South African Statistical Journal
Vol. 60, No. 1, pp. 1–24
https://doi.org/10.37920/sasj.2026.60.1.1

Computational methods for optimal stratified sampling of ECE outcomes, fees and household income in South Africa

Georgi Borros[1,2], *Şebnem Er*[2] *and Sulaiman Salau*[2]

[1]Southern Africa Labour and Development Research Unit, University of Cape Town
[2]Department of Statistical Sciences, University of Cape Town

In South Africa, the country with the highest income inequality in the world and an unemployment rate of 32.9%, survey research forms a crucial enabler for evidence-based policymaking to drive inclusive growth (Francis and Webster, 2019; Statistics South Africa, 2025a). For decades, nationally representative household surveys have provided the country with key information on livelihoods and areas for targeted decision making by capturing measures such as household income, food expenditure and employment status. More recently, notable survey research has been undertaken in early childhood education (ECE) – seen through the *Thrive by Five Index*, a nationally representative survey of child outcomes using ECE tools developed and validated for the South African context. There are several computational methods developed in the literature offering solutions for optimum boundary determination and sample size allocation in the stratified sampling approach to survey research. The uptake of computational methods in the South African context, however, remains limited. Our study offers the first quantitative evaluation of more common stratification approaches used in South Africa in comparison to five prominent computational methods in the literature – random search, genetic algorithm, biased random key genetic algorithm, grouping genetic algorithm, and variable neighbourhood search. The findings indicate that substantial precision gains can be realised when adopting these novel methods. Additionally, this study is the first application of the methods to South African datasets, contributing to the literature using notable, recent research use cases in the country: the Thrive by Five Index 2021, ECD Census 2021, and General Household Surveys of 2023 and 2024. Through comprehensive evaluation, the work offers insights for performing stratified sampling in these applied contexts using existing methods available in the R programming language.

Keywords: Sampling, Stratification, Survey methodology

1. Introduction

Relative to simple random sampling, where sampled units are drawn from a population with equal probability of selection, stratified sampling can enhance estimator efficiency by dividing a heterogeneous finite population into homogeneous sub-populations (called strata) according to one (univariate

Corresponding author: Georgi Borros (georgina.borros@uct.ac.za)
MSC2020 subject classifications: 62D05, 62P25

stratification) or multiple characteristics (multivariate stratification). These efficiency gains are especially important in resource constrained contexts, as a smaller sample (n), typically associated with lower research cost, is required to reach a desired level of precision, often measured as the variance of the mean estimator under univariate stratified sampling (Cochran, 1977).

Across critical research areas, stratified sampling offers an opportunity to make research budgets go further, provided that efficiency gains can be realised through optimisation. In the process of optimisation, two challenges emerge: the delineation of strata boundaries and the allocation of sample sizes across strata. Boundary determination is an especially complex problem when the stratification variable is continuous, as the range of potential stratification points becomes large. As such, extensive research continues to be conducted, offering various methodologies and computational methods written in the R programming language to optimise these decisions (Kozak, 2004b; Keskintürk and Er, 2007; Brito et al., 2019; O'Luing et al., 2018; Ballin and Barcaroli, 2020; Reddy and Khan, 2020; Brito et al., 2021).

While various methodologies have been published to date for both univariate and multivariate problems, implementation remains limited in the South African context. When a continuous variable is to be stratified[1], this process tends to use distributional grouping (a percentile-based approach), prior categorisation of the continuous variable (for example, construction of income brackets for a continuous income variable) and/or k-means clustering (Kerr et al., 2020; Statistics South Africa, 2025a, 2002). Regardless of stratification variable type (continuous or categorical), sample allocation across strata is then done proportionally to stratum size (Statistics South Africa, 2024, 2025a).

As such, the abundance of complex methodologies and corresponding software available for optimising stratified sampling is yet to be leveraged for South African use cases and, in the event that a new method is applied, the selection process becomes an increasingly infeasible task given an array of options, further disincentivising sampling practitioners to deviate from current techniques. To the best of the authors' knowledge, there has been no application of these methods on South African datasets to date nor in the early childhood education (ECE) sector more broadly. Consequently, the generalisability of prior findings and suitability of complex methods in such contexts remain unknown.

This paper aims to demonstrate the benefits of utilising novel computational methods for stratification relative to more commonly used approaches in South Africa. We additionally provide guidance on method choice by conducting a comprehensive evaluation of computational methods applied to the stratification of ECE learning outcomes, early learning programme (ELP) fees and household income in South Africa. What follows is a formal presentation of the optimisation problem inherent in stratified sampling, focusing on continuous stratifiers for mean estimation in the univariate case. Thereafter, an overview of prominent stratification methodologies is provided along with the parameters for computational testing, establishing a basis for evaluation. Finally, the paper concludes by presenting and discussing the results obtained. Here, we consider continuous stratifiers likely to be applied in ECE or household survey sampling design, offering readers examples of optimised strata formation and sample allocation for these measures.

[1] Continuous stratification variables are not commonly used in large-scale South African surveys, in favour of categorical variables for stratification (Giese et al., 2022; Statistics South Africa, 2025a, 2024).

2. Optimisation problem

In this section we formally outline the statistical optimisation problem facing the sampling practitioner when designing a stratified sample. We adopt notation from Cochran (1977)[2]:

Y	Stratification variable
N	Population size
n	Total sample size
L	Number of strata
N_h	Population size in stratum h
n_h	Units in the sample in stratum h
$\bar{Y}_h$	True mean of Y in stratum h
$\bar{y}_{st}$	Mean under stratification
s_h^2	True variance of Y in stratum h

Optimisation in stratified sampling refers to stratification that minimises the variation of the population parameter of interest for a given sample size. Much of the literature to date considers this parameter to be the mean of the population variable of interest, denoted as $\bar{y}_{st}$ under stratified sampling, and therefore we follow from this for consistency (Cochran, 1977; Kozak, 2004b; Gunning and Horgan, 2004; Keskintürk and Er, 2007). Equivalently, optimisation can be considered as stratification that minimises the sample size for a given variance of $\bar{y}_{st}$.

Two key measures under stratified sampling are therefore (Cochran, 1977):

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{Y}_h}{N}, \tag{1}$$

and

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h^2 \frac{s_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right), \tag{2}$$

where $\bar{y}_{st}$ is the mean under stratification and $v(\bar{y}_{st})$ is the design variance of $\bar{y}_{st}$.

Equation 2 demonstrates that the variance of the mean depends on the sample allocation in the stratum, n_h, as well as the variance within the stratum, s_h^2, therefore subsequently relying on the determination of strata boundaries (Dalenius, 1950). Hence, finding the optimal strata boundaries (OSB) as well as the optimal sample allocation across strata (OA) form the crux of solving the problem of optimal stratification.

2.1 Optimal strata boundaries

The determination of OSB involves finding a set of boundaries, $b_1, ..b_{L-1}$, such that the variance of the mean, $\bar{y}_{st}$, is minimised. Dalenius (1950) developed a relation where the set $\{b_h\}_{h=1}^{L-1}$ of stratification

[2] Datasets considered in this study are treated as populations rather than samples to reflect the process of drawing a sample from a full population frame. As such, $\bar{Y}_h$ is taken to be the true mean in stratum h, and s_h^2 the true variance in stratum h, rather than unbiased estimates as seen in Cochran (1977).

points correspond to the minimum among the variances of all possible boundary combinations, given by

$$\frac{s_h^2 + (b_h - \bar{Y}_h)^2}{s_h} = \frac{s_{h+1}^2 + (b_h - \bar{Y}_{h+1})^2}{s_{h+1}}. \tag{3}$$

While Equation 3 is perhaps mathematically pleasing, it is unusable in practice as both $\bar{Y}_h$ and s_h^2 depend on b_h (Horgan, 2010; Er, 2011). Researchers have consequently derived various solutions in order to approximate the optimal boundaries as specified in this equation. Early approximations include the cum $\sqrt{f}$ rule (Dalenius and Hodges, 1959). Thereafter, Lavallée and Hidiroglou (1988) developed their iterative method with a take-all stratum, and other methods ensued including the geometric method (Gunning and Horgan, 2004), mathematical programming (Khan et al., 2008), random search (Kozak, 2004b) and various genetic algorithms (Keskintürk and Er, 2007; O'Luing et al., 2018; Brito et al., 2019). Since the early 2000s, research involving random search (RS), neighbourhood search, mathematical programming and genetic algorithms (GA) have all claimed optimality and have indeed shown superiority in some scenarios and publications (Er, 2011; Kozak, 2014; Ballin and Barcaroli, 2013; Reddy and Khan, 2016; Brito et al., 2019, 2021). An overarching 'optimal' method, however, is yet to be reached – as heuristic methods are not able to guarantee a global optimum while deterministic methods have been solved for select cases which rely on prior knowledge of the distribution of the population variable of interest (Er, 2011; Kozak, 2014; Reddy and Khan, 2016, 2020). With an unclear 'optimal' method, the sampling practitioner is left with an abundance of choice and relative uncertainty about which to use.

