Spring 2011

Counting the Impossible: Sampling and Modeling to Achieve a Large State Homeless Count

Jennifer L. Priestley
Kennesaw State University, jpriestl@kennesaw.edu

Jane Massey
Massey Consulting

Recommended Citation

Priestley, J. L., & Massey, J. (2011). Counting the impossible: Sampling and modeling to achieve a large state homeless count. Journal of Public Management & Social Policy, 17(1), 85-106.

Counting the Impossible: A Research Note on Using Inferential Statistics to Enumerate Homelessness in Georgia

Jennifer Lewis Priestley
Kennesaw State University

Jane Massey
Massey Consulting

Objective: Using inferential statistics, we develop estimates of the homeless population of a geographically large and economically diverse state – Georgia. Methods: Multiple independent data sources (2000 U.S. Census, the 2006 Georgia County Guide, Georgia Chamber of Commerce) were used to develop Clusters of the 159 Georgia counties. These clusters were used as "strata" to then execute stratified sampling. Homeless counts were conducted within the sampled counties, allowing for multiple regression models to be developed to generate predictions of homeless persons by county. Results: In response to a mandate from the US Department of Housing and Urban Development, the State of Georgia provided an estimate of its unsheltered homeless population of 12,058 utilizing mathematically validated estimation techniques. Conclusions: Utilization of statistical estimation techniques allowed the State of Georgia to meet the mandate of HUD, while saving the taxpayers of Georgia millions of dollars over a complete state homeless census.

Most major cities, every state in the United States and most developed countries make an attempt to quantify the population of homeless individuals within their boundaries. However, almost every report on these counts includes caveats regarding the inherent difficulties and admitted potential for under or over representing the population in question (i.e., Shapcott 2006; South Carolina Council on Homelessness 2007; Berk et al. 2008). The reasons for these difficulties are well documented and include physical difficulties of actually finding unsheltered individuals (particularly in warm climates and in large areas), requirements of different approaches in urban versus rural environments, inconsistent definitions of periodic and chronic homelessness, and limited training of data collectors and volunteers. The frustrations experienced by homeless advocates and those engaged in managing and reporting these counts can be summarized by David Hulchanski of the University of Toronto's Centre for Urban and Community Studies:

> We need to concede that all attempts at counting the homeless are doomed to failure…and attempts to count are never provided enough resources to produce a somewhat defensible number (Hulchanksi 2000, 5).

But in spite of the seemingly impossible task of quantifying the size of the homeless population, policy makers, service providers, and advocates continue to believe that the attempt to count is important. Without some reasonable quantification, there is no way to gauge resources needed nor is it possible to look longitudinally at the results of programs and policies aimed at reducing homelessness. That impulse to measure is clearly reflected in statements by Philip Mangano, Executive Director of U.S. Interagency Council on Homelessness:

> We seek visible, measurable, quantifiable change on our community's streets, within our homeless programs, and, most especially, in the lives of our most vulnerable neighbors. …We are not content to manage the crisis, or to maintenance the effort, or to accommodate the response. We were called to one goal, one objective, one mission – to abolish homelessness (Mangano 2008).

The position of the current paper is that all attempts at reliable enumeration of the homeless population need not be "doomed to failure." Recognizing that traditional approaches have limitations, more inferential approaches might provide an alternative. The methodology used in a large U.S. state count will be explored here as a case study to show how the application of traditional sampling and modeling tools can move us closer to the hope of tracking "measurable, quantifiable change on our community's streets" (Mangano 2008).

Homelessness and the Data Imperative

The need to reliably quantify the homeless problem in the United States has its roots in the emergence of homelessness as a major public concern in the 1970s and 1980s. Public awareness of the problem of homelessness grew as skid rows were eliminated through urban revitalization; enforcement of public drunkenness and status crimes was relaxed; and the mentally ill were deinstitutionalized (Rossi 1989). Rising housing costs and changing social norms related to family structure and the status of women were also in play during the same period. Whatever the root causes, the reality was that people were sleeping on the nation's streets. The homeless were even sleeping on the national Mall, under the shadow of Capitol building. In the late 1970s, Mitch Snyder, a leading advocate for the homeless, estimated that a million Americans were homeless. By 1980, his estimate had grown to between two and three million (Jencks 1994). So it is not surprising that when the U.S. Department of Housing and Urban Development (HUD) issued an estimate of only 250,000 to 350,000 homeless Americans in 1984, the numbers controversy became the stuff of the evening news (Wright et al. 1998).

The Homeless Person's Survival Act was introduced in 1986 and grew from a national recognition that something had to be done about homelessness in America. The emergency relief portions of the bill were passed in 1987 and signed into law by Ronald Reagan in July of that year. The Act was renamed following the death of its chief Republican sponsor, Representative Stewart B. McKinney from Connecticut. The Act has since been amended four times and is now known as the McKinney-Vento Act. In the twenty years since it was first adopted, the McKinney-Vento Act and its attendant program requirements have been marked by increasingly stringent demands for data collection and reporting (National Coalition for the Homeless 2008). In particular, Congress has taken action to require two major data collection efforts by HUD and the nation's homeless service providers. First, in 2001 Congress directed HUD to require implementation of a Homeless Management Information System (HMIS) and in 2003 Congress directed that beginning in 2004 a regular homeless census be conducted at least every two years.

