

Using Concomitant Information in Designing Cost-Effective Environmental Sampling

G. D. Johnson¹, B. D. Nussbaum², G. P. Patil¹ and N. P. Ross²

¹Center for Statistical Ecology and Environmental Statistics
Department of Statistics
Pennsylvania State University
University Park, PA 16802

²Environmental Statistics and Information Division
Office of Policy, Planning, and Evaluation
United States Environmental Protection Agency
Washington, DC 20460

Prepared with partial support from the Statistical Analysis and Computing Branch, Environmental Statistics and Information Division, Office of Policy, Planning, and Evaluation, United States Environmental Protection Agency, Washington, DC under a Cooperative Agreement Number CR-821531. The contents have not been subjected to Agency review and therefore do not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Introduction

Consider a hazardous waste site that is contaminated with polychlorinated biphenyls (PCB's), a class of anthropogenic chemicals that were once widely used in industry but have since been banned due to evidence of toxicity and cancer causing ability in mammals. A property that made PCB's desirable, namely their inherent stability, also causes increased concern over health risks because these chemicals persist in the environment and accumulate in fat tissue.

In order to evaluate the risk posed by such a site, investigators obtain a sample of 60 soil core measurements to estimate the average PCB concentration in order to compare to a risk-based standard. Now consider that the mean could have been estimated with equal precision using only 26 measurements. Such a favorable economic scenario can potentially be realized with an approach that exploits any suitable concomitant information about the variable of interest (PCB concentration), thus allowing judicious selection of soil cores for actual measurement. With the very high cost of measuring organic chemical concentrations in soil, site investigators should certainly welcome a sampling design that reduces the demand on analytical chemistry.

In general, responding to concerns about pollution effects on ecological and human health, along with the needs of natural resource management requires a large amount and wide variety of environmental monitoring. While applications range from characterizing soil contamination at hazardous waste sites to regional assessments of forest inventories, a common factor which limits the amount of sampling in all these cases is the analytical cost—the cost of obtaining the actual measurements on sample units. Meanwhile, investigators need data of sufficient quantity and quality for defending decisions that often balance the risk of undue liability with the risk of adverse human and ecological health impacts.

When data demands conflict with budgetary constraints, we need to devise sampling plans for estimating population features that maintain a required level of precision while reducing the number of sample units that require expensive and/or destructive measurement. Sampling innovation that helps achieve these otherwise conflicting goals is what we call “observational economy”, which is obtainable when the identification and acquisition of sample units are relatively much less expensive than final measurements on the sample units.

Now let us share a particularly appealing method for exploiting any read-

ily available concomitant information about the variable of interest that can help in selecting a sample that is more representative of the actual population. Say, for example, we wish to estimate the mean height of students at a university from a random sample of three students. Furthermore, in order to acknowledge the uncertainty of such an estimate, we need to include a confidence interval.

The simplest way to obtain our sample is to randomly select three students from the university's population, then measure their heights. While the arithmetic average of the three heights is an unbiased point estimate of the population mean, the associated confidence interval can be very large, reflecting the high degree of uncertainty with estimating the mean of a very large population with only three measurements. We may happen to grab two very short people and one very tall; or we may grab three very tall people or three very short people. The only way to overcome such a problem with a simple random sample (SRS) is to increase the sample size.

For another option, consider a ranked set sample (RSS). To do this, we randomly invite three students to breakfast and visually rank them with respect to height. We then select the student we believe is shortest and actually measure his or her height. Repeating this process with lunch, we then select

the middle ranked person, and, as such, select the tallest ranked person at dinner. The resulting three measurements of student heights constitute a ranked set sample. As with the SRS measurements, the arithmetic average of the RSS measurements provides an unbiased point estimate of the population mean; however, the associated confidence interval can potentially be much smaller than that obtained with SRS measurements, thus reflecting decreased uncertainty. This encouraging feature results because measurements obtained through RSS are likely to be more regularly spaced than those obtained through SRS, as visualized in Fig. 1, and therefore are more representative of the population.