2.2 Optimal allocation

OA in stratified sampling has featured prominently in the sampling literature since the 1920s (Tschuprow, 1923; Bowley, 1926). The seminal work of Neyman (1934) demonstrates a solution to this problem when unit cost is equal across strata, showing that sample allocation proportional to stratum standard deviation multiplied by stratum size leads to the most efficient allocation. Cochran (1977) provides generalised versions for the univariate case, where the optimal sample size for a desired level of precision is given as

$$n = \frac{\left(\sum_{h=1}^{L} W_h s_h \sqrt{c_h}\right) \sum_{h=1}^{L} \left(\frac{W_h s_h}{\sqrt{c_h}}\right)}{V + \frac{1}{N} \sum_{h=1}^{L} W_h s_h^2}, \tag{4}$$

where $W_h = N_h/N$, c_h is the survey cost in stratum h, and V is the desired variance (precision).

Equivalently, given a total survey cost (C), of which c_0 is a fixed overhead cost, the corresponding n that optimises precision is defined as

$$n = \frac{(C - c_0) \sum_{h=1}^{L} \left(\frac{N_h s_h}{\sqrt{c_h}}\right)}{\sum_{h=1}^{L} \left(\frac{N_h s_h}{\sqrt{c_h}}\right)}. \tag{5}$$

Later developments in OA include using GA and integer programming (IP), which have been found to be superior in certain instances (Keskintürk and Er, 2007; Er, 2011; Brito et al., 2015). That being said, the allocation presented in Neyman (1934) is still used extensively in the methods under review in this paper (Kozak, 2004b; Ballin and Barcaroli, 2020).

Table 1. Methods under evaluation.

R Package	Boundaries (OSB)	Allocation (OA)
`stratification` (Rivest and Baillargeon, 2007)	RS* (Kozak, 2004b)	Neyman* (Neyman, 1934)
`GA4Stratification` (Er et al., 2010)	GA (Keskintürk and Er, 2007)	GA* (Keskintürk and Er, 2007)
`stratbr` (Brito et al., 2017)	Biased Random Key GA (Brito et al., 2019)	Integer Programming (Brito et al., 2015)
`SamplingStrata` (Barcaroli et al., 2018)	Grouping GA (O'Luing et al., 2018; Ballin and Barcaroli, 2020)	Neyman* (Neyman, 1934)
`stratvns` (De Lima et al., 2020)	Variable Neighbourhood Search (Brito et al., 2021)	Integer Programming (Brito et al., 2015)
`stats [quantile]` (R Core Team, 2024)	Percentile-based method	Proportional
`stats [kmeans]` (R Core Team, 2024)	k-means clustering (Hartigan and Wong, 1979)	Proportional

*Other methods available in the package, but not used in the quantitative evaluation.
Methods are listed in order of the date of development (earlier methods listed first).

3. Methods

This section introduces the methods under evaluation. We begin by discussing prominent methods in the literature highlighted in an earlier review by Barcaroli and Ballin (2018), after which we consider approaches that have been used in South African surveys (Table 1).

3.1 Computational methods for stratified sampling

In parallel with literature discussed in Section 2, various computational methods have been developed to tackle the problem of optimal stratification (Kozak, 2004a; Keskintürk and Er, 2007; Brito et al., 2015, 2019; Ballin and Barcaroli, 2020; Brito et al., 2021). We evaluate prominent contributions that address the OSB and OA problems jointly, as recommended by Khan and Sharma (2015) for enhanced efficiency, given that the boundaries directly influence the allocation.

To date, OSB is most often treated as the computationally expensive optimisation problem requiring an algorithmic approach for solving, while the OA problem in the univariate case is generally deterministic[3] (Neyman, 1934; Kozak, 2004b; Rivest and Baillargeon, 2007; Brito et al., 2015, 2019, 2021). Consequently, novel computational methods have largely been developed surrounding the

[3] An exception to this trend is the GA used by Keskintürk and Er (2007) to solve for both OSB and OA.

OSB problem, while the OA problem is incorporated by way of a deterministic process set by the researcher.

We consider novel methods explored in a review by Barcaroli and Ballin (2018), with concurrent packages written in the R programming language[4]. The RS method in the `stratification` package seeks to refine the set of stratum boundaries through perturbation and subsequent evaluation against a fitness function, arriving at a solution once the stopping criteria is reached (Kozak, 2004b). The RS process is complemented by Neyman (1934) allocation for every stratification formed. RS often serves as a benchmark for other newly developed algorithms in the stratification literature (Kozak, 2004b, 2014; Kozak et al., 2007; Er, 2011).

Alongside RS, the GA is another widely used computational method. A biased random key GA (BRKGA) is used in `stratbr` to search for the OSB (Brito et al., 2019). In comparison to a traditional GA, the representation structure of the BRKGA is different in that candidate solutions are encoded and represented by vectors. The "biased" portion of the algorithm selects the best solutions and adds them to the population to be considered in the following iteration of the GA (Brito et al., 2019). The `stratbr` package concurrently makes use of IP to solve the allocation problem, a method developed in response to other techniques which can give rise to non-integer solutions for sample allocation (Brito et al., 2015).

Another variation of the traditional GA is the grouping GA (GGA) used in `SamplingStrata`, complemented by Neyman (1934) allocation (O'Luing et al., 2018; Ballin and Barcaroli, 2020). The fundamental shift from GA to GGA considers stratification as a grouping problem rather than a boundary definition problem. Here, an initial stratification is generated using k-means clustering as a starting point, after which selection, crossover and mutation are applied to the group structure.

A later development in search procedures, Variable Neighbourhood Search (VNS), is employed in `stratvns` to form the strata, after which IP is used to best allocate n (Brito et al., 2021)[5]. VNS systematically changes the neighbourhood in each search, without following a trajectory but considering different neighbourhoods and moving from one solution to another if an improvement has been made (Hansen and Mladenović, 2001).

In addition to the methods examined by Barcaroli and Ballin (2018), we consider the `GA4Stratification` package, the first to use a GA to solve the OSB and OA problems jointly (Keskintürk and Er, 2007). While `GA4Stratification` is no longer actively maintained (although it can be accessed from the R archive), inclusion of this package is prudent as a traditional GA benchmark in comparison to the modified GAs offered by later authors. An especially novel contribution in `GA4Stratification` is the use of the GA for OA, which has been found to be more efficient in comparison to Neyman (1934) and proportional methods (Keskintürk and Er, 2007). A further deviation from the methods considered by Barcaroli and Ballin (2018) is the exclusion of the dynamic programming method from our reported output, available through the `StratifyR` package (Reddy and Khan, 2016). Unfortunately, this method did not reach a solution across the datasets used in this

[4]While Barcaroli et al. (2018) included similar methods in their discussion, the authors focused their quantitative review solely on the packages 'stratification' and 'SamplingStrata'. This study aims to quantitatively evaluate all methods, with application to South African ECE and household survey data.

[5]As the high computational requirements of VNS can cause practical issues relating to extensive computational cost, `stratvns` makes use of a reduced VNS procedure.

study[6].

Earlier approximate methods, such as the cum$\sqrt{f}$ rule, geometric method and simplex method, are not included in this comparison, as evidence from earlier comparative studies shows that later methods have been found to be superior in terms of efficiency. Keskintürk and Er (2007) demonstrate that the geometric method fails to reach a solution for normal and negatively skewed distributions, while the GA tends to reach a lower variance of the estimate in comparison to both the cum$\sqrt{f}$ rule and geometric method. Kozak and Verma (2006) find that the RS algorithm achieves a smaller coefficient of variation than both geometric and simplex methods.