Both the federal data collection requirements and funding for homeless programs are directed through a Continuum of Care System in the United States. The Continuums of Care vary nationally in size and complexity. Some Continuums represent a single, relatively manageably sized community. Other Continuums were created as a local authority or by a coalition of service providers. And in some cases, the Continuum is a state coalition or state agency representing the "balance of state" – areas not covered by the single community Continuums. The widely varying Continuum of Care model means that "balance of state" Continuums, which often cover largely rural and suburban areas, must conduct homeless counts just as the single jurisdiction or urban areas routinely do. The $1.6 billion of homeless program assistance is tied to the count requirement, among other program mandates. For the Continuums nationwide, the need to conduct a point-in-time homeless count, at least once every two years, is not just a program ideal, it is a funding necessity. But while the count requirement is a program mandate, no funding is available from the federal government for this potentially expensive proposition. Since the count requirement is an unfunded mandate, the quality of the counts around the county varies reflecting local funding capacity.

Georgia's Early Response to the Count Mandate

The State of Georgia is somewhat unique. It is a moderately large state (59,425 square miles – 24th in total land mass), with a large population (9.5mm – 9th in the country), with a densely populated metropolitan area (Atlanta – 5.3 million), and sparsely populated rural areas. It also includes a large number of counties, given its size – 159. The Georgia Department of Community Affairs (DCA), through its State Housing Trust Fund for the Homeless, serves as Georgia's Balance of State Continuum of Care, covering 152 of the state's 159 counties.

When the count mandate was first put in place in 2003, to begin no later than 2004, DCA responded by using estimates based on local counts and national studies. Even after the mandate had been in place for three years, Georgia's Balance of State 2007 Continuum of Care Plan continued to rely on very simplistic estimations based upon anecdotal information (Georgia Department of Community Affairs 2008). Grappling with the count mandate for the balance of the state was daunting. Not only was the sheer size of the geography an obstacle, many of the counties covered by the Balance of State Continuum had few homeless service providers. The absence of service providers meant that in many counties there was not a local organizational infrastructure to conduct counts. At the same, a full state count conducted by state employees or contractors looked to be prohibitively expensive. Consequently, counting the homeless population in Georgia seemed an almost Herculean task — a physical census would have been financially impossible and would have almost assuredly resulted in an undercount. After investigating count approaches used by large locally-based Continuums, DCA staff determined that some type of sampling and modeling approach would be necessary.

A Multi-Step Inferential Statistics Approach

In 2007, work began on a count methodology using a statistical approach. After months of work, it was determined that to count Georgia's homeless population, a multi-step inferential statistics approach, incorporating Cluster Analysis, Stratified Sampling and Multiple Regression Analysis, would be a potential solution.

Initially, the researchers wanted to develop a predictive regression model, using the almost limitless number of potential economic and demographic data available from the U.S. Census, *the Georgia County Guide* and the Georgia Chamber of Commerce, to predict the number of homeless by county. However, to create a regression model a dependent variable was required. Specifically, some estimate of homelessness for every county in GA was needed, prior to model development. At the time the study was commissioned, reliable estimates for fewer than 10 of the 159 counties were available. This left the statisticians with a problem – how to generate a prediction without a complete dependent variable.

Step One – Cluster Analysis.

The first step to developing an initial homelessness estimate for each county was to assign each county into a cluster. Cluster analysis is an unsupervised (lacking a dependent variable) multivariate technique, which assigns observations into similar groupings or clusters, based upon the observational values across several variables of interest (Romesburg 1984). In the present context, the 159 counties in GA represented 159 observations. From these 159 counties, 9 clusters were generated (all analysis was conducted using SAS version 9.1), based upon 300 variables taken from the 2000 U.S. Census. These variables represented effectively all of the complete variables available for analysis from the Census at the county level. Because clustering of the counties was a new concept, the researchers had no preconceived notion regarding the most relevant variables for natural groupings. Therefore a data-mining approach was used, allowing the significant variables to emerge naturally. Cluster size ranged from a low of 5 counties to a high of 35 counties. See Table 1 for cluster summaries.

Mathematically-derived clusters are considered to be internally similar and exter-

Table 1. Summary of Georgia County Assignments into Nine Clusters

Cluster*	Total Cluster Population (2000 US Census)	Total Number of Counties in Cluster	Required Population Sample Size
1	1,434,613	5	229,645
2	1,060,599	15	70,700
3	308,797	29	87,372
5	746,826	24	63,179
6	150,944	14	87,372
7	626,520	35	87,372
8	197,805	9	87,372
9	3,331,052	19	291,384
10	329,297	9	71,527

**Cluster 4 was collapsed into Cluster 1*

nally different; counties within a particular cluster would be expected to demonstrate similar demographic and economic characteristics, while counties outside of a particular cluster would be expected to demonstrate very different characteristics. As a result, a county pulled from one cluster for example, would be expected to be similar to other counties in the same cluster– they would be, in effect, interchangeable in everything except population size. It is for this reason that cluster analysis is heavily used by organizations that collect data from individuals or groups spread across large areas, such as the U.S. Census Bureau.

The majority of the economic and demographic data in the present study were taken from the 2000 U.S. Census. From the thousands of variables available, approximately 300 were retained based upon their completeness for analysis and then scaled to a common set of units – percent of the county population. This was an important step, as the cluster analysis technique uses a Euclidean Distance algorithm to assign observations. These distances are calculated using the original units of the variable in question. As a result, variables scaled differently would generate wildly different distances and would therefore dominate or be dominated in the final cluster solution (Romesburg 1984). For this reason, all variables were expressed in common units of percent of county population.