The RSS procedure induces stratification of the whole population at the sample level; in effect, we are randomly sampling from the subpopulations of predominantly short, medium and tall students without having to construct the subpopulation strata. Each subpopulation has its own distribution, as visualized in Fig. 2, where we see how the parent population gets effectively partitioned into subpopulations. If we are terrible at assessing height, the worst result is a simple random sample.

Ranked set sampling is a form of double sampling, whereby an initial large sample of the population is taken for measuring a concomitant variable,

followed by obtaining a smaller sample for measuring the more expensive variable of interest. The second, smaller, sample is often a subsample of the initial larger sample.

More conventional double sampling estimators require the concomitant variable to be quantitative, whether numerical or categorical. One approach, known as double sampling for stratification, is to use values of a concomitant variable to stratify the population into more homogeneous strata prior to obtaining a random sample of the variable of interest within each stratum. Other approaches, like double sampling for regression or for ratio estimation, use values of a concomitant variable to adjust the estimator based on the variable of interest.

Besides requiring that the concomitant variable be quantitative, some of the conventional double sampling methods also depend on assumptions about how the variable of interest and the concomitant variable are jointly distributed.

In 1952, G. A. McIntyre recognized the value of incorporating purely judgmental, albeit expert, opinion into designs for sampling pastureland. He proposed a robust, non parametric method that was later coined ranked set sampling (RSS). For a historical development of RSS, see Patil, Sinha and

Taillie (1994). RSS can also utilize quantitative or categorical concomitant variables if they are available, and often still performs better than other double sampling estimators.

Method of Ranked Set Sampling

In order to create ranked sets, we must partition the large first phase sample into sets of size m which are typically small, around 3 to 5, to minimize ranking error. Thus, we proceed as follows:

- step 1: Randomly select m^2 sample units from the population.
- step 2: Allocate the m^2 selected units as randomly as possible into m sets, each of size m .
- step 3: Without yet knowing any values for the variable of interest, rank the sample units within each set based on a perception of relative values for this variable. This may be based on personal judgment or done with measurements of a covariate that is correlated with the variable of interest.
- step 4: Choose a sample for actual analysis by including the smallest ranked unit in the first set, then the second smallest ranked unit in the

second set, continuing in this fashion until the largest ranked unit is selected in the last set.

- step 5: Repeat steps 1 through 4 for r cycles until the desired sample size, $n = mr$, is obtained for analysis.

As an illustration, consider the set size $m = 3$ with $r = 4$ cycles. This situation is illustrated in Fig. 3 where each row denotes a judgment-ordered sample within a cycle, and the units selected for quantitative analysis are circled. Note that 36 units have been randomly selected in 4 cycles; however, only 12 units are actually analyzed to obtain the ranked set sample of measurements.

Obtaining a sample in this manner results in maintaining the unbiasedness of simple random sampling; however, by incorporating “outside” information about the sample units, we are able to contribute a structure to the sample that increases its representativeness of the true underlying population.

If we quantified the same number of sample units, $mr = 12$, by a simple random sample, we have no control over which units enter the sample. Perhaps all the 12 units would come from the lower end of the range, or perhaps

most would be clustered at the low end while one or two units would come from the middle or upper range.

Unequal Allocation

For populations that are known or suspected to have very skewed distributions, the number of sample units allocated into each rank may be proportional to the expected variance within each rank, as with Neyman allocation for optimal stratified random sampling designs. For the same number of measurements chosen in the previous example using equal allocation, $mr = 12$, unequal allocation may result in a design like in Figure 4.

Ranking Criteria

A real key to success lies with step 3 in the above procedure—ranking. This may be based on visual inspection or other expert opinion about the sample units. For example, a field-seasoned range scientist or forester may readily be able to rank three or four quadrats of grass with respect to overall volume or mass. Meanwhile a hazardous waste site inspector may be able to reliably rank areas of soil with respect to concentrations of a toxic contaminant, based on features like surface staining, discoloration or the appearance of stressed

vegetation.

More importantly, if another variable is available that is highly correlated with the variable of interest but costs much less to obtain, then we may rank by the values of such a covariate. For example, reflectance intensity of near-infrared electromagnetic radiation, as recorded in a remotely sensed digital image, is directly proportional to vegetation concentration on the ground. Another example might be to measure total organic halides (TOX) in soil in order to rank soil sampling units with respect to the concentration of volatile organic solvents. As an indicator variable, TOX is much less expensive to measure than specific organic compounds.

