Published in final edited form as:
Ann Epidemiol. 2011 April ; 21(4): 290–296. doi:10.1016/j.annepidem.2010.11.016.

A Simple Method to Generate Equal-Sized Homogenous Strata or Clusters for Population-Based Sampling

Michael R. Elliott[1,2]

[1] Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI 48109

[2] Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48109

Abstract

Purpose—Statistical and cost efficiency can be achieved in population-based samples through stratification and/or clustering. Strata typically combine subgroups of the population that are similar with respect to an outcome. Clusters are often taken from pre-existing units, but may be formed to minimize between-cluster variance, or to equalize exposure to a treatment or risk factor. Area probability sample design procedures for the National Children's Study required contiguous strata and clusters that maximized within-stratum and within-cluster homogeneity while maintaining approximately equal size of the strata or clusters. However, there were few methods that allowed such strata or clusters to be constructed under these contiguity and equal size constraints.

Methods—A search algorithm generates equal-size cluster sets that approximately span the space of all possible clusters of equal size. An optimal cluster set is chosen based on analysis of variance and convexity criteria.

Results—The proposed algorithm is used to construct 10 strata based on demographics and air pollution measures in Kent County, MI, following census tract boundaries. A brief simulation study is also conducted.

Conclusions—The proposed algorithm is effective at uncovering underlying clusters from noisy data. It can be used in multi-stage sampling where equal-size strata or clusters are desired.

Keywords

Sample Design; Stratification; Clustering; epsem; National Children's Study

Introduction

Population-based sample surveys often use stratification for statistical efficiency and cluster-based sampling for cost efficiency. Stratification is a general tool that allows samplers to take advantage of known homogenous groupings in a population to reduce variance. Clustering is often enforced by natural cost efficiencies arising when real-world data collection is required. Thus area-probability samples are almost always required to make data collection feasible in face-to-face surveys, allowing interviewers to make trips to an

order-of-magnitude fewer neighborhoods than the total number of interviews obtained. Clustering may also be necessary when local exposure data is required that is not available from administrative sources such as the Census, for example neighborhood disorder measures or environmental samples. An additional feature of strata and cluster formation in many epidemiologic survey settings is the need to construct geographically compact strata or clusters.

Although methods have long been available for analyzing data obtained from stratified and/or clustered sample designs (Cochran 1977), these strata or clusters are usually taken from existing physical or governmental entities, such as schools or Census tracts. When the clusters themselves can be constructed as part of the sample design, we have the opportunity to build desired features into the sample design, akin to traditional methods of stratification, where we can use our knowledge of the structure of the variability in the population to reduce variability in our sample. There are many techniques available to discover parsimonious structures in complex datasets. These include nonparametric algorithmic approaches such as dendrograms or classification and regression trees (LaVarnway 1988) or aggregation procedures such as k-means algorithms (MacQueen 1967), and parametric statistical procedures such as Gaussian mixture models (McLachen and Peel 2000). However, there appears to be little formal survey sampling literature in this area. One exception is Cantwell (1990), who describes the US Census Bureau method of constructing equal characteristic clusters (ECCs) in which the between-cluster variance is reduced by ordering observed covariates X that are proxies for the outcome of interest Y. Units with the largest and smallest values are paired together, then units with the second-largest and second-smallest values, and so forth; this process is then repeated $(M-1)$ times to form clusters of size 2^M. This decreases the between-cluster variance S_b^2 and yields small values of $\rho = \frac{S_b^2}{S_b^2 + S_w^2}$, the intra-cluster correlation that determines the statistical efficiency of the design. Since larger values of ρ are associated with reduced efficiency (e.g., if samples of equal size n are drawn from clusters of equal size N, the efficiency of the mean estimator is given by $[1 + \rho(n-1)]^{-1}$), the ECC method should improve efficiency, although in practice gains appears to be very limited (Cantwell 1991). In neither the Cantwell papers nor the much broader class of cluster analysis literature are there methods that enforce equal size constraints or geographic continuity.