3.2 Methods used for stratified sampling in South Africa

While the optimisation problem for stratification has been extensively developed in the sampling literature, application of the aforementioned computational methods in the South African survey context remains limited. For the OSB problem, especially relevant when considering continuous stratifiers, sampling practitioners tend to make use of several distinct approaches or a combination thereof (Maremba, 2019; Kerr et al., 2020; Statistics South Africa, 2024, 2025a):

- Categorisation of the continuous stratification variable using predefined categories.
- A percentile-based approach to stratification, whereby strata are formed based on equal distributional cut-offs (e.g. deciles for 10 strata or quintiles for 5 strata).
- k-means clustering to form the strata.

Regardless of stratification approach, OA is frequently based on proportional allocation across strata (Statistics South Africa, 2024, 2025a).

It is important to note that lack of implementation of novel computational methods may be due to sampling goals that are starkly different. For literature-based methods, the aim is to maximise precision of a given estimator. However, for surveys in practice, generating a sample that is broadly representative of the population for various subgroups may be a more prominent concern. Additionally, most novel computational methods are univariate in nature and large-scale surveys do not generally consider one outcome variable for estimation, but rather multiple outcomes (especially considering the large investment needed to support data collection activities). A risk in implementing novel univariate optimisation approaches is that while these methods might be optimal for one variable, they may be suboptimal for others. Multivariate optimal methods in stratification are therefore a crucial area for further research, although beyond the scope of this paper[7].

For instance, the stratified sample used for the National Income Dynamics Study (NIDS) Coronavirus Rapid Mobile Survey (CRAM), NIDS-CRAM, demonstrates sample goals that are not purely based on maximising precision for a single estimator. Here, the survey is stratified according to household per capita income decile, race, age and urban/rural geography type, with the aim to be "broadly representative" of the South African population in 2017 (Kerr et al., 2020). While household per capita income is a continuous stratifier that is stratified using a percentile-based approach,

[6] All code used was submitted alongside this article for reproducibility. Version 1.0-4 of this package was used.
[7] Readers may refer to Ballin and Barcaroli (2020) for a multivariate application.

the goal of the sample is not to maximise precision in measurement of this variable, but rather to generate a sample that has adequate coverage of the *general* population across multiple measures of interest.

Nevertheless, for the purposes of this study we compare two methods used in practice, the percentile-based approach as well as k-means clustering, to optimisation methods in the literature. We consider one variable of interest with a goal of maximising precision for that variable, demonstrating the possibility for efficiency gains in practice while noting the need for multivariate methods for wider implementation and with adequate representation across multiple measures of interest.

4. Datasets

Early stratification literature has highlighted the importance of testing methods across distributions, especially skewed data, as some methods are especially sensitive in this regard (Lavallée and Hidiroglou, 1988; Gunning and Horgan, 2004). As such, without application to South African data there is limited evidence on best practices for stratified sampling in this context. Consequently, the datasets used in this study cover major primary research contributions in the country, summarised in Table 2, particularly in the ECE field. Here, we treat each dataset as its own population sampling frame. For each population, we budget for a total sample size n equivalent to approximately 10% of the population size N.

TB5 The Thrive by Five (TB5) Index 2021 forms the nationally representative baseline for a series of surveys that monitor trends in preschool children (aged 50-59 months) attending ELPs (Giese et al., 2022). TB5 is the largest survey in South Africa to focus on child outcomes, measured using the early learning outcomes measurement (ELOM). The ELOM was developed to be responsive to the multi-cultural context of South Africa and has since been extensively validated (Snelling et al., 2019; Anderson et al., 2021). For this study, we make use of the ELOM scores from the *Thrive by Five Index and ECD Baseline Audit 2021* (DataDrive and Department of Basic Education, 2022). The total ELOM score is used as the stratification variable.

ECD The Early Childhood Development (ECD) Census of 2021 is a national census of ELPs in every ward in South Africa, mapping the sector with insight into ELP characteristics and services offered (Department of Basic Education, 2022). We focus on the fees variable captured here, as

Table 2. Datasets.

Dataset	**Area**	*Y*	Range	*N*	*n*
TB5	Education	ELOM score	[6.37; 96.53]	5,222	523
ECD	Education	ELP fees	[1; 9000]	37,103	3,711
GHS23	Socioeconomic	Household income	[0; 1,460,000]	20,507	2,051
GHS24	Socioeconomic	Household income	[0; 1,921,583]	20,545	2,055

fees charged by ELPs[8] have been shown to be a significant socio-economic factor related to child outcomes (Henry and Giese, 2023). In considering fees as a stratification variable, however, the continuous range makes it a complex task to determine the best stratification. We contribute by rigorously determining the stratification points across computational methods.

GHS Finally, the General Household Surveys (GHS) are annual, nationally representative household surveys run by Statistics South Africa, with a mandate to track the progress of development in the country (Statistics South Africa, 2025a). Household income is one such measure not only used for tracking development progress, but also adjacent measures such as poverty and inequality (Mdluli and Dunga, 2022). Studies in these areas use household income as a basis for measurement, making this variable an important one to consider during the sampling phase. To offer insight into how best this might be done across the prominent computational methods available, we use total household income from the GHS of 2023 and 2024 as our stratifier (Statistics South Africa, 2024, 2025b).

Figure 1 demonstrates the distribution of variables corresponding to the different datasets. Notably, the GHS datasets display extreme skewness, with the majority of the population below R6,000 per month. ELP fees follow this pattern and display a positive skew distribution. These types of distributions are common in the South African context, where the majority of the population earns below R10,000 per month. The TB5 ELOM score data, however, shows a relatively normal distribution, an expected result for a psychometrically validated assessment tool.

5. Evaluation methodology

The methods considered in Table 1 will be evaluated based on the optimisation parameter, stability and runtime. The method that performs best in all cases will be recommended for use by sampling practitioners working with similar datasets.

5.1 Optimisation

As outlined in Section 2, optimisation is the minimisation of $v(\bar{y}_{st})$ with respect to n or, equivalently, minimisation of n subject to a desired level of precision. Most computational methods in this study use a mean-weighted version of $v(\bar{y}_{st})$ for optimisation, the coefficient of variation $cv(\bar{y}_{st})$, enabling interpretation across different measurement units. Following stratified sampling literature, which frequently uses $cv(\bar{y}_{st})$ as a measure for comparison across methods, we further use this metric to evaluate our optimisation results (Er, 2011; Brito et al., 2019). The coefficient of variation is

$$cv(\bar{y}_{st}) = \sqrt{v(\bar{y}_{st})}/\bar{y}_{st}. \tag{6}$$

Table 3 outlines the optimisation approach adopted by each method, with the 'min' column showing the objective for minimisation and the 'given' column outlining the value that the optimisation process will be subject to. Here, the percentile method does not follow an optimisation process as it is deterministic and included for comparison purposes. Once the percentile cut-offs are established,

[8] This variable is technically the 'maximum fee charged' as ELP fee models are generally dynamic in response to their target market.

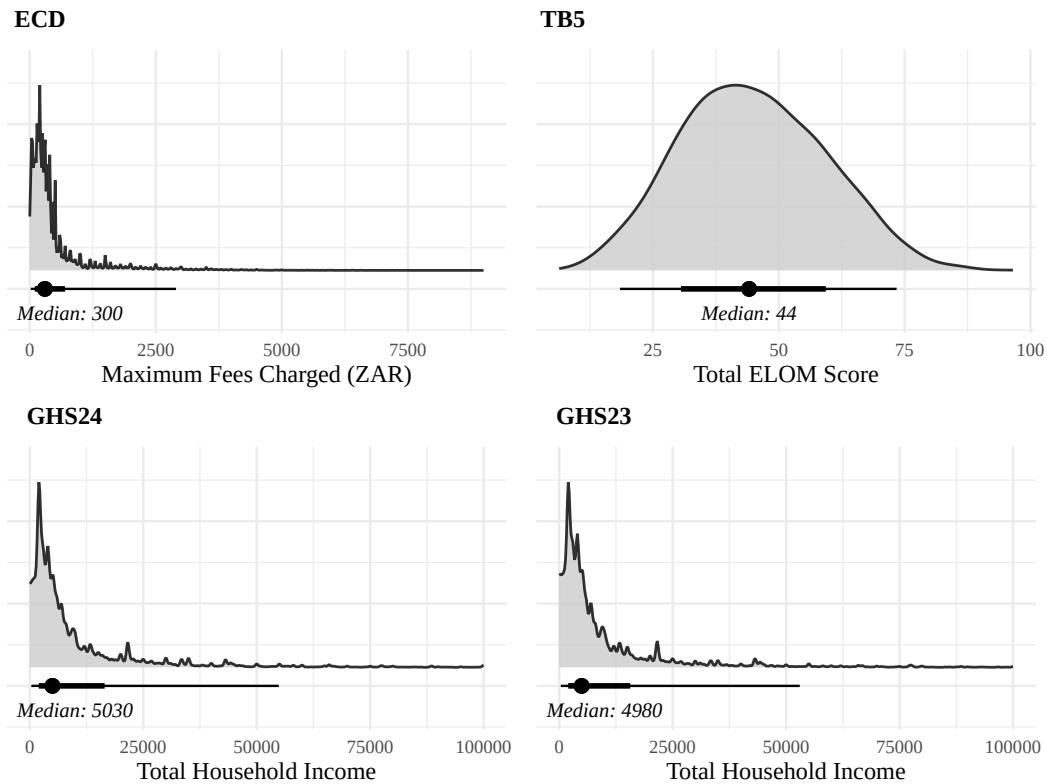

Figure 1. Distribution of stratification variable, by dataset.