Robustness of the Procedure

Several questions may now arise, such as:

1. What if the distribution of sample measurements is not uniform or near-uniform? or essentially unknown?
2. What if the sample units are not randomly allocated into sets?
3. How does error in ranking affect results?

First of all, while independent (random) and identically distributed sample measurements obtained through perfect ranking may lead to *optimum* performance of ranked set sampling, no matter how much these desirable characteristics are deviated from, the sampling efficiency will never be worse than with simple random sampling, using the same number of quantifications. In fact, when efficiency is expressed as the relative precision (RP) such that

$$RP = \frac{\text{variance of sample average with simple random sampling}}{\text{variance of sample average with ranked set sampling}},$$

it can be shown that the bounds of this relative precision are

$$1 \leq RP \leq (m + 1)/2,$$

where m is the set size. Since RP can not be less than one, the RSS protocol can not be worse than the simple random sampling protocol. (see Patil, Sinha and Taillie, 1994).

Unequal allocation can actually increase the performance of RSS above and beyond that achievable with standard equal allocation; however, if not properly applied, the performance of RSS can be worse than the performance

of simple random sampling. Actually, the bounds on relative precision with unequal allocation become

$$0 \leq RP \leq m.$$

indicating that, with appropriate unequal allocation, the relative precision may even increase to a level of m , and not just $(m + 1)/2$ as in the case with equal allocation.

Although an optimal RSS design would allocate samples into ranks in direct proportion to the rank standard deviations, we rarely know the standard deviations beforehand. We do know, however, that the distributions of many environmental and ecological variables are skewed towards the right, meaning that while most values are clustered around a median, a few much larger values are usually present. This skewness can actually be exploited to increase the precision beyond that obtained with ranked set sampling under equal allocation because standard deviations usually tend to increase with increasing rank values for right-skewed distributions.

Vegetation Research

Although ranked set sampling was formally proposed in 1952, apparently no applications were reported until 1966 when Halls and Dell discovered RSS to be considerably more efficient than SRS for estimating the weights of browse and herbage in a pine-hardwood forest of east Texas.

In their study, sets of three closely grouped quadrats were formed on a 300-acre tract. At select locations, metal frames of 3.1 square feet were placed at three randomly selected points within a circle of 13 foot radius as seen in Fig. 5. Quadrats were then ranked as lowest, intermediate and highest according to the perceived weight of browse and, separately, of herbage. Then, after clipping and drying, the separate weights of browse and herbage were determined for each quadrat. This was repeated for 126 sets for estimating browse and 124 sets for estimating herbage.

In order to simulate the SRS estimator for the mean weight of browse, one quadrat was randomly selected from each set without considering its rank. Since actual values were known for each quadrat, the RSS estimator was obtained by randomly choosing the ranks to be quantified for each set, resulting in 37 lowest ranks, 46 intermediate ranks and 43 highest ranks.

Halls and Dell also examined the effect of unequal allocation of sample units into sets. Since the standard deviations for the ranks were 7, 13 and 27.7 for the low, intermediate and high yield, respectively, (ratio of 1:2:4) they selected 14 quadrats in the low group, 40 in the intermediate group and 72 in the high group. Note that perfect ranking was obtained for both RSS protocols because the actual values were already known for each quadrat.

Results of these three sampling protocols are reported in Table 1. As expected under perfect ranking, precision due to RSS with approximately equal allocation increased, more than doubling for browse estimates. Furthermore, when allocation was proportional to the rank standard deviations, the precision increased still further, thus supporting McIntyre's contention.

Another very valuable aspect of this study was that two observers independently ranked the quadrats, one a professional range man and the other a woods worker. There was practically no difference in the ranking results between the two observers.

A 1967 master's thesis from Louisiana State University, School of Forestry and Wildlife Management showed that RSS increased the precision of sampling seedling counts of Longleaf Pine (*Pinus palustris*), when compared to simple random sampling.