The methodology proposed here is motivated by the sample design for the National Children's Survey (NCS). The NCS is a prospective cohort study of the antecedents of pediatric health and disease in a probability sample of 100,000 US births, to be followed prenatally through age 21 (Landrigan et al. 2006). The NCS is currently designed as a multi-stage area probability sample of births, with the first stage consisting of a probability-proportional-to-size (PPS) sample of 105 primary sampling units or PSUs (mainly US counties, but sometimes groups of very small counties or parts of very large counties). Within each of these 105 PSUs 10–15 approximately equal-size strata are to be constructed, using a size measure based on the predicted number of births over a four-year recruitment period. Within each stratum approximately equal-size sample segments are constructed sufficient to yield a total of 250 births per year within the PSU. One segment within each of the strata will then be sampled, with recruitment attempted for all births in the sample segment. In urban PSUs, an intermediate stage of sampling will occur within each stratum, with equal-sized "geographic unit" (GU) clusters constructed and a single GU sampled within which the sample segments will be constructed and sampled. This reduces the need to create a very large number of sample segments when only 10–15 will actually be used. The goal of the stratification is, as usual, to reduce within-stratum variance to improve statistical efficiency. In contrast to most survey sample settings, however, GU and segment formation

also attempts to minimize within-cluster variance. This is done for two primary reasons: to maximize community outreach efforts within reasonably homogeneous communities, and to allow for accurate estimation of local environmental exposures.

A key distinguishing feature in the NCS sample design is the need to maintain equality of size while forming strata and clusters that are geographically compact and homogeneous with respect to measures that are predictive of pediatric health outcomes of interest. Geographic compactness greatly decreases the cost of recruitment and household data collection, while maintaining approximate equality in size at the stratum level maintains an approximately equal probability of selection sample design, reducing design effects due to selection weights. A second-stage PPS design could be implemented at the GU or segment level with unequal cluster sizes, but for quality control reasons an (approximately) equal-size requirement was maintained here as well. Traditional nearest-neighbor clustering algorithms such as k-means do not provide ways to generate clusters of equal size. Standard methods for forming equal-size strata typically do not incorporate geographic or other forms of distance constraints. Hence we developed a simple clustering algorithm for constructing reasonably compact candidate clusters that maximize the between-to-within cluster variance for proxy measures of pediatric health. We describe the algorithm and apply it to create equal-size clusters within Kent County, Michigan. (Kent County is not one of the PSUs sampled for the NCS; for confidentiality reasons we do not show results for the NCS counties for which this method was developed). We also consider a simulation study to illustrate the effectiveness of the algorithm.

Methods

The proposed algorithm generates clusters of approximately equal size by trying all possible areas as initial “seeds” around which local areas are attached until the proper size is obtained. The areas used to start the seeds are varied from being near to distant from each other. The goal is not to explore the entire space of possible clusters, but to search through an approximation of that space using a reasonable set of seed values to grow the clusters.

Let K be the number of clusters desired from a set of n geographic units with a total population of N:

1. Order geographic units from 1 to n (any method for ordering is acceptable).
2. Compute an $n \times n$ distance matrix between units i and j using centroid longitude and latitude points.
3. Use the first unit as a “seed” to construct the first candidate cluster. Add tracts to this cluster in the order of their distance from the seed tract, until the population of the cluster is greater than $\lfloor N/K - N/2n \rfloor$, where $\lfloor x \rfloor$ is the integer part of a real-valued number x.
4. Choose the seed tract for the second candidate cluster as the next closest tract to the first seed cluster that was not previously included in the first cluster. Add tracts not already assigned to the first cluster to this second cluster in the order of their distance from the second seed tract, until the population of this second cluster is greater than $\lfloor N/K - N/2n \rfloor$.
5. Repeat 4), this time choosing as the second seed tract the second-closest tract, third-closest tract, and so forth, through all tracts not included in the first candidate cluster. For all pairs of candidate clusters, conduct a one-way analyses of variance using the two candidate clusters and the outcome proxy, treating the remaining clusters as a residual third cluster. A multivariate analysis of variance can be conducted if the outcome proxy is multivariate. Choose the second seed tract that

produces the maximum R^2 value (if between-to-within cluster variance is to be maximized) or the minimum R^2 value (if between-to-within cluster variance is to be minimized) to construct the second cluster.

