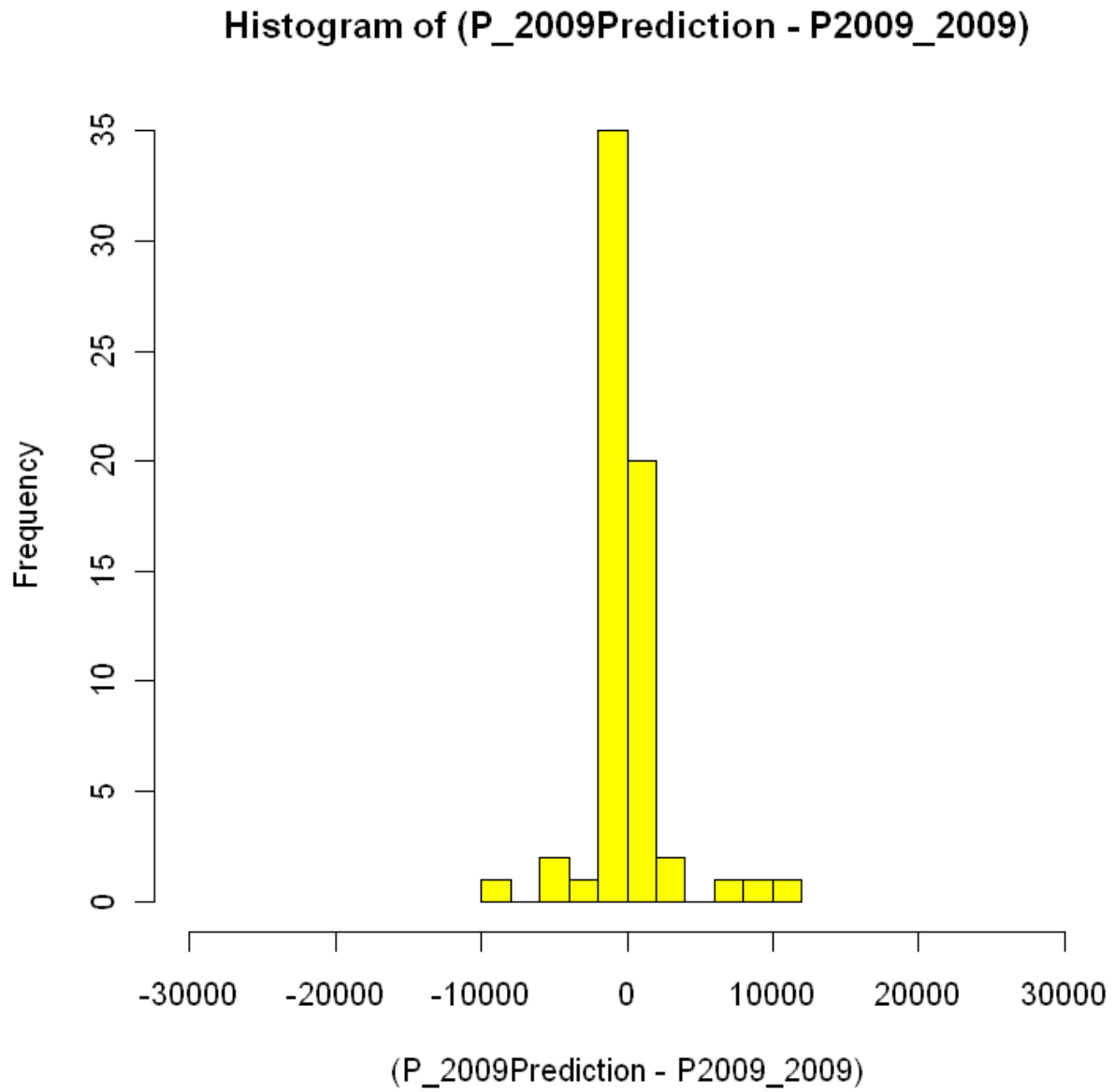


Figure 7: A histogram of the differences between the SDO estimates for 2009, and the 1990-2000 model estimates for 2009



