Is Anybody Home?
Small Area Estimates of Housing Vacancy Rates

By
Warren A. Brown
Applied Demography Program
Carl Vinson Institute of Government
University of Georgia
Email: brown@cviog.uga.edu

Presented at the 2011
Meeting of the Population Association of America
Washington, DC

This paper reports on preliminary findings regarding the use of administrative records from the US Postal Service (USPS) to estimate the number of housing units and vacancy rates in counties of the United States. The larger goal of the research effort is to develop improved estimates of some of the components used in the Housing Unit Method for estimating resident household population, specifically county level estimates of housing units and occupancy rates. The Census Bureau population estimates program produces annual estimates of housing units using local government records on building permits and their own estimates of mobile home placements and loss of housing units. We are using the recently released results of the 2010 Decennial Census counts of housing units for counties in Louisiana, Mississippi, New Jersey, and Virginia to compare with the USPS data. The USPS data are the Aggregated USPS Administrative Data on Address Vacancies files distributed by the U.S. Department of Housing and Urban Development. These data are available nationally at the census tract level of geography.

The Housing Unit Method (HUM) for estimating residential household population is based on a fundamental accounting identity in demography.

$$\mathbf{HHPOP = HU * OCC * PPH}$$

Where:

HHPOP – Household population

HU – Housing Units

OCC – Occupancy rate for housing units (complement of the vacancy rate)

PPH – Persons per household, also referred to as average household size

The accuracy of estimates derived by the HUM is a function of the net accuracy of estimates of the three components: number of housing units; occupancy rate; and persons per household. In some cases an improvement in the accuracy of one of the components produces greater overall error. For example if the number of housing units is underestimated and the occupancy rate is over estimated then the errors may result in an estimate of occupied housing units/households which is more accurate than if the estimated number of housing were more accurate and the occupancy rate were still overestimated. Hence the importance of specifying that ultimately an evaluation of improvements in estimating components used in the Housing Unit Methods must be based on net accuracy.

Administrative Records as a Source of Statistical Data

Three main sources of statistical data for demographic analysis are censuses, sample surveys and administrative records. Statistical agencies in Scandinavian countries have a history of using administrative data gathered from their population registers and similar sources for estimating detailed

characteristics of their populations (Wallgren and Wallgren, 2007:1). The U.S. Census Bureau has also identified the need to increase their use of administrative records as a source of statistical data in their efforts to better serve their customers. The rationale was that the use of administrative records could lead to reduced data collection costs, increased data quality, and reduced respondent burden. (Prevost and Leggieri, 1999:2).

Survey organizations have made use of the USPS Delivery Sequence File (DSF) as a frame for their sample surveys. The U.S. Census Bureau uses the USPS DSF in the development of their Master Address File (MAF) for conducting the decennial censuses and the American Community Survey. Evaluations of the completeness of the DSF as a frame for housing units demonstrates that in general the DSF performs well in urban and suburban areas but is deficient in rural areas due to the large proportion of non-city style addresses such as rural delivery routes and P.O. boxes instead of home delivery. The Washington State Office of Financial Management has conducted evaluations of the USPS address files that are the most relevant to our objectives. They too were interested in using data on the address files to improve inputs to their housing unit method for estimating population. In a couple of research briefs (Office of Financial Management Population Section, 2003a; 2003b) they point out the differences in concepts and measurements for housing units and vacancy status between the USPS administrative series and the Census Bureau estimates based on decennial census counts.