6. Repeat 4) and 5) for the third through $K-1$ clusters, each time choosing the seed tract that maximized(minimized) the between-to-within cluster variance. The unassigned tracts become the K^{th} cluster.

7. Repeat 3)–6), starting with the second unit in the ordering defined in 1) as seed for the construction of the first cluster, then the third ordered unit as a seed for the construction of the first cluster, and so forth through all n units.

8. Conduct n one-way analyses of variance using the n sets of candidate clusters defined in 7) and the outcome proxy. Compute n compactness measures as the mean of the squared distances between all of the areas in each of the candidate clusters defined in 7): $D=\frac{\sum_k \sum_{i,j\in k, j>i} d_{ij}^2}{\sum_k m_k}$, where k indexes cluster assignment for the ith and jth areas and m_k is the number of area pairs within the kth cluster.

9. From 8), choose the cluster set that provides the largest R^2 value (if between-to-within cluster variance is to be maximized) or the smallest R^2 value (if between-to-within cluster variance is to be minimized), or, if compactness is of interest, a plot of the residual variance against the distance measure can be examined and the cluster set that is at the upper- or lower-left of the plot chosen.

A sketch of R code for this algorithm is provided at the author's home page at http://www.sph.umich.edu/~mrelliot/.

Results

Application to Kent County, Michigan

We consider constructing 10 approximately equal size strata out of the 126 Census Tracts in Kent County, Michigan. Kent County is a large county in western Michigan (2000 population 574,335; area 872 square miles). The county seat is Grand Rapids, and while it dominates the county's population, over half of the county's area is exurban or rural. Our measure of size is the number of children under the age of 5 in each tract as measured in the 2000 Census.

We begin by conducting a factor analysis as a preliminary data reduction step. The factors on which we wish to cluster include the following racial, socio-economic status (SES), and dwelling unit factors from the 2000 Census

- Percent African-American
- Percent Hispanic
- Percent with less than a high school education
- Percent with 4 years of college or more
- Percent employed
- Per capita income
- Percent below the poverty level
- Median year structure built (before 1940 coded as 1940)

- Median owned housing unit value
- Percent of housing units vacant

as well as summary measures of neurological and respiratory air pollution from the 2002 National-Scale Air Toxics Assessment (http://www.epa.gov/ttn/atw/nata2002/tables.html). These factors were chosen to mimic the NCS as being predictive of pediatric health outcomes, although any relevant factors to a study of interest could be used. The value .001 was added to all proportions, and all variables except for median year of construction were then log-transformed. These transformed variables were standardized to have mean 0 and variance 1. A factor analysis was then performed to summarize the factors as parsimoniously as possible. Three factors were determined to be sufficient on the basis of a screen plot; they are defined in Table 1. These three factors explained 81% of the variance in the Kent County 2000 Census and 2002 National-Scale Air Toxics Assessment variables. The first factor is an “education” factor that loads heavily positively on college education and per capita income, and negatively on a less than high school education; the second factor is a “pollution” factor that loads heavily on the air pollution measures; the third factor is a “disadvantage” measure that loads heavily on poverty, vacancy rates, and percent African American.

To assist is choosing the best set of the proposed clusters, Figure 1 plots the total residual variance from the ANOVA of the three factors on the 126 proposed clusters against mean squared distance between census tracts within proposed clusters. The circled cluster has the best combination of small residual variance and small within-cluster distances among the census tracts within each cluster.