Note: Values above R100,000 in GHS24 and GHS23 (comprising $\approx$ 1% of each dataset, respectively) are omitted from the figure for readability.

n is then allocated proportionally, the commonly used allocation approach in South Africa (Statistics South Africa, 2025a). Similarly, the k-means approach does not follow a constrained optimisation process (optimisation subject to a given n) and, once clusters are formed, n is allocated proportionally.

As three of the methods minimise variation, $cv(\bar{y}_{st})$ or $v(\bar{y}_{st})$, for a given n, while two minimise n for a given cv, the former will be referred to as *variation minimising methods* and the latter *n minimising methods*. For *variation minimising methods* as well as the percentile and k-means methods, the prescribed sample size is given as $n_{10\%} \approx N/10$. To facilitate comparison between *variation minimising methods* and *n minimising methods*, the minimum cv achieved across *variation minimising methods*, cv_{min}, is used as the prescribed cv for the *n minimising methods*[9]. When one of the *n minimising methods* is able to reach less than $n_{10\%}$ given cv_{min}, this method is deemed superior in terms of optimisation.

[9] For *GA4Stratification*, we simply convert $v(\bar{y}_{st})$ to $cv(\bar{y}_{st})$ using Equation 6.

Table 3. Optimisation process by method.

Method	R Package	Min	Given
RS	`stratification`	$cv(\bar{y}_{st})$*	$n_{10\%}$
GA	`GA4Stratification`	$v(\bar{y}_{st})$	$n_{10\%}$
BRKGA	`stratbr`	$cv(\bar{y}_{st})$	$n_{10\%}$
VNS	`stratvns`	n	cv_{min}
GGA	`SamplingStrata`	n	cv_{min}
Percentile-based approach	`stats [quantile]`	*N/A*	*N/A*
k-means clustering	`stats [kmeans]`	ss_w**	*N/A*

*In *stratification*, either n or cv can be used as the objective for optimisation. For this study, we choose to minimise cv given n.
**k-means minimises the within-cluster sum of squares (ss_w) (Hartigan and Wong, 1979).

Number of strata

Five strata, $L = 5$, were used across datasets and methods. This number was informed by Cochran (1977)'s observation of diminishing returns (in terms of gains in precision) as the number of strata is increased. Using elbow plots of the total within sum of squares based on k-means clustering to form groups (in our case, strata), we determined that the gains in further grouping beyond 5 strata would be minimal across all datasets[10].

Algorithm parameters

For comparability, it is important that algorithmic parameters are kept consistent as far as possible. Across GA methods, we specified 5,000 iterations, a generation size of 50 and mutation rate of 0.2. Parameters applied to the VNS method were set as follows: 5 as the maximum number of VNS iterations, 3 neighbourhoods, 5 cut points maximum per neighbourhood and 10 initial solutions.

5.2 Stability

Each method tested, other than the percentile method, has a randomised component and hence the stability across runs is an important performance metric. This is because a method prone to high fluctuation across runs might lead an end user to incorporate a suboptimal stratified sample. As such, all tests were simulated over 10 runs. This number of runs was chosen using the RS method as a benchmark, for which 10 runs was sufficient to produce a margin of error of at most 1.03%, relative to the estimated mean, across datasets. Consequently, we evaluate stability using the percentage margin of error relative to the estimated mean ($ME_{\%}$) associated with each method across runs, where methods offering a lower $ME_{\%}$ are preferred for practical implementation.

We follow from Sapra (2022) in deriving relative margin of error, $ME_{\%}$. Since $ME_{\%}$ is computed in relation to the mean estimate (either $\bar{cv}_{st}$ or $\bar{n}$), it is calculated based on the result of optimisation.

[10] See Appendix A for a visualisation of the total within sum of squares by number of strata.

Table 4. Summary of packages used.

Package	Created	Version	CRAN
`stratification`	2007	2.2-7	Yes
`GA4Stratification`	2007	1.0	No
`SamplingStrata`	2011	1.5-4	Yes
`stratbr`	2017	1.2	Yes
`stratvns`	2017	1.1	No
`stats [kmeans]`	2000	4.6.0	Yes
`stats [quantile]`	2000	4.6.0	Yes

In the case of *n minimising methods*, $ME_{\%}$ is derived from the standard error of the resulting n in relation to the mean, $\bar{n}$:

$$ME_{\%} = \frac{t_{R-1} \times se_n}{\bar{n}} 100, \tag{7}$$

where R is the number of runs, t is the critical value corresponding to a 95% confidence interval with $R-1$ degrees of freedom and se_n is the estimated standard error for $\bar{n}$ across R runs. For *variation minimising methods*, $ME_{\%}$ was calculated as

$$ME_{\%} = \frac{t_{R-1} \times se_{cv}}{\bar{cv}_{st}} 100, \tag{8}$$

such that se_{cv} is the standard error of $\bar{cv}_{st}$ over R runs.

5.3 Computation time

All computations were performed using R version 4.5.1 on a Windows 10 machine with an Intel Core i7 processor, 32GB of RAM and 12 cores. A summary of each computational package used is presented in Table 4. At the time of writing, *stratvns* and *GA4Stratification* were not stored on the Comprehensive R Archive Network (CRAN) and archived versions were used (R Core Team, 2024).

Finally, as optimisation methods for stratified sampling can be complex and subject to significant computation time, researchers have begun to consider runtime when developing computational methods (O'Luing, 2022). We therefore evaluate the system time taken to execute each computational method as an aspect of usability for consideration.

6. Results

The results section is presented in two parts. The first covers the method evaluation based on the metrics discussed in Section 5, highlighting which methods are top performers in the given domains. The second part unpacks the OSB and OA results for the top performing methods relative to the more traditional methods.

6.1 Method evaluation

Table 5 outlines the optimisation results for each computational method and dataset combination. $\bar{cv}_{st}$ corresponds to the average cv over the 10 runs, $cv_{st,min}$ is the minimum achieved across the runs, $\bar{n}$ is the average sample size across runs and n_{min} is the minimum sample size achieved across runs. It can be seen that for *n minimising methods*, all cv figures remain constant, as this was the prescribed cv given from the minimum achieved across *variation minimising methods*. Similarly, n remains constant for *variation minimising methods*. $ME_{\%}$ in relation to the mean estimate (either $\bar{cv}_{st}$ or $\bar{n}$) is shown in the '$ME_{\%}$' column, calculated based on the result of optimisation. Finally, the 'time' column indicates average the system time taken (in minutes) to run the algorithm.

6.1.1 Optimisation and stability

In terms of optimisation and stability, the computational methods performed almost equivalently. In some instances, the GGA and VNS methods require a few more sampled units (n) to reach a given cv. The largest of these additional sample requirements is observed for the GGA for the ECD data, where, on average, GGA required $n = 3739$ to achieve the same cv as other methods with a prescribed $n_{10\%} = 3711$. In other cases, there were slight differences among *variation minimising methods*, but only resulting in a maximum $\bar{cv}$ difference of 0.02 percentage points between the BRKGA and the traditional GA for the GHS23 dataset. As such, all computational methods produce a high degree of precision and would generate a generally optimised stratified sample consistently. At a marginal level, the BRKGA method was the top performer in this domain, most consistently achieving the lowest cv with minimal margin of error. The superiority of BRKGA was observed for the highly skewed income data from the GHS datasets. Here, BRKGA may have allowed more diversity in its solution generation in comparison to the other GAs, as the mutation operator used introduces a new random solution vector at each new generation - rather than applying a mutation function to an already generated individual (Brito et al., 2019). The integer programming method for optimal allocation could also see marginal gains for BRKGA relative to RS in instances where Neyman (1934) allocation did not reach an integer solution.