The USPS administrative series is based on the USPS Delivery Sequence File (DSF) which is developed and maintained with the objective of more efficient mail processing and delivery. For the Census estimates, a housing unit is a physical structure intended for residential occupancy by a single household; and for the USPS DSF file the equivalent of a housing unit is a residential postal delivery point that is an active or potential delivery stop where people reside. There are differences between the Census estimates and USPS DSF with respect to coverage, adding and deleting units, and units for seasonal or occasional use. The OFM compared state and county totals for Washington between the 2000 Census and the USPS DSF for 2000. In general, including all residential address delivery points in the DSF results in over-estimation of housing units, and removing P.O. Boxes from the tally results in under-estimation. However the “absolute errors” were less for metropolitan counties when P.O. Boxes were removed and for nonmetropolitan counties when P.O. Boxes were included. The USPS DSF tally of “vacant” units were substantially lower than census counts, especially in areas with housing units held for seasonal and occasional use. Persons can actively receive mail at more than one address, while the census count of occupied units is based on a single housing unit as the “usual place of residence.” The correspondence of postal and census vacancy rates were closer in metropolitan areas.

In a finding that was encouraging regarding the use of postal vacancy rates for updating census rates, OFM observed that the direction of changes in the postal vacancy rates between 2000 and 2002 for counties were consistent with what was expected due to population growth and economic conditions. One of the major benefits from the OFM research is that the classification of geographic areas into categories such as metropolitan or non-metropolitan, and seasonal or year-round, may help to improve the quality of post-censal estimates of housing units and vacancy derived from the USPS DSF files.

Data Series Evaluated

The data we used in our evaluation are the HUD Aggregated USPS Administrative Data on Address Vacancies. They are derived from a second generation version of the USPS DSF that the USPS refers to as DSF2 (the superscript "2" referring to second generation). The HUD Aggregated USPS Administrative Data on Address Vacancies files are publicly available on HUD's HUDUSER website. The files provide key variables from DSF that are aggregated by census tract on a quarterly basis since the quarter for October 1 – December 31, 2005. The lag time is very short. The data file used in this analysis, covering the second quarter April 1 – June 30, 2010 was available by October 1, 2010. HUD and the USPS have been enhancing the utility of these data in ways that benefit our research objectives. Beginning with June 30, 2008 the files have fields distinguishing address counts between residential, business and other addresses. The fields of primary interest are: Total Count of Addresses –Residential; and Total Count of Vacant Addresses –Residential. Unfortunately these data from HUD are not available for April 1, 2000. We then could have fit the USPS housing unit and vacancy data to the 2000 Census results and developed a model to estimate annual housing units and vacancy rates and test it against the 2010 Census.

For this preliminary analysis we focused on county summaries of the total count of residential addresses for Louisiana, Mississippi, New Jersey and Virginia. These were the first four states for which the Census Bureau published 2010 Census redistricting data. We compared the counts of residential addresses and occupancy rates from the HUD-USPS files with the 2010 Census redistricting data counts of total and occupied housing units. There are 301 counties and county equivalents in the four states. The range in number of housing units, based on the 2010 Census Bureau, from Fairfax County, Virginia with 407,998 housing units to Highland County, Virginia with 1,837 units. The data for the 2nd Quarter, 2010 (April 1 – June 30, 2010) from the HUD-USPS files were summarized by county. This quarter most closely matches the collection period covered in the 2010 Census.

Total Residential Housing Units

The 2010 Census data are treated as the standard and the differences between the census counts and the postal addresses are treated as errors. The HUD-USPS count of residential addresses were subtracted from the 2010 Census count of housing units and then divided by the census count to express relative error. A percentage difference greater than zero means that the USPS residential addresses are less than the Census count of housing units, and a negative percentage difference means the USPS residential addresses are greater than the Census count. Issaquena County, Mississippi has been omitted from analyses because while the 2010 Census reports there were 560 housing units, the HUD-USPS Address file reported there were no residential addresses. This discrepancy merits further investigation and while it is an extreme case it may yield clues as to important differences in the way that the USPS address files are created and what they actually contain. For the 300 counties in the four states, the Mean Algebraic Percent Error (MALPE), a measure of bias in the estimates, was 4.0% indicating a negative bias in which the HUD-USPS data were lower on average than the Census counts. The Mean Absolute Percent Error (MAPE), a measure of the variability in the estimates, was 10.7%. The larger MAPE indicates that the lower overall net bias was due to offsetting errors of under and over estimation.