Figure 2 shows a selection of the geographic distribution of some of the variables that where used to create the factor scores: the proportion of the Kent County population that is African-American; the proportion 25 and older that has less than a high-school education; the proportion living below the poverty line; and the distribution of respiratory pollutants. (All levels are log-transformed to reduce skewness.) Figure 3 shows the proposed 10 clusters based on the “education,” “pollution,” and “disadvantage” factors, compared against a street map of Kent County.

The residual variance from the 10 cluster mapping in Figure 3 is 1.11. The total variance across the three factor scores is 2.57, indicating that the clusters absorb about 57% of the variance in the factors across the Census tracts. The clusters are nearly equal in size as well, with the smallest (4,279) being 94% the size of the largest (4,554), and a coefficient of variation of 2%. The algorithm appears to have clearly delineated Grand Rapids from the remainder of the county, and has separated out the inner-city core from the suburban neighborhoods. Further divisions appear between the suburban areas and the rural/forested regions of the county. The proposed clusters are reasonably compact.

Simulation Study

To explore the characteristics of the proposed algorithm in a controlled setting, we simulated a scalar outcome variable in a 9 × 9 map of equally spaced “tracts.” These tracts were structured as a 3 × 3 set of cells, each consisting of a 3 × 3 set of tracts of equal means, with the means in each set of cell increasing by three units from the upper left to the lower right of map (see Table 2). We then considered three sets of simulations by generating the outcome variable for clustering by adding normally-distributed error terms with means of 0 and variances of 1, 10, and 100. For each simulation, we used the proposed algorithm to create 9 clusters. Each tract was assumed to have a population of 1,000. A total of 50 simulations were generated for each set.

Figure 4 reports the results of this simulation by plotting an intensity map of the outcome data and the associated clusters derived using the proposed algorithm for the simulation associated with the 10^{th} percentile of residual variance and for the simulation associated with the 90^{th} percentile of residual variance. When the signal-to-noise ratio is low (residual variance equal to 1) and thus clusters are clearly defined visually, the algorithm always finds the correct cluster formation. As the residual variance increases and the clusters become visually muddled (residual variance equal to 10), the algorithm continues to approximate the underlying correct cluster formation. Even when the signal-to-noise ratio is extremely high (residual variance equal to 100) and the visual image of the clusters is completely lost, the algorithm continues to pick up some semblance of the true underlying clusters.

Conclusions

This manuscript proposes an algorithm that generates either homogenous strata and clusters or heterogeneous strata and homogeneous clusters of geographic units of approximately equal population size, for use in complex sample designs. By using a “seeding” mechanism to generate clusters of a minimum size and using a search algorithm that maximizes or minimized variance reduction for each generated cluster, approximately equal-size clusters strata are generated. For strata, we will generally want to maximize variance reduction to produce homogenous strata: in the Kent County application, this algorithm produced strata that absorbed more than one-half of the variability of a set of SES and pollution variables. For clusters, we might want to also maximize variance reduction to produce homogenous clusters, as for the NCS, or to minimize variance reduction to produce heterogeneous clusters, to minimize design effects from clustering.

Other homogeneity features can be incorporated in the design by “fixing” geographic units as seed. In the Kent County application, Census tracts containing point sources of pollution such as toxic waste dumps or incinerators could be fixed as seeds, forcing the point source into the approximate center of the cluster. The remainder of the county would then be assigned to clusters using the standard algorithm. Such restrictions in the cluster construction would typically reduce the amount of variance absorbed into the clusters.

The proposed method is not without limitations. Large variations in size of the underlying clusters can limit the flexibility in cluster construction; similarly, if the number of clusters is large relative to the number of available units used to construct the cluster (so that the average number of units per cluster is less than five), there may be more variability in the size of clusters. If the true clustering in the data is relatively weak, or conversely if the number of underlying clusters in the data are large relative to the number of clusters required by the sample design, the convexity criterion enforced by the distance measure may not be sufficient to generate completely contiguous clusters. An example of this can be seen in the simulation study with high variance: see the maps of clusters in Figure 4(c). If contiguous clusters are required, a final stage of cleaning “by hand” will be required to switch cluster membership among the geographic units.