Table 1: Summary statistics for browse and herbage estimates

	<u>browse</u>		<u>herbage</u>	
	mean	Variance of mean	mean	Variance of mean
Unranked: random	14.9	4.55	7.3	1.00
Perfect ranking: near equal allocation	13.2	2.18	7.0	.73
Perfect ranking: proportional allocation	12.9	1.91	7.2	.58

(Source: Halls and Dell (1966), "Trial of ranked set sampling for forage yields," *Forest Science*, 12,22-26.)

A 1985 study at Hurley (UK), compared RSS to simple random sampling for estimating herbage mass in pure grass swards and both herbage mass and clover content in mixed grass-clover swards. This study assessed the effects of the following factors on RSS: (i) imperfect ranking within sets, (ii) greater variation between sets than within sets, and (iii) asymmetric distribution of the quantified values.

The first two experiments were conducted by randomly selecting 15 locations, followed by randomly selecting three quadrats at each location and having several observers rank the quadrats within each set. For the last two experiments, 45 quadrats were drawn at random from the entire target area. This allowed an assessment of the effects of both spatial variation and

Table 2: Relative precisions (RP) \pm s.e. of the worst and the best observers, and under perfect ranking; and the between and the within set variances while estimating herbage mass (grass and mixture) and clover contents.

Experiments	Relative Precisions (R P)			Variances	
	Worst	Best	Perfect	Between	Within
1 (Grass)	1.11 \pm 0.09	1.23 \pm 0.14	1.31 \pm 0.17	0.24	0.31
2 (Mixture)	1.11 \pm 0.09	1.27 \pm 0.10	1.40 \pm 0.16	0.07	0.09
3 (Grass)	————	————	1.66 \pm 0.17	0.00	1.58
4 (Mixture)	1.36 \pm 0.14	1.51 \pm 0.15	1.55 \pm 0.16	0.11	0.66
2 (Clover)	1.15 \pm 0.12	1.34 \pm 0.15	1.44 \pm 0.16	16.3	34.4
4 (Clover)	1.36 \pm 0.19	1.62 \pm 0.18	1.72 \pm 0.20	16.2	71.6

(Source: Cobby, J.L., et al. (1985), "An investigation into the use of ranked set sampling on grass and grass-clover swards," Grass and Forage Science, 40, 257-263.)

ranking errors within sets.

Their results are reproduced in Table 2, where RP of both the worst and best observers are compared to the RP under perfect ranking, and the between and within set variances are presented for assessing spatial variation. These authors determined the main adverse factor to be within set clustering, and they recommend spacing quadrats within sets as far apart as possible when local spatial autocorrelation exists. With this in mind, they recommend RSS over SRS for sampling grass and grass-clover swards.

Combining with Line Intercept Sampling

A common field sampling method for ecological assessments is to include sample units that one encounters along a line (transect) that is randomly selected within a two-dimensional area of interest. Units are typically members of a plant or animal species.

Often the number of sample units identified are too numerous to select every one for quantification, especially if measurements are destructive, such as with cutting vegetation for weighing. If the initially identified sample units are treated as a larger first phase sample, $n' = m^2r$, then the RSS protocol can be applied to select a smaller subsample, $n = mr$, for actual quantification. For example, consider a single sampling cycle when the set size, m equals three for estimating the biomass of shrubs in a given area. A line transect for such a situation may be visualized as in Fig. 6.

Such an RSS-based line intercept sample has been found to produce more precise, and still unbiased, estimators of the population mean, size, total and cover, compared to the SRS-based line intercept sample.

In 1980, researchers were regularly employing the RSS technique at the Pastoral Research Laboratory, CSIRO at Armidale, N.S.W., Australia. A

plate with four holes is randomly thrown on a field and the pasture in each hole is ranked by eye, followed by selection of one hole for quantification of pasture. This application was mentioned by Yanagawa and Chen (1980), who also mention that RSS has been used to estimate rice crops in Okinawa, Japan.