Based on the findings in previous research, we expect to find the greatest errors among the more rural counties and the least error among the more urban and suburban counties. The measures of bias and variability for all counties, counties categorized by metropolitan status, and counties categorized by number of housing units are presented in Table 1. Metropolitan status is based on OMB's current Core Based Statistical Areas. The HUD-USPS address counts for Metropolitan counties had less error in terms of bias and variability than for Non-Metropolitan counties. For Metropolitan counties the MALPE was 0.4% and the MAPE was 7.5%. While the net error was small and there was little bias in the HUD-USPS Address counts, the level of gross error indicated by MAPE shows that there were still substantial errors in the positive and negative directions, though less than for all counties. For the Non-Metropolitan counties the number of HUD-USPS residential addresses, on average, were 7.4% lower than the 2010 Census counts of housing units. The variability in error as indicated by the MAPE at 13.7% was almost double that for Metro counties.

For counties categorized by number of housing units reported in the 2010 Census counts, the larger counties tended to have greater positive bias and the smaller counties more negative bias. That is, the address counts tended to over-estimate the number of housing units in the larger counties and

under-estimate in the smaller counties. There was also a tendency toward greater variability in the estimates based on address counts for the smaller counties.

In order for the HUD-USPS files of address counts to be useful in estimating the number of housing units it is necessary to find whether there is a consistent pattern to the errors that can be controlled and accounted for in a statistical model. The HUD-USPS address counts and 2010 Census counts for 300 counties are plotted in an x-y scattergram in Figure 1, showing the differences between the data. If there was a 1-to-1 relationship between address counts and housing units from the two sources, then all the counties would fall along the black diagonal line. The data from both files have been transformed to their natural log value and then plotted. As size increases there is a tendency for counties to fall near the diagonal. The greatest variability is evident for the smaller counties. The counties have also been coded by Metropolitan status. Size and metropolitan status appear to be important in explaining the degree of variability.

Occupancy Rates

In addition to correctly estimating the number of housing units, the Housing Unit Method for estimating household population requires estimates of the occupancy rate in order to estimate the number of households. The Census Bureau has been using vacancy rates from the prior decennial census as the rates to employ. The housing bubble and drastic swings in the housing market during the last decade have shown that occupancy rates can vary dramatically over time. We examine the HUD-USPS Address files to determine whether the data on occupied and vacant residential units can be used to estimate current occupancy rates. Addresses that letter carriers have identified as being vacant (not collecting their mail) for 90 days or longer are classified as vacant. The USPS has been in the process of improving the quality of the data on vacant units in their address database. The improvements began in March 2010 and so the quarterly file we are using has been affected by the early stages of these changes. The USPS notes that the reporting changes are having the effect of increasing the number of vacant units, which is lowering the occupancy rate. For the 2010 2nd Quarter (April – June) USPS Address file the lowest occupancy rate was 85.8% and the highest was 100.0%

The 2010 Census reports that among the 300 counties we are analyzing, the lowest occupancy rate was 41.5% for Cape May County, New Jersey. The USPS address file reported an occupancy rate of 93.2% for the same county. Cape May County is at the southern tip of New Jersey, a seaside resort area with many seasonal homes. Another major anomaly between the census and USPS is Orleans Parish in Louisiana. Orleans Parish is coterminous with the city of New Orleans and was ravished by Hurricane

Katrina. The 2010 Census reports an occupancy rate of 74.9% while the USPS Address file reports an occupancy rate of 96.1%. Differences in the definition of "occupied" between the USPS Address file and the 2010 Census are yielding vastly different data. In only one of the 300 counties we looked at did the 2010 Census report a higher occupancy rate than the USPS Address file and it was only slightly greater. For the 2010 Census if a housing unit was not occupied as a person's sole "usual place of residence" then it was classified as vacant. For the USPS Address file, if mail were being received by the person whose address was associated with the housing unit, then the unit was occupied. That is, occupied even if the mail were being forwarded to another address. In order to be classified as vacant a housing unit would have to not have mail received for 90 days. Mail that is forwarded is considered as having been received.