When compared to more traditional percentile and k-means approaches, however, pronounced differences in resultant optimisation are observed – particularly for the highly skewed datasets. For GHS23, the percentile approach had a cv 9.4x larger than BRKGA and k-means resulted in a cv more than 3.5x larger than BRKGA, showing an opportunity for efficiency gains through adoption of computational methods. It is seen that, relative to the percentile approach, k-means consistently required a smaller sample size. This is likely due to its algorithmic mechanism to minimise within-group sum of squares, while groupings formed using the percentile approach are pre-defined and unable to iterate to more internally homogenous alternatives. These differences are more formally captured in Table 6, showing the sample size required (n_{trad}) for traditional methods to achieve the same cv as the best performing computational approach ($cv_{brkga}{}^{\%}$).

6.1.2 Computation time

In contrast to the optimisation and stability metrics, runtime in minutes varied more substantially across computational methods. The GGA had the longest runtime, followed by either VNS or GA. RS had the lowest runtime among computational methods, followed (by some distance) by BRKGA. That being said, all computational methods were able to run on a personal computer within a reasonable

Table 5. Optimisation results across methods.

Method	Data	R	$\bar{cv}_{st}$%	$cv_{st,min}$%	$\bar{n}$	n_{min}	ME%	$time$
BRKGA	ecd	10	0.37	0.37	3711.0	3711	0.0624	76.515
GA	ecd	10	0.37	0.37	3711.0	3711	1.2422	133.369
RS	ecd	10	0.37	0.37	3711.0	3711	0.0010	0.082
ss	ecd	10	0.37	0.37	3738.7	3714	0.4450	472.446
VNS	ecd	10	0.37	0.37	3715.9	3711	0.1020	171.246
k-means	ecd	10	0.60	0.60	3711.0	3711	0.1348	0.006
percentile	ecd	1	1.32	1.32	3711.0	3711	NA	NA
BRKGA	tb5	10	0.34	0.34	523.0	523	0.0026	4.445
GA	tb5	10	0.34	0.34	523.0	523	0.0563	98.396
RS	tb5	10	0.34	0.34	523.0	523	0.0000	0.016
ss	tb5	10	0.34	0.34	526.8	524	0.3553	99.313
VNS	tb5	10	0.34	0.34	524.0	524	0.0000	3.260
k-means	tb5	10	0.35	0.35	523.0	523	0.0000	0.001
percentile	tb5	1	0.39	0.39	523.0	523	NA	NA
BRKGA	ghs24	10	0.49	0.49	2051.0	2051	0.0135	41.053
GA	ghs24	10	0.50	0.49	2051.0	2051	1.1307	115.453
RS	ghs24	10	0.49	0.49	2051.0	2051	1.0322	0.018
ss	ghs24	10	0.49	0.49	2066.1	2056	0.3826	322.356
VNS	ghs24	10	0.49	0.49	2064.0	2052	0.8588	36.747
k-means	ghs24	10	1.39	1.38	2051.0	2051	0.3555	0.004
percentile	ghs24	1	3.98	3.98	2051.0	2051	NA	NA
BRKGA	ghs23	10	0.49	0.49	2055.0	2055	0.0954	37.871
GA	ghs23	10	0.51	0.49	2055.0	2055	3.7226	115.604
RS	ghs23	10	0.50	0.49	2055.0	2055	0.5881	0.019
ss	ghs23	10	0.49	0.49	2072.2	2056	0.6923	322.032
VNS	ghs23	10	0.49	0.49	2067.5	2056	0.8300	42.906
k-means	ghs23	10	1.76	1.76	2055.0	2055	0.0000	0.004
percentile	ghs23	1	4.61	4.61	2055.0	2055	NA	NA

R: number of runs. $\bar{cv}$%: average percentage coefficient of variation across runs. cv_{min}%: minimum percentage CV achieved. $\bar{n}$: average sample size across runs. n_{min}: minimum sample size. ME%: percentage margin of error. $time$: average runtime in minutes.

Table 6. Sample size required (n_{trad}) for traditional methods to reach BRKGA precision (cv_{brkga}%).

Method	Data	R	cv_{brkga}%	$n_{10\%}$	n_{trad}
k-means	ecd	10	0.37	3711	8400
percentile	ecd	1	0.37	3711	20250
k-means	tb5	10	0.34	523	540
percentile	tb5	1	0.34	523	640
k-means	ghs24	10	0.49	2051	10300
percentile	ghs24	1	0.49	2051	18000
k-means	ghs23	10	0.49	2055	11800
percentile	ghs23	1	0.49	2055	18575

R: number of runs. cv_{brkga}: minimum cv% achieved by computational methods. $n_{10\%}$: 10% sample size used to generate earlier computational results. n_{trad}: sample size required to achieve cv_{brkga}%.

time frame (without the need of a high performance machine), suggesting that these methods can be applied by users with a conventional computing set up.

The traditional approaches, percentile and k-means, had almost negligible runtime requirements. The percentile-based approach is deterministic and runtime is not reported here as it is the time taken for R to execute the quantile function. k-means, however, is algorithmic and displays high computational efficiency, with the shortest runtime across all algorithmic methods used.

Based on the findings across this selection of South African datasets, researchers may want to consider incorporating a novel method such as BRKGA especially when working with highly skewed income or fee stratifiers.

6.2 Resulting sample: Computational vs. traditional methods

In Table 7, we provide an overview of strata boundaries and sample allocation resulting from the best result achieved for each method. It can be observed that boundaries for the TB5 data are more equally spaced regardless of method, likely because of the relative normality of the distribution. Conversely, the skewed datasets - ECD, GHS23, GHS24 - have potentially less intuitive cut points for the novel methods to make the strata more internally homogeneous.

For the GHS surveys across novel methods, the first boundary is placed close to the median (see Figure 1 for the median values), indicating that almost 50% of the distribution are placed in the first stratum. This result reflects less variation among those at the lower end of the income distribution, increasing only at higher percentiles. The ECD fee data, slightly less skewed, sees the median of the distribution (300) fall between the first and second boundaries, also showing that more variation between fee groups is found above the median, where boundaries 2, 3 and 4 are located.

It is further seen that optimal sample allocation for novel methods follows neither proportional nor equal allocation. For GHS, although the majority of the population are situated in the first stratum,

more sample is allocated to the fifth stratum. This outcome follows the optimal formula of Neyman (1934), with sample allocation proportional to population share as well as variance. Since more variation is located in the fifth stratum, more sample is needed in order to get a precise estimate of $\bar{y}_{st}$.

The differences in stratification outcome between novel methods and traditional methods are more clearly depicted in Figures 2 and 3. We use the BRKGA as the novel method and compare the resulting stratification and allocation to the percentile-based approach and the k-means clustering solution with proportional allocation. In Figure 2, we see that BRKGA allocates the largest sample size to the fifth stratum, in contrast to the traditional methods. The stratum boundary positioning is largely different too, with stratum 3, 4 and 5 determined by BRKGA spanning almost the same section of the distribution as the fifth stratum under the percentile method. Figure 3 shows the stratification for the more normally distributed TB5 ELOM data. Here, the strata allocation and boundary determination is more similar across methods, although BRKGA prioritises more sample allocation in the first and final strata, having the more variation rather than a greater portion of the distribution.

7. Conclusion and areas for future work

In this work, novel computational methods have been evaluated across various scenarios in South African research contexts, particularly those of child outcomes, ELP fees and household income. We find that the novel methods – RS, GA, GGA, BRKGA and VNS – arrive at very similar optimisation results, with the BRKGA method of Brito et al. (2019) leading the group marginally. In contrast, more traditional approaches used for stratified sampling in South Africa did not reach similar levels of precision for a given sample size. These differences were especially pronounced for highly skewed data and less so for the relatively normally distributed TB5 ELOM data. Across skewed datasets, we demonstrate that in order to reach the same level of precision as that achieved by novel methods, traditional approaches would generally require a total sample size between 2.5-8 times larger than novel methods. As highly skewed income data is common in the South African context, novel methods are therefore recommended for stratification of such variables, especially when the survey goal is to estimate the stratification variable with a high level of precision.