PCB Contamination Levels

As part of a settlement between the state of Pennsylvania and a gas pipeline company, the company was required to excavate soil contaminated with polychlorinated biphenyls (PCB's) at 19 compressor stations across the state (see Fig. 7). Using initial site characterization data from the Armagh site, which was one of the worst, Patil, Sinha, and Taillie (1994) evaluated the effectiveness of RSS for estimating the average soil PCB concentration. Since sample values were already known, this study served to compare equal and unequal allocation schemes under perfect ranking. Although data from all sampling grids were very skewed, grids A and C were purposely chosen to represent considerably different degrees of skewness (grid A coeff. of skewness = 9.27, and grid C coeff. of skewness = 4.48). Grids A and C consisted of 184 and 68 observations, respectively.

For equal allocation of soil cores into ranks, the relative savings (RS) of RSS is reported in Table 3, where

$$RS = \frac{\text{variance of SRS average} - \text{variance of RSS average}}{\text{variance of SRS average}}$$

We see that RSS performed better for the sampling grid that was less skewed, and this was consistent for set sizes $m = 2, 3$ and 4.

For evaluating the effectiveness of unequal allocation into ranks, two different proportional allocations were considered for each set-size. This has been done to show the impact of proportional allocation on both relative precision and relative savings accrued due to RSS over simple random sampling. The results are given in Table 4, where the relative savings are seen to be quite substantial for all set sizes in both the grids. When compared to the relative savings in Table 3, proportional allocation appears to be far superior for such skewed data sets.

Now one can see how the illustrative example used to introduce this paper is entirely possible. A relative precision of 2.3 is reported in Table 4 for a set size of $m = 3$ with samples allocated to ranks in a 1:4:25 proportion for grid A. Since the relative precision is based on the same sample size, $n = 60$, for both estimators, the same precision obtained with a simple random sample

Table 3: Relative savings (RS) considering all possible combinations of each set size under perfect ranking situation with equal allocation.

Set Size (m)	Grid	
	A RS	C RS
2	.04	.09
3	.07	.16
4	.10	.22

(Source: Patil, Sinha and Taillie (1994))

of size 60 may be achieved with a ranked set sample of size 26.

Table 4: Values of the sample mean, $\bar{X}_{(m)u}$, relative precision (RP), and relative savings (RS) under the perfect ranking protocol with unequal allocation of sample units into ranks. The actual number of sample units allocated to each rank is provided in parenthesis below the theoretically optimum proportional allocation.

Set Size m		Grid							
		A				C			
		Proportion of samples (Exact No.)	$\bar{X}_{(m)u}$	RP	RS	Proportion of samples (Exact No.)	$\bar{X}_{(m)u}$	RP	RS
2		1:10 (8,84)	205.9	1.724	.42	1:10 (3,31)	535.2	2.041	.51
2		1:15 (6,86)	203.1	1.818	.45	1:15 (2,32)	520.4	2.174	.54
3		1:4:20 (2,10,48)	203.6	2.174	.54	1:1.7:1.5 (5,8,8)	560.1	1.471	.32
3		1:4:25 (2,8,50)	201.1	2.326	.57	1:2:7 (2,4,15)	615.2	1.923	.48
4		1:3:5:16 (2,5,9,28)	247.1	1.695	.41	1:2:3:4 (2,3,5,6)	576.6	2.083	.52
4		1:3:9:27 (2,2,10,30)	226.1	1.316	.24	1:1:3:5 (2,2,4,8)	802.4	1.449	.31

(Source: Patil, Sinha and Taillie (1994))

Using Geographic Information Systems to Exploit Auxiliary Data

With the availability of computerized Geographic Information Systems (GIS), ranking prospective sample locations across a landscape may be done rapidly prior to expensive field visits, thus allowing RSS to be applied to large scale surveys to obtain a more precise estimate at reduced cost. If prospective locations are selected at random from across a region and allocated to a set, then each location can be referenced to data “layers” in a GIS and, based on a derived ranking index, each member of the set can be ranked relative to each other.

Following a catastrophic event such as flooding or fire, those in charge of management and planning for natural or cultural resources need rapid assessments of the spatial extent and magnitude of damage. Such a situation may be well served by the combination of RSS and GIS, which can result in rapid mobilization of available information to design a very efficient field sampling strategy. This merger of GIS and RSS has been recommended by Myers, Johnson and Patil (1994).