In its present form it appears that the USPS Address file is not useful for estimating the number of housing units that conform to a decennial census definition of vacancy status. Figure 2 is a graph of the proportion of a state's counties by quintile of difference between the USPS Address file occupancy rate and that of the 2010 Census. The greatest differences are found in Louisiana where about one-third of the parishes fall in the lowest quintile where the percentage differences between the USPS and 2010 Census data fall between -52% and -15%. The proportion Virginia's counties and independent cities in this lowest quintile were almost as great. Two of New Jersey's 21 counties were in this lowest quintile. They were Cape May and Ocean Counties, both seaside resort areas with high proportions of seasonal homes.

Next Steps

Are the HUD Aggregated USPS Address files useful administrative records for estimating the number of housing units and their occupancy rate? Yes and no, at least not in their present form. Yes, that the count of residential addresses in the USPS Address files that we examined did correspond to the number of housing units reported in the 2010 Census. Once the 2010 Census data are available for all counties we will proceed to carry out an investigation with all the data. Given the data for the first four states it does appear that for more populous and metropolitan counties that there is a fairly close correspondence. For more rural and non-metropolitan counties there is far less correspondence between the USPS Address data and the 2010 Census.

Regarding the classification of housing units as vacant or occupied, the USPS Address files in their present form are not useful. My suspicion is that counties which have been hit by disasters forcing

residents to relocate and recreation/resort areas with a great number of seasonal homes are two types of counties with major discrepancies. A series of next steps are in order to see if it is possible to use additional fields of information in the USPS Address file to derive a classification of addresses that more nearly conforms to the decennial census definition of occupancy status. The HUD statement regarding changes in the manner in which letter carriers enter status updates which are yielding higher vacancy rates needs to be understood in detail. Also we need to access a list of all the data fields and codes in the USPS Delivery Sequence File (DSF) to determine if a field exists denoting that mail is being forwarded and not delivered to the house.

References

Office of Financial Management Population Section, State of Washington

2003a Developing Postal Delivery Data for Use in Population Estimates. Research Brief No. 17. January 2003.

_____

2003b Use of Postal Delivery Data in the Population Estimate Process. Research Brief No. 18. August 2003.

Prevost, Ron and Charlene Leggieri

1999 Expansion of Administrative Records Uses At The Census Bureau: A Long-Range Research Plan. Presented at the November 1999 Meeting of the Federal Committee on Statistical Methodology. Washington D.C.

Wallgren, Anders and Britt Wallgren

2007 *Register-based Statistics: Administrative data for Statistical Purposes* Anders Wallgren, Britt Wallgren. Wiley. West Sussex, England

Table 1. Measures of Bias and Reliability in Housing Units Difference
2010 Census vs HUD-USPS

	# of Counties	MALPE	MAPE
All Counties of LA, MS, NJ, VA	300	4.0%	10.7%
Metro	146	0.4%	7.5%
Non-Metro	154	7.4%	13.7%
Size - # of Housing Units			
100,000 and more	27	-1.9%	2.8%
50,000 - 99,999	20	-2.8%	5.2%
25,000 - 49,999	38	0.5%	4.2%
10,000 - 24,999	93	1.2%	10.1%
5,000 - 9,999	82	10.3%	15.7%
Less than 5,000	40	8.2%	16.0%