While this study is the first to quantitatively evaluate all listed novel methods concurrently and in the South African context, limitations include the fact that only univariate scenarios are considered. While there is a growing literature on multivariate stratification, including dimension reduction techniques, compromise allocation and mathematical programming, computational methods for public implementation remain limited (Mulvey, 1983; Bethel, 1985; Pla, 1991; Díaz-García and Garay-Tápia, 2007; Khan et al., 2010; Iftekhar et al., 2015; Ballin and Barcaroli, 2020; Alshqaq et al., 2022; Borros et al., 2025). Out of the methods evaluated, only the `SamplingStrata` package currently supports multivariate stratification, an area for future research. Costs are an additional element to examine. In the formulation of the stratification problem the survey cost is either used as the object of minimisation or as the constraint factor. In the literature to date, n is often used as the measure of cost and in the case where a cost function is included it is assumed to be linear (Kokan and Khan, 1967; Brito et al., 2015; Barcaroli, 2015). While these assumptions are likely to hold in most cases, further research into the nature of the cost could have a substantial impact on

Table 7. Strata boundaries and sample size allocation by method and dataset

Method	**Data**	b_1	b_2	b_3	b_4	n_1	n_2	n_3	n_4	n_5
RS	ecd	225.5	475.0	1039.5	2168.0	705	594	664	557	1191
BRKGA	ecd	223.2	476.7	1047.6	2165.5	704	595	665	557	1190
GA	ecd	230.0	525.0	1120.0	2235.0	726	838	492	544	1111
VNS	ecd	220.0	470.0	1034.0	2170.0	705	594	664	559	1189
`ss`	ecd	226.0	450.0	1055.0	2160.0	706	586	696	534	1192
percentile	ecd	120.0	220.0	350.0	600.0	700	782	666	804	759
k-means	ecd	315.0	890.0	1870.0	3300.0	2229	1015	254	153	60
RS	tb5	30.5	40.5	50.0	60.9	119	95	93	92	124
BRKGA	tb5	30.4	40.4	50.0	60.9	119	95	93	92	124
GA	tb5	30.4	40.5	50.0	60.9	118	96	94	92	123
VNS	tb5	30.2	40.5	50.3	61.1	115	100	98	90	121
`ss`	tb5	30.5	40.4	50.4	61.0	121	94	100	86	123
percentile	tb5	32.2	40.3	48.2	57.7	105	104	105	104	105
k-means	tb5	29.0	40.5	51.5	63.7	72	139	141	113	58
RS	ghs24	4649.5	11613.5	25716.5	51372.5	341	323	341	293	753
BRKGA	ghs24	4633.4	11598.8	25792.9	51660.2	328	327	347	298	751
GA	ghs24	4650.0	11680.0	26500.0	58154.0	351	334	370	391	605
VNS	ghs24	4608.0	11560.0	25500.0	50124.0	337	323	335	283	774
`ss`	ghs24	4547.0	11640.0	26570.0	53058.0	340	331	366	287	732
percentile	ghs24	2008.0	3980.0	6790.0	14554.4	408	404	418	411	410
k-means	ghs24	16220.0	48427.0	113990.0	370583.0	1679	288	66	15	3
RS	ghs23	4658.5	11588.5	25879.0	51845.0	370	333	354	282	716
BRKGA	ghs23	4660.4	11730.4	26011.9	52316.9	359	342	368	275	711
GA	ghs23	4657.0	11630.0	26450.0	56287.0	378	342	376	339	620
VNS	ghs23	4522.0	11525.0	25820.0	51245.0	363	337	354	279	723
`ss`	ghs23	4690.0	11540.0	25908.0	53762.0	375	331	359	301	690
percentile	ghs23	2000.0	3950.0	6500.0	13990.0	394	427	411	411	412
k-means	ghs23	18246.0	61418.0	191304.0	900000.0	1734	266	50	3	2

Columns $b_1, \ldots, b_4$ indicate the boundary cut points for each of the 5 strata, while columns $n_1, \ldots, n_5$ indicate the allocation of sample across strata.

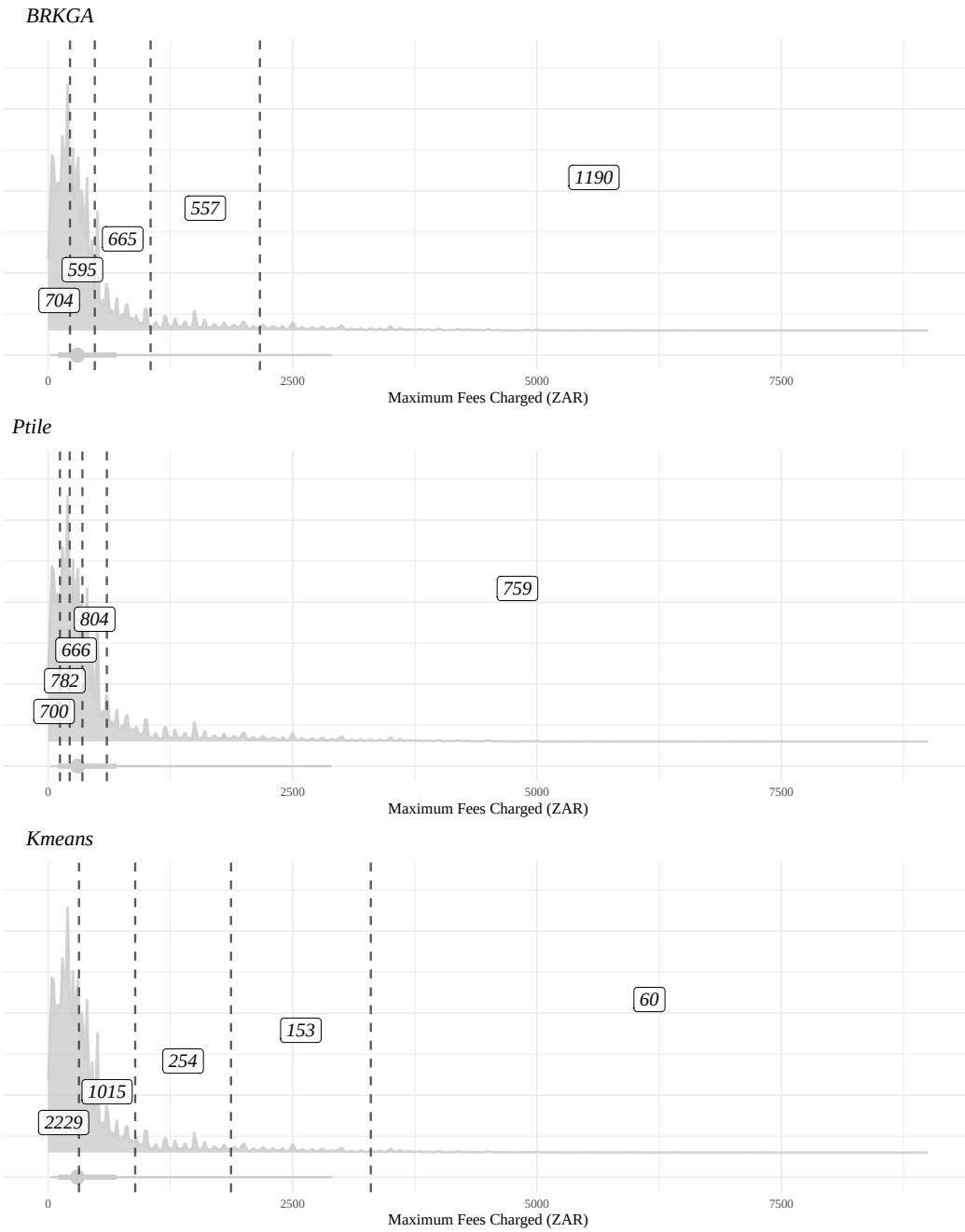

Figure 2. Comparison of boundaries and sample allocation using BRKGA, percentile and k-means methods for ECD Census 2021 Fee Data. Dashed lines indicate stratum boundaries. Sample allocation per stratum is given inside the rectangular labels.