Additional Reading

McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets. *Austral. J. Agricultural Res.* 3, 385-390.

Myers, W., Johnson, G. D., and Patil, G. P. (1994). Rapid mobilization of spatial/temporal information in the context of natural catastrophes. Invited paper presented at 1994 Spring Statistical Meetings in Cleveland, Ohio. American Statistical Association, Proceedings on Statistics and the Environment (to appear).

Patil, G. P., Sinha, A. K., and Taillie, C. (1994). Ranked set sampling. In *Handbook of Statistics, Volume 12: Environmental Statistics*, G. P. Patil and C. R. Rao, eds. North Holland/Elsevier Science Publishers, Amsterdam, 167-200.

Stokes, S. L., (1986). Ranked set sampling. In *Encyclopedia of Statistical Sciences* S. Kotz, et. al., (eds.). Wiley, New York, 585-588.

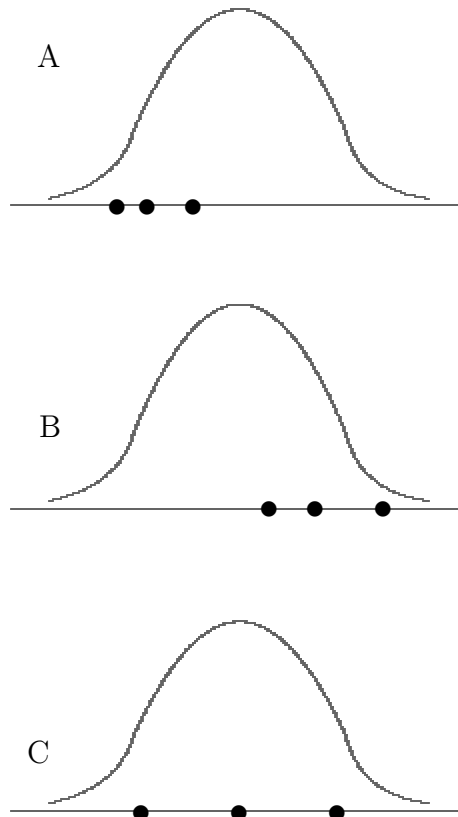


Figure 1: Possible location of three sample units along a frequency distribution. A and B show bunching up that may occur with simple random sampling (SRS), while C shows more of a regular spacing that results from ranked set sampling (RSS) when premeasurement ranking is accurate—a schematic diagram.

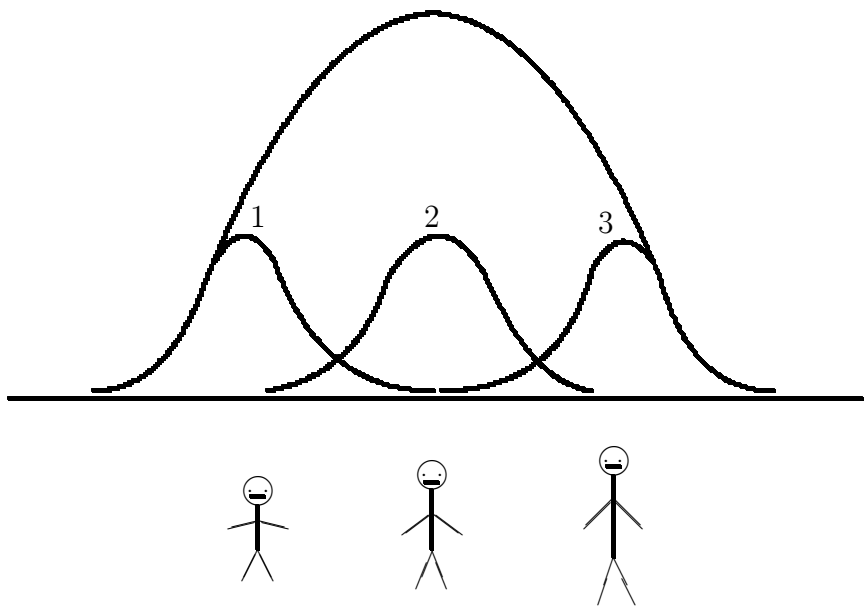


Figure 2: Frequency distributions of heights of different ranks superimposed on population frequency distribution of all heights—a schematic diagram.

cycle	rank		
	1	2	3
1	⊙	·	·
	·	⊙	·
	·	·	⊙
2	⊙	·	·
	·	⊙	·
	·	·	⊙
3	⊙	·	·
	·	⊙	·
	·	·	⊙
4	⊙	·	·
	·	⊙	·
	·	·	⊙

Figure 3: A ranked set sample design with set size $m = 3$ and the number of sampling cycles $r = 4$. Although 36 sample units have been selected from the population, only the 12 circled units are actually included in the final sample for quantitative analysis.