Because population effects were not included for cluster generation, some county assignments may appear initially counter-intuitive for readers familiar with the State of Georgia. For example, Fulton County (the primary county housing the City of Atlanta) was found to be more similar to counties like Richmond (the county which houses the City of Augusta) and Houston (an urbanized county in middle Georgia, with a significant military presence and proximity to Macon), than counties like Cobb and DeKalb (which are large, relatively affluent suburban counties that are part of the Atlanta Metropolitan area, and border Fulton County).

An important method for validating cluster assignment is through the use of ANOVA (analysis of variance) using variables not included in the original clustering process. For example, Graph 2 depicts the differences among the clusters, based upon the percentage of students qualifying for a free lunch – a variable not included in the clustering process. Importantly, the differences among the clusters is clear – providing further evidence for the existence of natural groupings of counties within the state (ANOVA $p<.0001$, $F=19.00$).

Graph 3 is a second visualization of the differences among the clusters with a variable not included in the clustering process – Median Household Income (ANOVA $p<.0001$, $F=44.80$).

These examples provide additional evidence for the internal similarity and externally differences amongst the clusters. As a result, there is an inherent assumption, that any county selected within a cluster, is reasonably similar in economic and demographic characteristics to every other county within that cluster (Romesburg 1984; Hair et al. 1998). Therefore, once the county assignments to the clusters were complete and validation using alternative data sources was complete, the second step in the analysis could begin – selecting a sample of counties from within each cluster to conduct a homeless census.

Step two – Stratified Sampling

Recall that the methodological objective was to create a predictive model, which required a complete dependent variable. If a sample of counties is to be taken for a homeless census to be conducted and then make inferences from those results onto other counties to complete the dependent variable, this may raise the question "Why not simply take a homeless census across a random sample of counties in the state? Why was the interim clustering process required?" A simple random sample would have entailed randomly selecting some number of counties – say 10 – from the total of 159, where every county has a 10/159 probability of selection (approximately 6.3 percent).

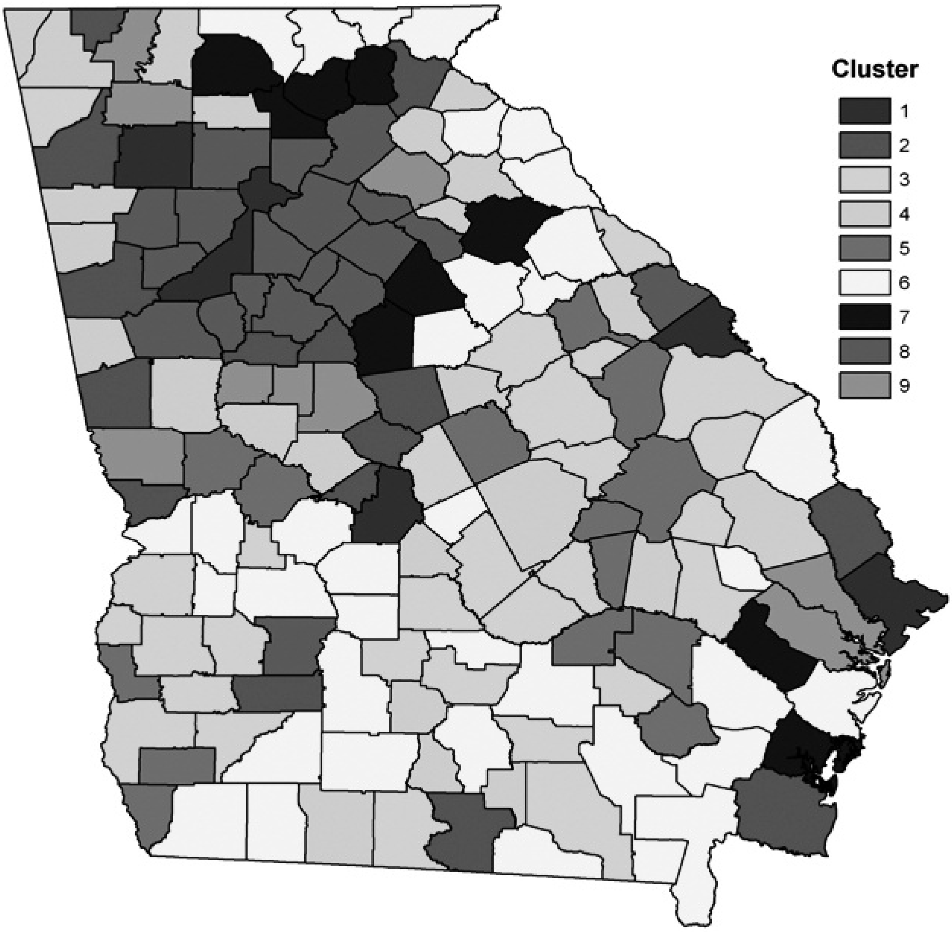
The problem with this simplistic sampling approach in the current context was two-fold:

First, simple random sampling is subject to sampling error – "the luck of the draw"… Without any pre-specified conditions (e.g., strata) a random selection of counties could have resulted in a selection of very similar counties where certain demographic or economic factors were not accommodated. For example, a simple random sample of 10 counties could have yielded 10 rural counties in the southern part of the state with very small populations and a high percentage of the local economies based on agriculture. Alternatively, a sample of 10 counties could have resulted in 10 counties from the Atlanta metropolitan area. Although both of these results would have been possible using a simple random sampling methodology, neither of these samples would have been considered to have been representative of the state as a whole – for the purposes of homeless enumeration or any other objective.