While equal-size cluster sample designs are commonly discussed in the survey statistics literature because of they simplify computation of variance estimates, their use in general is limited, although examples outside of the NCS do exist (Thapa et al. 1987; Garrouste 2010). Part of the reason for this limited use may be due to the fact that, despite a large data mining literature on cluster construction, methods that constrain clusters to be contiguous and to be of equal size are lacking. Although this method was inspired by work on the National Children’s Study, the methods proposed here could be used in any area probability sample or other multi-stage sample where equal-size strata or clusters are desired.

Acknowledgments

The author thanks Dr. Nigel Paneth and Dr. Christine Joseph, along with the editor and two anonymous reviewers, for their helpful review and comments.

References

1. Cochran, WG. Sampling Techniques. 3. New York: Wiley; 1977.
2. LaVarnway, GT. An introduction to CART: Classification and regression trees. In: Wegman, Edward J., editor. Computer Science and Statistics: Proceedings of the 20th Symposium on the Interface. Alexandria, VA: American Statistical Association; 1988. p. 298-301.
3. MacQueen, JB. Proceedings of the Fifth Symposium on Match, Statistics, and Probability. Vol. 1. Berkeley, CA: University of California Press; 1967. Some methods for the classification and analysis of multivariate observations; p. 281-297.
4. McLachen, G.; Peel, D. Finite mixture models. New York: Wiley; 2000.
5. Cantwell, PJ. Equal Characteristic Clustering. Proceedings of the American Statistical Association, Survey Methods Section. Alexandria, VA: American Statistical Association; 1990. p. 231-236.
6. Cantwell, PJ. Clustering in Demographic Surveys: Long-Term Results. Proceeding of the American Statistical Association, Survey Methods Section. Alexandria, VA: American Statistical Association; 1991. p. 550-555.
7. Landrigan PJ, Trasande L, Thorpe LE, Gwynn C, Lioy PJ, D'Alton ME, Lipkind HS, Swanson J, Wadhwa PD, Clark EB, Rauh VA, Perera FP, Susser E. The National Children's Study: a 21-year prospective study of 100,000 American children. Pediatrics. 2006; 118:2173–2186. [PubMed: 17079592]
8. Thapa S, Abeywickrema D, Lynne R. Effects of Compensatory Payments on Vasectomy Acceptance in Urban Sri Lanka: A Comparison of Two Economic. Studies in Family Planning. 1987; 18:352–360.
9. Garrouste C. Explaining Learning Gaps in Namibia: The role of language proficiency. International Journal of Educational Development. 2010 in press.

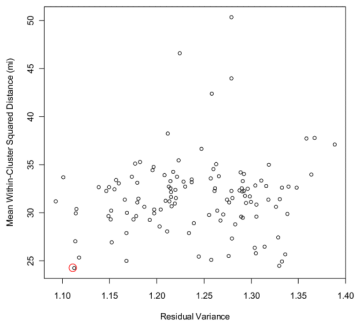

Figure 1.
Plot of total residual variance from ANOVA of three factors on proposed clusters against mean squared distance in miles between census tracts within proposed clusters.

Figure 2.
Log-percent of Kent County, MI population (a) African-American, (b) less than high school education (25 and older), (c) in poverty; (d) log-level of respiratory pollutants: by Census tract. Low levels are blue; high levels are pink.

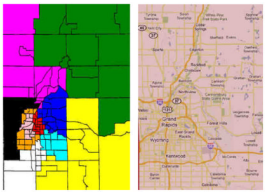

Figure 3.
Proposed 10 clusters for Kent County based on 3-level factor score, together with street map (from http://michigan.hometownlocator.com/mi/kent/).