Sets	Units	No. of sets
1	$\odot \cdot \cdot$	3
2	$\odot \cdot \cdot$	
3	$\odot \cdot \cdot$	
4	$\cdot \odot \cdot$	4
5	$\cdot \odot \cdot$	
6	$\cdot \odot \cdot$	
7	$\cdot \odot \cdot$	
8	$\cdot \cdot \odot$	5
9	$\cdot \cdot \odot$	
10	$\cdot \cdot \odot$	
11	$\cdot \cdot \odot$	
12	$\cdot \cdot \odot$	

Figure 4: Ranked set sampling with unequal allocation: circles indicate sample units chosen for quantification.

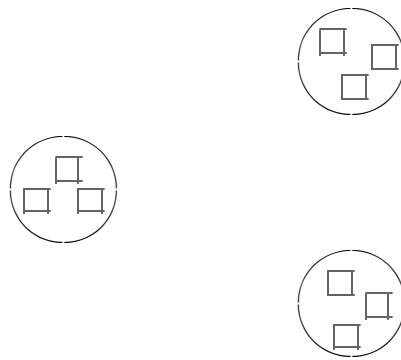


Figure 5: Within each circle, quadrats are randomly placed, followed by ranking and analysis of one appropriate quadrat. (not to scale)

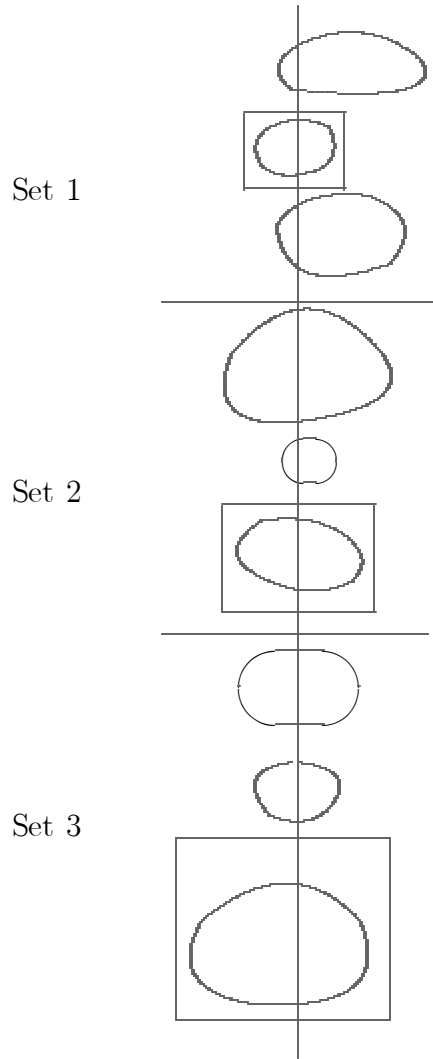


Figure 6: Aerial view of a line transect intercepting shrubs. For set size $m = 3$, nine shrubs are partitioned into 3 sets of 3. Using apparent shrub size for ranking with respect to biomass, the shrubs taken for analyses include the smallest ranked in the first set, the second smallest ranked in the second set and the largest ranked in the third set. (Source: Muttlack, H. A. and McDonald, L. L. (1992). Ranked set sampling and the line intercept method: A more efficient procedure. *Biometric Journal*, 34, 329-346.)

Figure 7: Above: Location of pipeline and compressor stations across Pennsylvania. Below: Part of the Armagh site that included sampling grids A and C that were used in this study. (courtesy of the Pennsylvania Department of Environmental Resources)