A random sample based on the most "convenient" counties (selection based upon relationships or facilities) would not be appropriate either. This non-probabilistic approach would be analogous to "mall intercepts" or, in statistical terms "convenience sampling." While directionally interesting, the results are not statistically valid and should not be used for inference.

Second, homelessness is, by definition, heavily correlated with the size of the population – if there are more people, one would expect more homeless individuals. For ex-

Graph 1. Cluster Map of 159 Georgia Counties

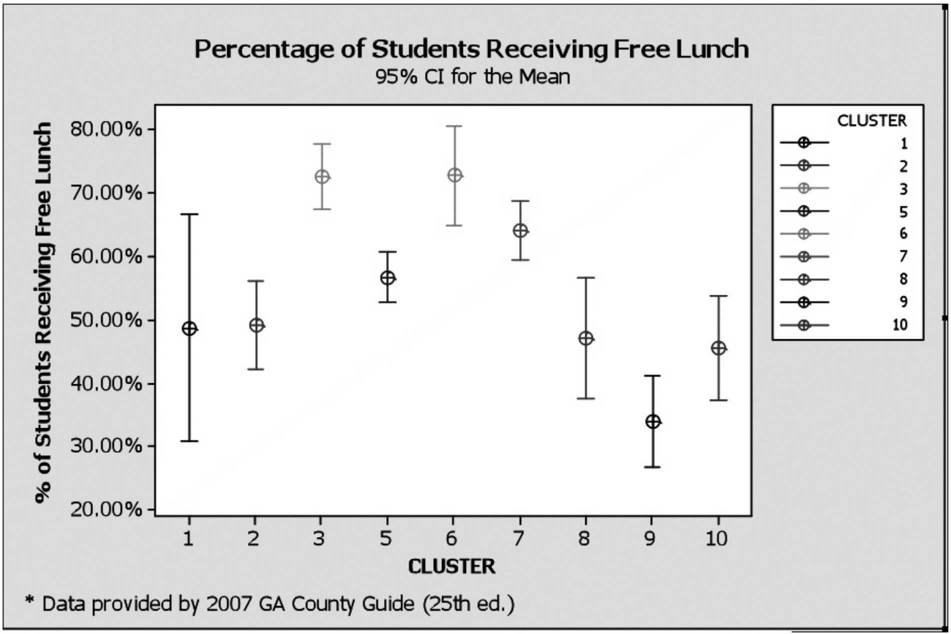


ample, it would be not be a surprise cto learn that there are more homeless people in Fulton County (population 915,623[1]) than in Lowndes County (population 92,115[1]). However, what might be somewhat of a surprise is that, based on local counts in 2007, there were <u>proportionately</u> more homeless in Lowndes County (0.27 percent of the population[2]) than in Fulton County (0.23 percent of the population). A random sample based on population would not accommodate this subtle, but very important, difference.

In an effort to address these issues, the cluster procedure outlined above was an important first step. These clusters provided the researchers with the necessary "strata" to gen-

[1] Source: 2007 Georgia County Guide. 2005 Population Estimate.
[2] Source: Unsheltered Homeless Count as provided by DCA, divided by the 2005 Population Estimate.

Graph 2. Percentage of Students Receiving a Free Lunch by Cluster

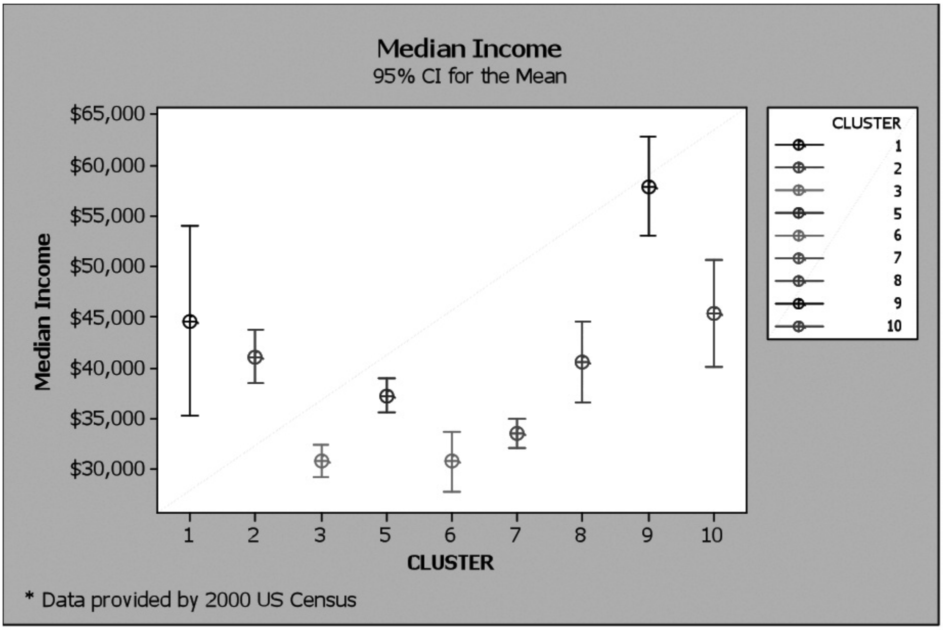


erate a stratified sample, where the different "types" of counties would be represented. This stratified sampling approach is common to pollsters who work to ensure an accurate representation of genders, races, incomes, education levels and other demographic characteristics within a population of interest. While some sampling error could still be present using a stratified approach, the potential for bias and error is somewhat mitigated by ensuring that individual samples are drawn from each uniquely defined cluster, thereby capturing a broad representation of demographic and economic characteristics.