Comparison of Colorado county population estimates for April 1, 2010*

County	Census Estimates	Percent Error	MAPE: 4.46	Regression Estimates	Percent Error	MAPE: 4.97	Naïve Estimates	Percent Error	MAPE: 13.25	2010 Census
ADAMS COUNTY	444,006	0.54		440,088	0.34		406,806	7.88		441,603
ALAMOSA COUNTY	15,221	1.45		15,238	1.34		17,496	13.28		15,445
ARAPAHOE COUNTY	566,538	0.96		565,564	1.13		571,528	0.08		572,003
ARCHULETA COUNTY	12,227	1.18		11,762	2.66		11,569	4.26		12,084
BACA COUNTY	3,633	4.09		4,225	11.54		5,281	39.41		3,788
BENT COUNTY	6,851	5.42		6,615	1.78		7,012	7.89		6,499
BOULDER COUNTY	301,989	2.52		302,847	2.81		315,368	7.06		294,567
BROOMFIELD COUNTY	56,172	0.51		54,705	2.12		45,818	18.02		55,889
CHAFFEE COUNTY	17,060	4.21		17,678	0.74		18,987	6.61		17,809
CHEYENNE COUNTY	1,700	7.41		1,971	7.35		2,608	42.05		1,836
CLEAR CREEK COUNTY	8,448	7.04		8,104	10.83		10,898	19.92		9,088
CONEJOS COUNTY	7,647	7.38		8,229	0.33		9,820	18.94		8,256
COSTILLA COUNTY	3,037	13.82		3,410	3.23		4,282	21.51		3,524
CROWLEY COUNTY	6,254	7.40		6,146	5.55		6,451	10.78		5,823
CUSTER COUNTY	3,992	6.18		3,779	11.19		4,095	3.76		4,255
DELTA COUNTY	31,298	1.12		31,782	2.68		32,539	5.13		30,952
DENVER COUNTY	615,610	2.57		618,813	3.11		647,283	7.85		600,158
DOLORES COUNTY	1,906	7.66		1,943	5.86		2,156	4.46		2,064
DOUGLAS COUNTY	289,926	1.56		290,094	1.62		205,476	28.02		285,465
EAGLE COUNTY	54,203	3.84		51,413	1.50		48,719	6.66		52,197
ELBERT COUNTY	23,259	0.75		22,049	4.49		23,231	0.63		23,086
EL PASO COUNTY	603,900	2.95		620,147	0.34		604,312	2.88		622,263
FREMONT COUNTY	47,328	1.08		46,046	1.66		53,945	15.21		46,824
GARFIELD COUNTY	56,686	0.53		55,438	1.69		51,193	9.21		56,389
GILPIN COUNTY	5,806	6.71		5,168	5.02		5,577	2.50		5,441
GRAND COUNTY	13,840	6.76		13,085	11.84		14,545	2.01		14,843
GUNNISON COUNTY	15,256	0.44		14,966	2.34		16,315	6.47		15,324
HINSDALE COUNTY	813	3.56		933	10.68		924	9.61		843
HUERFANO COUNTY	7,312	8.96		6,487	3.34		9,191	36.95		6,711
JACKSON COUNTY	1,374	1.43		1,438	3.16		1,844	32.28		1,394
JEFFERSON COUNTY	532,997	0.29		535,857	0.25		614,128	14.89		534,543
KIOWA COUNTY	1,198	14.31		1,251	10.52		1,896	35.62		1,398
KIT CARSON COUNTY	8,502	2.81		8,696	5.15		9,365	13.24		8,270
LAKE COUNTY	7,982	9.19		7,517	2.83		9,132	24.92		7,310
LA PLATA COUNTY	51,446	0.22		51,108	0.44		51,378	0.09		51,334
LARIMER COUNTY	298,777	0.28		299,891	0.09		293,995	1.88		299,630
LAS ANIMAS COUNTY	15,911	2.61		16,477	6.26		17,777	14.64		15,507
LINCOLN COUNTY	5,076	7.15		5,916	8.21		7,116	30.16		5,467
LOGAN COUNTY	20,556	9.48		20,257	10.80		24,052	5.91		22,709
MESA COUNTY	147,396	0.46		148,453	1.18		136,701	6.83		146,723
MINERAL COUNTY	873	22.61		949	33.29		971	36.38		712
