

An integration of stratified sampling designs and Geographic Information Systems

- An application in Educational Research

Satharasinghe A, Ranjith Premalal De Silva, Thattil R.O., Samitha S.

Postgraduate Institute of Agriculture

University of Peradeniya

SRI LANKA

Email: rpdesilva@pdn.ac.lk

Tel: +94-777-801712

Fax: +94-8-387216



II. INTRODUCTION

The standards of education in schools are highly diverse in Sri Lanka. Most of the facilities have been concentrated to the schools located in the major cities and suburbs. A government initiative for a major restructuring programme for education is underway aiming at improving national standards of education and minimizing the discrepancies on the quality of teaching offered at school level. This brings about the need for the evaluation of performance of schools.

In this exercise, various parameters that estimate the performances of schools and competence levels of students need to be defined. In general, these parameters could be enumerated using sample surveys and accordingly, the need for a proper sampling framework to derive the correct inferences is emphasized.

The most commonly used sampling strategy is based on the stratified sampling design. Stratified sampling designs are frequently used in demographic and socio-economic surveys. It may be possible to divide a

heterogeneous population into sub populations each of which is internally homogenous. Geographical regions are generally used as strata in most of the stratified sampling surveys. If each stratum is homogenous, a precise estimate of any stratum parameter can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population. In this context, the value of characteristic being measured is the most suitable variable to be used for the stratification of elements in a population. However, in general, this information is not available. Therefore, other variables that are highly correlated to the characteristics of interest are used.

Further, sub populations identified based on geographical regions of populations may not be homogenous due to two reasons. The first reason is that no spatial information other than rigid regional boundaries is used in the stratification. Secondly, the factors that influence the performance of schools within a selected administrative boundary are very diverse and vary widely.

Further, it is obvious that inferences cannot be made about individual behaviour from sample data and it is often both necessary and relevant to estimate parameters at the aggregate level. In most of the surveys, these aggregate levels are geographic regions including administrative boundaries. Hence, it is required to identify the suitable variables, which influence the performance and then determine the spatial boundaries for the process of stratification.

Many factors such as characteristics of students, parents, teachers and learning environment within the school, household status, community background influence the performance and students and their schools. Some of these variables can be described as spatial variables and some others as non-spatial, attribute variables. Spatial variables are used to stratify the areas in this study while non-spatial, attribute variables are used to stratify schools. Classification of schools by type cannot be justified, as type does not account for the facilities available in the schools and the other characteristics, which influence the performance of schools. Strata determined by the spatial and non-spatial variables are used in the first stage and second stage in the two stage stratified sampling design.

Accordingly, the objective of this study is to develop and propose a suitable sampling strategy for the evaluation of the performance of schools in the Southern province of Sri Lanka.

III. METHODOLOGY

IV. COMPILATION OF SPATIAL DATA

Digital thematic coverage showing the boundaries of Southern province, administrative boundaries such as Divisional Secretary (DS) divisions, distribution of Type A, B, C and D type roads, railway stations, hospitals, police stations and schools were obtained for the study. Locations of the schools were recorded from an extensive GPS survey carried out in the province as there were no proper records of the localities of schools available.

In addition, statistics available for the poverty status and population density were transferred to corresponding maps where DS divisional boundaries form the basic spatial unit for data integration.

The first task was to identify a suitable grid resolution where all these spatial variables can be presented in the proper spatial context. In this process, it was noted that the grid resolution should be sufficiently high to include a maximum of one schools in a grid cell in order to avoid the spatial aggregation of variables derived for each school.

A key spatial variable identified in this evaluation process is the distance to schools from A roads. In Sri Lanka, roads are categorized into four groups such as Type A, B, C and D. Type A roads are wide tarred main roads. Most of the resources are thought to be distributed around the main roads. A one km line buffers were created in order to identify the distance of each grid cell from the Type A roads. Further, distance to other infrastructure facilities were also identified based on point and line buffers and corresponding values for grid cells were identified. The calculated distances show the areal distances, however, in reality, on-road distances in terms of other types of connecting roads need to be considered and network analysis (Green et al, 1998, Naude, 1995) needs to be employed to get the true distance to reach a destination on roads.

Distribution of poverty and population densities was also transferred upon the grid maps. An addition, average year 5 scholarship examination marks were also computed and transferred to the cells where schools are located. Year 5-scholarship examination is a national level examination to assess the comparative performance of students at a common numerical scale.

The thematic coverage of each variable was transferred to a common scale to identify the homogenized areas in terms of each variable. These processed grid values were then converted to a database file with the corresponding location identities for the multivariate stratification.

Attribute Variables

The database files from the Ministry of Education were acquired for the attribute data covering the information on the availability of libraries, computing facilities, drinking water, electricity and sanitary facilities, total revenues received, qualification of teachers and their professional background, number of students in year 5 grade and in all other grades and the total number of classes of year 5 grade and all other grades. Travel time from schools to the nearest urban centre was also considered as an attribute variable although it is truly a spatial variable.

Stratification Methodology

Traditionally, stratification is done with respect to one or two variables because of the administrative and computational convenience. However, this does not guarantee that stratification covers all important effects that could influence the variable under study. In this study, two methods of multivariate stratification techniques were employed without affecting the characteristics of stratified sampling.

V. STRATIFICATION PROCEDURE – MULTIVARIATE ANALYSIS

Both factor analysis and cluster analysis were employed in this study. After analyzing the correlations, variables proposed to be used in the stratification were submitted to factor analysis, which identified the variables explaining the patterns of correlations within a set of observation variables. Factor analysis is commonly used in data reduction by identifying a small, number of factors which could account for the most of the variations observed in a much larger set of variables (Unwin et al., 1996). A sufficient number of factors were extracted and factor scores were computed and submitted for cluster analysis which identified relatively homogenous groups of cases based on factor scores. These groups were then used as strata. Application of these procedures for spatial variables and attribute data, the Southern province and the schools were stratified.