Required sample sizes (to be taken from the general population) were determined for each cluster, to ensure an acceptable maximum margin of error. Once a minimum sample size was determined for each cluster (see Table 1), a random sample of counties was selected from within each cluster, where the summation of the county populations met or surpassed the minimum sample (population) size requirement for the cluster.

The homeless census counts, which were taken during the last week in January 2008 within the sampled counties, utilized the sampling methodology outlined above. For the participating counties, the homeless counts involved enumerating two distinct subpopulations. The first of the two enumerations involved a census of homeless individuals who were in emergency shelters and transitional housing on a single night, representing the count's point-in-time (PIT). Typically, all known shelters and transitional programs are canvassed about the numbers of persons (along with demographic information about those families and individuals) in residence on that night. The Georgia Balance of State count used January 27, 2008 as the PIT. The second enumeration identified and counted persons who

Graph 3. Median Income by Cluster



were unsheltered on the point-in-time night. People are considered to be unsheltered if they sleep in cars, abandoned buildings, on the street or other outdoor locations, or are staying in places not meant for human habitation.

HUD recognizes two basic methodologies for counting persons who are unsheltered (HUD 2004). First many communities use the street count, with variations on time of day and extent of geographical coverage: The street count essentially uses staff or volunteers to canvass an area and count the number of homeless people that they encounter. A good example of this approach is the Metropolitan Atlanta Tri-Jurisdictional count. The Atlanta Tri-J uses some 500 staff and volunteers to cover the City of Atlanta, Fulton County and DeKalb County from midnight to 5 a.m. on a single night. Teams of enumerators use tally sheets to literally count the number of persons found to be sleeping in unsheltered locations. Because the street count may not be an effective method of counting homeless persons in rural and suburban areas, the second method uses surveys or interviews at service provider locations, as well as at places where people are likely to stay or congregate during the day. The Georgia Balance of State count used this service-based survey method to count unsheltered homeless persons in its sample counties. A short (10 minute or less) survey was developed that asked respondents where they slept on the point-in-time night, January 27, 2008, and if they had family members with them. The survey also asked about demographic characteristics, length and frequency of homeless episodes, disabilities, and related topics. The survey was designed to be administered by service providers in the week following the PIT.

Table 2. Variables Retained from the Final Regression Model

Variable Name	Parameter Estimate	t-value	p-value
Intercept	0.00051273	.93	0.3523
Per Capita Income in 1999	.00000007	2.98	0.0034
Percent of HHs with Income <1.5 of the Poverty Level	0.01422	3.37	0.001
Percent of Housing Units for Rental, Not Occupied	0.01422	1.48	.1043
Percent of Housing Units with a Mortgage or other contract to purchase	.00342	-7.17	<.0001
Percent of Housing Units with no Mortgage contract	-0.01525	5.67	<.0001
Percent of Housing Units Lacking Complete Plumbing Facilities	-0.12188	-2.72	0.0073

Survey respondents' initials and date of birth were used to screen for duplicate surveys.

The census counts were not needed broadly. Given the principles of cluster analysis, only one or two counties in each cluster needed to be canvassed. Once a reliable count was available for the one or two counties, the proportion of homeless in those counties was applied to the rest of the counties in that cluster, finally providing the researchers with a complete "dependent variable" for developing a prediction model for the number of homeless individuals in each county. This process also allowed for an interim "prediction" – those proportions were applied against each county's population, generating an initial estimate of homelessness by county, and then aggregated for the state.

Step Three – Regression Model

Once a complete dependent variable was available (proportion of homeless by county), multiple regression models[3] were developed using a combination of variables from the 2000 US Census and the 2007 Georgia County Guide. Again, not using any preconceived notions regarding predictors, the analysts allowed the relevant variables to emerge naturally, using a stepwise selection methodology. The final model retained six variables as presented in Table 2.

[3] While ordinary least squares regression is a preferred technique when modeling a single, quantitative dependent variable, if the dependent variable in question is fractional and limited to a logical range of 0 to 1, other techniques such as the generalized linear model may need to be considered. This is true because OLS regression may generate non-constant variance of the residuals (heteroscedasticity) and the predictions may fall outside of the logical range. The present study did use OLS without any issues related to illogical predictions or heteroscedasticity, however GLM could have been used if these issues were detected.

Using these variables, the prediction of the proportion of homeless by county was made. These ratios were then applied against the most recently published population values to generate homeless predictions by county. The full listing of predictions, including confidence intervals, by county, can be seen in Appendix A.

It is worth noting that the overall estimate of unsheltered homeless individuals in Georgia was 12,058. Based on a 2007 total population of 9,544,750, this prediction would indicate that approximately 0.13 percent of individuals in the State of Georgia are homeless and unsheltered. In the February 2007Annual Homeless Assessment Report to Congress (HUD, 2007), reported the number of unsheltered homeless in the US was 338,781 (page iii in the Executive Summary). Based upon a U.S. population of 296,507,061 in 2005 (when the unsheltered data was collected), this would indicate that the U.S. rate of unsheltered homelessness was approximately 0.11 percent. The prediction generated using the current model for the State of Georgia is higher than the HUD prediction for the country by about 0.012 percent. It is not an unexpected result that the predictions from the current study were slightly higher than the overall predictions for the U.S., as Georgia's climate is more hospitable than that of many other U.S. states.