MOFFAT COUNTY	13,925	0.94		14,095	2.17		15,413	11.73		13,795
MONTEZUMA COUNTY	25,217	1.25		24,705	3.25		27,858	9.10		25,535
MONTROSE COUNTY	41,589	0.76		39,700	3.82		39,083	5.31		41,276
MORGAN COUNTY	27,737	1.50		28,143	0.06		31,764	12.80		28,159
OTERO COUNTY	18,463	1.95		18,113	3.81		23,744	26.09		18,831
OURAY COUNTY	4,556	2.71		4,589	3.45		4,375	1.38		4,436
PARK COUNTY	16,382	1.09		14,518	10.42		16,978	4.76		16,206
PHILLIPS COUNTY	4,435	0.16		4,571	2.90		5,237	17.90		4,442
PITKIN COUNTY	16,162	5.75		17,334	1.08		17,386	1.39		17,148
PROWERS COUNTY	12,817	2.12		12,909	2.85		16,931	34.90		12,551
PUEBLO COUNTY	156,188	1.81		158,053	0.63		165,385	3.97		159,063
RIO BLANCO COUNTY	6,614	0.78		6,928	3.93		6,998	4.98		6,666
RIO GRANDE COUNTY	11,401	4.85		12,306	2.70		14,511	21.11		11,982
ROUTT COUNTY	23,553	0.19		22,574	3.98		23,016	2.10		23,509
SAGUACHE COUNTY	7,068	15.72		5,816	4.78		6,917	13.24		6,108
SAN JUAN COUNTY	536	23.32		559	20.03		652	6.72		699
SAN MIGUEL COUNTY	7,488	1.75		7,266	1.26		7,709	4.76		7,359
SEDGWICK COUNTY	2,281	4.12		2,628	10.47		3,211	34.97		2,379
SUMMIT COUNTY	27,285	2.53		25,695	8.21		27,528	1.66		27,994
TELLER COUNTY	21,469	8.06		21,016	10.00		24,029	2.91		23,350
WASHINGTON COUNTY	4,327	10.12		5,174	7.48		5,759	19.63		4,814
WELD COUNTY	255,998	1.26		245,261	2.99		211,428	16.37		252,825
YUMA COUNTY	9,699	3.43		10,708	6.62		11,504	14.55		10,043

*Notes:

1. This is a comparison of allocation, so all are controlled to the 2010 Census statewide number for Colorado (5,029,196). All comparisons (error calculations) are to the actual 2010 Census counts.
2. The Census Estimates were made by taking the July 1, 2008 and July 1, 2009 (both vintage 2009) Census Bureau estimates, then simply extrapolating the change between them to April 1, 2010 (difference times .75, plus July 1, 2009), and proportionally controlling to the April 1, 2010 Census count for Colorado.
3. The Regression Estimates were made with a ratio-correlation model that has Births, Public School Enrollment, Vehicle Registrations and Voter Registrations as the independent variables, and Household Population as the dependent variable. The Group Quarters population estimate from July 1, 2009 was added to the Household Population. They were proportionally controlled to the April 1, 2010 Census count for Colorado.
4. The Naive estimates were made by simply taking the year 2000 (Census Estimates base) county proportions (of state) and multiplying them by the 2010 Census state number-- simply a benchmark for whether we know more than nothing.
5. 2010 Census is the actual counts for the counties from the 2010 Census.
6. MAPE is the Mean Absolute Percent Error.