* Issaquena County, Mississippi omitted from analysis

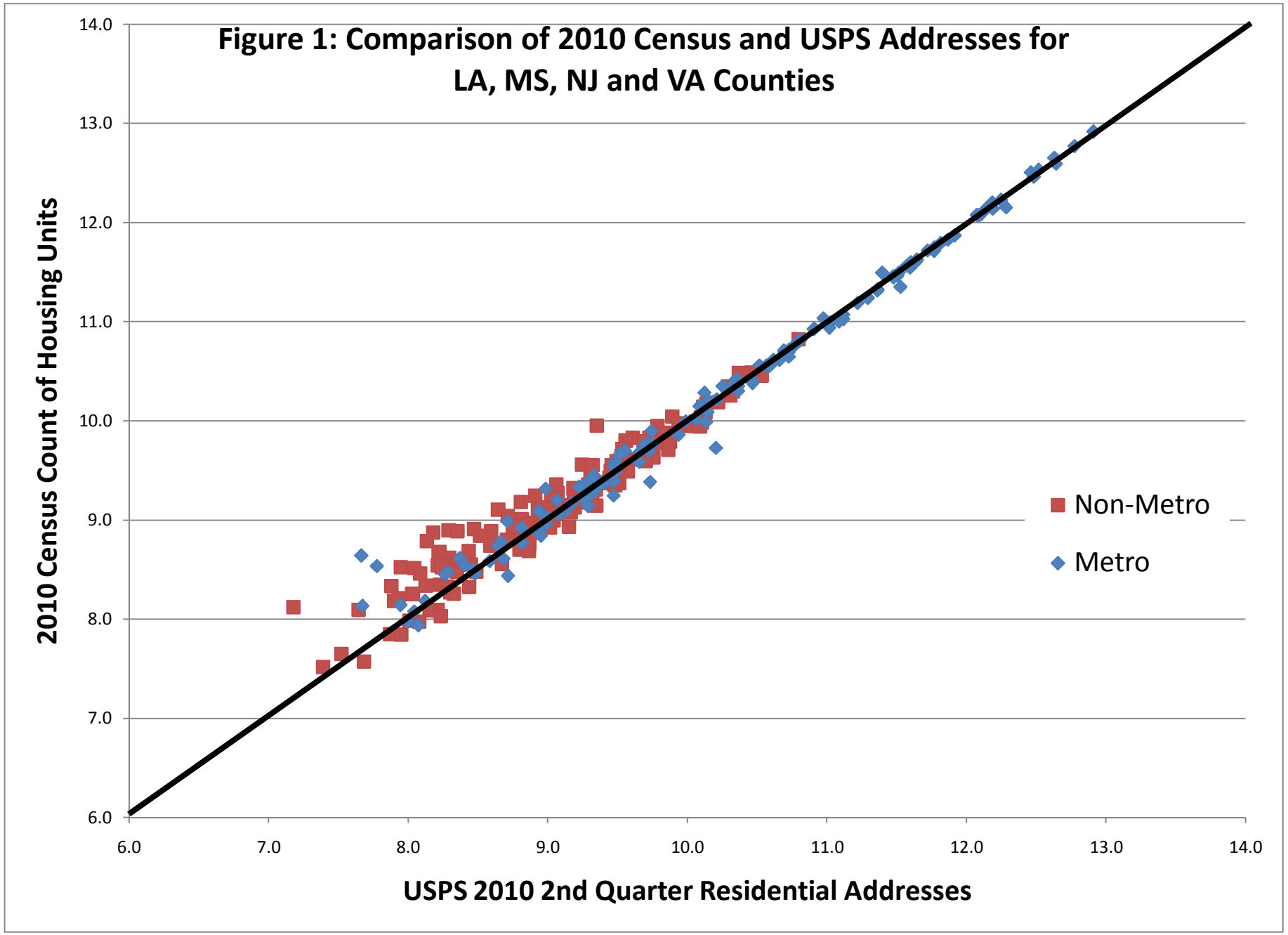
Figure 1: Comparison of 2010 Census and USPS Addresses for LA, MS, NJ and VA Counties
2010 Census Count of Housing Units
USPS 2010 2nd Quarter Residential Addresses
Non-Metro
Metro
14.0
13.0
12.0
11.0
10.0
9.0
8.0
7.0
6.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
14.0