Limitations

The great statistician George Box is attributed with the quote *"Essentially, all models are wrong, but some are useful"*.

The sampling methodology and prediction model provided in the current study are acknowledged to have limitations, but should provide other state policy makers and their analysts with an alternative approach to a common problem.

First, the clustering procedure was primarily based upon the data from the 2000 U.S. Census. More than any other data source, the Census data provided a vast, rich assortment of demographic and econometric variables, which are not readily available through any other source. These data also exists at many different levels of aggregation, including county and census tract, making these data particularly attractive for finding natural groupings. However, at a minimum, these data are eight years old. As populations migrate and economies change, the data from the 2000 Census becomes outdated. As a result, the clusters defined in the present report might be differently configured with the 2010 Census.

Second, anyone who has engaged in statistical modeling at any point would have expected to see an "R^2" value reported with the regression model. In brief, this value is the primary metric used to understand how well the independent variables explain the variation in the dependent variable, in this case, proportion of unsheltered homeless. This value was not discussed for the present model, because the "actual" number of unsheltered homeless was only provided for 23 (those selected in the stratified sampling methodology) out of 159 counties. The small number of counties with an actual count to be used as the dependent variable reflects the realities of implementing a count protocol in a geographically large and economically diverse state. The remainder of the dependent variable starting values – 136 of them – were estimates based upon the proportions derived from the clusters.

The eventual model did generate an adjusted R^2 value of 0.7 percent. This would indicate that the model can explain 70 percent of the change in the proportion of homelessness as estimated. If the starting estimates used, for example, actual counts from the county, the model and the R^2 value would change.

As stated previously, homeless individuals are difficult to count; the confidence in some of the "actual" (the initial estimations based upon the proportions) values may be low. As a result, the accuracy of the predictions from the model becomes somewhat of a moving target. For example, if the "actual" count for a county is 100 but the model predicted 150 for the county, there is a possibility that, given the characteristics of the county, the reported count is an under representation of the actual homeless individuals. It is for this reason that positive variances (over predictions) were preferred to negative variances (under predictions). These issues make the traditional R^2 metric less meaningful in the present study.

Discussion: The Future of Counting the Impossible

The 2008 project to enumerate homeless in the state was only the first iteration of implementation of a protocol and use of inferential approaches described here. For the first time, organizations concerned about Georgia's homeless have some parameters within which to consider needs and programmatic responses. After the 2008 initiative, DCA issued a report, *Homeless in Georgia 2008*. The report looked at many aspects of homelessness in the state and, for the first time, reported a statewide homeless number using the results of the predictive model. The revelation that over 20,000 Georgians were literally homeless on a single night in January peaked interest in the issue and generated local interest around the state in participating in future projects. A second count using the methodology outlined here, was conducted in late January 2009 and allowed for additional refinement of the nascent methodological approach. Also, under mandate from the federal government, DCA is expected to continue to refine this new approach in coming years.

The use of the multi-step inferential statistics approach described here provided a satisfactory solution to an applied research problem. More importantly, it gave the state agency and participating communities a way to begin thinking about and planning for programs and services for the state's most vulnerable citizens. The use of cluster analysis, followed by a count of sample counties and multiple regression analysis produced an estimate that only differed from the national prediction by about 0.012 percent – adding intuitive validity to its methodological strength. When faced with an extremely complex data collection task, the statistical approach outlined here performed exceedingly well.

Faced with budget constraints, sure to grow worse in the short term as the nation grapples with an economic down turn that turned critical in 2008, the use of these relatively common statistical tools made the impossible task of an accurate enumeration of the state's unsheltered homeless population possible, even in its most rural communities. But its value extends well beyond the confines of service providers, administrators and policy makers concerned specifically about homelessness and its use will continue to be important even after the current economic crisis wanes. The reality is that much of the public policy in the United States faces analogous issues. Federalism at the national level, along with home rule at the state level, means that issues handled by the public sector include the prospect of multiple actors, at multiple levels of government (federal, state, and local), operating with varying degrees of independence. Additionally, issues handled by the public sector are often complex and include populations that are difficult to track. Even seemingly straight forward questions like how many children graduate from high school seem to be fraught with measurement controversies (The Center for Public Education 2008). The Georgia experience with its unsheltered homeless estimates reinforces the importance of using statistical tools creatively to meet the complex needs of governance.

Jennifer Lewis Priestley is an Associate Professor of Statistics in the Department of Mathematics and Statistics at Kennesaw State University. Her research interests lie in Binary Classification, Sampling, Segmentation, Application of Statistical Methodologies for Problem Solving. She has authored several textbook manuals for Excel, SPSS, SAS and Minitab. She can be contacted at: jpriestl@kennesaw.edu.

Jane Massey is a former professor of Public Administration and Governmental Relations. She served as the Director of the Office of Research and Analysis for the Georgia Department of Community Affairs. Jane currently consults with governmental and non-profit organizations including project management, strategic planning, research, data analysis, and special projects.

References

Berk, R.,B. Kriegler, and D. Ylvisaker. 2008. Counting the homeless in Los Angeles County. *Probability and Statistics: Essays in Honor of David A. Freedman*. Nolan, Debbie and Speed, Terry, eds. Beachwood, Ohio:Institute of Mathematical Statistics.

Georgia Department of Community Affairs. 2008. 2007 Balance of State Continuum of Care Plan. Chart L. http://www.dca.state.ga.us/housing/HousingDevelopment/programs/continuumofcare.asp

Georgia Department of Community Affairs. July 2008. *Homeless in Georgia 2008*.