VI. STRATIFICATION PROCEDURE – EXTENDED EKMAN RULE

Ekman rule is extensively used for univariate stratification. It is used to stratify a population according to the values of the stratification variable minimizing the standard error of the estimate. Dan (2000) extends this rule to apply for the multivariate stratification.

A linear combination of several stratification variables instead of a single variable as the stratification variable was used here. Correlation analysis and regression analysis were used to identify the factor that is

most significantly related to the performance of schools. This factor was then used as the stratification variable and stratum boundaries were determined by extended Ekman rule. This algorithm available in SAS software was used to determine the boundaries.

VII. RESULTS AND DISCUSSION

VIII. GRID RESOLUTION

Number of schools falling into a cell decreases rapidly with the increase of grid resolution. In this study, a grid resolution of 1 km was chosen as this resolution results maximum of one school per cell. Further improvements of grid resolutions were not required as spatial segregation would not contribute for any further improvement of spatial scale of information into the process.

Stratification Scenarios

The results of the correlation analysis revealed that the average marks are highly correlated with the almost all the spatial variables. Spatial variables are significantly correlated with each other at 95% level of confidence although a few cases of exceptions were noted. Accordingly, the spatial variables were submitted for actor analysis. The variable of distance from the forest was excluded based on the commonalities and factor loading statistics.

There are two factors identified with eigen values greater than one and accounting for 66% of the variation of the original five variables. These two factors were extracted using Principal Component Analysis method together with Varimax Rotation. The Factor one mainly represent the distance to the nearest police station, hospital and railway station while Factor 2 include the distance to the nearest Type A road and to the town. Visual assessment of the component plot in rotated space justifies the use of Principal Component Factor Extraction method and Varimax Rotation as factor loadings are concentrated at the ends of zero lines.

Factor scores were submitted to k-means cluster analysis procedure. ANOVA tables of factor scores using as groups and practical considerations supported the choice of 4 cluster classification. Grid cells with cluster membership status were transferred to the spatial domain and homogenous sub regions were identified as the first stage strata. Boundaries defined at the first stage are obviously different from the administrative boundaries of the province. Since these new regions are derived from the characteristics of performance evaluation, the location of these regions reflects the distance form the facilities that are likely to influence the performance of students in schools.

In view of the need for the stratification of schools, factor and cluster analysis were run on attribute data of schools. Different cluster sizes were compared and based on the average and variability of marks among schools five clusters were selected. The number of schools included in each cluster is 180, 411, 107, 149 and 140. It is noted that schools categorization in terms of Type 1AB, Type 1C, Type 1 and Type 2 are 80, 290, 457 and 339, respectively. These distributional differences are probably due to the fact that grouping schools by type, subjects taught and available highest class in school are considered while in the classification method, several other variables likely to influence the performance of students were incorporated. However, these clusters can be used as second strata.

IX. ASSESSMENT OF SAMPLING DESIGN

The proposed new stratification methodology needs to be tested and proven to be significantly better than the existing stratification scheme where the district and type of school are the only parameters. This study includes validation procedure for the proposed methodology through the comparison of estimates and associated standard errors derived from several other sampling designs.

The results of the year five scholarship examination were used as the performance evaluation variable. Samples were drawn according to the sampling design as detailed below. Standard errors were compared in order to propose the most suitable stratification system and sampling design.

- (a) Two-stage stratified sampling using the district and type of school as the first and second stage strata
- (b) Two-stage stratified random sampling using the strata defined by multivariate procedures
- (c) Two-stage stratified random sampling using the strata defined by Extended Ekman rule
- (d) One-stage stratified random sampling using the strata defined by multivariate procedures to pooled spatial data and attributes
- (e) One-stage stratified random sampling using the strata defined by Extended Ekman rule to pooled spatial data and attributes

Conclusions

This study highlights the fact that the selected sets of spatial and attribute variables result four homogenous geographic strata with respect to the factors that influence the performance of schools. These strata clearly extend across the district boundaries and include all five types of schools.

The one-stage stratified sampling design, using strata constructed under Extended Ekman Rule using linear

combination of several factors, yielded the highest precision of the estimate. The strata constructed by this facility-based classification of schools can be used to grade schools and such a grading system can be used to determine cut-off marks for university entrance and national level examinations.

Almost all-important features of sampling such as randomness, homogeneity of strata as well as the findings can be visualized by digital thematic maps and can be verified against field conditions easily, by user-friendly facilities available in GIS. Further, under this methodology once the system is established, samples could be drawn according to the desired requirements quickly and easily without any manual work including drawing of new geographic sub-regional boundaries and drawing sampling units according to random numbers manually.

X. REFERENCES

Dan, H. (2000). A Procedure for Stratification by an Extended Ekman Rule, *Journal of Official Statistics* 16: pp 15-29.

Green, C., Morojele, N. and Mantz, J. (1998). GIS Tools to Bring Services Closer to People, A Case Study on Planning Police Facility Locations in Kahayelitha, CISIR Transporterk, South Africa.

Naude, N. (1995). Planning Support Tools for Addressing Accessibilty and Related Rural Development Problems, A Case Study, CISIR Transporterk, South Africa.

Unwin, A., Unwin, D. and Fisher P. (1996). *Exploratory Spatial Data Analysis*, Department of Geography, University of Leicester, United Kingdom.