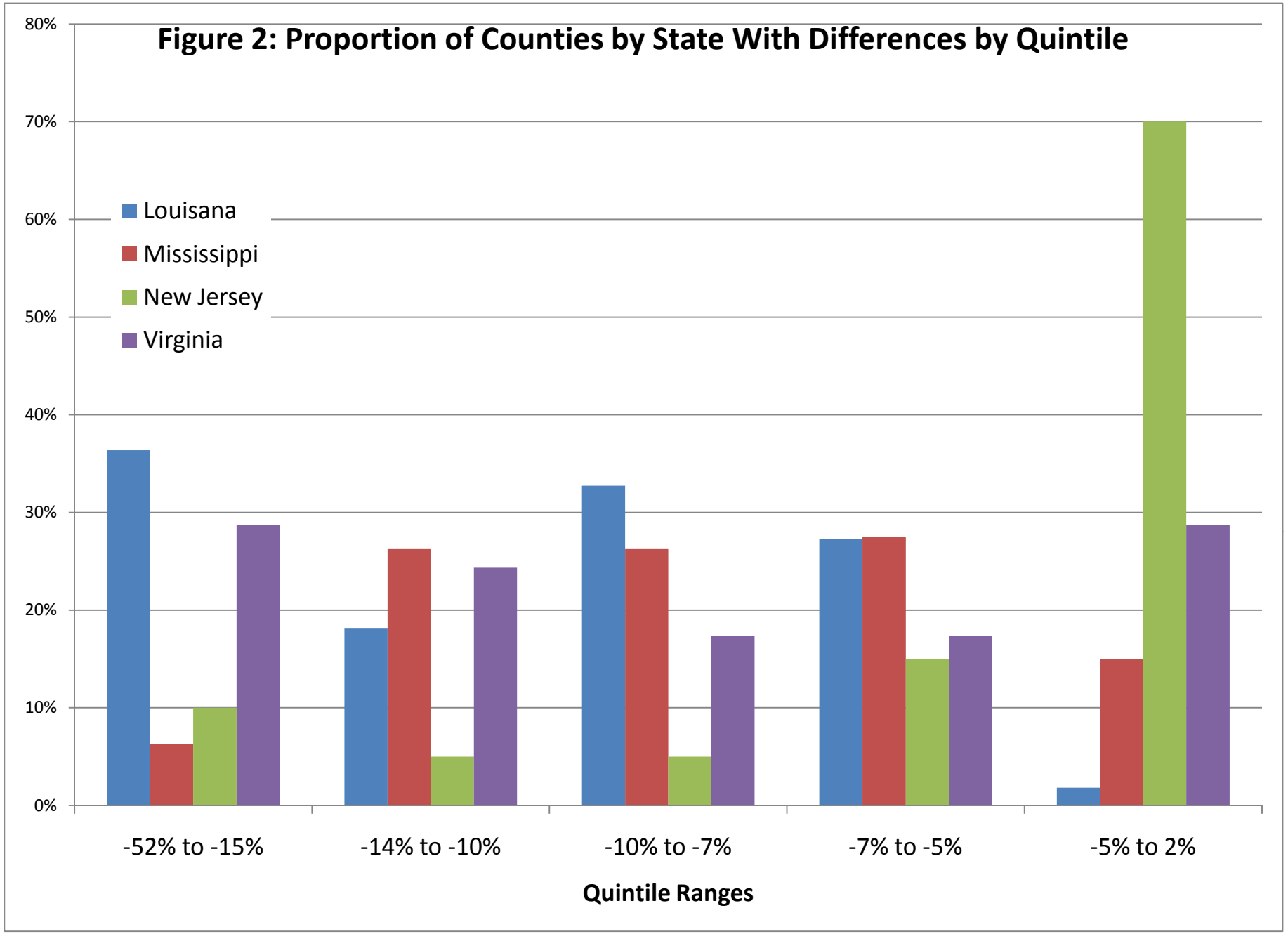
Figure 2: Proportion of Counties by State With Differences by Quintile
80%
70%
60%
50%
40%
30%
20%
10%
0%
Louisana
Mississippi
New Jersey
Virginia
-52% to -15%
-14% to -10%
-10% to -7%
-7% to -5%
-5% to 2%
Quintile Ranges