Hair, J., R. Anderson, R.Tatham, and W. Black 1998. *Multivariate Data Analysis.* New York, NY: Prentice Hall.

Hulchasnki, J. David. 2000. *A New Canadian Pastime? Counting Homeless People: Addressing and preventing homelessness is a political problem, not a statistical or definitional problem.* University of Toronto Centre for Urban and Community Studies.

Jencks, Christopher. 1994. The Homeless. Cambridge, MA: Harvard University Press.

National Coalition for the Homeless. 2008. NCH Fact Sheet #18. "McKinney-Vento Act." http://www.nationalhomeless.org/publications/facts/mckinney.html

Romesburg, Charles. 1984. *Cluster Analysis for Researchers*. USA: Wadsworth Inc.

Rossi, Peter H. 1989. Down and Out in America. Chicago: The University of Chicago Press

Shapcott, Michael. 2006. *Backgrounder: Counting Toronto's homeless people.* Toronto, Canada: Wellesley Institute.

South Carolina Council on Homelessness. 2007. The Faces of Homelessness: A Study of Homelessness in South Carolina. http://www.schomeless.org/count07_report.php

United States. Department of Housing and Urban Development, Office of Community Planning and Development. October 2004. A Guide to Count in Unsheltered Homeless People.

United States. Department of Housing and Urban Development. *Annual Homeless Assessment Report to Congress*. 1st ed. Washington, 2007.

United States. Interagency Council on Homelessness. 2008. Excerpts from the speeches of Phillip Mangano, Executive Director. http://www.ich.gov/mission/index.html

Wright, J., B. Rubin, and J. Devine. 1998. *Beside the Golden Door: Policy, Politics and the Homeless*. New York: Aldine DeGruyter.

Appendix A. Final Predictions of Homeless Individuals by Georgia County

County Name	2007 US Census Estimate	Cluster Assign-ment	Upper End of 95% Confidence Interval	Final Prediction	Lower End of 95% Confidence Interval
Bartow County	92,834	1	141	87	10
Chatham County	248,469	1	454	252	16
Fulton County	992,137	1	2,204	2,660	502
Houston County	131,016	1	156	63	12
Richmond County	197,372	1	378	100	5
Cluster Total	**1,661,828**			**3,251**	
Bibb County	154,709	2	369	225	78
Butts County	23,759	2	36	22	2
Camden County	48,689	2	82	45	3
Carroll County	111,954	2	186	137	26
Catoosa County	62,241	2	99	58	1
Dougherty County	95,693	2	235	144	54
Effingham County	50,728	2	67	43	2
Floyd County	95,618	2	186	109	20
Habersham County	42,272	2	78	52	9
Jones County	27,229	2	43	25	5
Lowndes County	101,790	2	241	171	67
Muscogee County	187,046	2	287	421	10
Peach County	25,672	2	43	23	8
Spalding County	62,826	2	121	72	12
Troup County	9,270	2	117	10	8
Cluster Total	**1,099,496**			**1,557**	

County Name	**2007 US Census Estimate**	**Cluster Assign-ment**	**Upper End of 95% Confidence Interval**	**Final Prediction**	**Lower End of 95% Confidence Interval**
Atkinson County	8,223	3	34	29	18
Bacon County	10,507	3	50	41	28
Barker County	16,556	3	14	43	6
Brooks County	16,340	3	48	32	17
Burke County	22,754	3	75	55	30
Calhoun County	6,098	3	16	10	3
Candler County	10,550	3	35	29	16
Chattooga County	26,797	3	78	56	29
Clinch County	6,992	3	31	25	16
Dodge County	20,042	3	83	67	44
Early County	11,836	3	46	33	21
Glascock County	2,771	3	12	11	7
Hancock County	9,568	3	45	32	21
Irwin County	9,934	3	34	25	15
Jenkins County	8,595	3	40	31	22
Johnson County	9,533	3	33	28	16
Lincoln County	8,098	3	47	35	24
Quitman County	2,666	3	25	20	12
Randolph County	7,294	3	27	18	11
Schley County	4,123	3	9	6	1
Stewart County	4,647	3	25	17	13
Tattnall County	23,179	3	83	64	40

County Name	2007 US Census Estimate	Cluster Assign-ment	Upper End of 95% Confidence Interval	Final Prediction	Lower End of 95% Confidence Interval
Telfair County	13,366	3	44	37	20
Terrell County	10,260	3	27	16	6
Turner County	10,280	3	32	25	13
Twiggs County	20,968	3	31	41	10
Washington County	20,937	3	57	37	16
Wheeler County	6,830	3	27	23	14
Wilcox County	8,613	3	32	24	15
Cluster Total	**338,357**			**909**	
Baldwin County	3,781	5	114	6	25
Banks County	46,057	5	42	93	15
Bulloch County	66,176	5	145	106	32
Clarke County	114,063	5	292	198	60
Cook County	16,432	5	45	31	13
Crawford County	12,483	5	32	21	9
Dade County	16,098	5	42	30	14
Harralson County	28,718	5	81	63	32
Heard County	11,387	5	38	29	17
Lanier County	7,947	5	26	21	12
Laurens County	47,520	5	148	112	62
Madison County	28,012	5	67	47	19
McDuffie County	21,551	5	64	44	22
Meriwether County	22,748	5	54	34	12