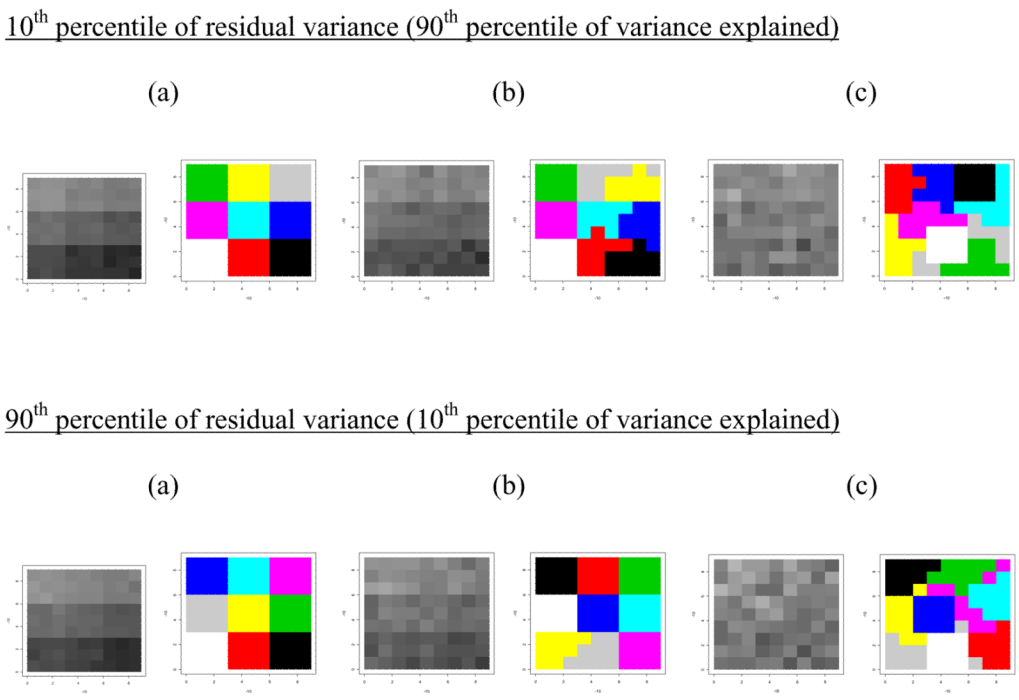

Figure 4.
Simulation study using means given in Table 3 with normally distributed errors with mean 0 and variance (a) 1, (b), 10, and (c) 100. First row gives density map of observed data and associated cluster results for 10th percentile of residual variance; second row gives equivalent results for the 90th percentile of residual variance. Results from 50 simulations.

Table 1

Factor Loadings for Racial, SES, Dwelling Unit, and Air Pollution Measures in the 126 Kent County Census Tracts.

	Factor 1 (“Education”)	Factor 2 (“Pollution”)	Factor 3 (“Disadvantage”)
% African-American	−.214	.476	.628
% Hispanic	−.642	.412	.349
% LT High School Education	−.810	.226	.339
% College Education	.994	−.060	−.191
% Employed	−.482	.187	.487
Per Capital Income	.784	−.256	−.453
% Below Poverty	−.498	.377	.652
Median Year Structure Built	.315	−.586	−.340
Median Housing Unit Value	.752	−.244	−.312
% Vacant	−.293	.037	.530
Neurological Toxins	−.227	.840	.171
Respiratory Toxins	−.071	.932	.086

Table 2

Mean Tract Values for Simulated Dataset.

		“Longitude”								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
“Latitude”	1	1	1	1	4	4	4	7	7	7
	2	1	1	1	4	4	4	7	7	7
	3	1	1	1	4	4	4	7	7	7
	4	10	10	10	13	13	13	16	16	16
	5	10	10	10	13	13	13	16	16	16
	6	10	10	10	13	13	13	16	16	16
	7	19	19	19	22	22	22	25	25	25
	8	19	19	19	22	22	22	25	25	25
	9	19	19	19	22	22	22	25	25	25