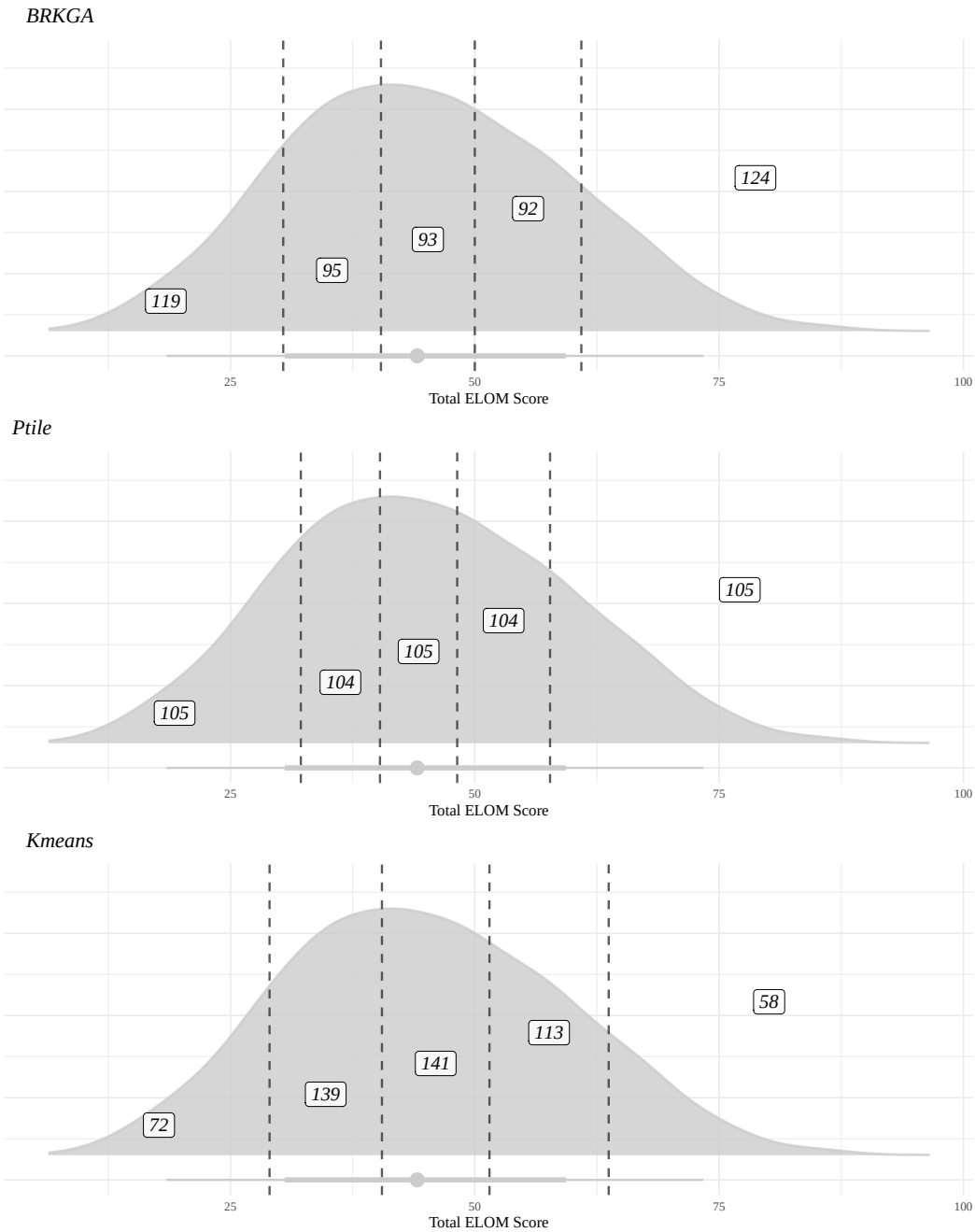

Figure 3. Comparison of boundaries and sample allocation using BRKGA, percentile and k-means methods for TB5 2021 ELOM Data. Dashed lines indicate stratum boundaries. Sample allocation per stratum is given inside the rectangular labels.

the optimisation process. Finally, spatial data are due for consideration under stratified sampling, as sampling frames can include this data type (Cajka et al., 2018). Without appropriate consideration of autocorrelation when using spatial data, variation will be misspecified (Getis, 2007). There are currently few prominent spatial stratification packages available in R, such as `SamplingStrata` and `rassta` (Ballin and Barcaroli, 2020; Fuentes et al., 2022).

As such, this study has provided a comprehensive evaluation of methods for stratified sampling in the South African ECE and household survey contexts, with areas for expansion including multivariate methods, survey cost and spatial data types.

References

Alshqaq, S., Ahmadini, A., and Ali, I. (2022). Nonlinear stochastic multiobjective optimization problem in multivariate stratified sampling design. *Mathematical Problems in Engineering*, **2022**, 2502346. doi:https://doi.org/10.1155/2022/2502346.

Anderson, K. J., Henning, T. J., Moonsamy, J. R., Scott, M., du Plooy, C., and Dawes, A. R. L. (2021). Test-retest reliability and concurrent validity of the South African Early Learning Outcomes Measure (ELOM). *South African Journal of Childhood Education*, **11**, a881. doi: https://doi.org/10.4102/sajce.v11i1.881.

Ballin, M. and Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, **39** (2), 369–394.

Ballin, M. and Barcaroli, G. (2020). R package SamplingStrata: New developments and extension to spatial sampling. *arXiv*. doi:https://doi.org/10.48550/arXiv.2004.09366.

Barcaroli, G. (2015). Optimization of sampling strata with the SamplingStrata package. *Istituto Nazionale di Statistica*.

Barcaroli, G. and Ballin, M. (2018). R packages for optimal stratified sampling: A review and compared evaluation.
URL: *https://www.researchgate.net/publication/327791484_R_packages_for_optimal_stratified_sampling_a_review_and_compared_evaluation*

Barcaroli, G., Ballin, M., Pagliuca, D., Willighagen, E., and Zardetto, D. (2018). SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys. R Package.
URL: *https://barcaroli.github.io/SamplingStrata/*

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. Technical report, US Department of Agriculture, Statistical Reporting Service, Statistical Research Division.

Borros, G., Er, S., and Salau, S. (2025). Integrating PCA with random search for variable importance in multivariate stratified sample allocation. *In Annual Proceedings of the South African Statistical Association Conference*. Stellenbosch, South Africa, 17–32.

Bowley, A. J. (1926). Measurement of the precision attained in sampling. *International Statistical Institute*, **Bulletin 22** (part 1:), 6–62.

Brito, J., De Lima, L., Henrique González, P., Oliveira, B., and Maculan, N. (2021). Heuristic approach applied to the optimum stratification problem. *RAIRO - Operations Research*, **55**, 979–996.

Brito, J., Silva, P., Semaan, G., and Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, **41**, 427–442.

Brito, J., Silva, P., and Veiga, T. (2017). Package 'stratbr': Optimal Stratification in Stratified Sampling. R Package.
URL: *https://cloud.r-project.org/web/packages/stratbr/stratbr.pdf*

Brito, J., Veiga, T., and Silva, P. (2019). An optimisation algorithm applied to the one-dimensional stratification problem. *Survey Methodology*, **45**, 295–315.

Cajka, J., Amer, S., Ridenhour, J., and Allpress, J. (2018). Geo-sampling in developing nations. *International Journal of Social Research Methodology*, **21**, 729–746.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, New York, NY.

Dalenius, T. (1950). The problem of optimum stratification. *Scandinavian Actuarial Journal*, **1950**, 203–213.

Dalenius, T. and Hodges, J. L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, **54**, 88–101.

DataDrive and Department of Basic Education (2022). Thrive by Five Index and ECD Baseline Audit 2021 [dataset]. Cape Town: DataFirst [distributor]. Cape Town and Pretoria: DataDrive and Department of Basic Education [producer].

De Lima, L., Brito, J., Gonzalez, P., and Oliveira, B. (2020). Package 'stratvns': Optimal stratification in stratified sampling. R Package.
URL: *https://www.rdocumentation.org/packages/stratvns/versions/1.1*

Department of Basic Education (2022). Early Childhood Development Census 2021 [dataset]. Cape Town: DataFirst [distributer]. Pretoria: Department of Basic Education [producer].

Díaz-García, J. A. and Garay-Tápia, M. M. (2007). Optimum allocation in stratified surveys: Stochastic programming. *Computational Statistics & Data Analysis*, **51**, 3016–3026.

Er, S. (2011). Computational methods for optimum stratification: A review. *In Proceedings of the 58th World Statistical Congress*. International Statistical Institute, Dublin, Ireland, 3304–3312.