County Name	2007 US Census Estimate	Cluster Assign-ment	Upper End of 95% Confidence Interval	Final Prediction	Lower End of 95% Confidence Interval
Murray County	40,664	5	113	88	43
Pickens County	30,488	5	66	57	20
Polk County	41,460	5	116	87	43
Pulaski County	9,843	5	29	21	11
Stephens County	25,268	5	81	55	29
Thomas County	45,237	5	99	61	16
Thomaston-Upson County	41,610	5	66	61	15
Tift County	27,820	5	88	38	16
Toombs County	10,894	5	86	26	36
Walker County	64,554	5	158	108	45
Cluster Total	**780,811**			**1,438**	
Appling County	17,946	6	88	72	51
Clay County	3,207	6	24	18	13
Emanuel County	22,469	6	94	75	51
Jeff Davis County	13,291	6	53	43	28
Jefferson County	16,454	6	49	31	16
Miller County	6,163	6	25	18	12
Montgomery County	9,060	6	30	25	14
Pierce County	17,881	6	52	43	22
Seminole County	9,081	6	46	34	23
Talbot County	6,607	6	24	18	11
Taylor County	8,738	6	41	31	22

County Name	2007 US Census Estimate	Cluster Assign-ment	Upper End of 95% Confidence Interval	Final Prediction	Lower End of 95% Confidence Interval
Treullen County	63,535	6	27	189	13
Warren County	5,908	6	25	18	12
Wilkinson County	10,064	6	32	22	12
Cluster Total	**210,404**			**637**	
Ben Hill County	17,650	7	54	39	21
Berrien County	16,722	7	63	48	31
Bleckley County	12,306	7	30	20	8
Brantley County	15,440	7	61	50	32
Charlton County	10,609	7	34	26	14
Chattahoochee County	9,430	7	22	5	0
Coffee County	40,085	7	141	109	62
Colquitt County	44,814	7	130	97	51
Crisp County	22,125	7	67	47	25
Decatur County	28,554	7	82	56	28
Dooly County	11,592	7	36	26	14
Echols County	4,093	7	22	20	13
Elbert County	20,525	7	64	45	25
Evans County	11,505	7	32	24	11
Fannin County	22,580	7	108	94	56
Franklin County	21,793	7	69	54	30
Grady County	25,042	7	81	62	36
Greene County	15,662	7	63	51	30

County Name	**2007 US Census Estimate**	**Cluster Assign-ment**	**Upper End of 95% Confidence Interval**	**Final Prediction**	**Lower End of 95% Confidence Interval**
Greene County	15,662	7	63	51	30
Hart County	24,240	7	84	64	37
Macon County	13,524	7	31	17	4
Marion County	7,024	7	24	17	10
McIntosh County	11,420	7	54	43	26
Mitchell County	24,139	7	58	35	11
Putnam County	20,251	7	106	84	49
Rabun County	16,519	7	136	112	66
Screven County	15,037	7	63	47	32
Sumter County	32,532	7	57	25	0
Taliaferro County	1,884	7	9	6	4
Towns County	6,938	7	75	42	36
Union County	27,562	7	102	122	50
Ware County	35,831	7	113	80	45
Wayne County	29,046	7	78	59	28
Webster County	2,245	7	10	7	4
Wilkes County	10,262	7	43	32	22
Worth County	21,285	7	72	50	30
Cluster Total	**650,256**			**1,711**	
Dawson County	21,484	8	30	105	0
Gilmer County	28,389	8	115	136	57
Glynn County	74,932	8	193	36	52

County Name	2007 US Census Estimate	Cluster Assign-ment	Upper End of 95% Confidence Interval	Final Prediction	Lower End of 95% Confidence Interval
Jasper County	13,660	8	27	31	4
Long County	11,300	8	38	35	18
Lumpkin County	26,554	8	47	17	7
Morgan County	18,165	8	29	27	0
Oglethorpe County	13,963	8	36	49	12
White County	25,020	8	60	**436**	18
Cluster Total	**233,467**				
Barrow County	67,139	9	89	117	144
Cherokee County	204,363	9	214	85	12
Clayton County	272,217	9	299	339	10
Cobb County	691,905	9	884	50	12
Columbia County	109,100	9	73	59	45
Coweta County	118,936	9	127	117	8
DeKalb County	797,093	9	760	79	54
Douglas County	124,495	9	145	26	8
Fayette County	106,144	9	111	141	23
Forsyth County	158,914	9	185	129	73
Gwinnett County	776,380	9	661	115	43
Hall County	180,175	9	220	54	13
Henry County	186,037	9	146	12	19
Lee County	33,050	9	32	49	0
Newton County	96,019	9	90	12	0

County Name	**2007 US Census Estimate**	**Cluster Assign-ment**	**Upper End of 95% Confidence Interval**	**Final Prediction**	**Lower End of 95% Confidence Interval**
Oconee County	31,367	9	34	84	0
Paulding County	127,906	9	131	30	16
Rockdale County	82,052	9	91	59	0
Walton County	83,144	9	99	**1,620**	3141
Cluster Total	**4,186,436**				
Bryan County	30,132	10	51	38	7
Gordon County	52,044	10	102	71	17
Harris County	29,073	10	63	50	17
Jackson County	59,254	10	93	79	17
Lamar County	16,961	10	30	16	0
Liberty County	60,503	10	134	72	12
Monroe County	25,145	10	39	21	0
Pike County	17,204	10	22	12	0
Whitfield County	93,379	10	204	138	43
Cluster Total	**383,695**			**497**	
Georgia	**9,544,750**			**12,058**	

Some Predictions were replaced with actual census counts where available.