Er, S., Keskintürk, T., and Daly, C. (2010). Package 'GA4Stratification': A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. R Package.
URL: *https://rdrr.io/cran/GA4Stratification/*

Francis, D. and Webster, E. (2019). Poverty and inequality in South Africa: Critical reflections. *Development Southern Africa*, **36**, 788–802.

Fuentes, B. A., Dorantes, M. J., Tipton, J. R., Hijmans, R. J., and Brown, A. G. (2022). rassta: Raster-based spatial stratification algorithms. R pacakge.
URL: *https://cran.r-project.org/web/packages/rassta/rassta.pdf*

Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*, **37**, 491–496.

Giese, S., Dawes, A., Tredoux, C., Mattes, F., Bridgman, G., van der Berg, S., Schenk, J., and Kotzé, J. (2022). Thrive by Five index report revised August 2022. Technical report, Innovation Edge, Cape Town.

Gunning, P. and Horgan, J. M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Statistics Canada*, **30**, 159–166.

Hansen, P. and Mladenović, N. (2001). Variable neighborhood search: Principles and applications. *European Journal of Operational Research*, **130**, 449–467.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **28**, 100–108.

Henry, J. and Giese, S. (2023). Reviewing the Socio-Economic Gradient in Learning Outcomes for Children who Participated in the Thrive by Five Index. Technical Paper, DataDrive, Cape Town.

Horgan, J. M. (2010). Choosing the stratification boundaries: The elusive optima. *Istanbul University Journal of the School of Business Administration*, **39**, 195–204.

Iftekhar, S., Ahsan, M., and Mazhar, A. Q. (2015). An optimum multivariate stratified sampling design. *Research Journal of Mathematical and Statistical Sciences*, **3**, 10–14.

Kerr, A., Ardington, C., and Burger, R. (2020). Sample design and weighting in the NIDS-CRAM survey. *SALDRU Working Papers*, **267**.
URL: *http://hdl.handle.net/11090/983*

Keskintürk, T. and Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, **52**, 53–67.

Khan, M., Maiti, T., and Ahsan, M. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, **26**.

Khan, M., Nand, N., and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey methodology*, **34**, 205–214.

Khan, M. G. and Sharma, S. (2015). Determining optimum strata boundaries and optimum allocation in stratified sampling. *Aligarh Journal of Statistics*, **35**, 23–40.

Kokan, A. R. and Khan, S. (1967). Optimum allocation in multivariate surveys: an analytical solution. *Journal of the Royal Statistical Society. Series B (Methodological)*, **29**, 115–125.

Kozak, M. (2004a). Method of multivariate sample allocation in agricultural surveys. *In Colloquium Biometryczne*, volume 34. 241–250.

Kozak, M. (2004b). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, **6**, 797–806.

Kozak, M. (2014). Comparison of random search method and genetic algorithm for stratification. *Communications in Statistics - Simulation and Computation*, **43**, 249–253.

Kozak, M. and Verma, M. R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, **32**, 157–164.

Kozak, M., Verma, M. R., and Zieli, A. (2007). Modern approach to optimum stratification: review and perspectives. *Statistics in Transition*, **8**, 223–250.

Lavallée, P. and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey methodology*, **14**, 33–43.

Maremba, T. A. (2019). *Computation of optimal estimates in a complex survey sample design*.

Master of Science in Statistics, University of Limpopo.

Mdluli, P. and Dunga, S. (2022). Determinants of Poverty in South Africa Using the 2018 General Household Survey Data. *Journal of Poverty*, **26**, 197–213.

Mulvey, J. M. (1983). Multivariate stratified sampling by optimization. *Management Science*, **29**, 715–724.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558–625.

O'Luing, M. (2022). *Metaheuristics and machine learning for joint stratification and sample allocation in survey design*. Ph.D. thesis, University College Cork.

O'Luing, M., Prestwich, S., and Tarim, S. A. (2018). A grouping genetic algorithm for joint stratification and sample allocation designs. *arXiv*. doi:https://doi.org/10.48550/arXiv.1709.03076.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, **47**, 1409.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reddy, K. G. and Khan, M. G. M. (2016). Optimal stratification of univariate populations via stratifyR Package. *In JSM Proceedings:, Statistical Computing Section*. Alexandria, VA: American Statistical Association. 1121-1131.

Reddy, K. G. and Khan, M. G. M. (2020). stratifyR: An R package for optimal stratification and sample allocation for univariate populations. *Australian & New Zealand Journal of Statistics*, **62**, 383–405.

Rivest, L.-P. and Baillargeon, S. (2007). Stratification: Univariate stratification of survey populations. R pacakge.

Sapra, R. L. (2022). How to calculate an adequate sample size? *In How to practice academic medicine and publish from developing countries? A practical guide*. Springer Nature, Singapore, 81–93.

Snelling, M., Dawes, A., Biersteker, L., Girdwood, E., and Tredoux, C. (2019). The development of a South African Early Learning Outcomes Measure: A South African instrument for measuring early learning program outcomes. *Child: Care, Health and Development*, **45**, 257–270.

Statistics South Africa (2002). Sampling Methodology for Economic Statistics. Technical report, Statistics South Africa, Republic of South Africa, Pretoria, South Africa.

Statistics South Africa (2024). General Household Survey 2023 [dataset]. Cape Town: DataFirst [distributor]. Pretoria: Statistics South Africa [producer].

Statistics South Africa (2025a). General Household Survey 2024. Statistical Release P0318, Statistics South Africa, Republic of South Africa, Pretoria.

Statistics South Africa (2025b). General Household Survey 2024 [dataset]. Cape Town: DataFirst [distributor]. Pretoria: Statistics South Africa [producer].

Tschuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observation. *Metron*, **2**, 461–500.

A. Additional results

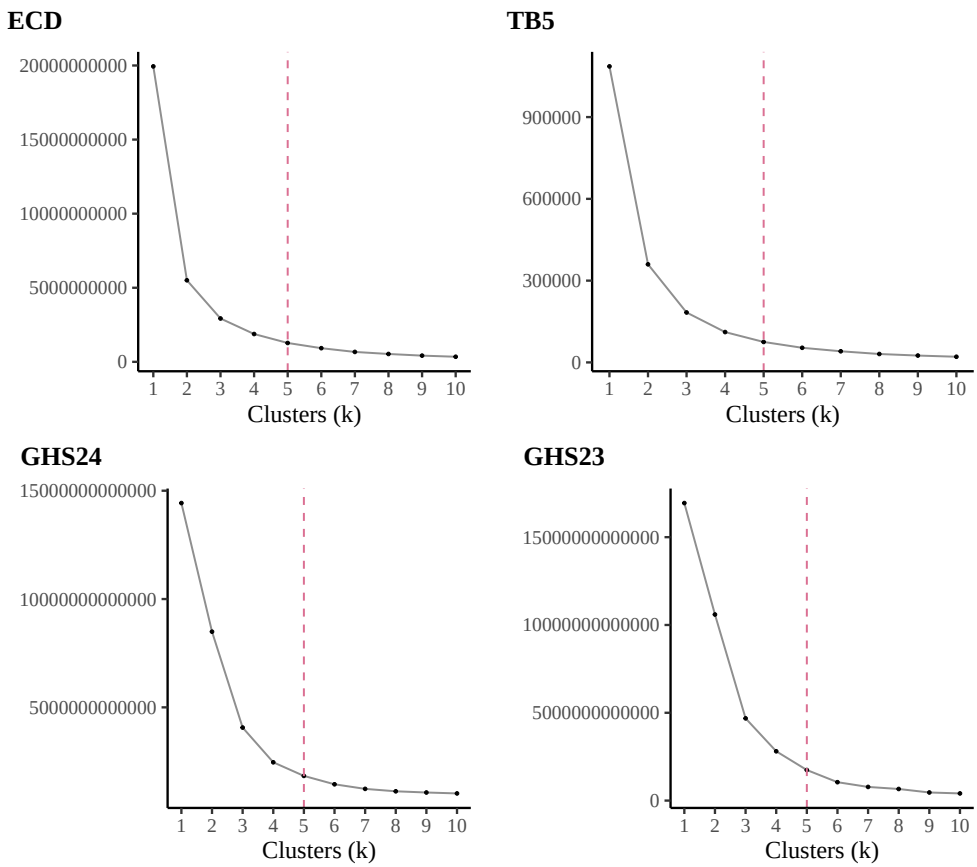

Figure 4. *k*-means elbow plot of total within sum of squares, by dataset.

Manuscript received 2025-08-21, revised 2026-02-23, accepted 2026-02-23.