# AN APPLICATION OF NETWORK ANALYSIS TO A SAMPLE ALLOCATION PROBLEM

Judith C. Schwenk and Theodore S. Glickman
U.S. Department of Transportation, Transportation Systems Center

## INTRODUCTION

This paper describes the research involved in the use of network analysis to solve a sample allocation problem, motivated by the sample design for the U.S. Department of Transportation's annual General Aviation Activity and Avionics Survey. The paper begins with the presentation of the general sample allocation problem and its formulation as a network model. It continues with a technical discussion of the technique chosen for solution, the out-of-kilter algorithm, describing the general procedure and reporting results of various tests of sensitivity performed on the model. It concludes with some specific applications to the General Aviation Survey, and to another related sample allocation situation.

## THE PROBLEM

The need for a two-dimensional stratified sample design arises when the main products from a survey are estimates for a particular population characteristic classified by two independent criteria. In two-dimensional stratification the population is divided into mutually exclusive cells, and sample units are then allocated to the cells for the purpose of sample selection. The populations and sample sizes can be displayed in a two-way table as in Table 1.

| A \ B | Level 1 | Level 2 | --- | Level J | Total |
|---|---|---|---|---|---|
| Level 1 | $N_{11}$ $n_{11}$ | $N_{12}$ $n_{12}$ | --- | $N_{1J}$ $n_{1J}$ | $N_{1\cdot}$ $n_{1\cdot}$ |
| Level 2 | $N_{21}$ $n_{21}$ | $N_{22}$ $n_{22}$ | --- | $N_{2J}$ $n_{2J}$ | $N_{2\cdot}$ $n_{2\cdot}$ |
| ⋮ | ⋮ | ⋮ | ⋱ | ⋮ | ⋮ |
| Level I | $N_{I1}$ $n_{I1}$ | $N_{I2}$ $n_{I2}$ | --- | $N_{IJ}$ $n_{IJ}$ | $N_{I\cdot}$ $n_{I\cdot}$ |
| Total | $N_{\cdot 1}$ $n_{\cdot 1}$ | $N_{\cdot 2}$ $n_{\cdot 2}$ | --- | $N_{\cdot J}$ $n_{\cdot J}$ | $N$ $n$ |

I = number of levels for criterion A
J = number of levels for criterion B
$n_{ij}$ = cell sample size  $N_{ij}$ = cell population
$n_{i\cdot}$ = marginal sample size for level i of A  $N_{i\cdot}$ = marginal population for level i of A
$n_{\cdot j}$ = marginal sample size for level j of B  $N_{\cdot j}$ = marginal population for level j of B
$n$ = total sample size  $N$ = total population

TABLE 1

This paper focuses on the type of sample design where estimates are not required for populations within individual cells, but are required only for marginal populations. Nevertheless sampling units are allocated to individual cells prior to sample selection to insure that the desired marginal sample sizes are achieved.

Generally, the allocation of sample units to individual cells takes place in two steps. First, the marginal sample sizes are determined. Sometimes the marginal sizes are proportional to the populations of the rows and columns, as when 20% of the population is to be sampled and the marginal sample sizes are simply 20% of the marginal populations. At other times the design calls for disproportional sampling from the marginal populations, as when the marginal sample sizes are determined optimally according to the variances of the marginal populations.

The second step is the allocation of the marginal sample sizes across individual cells. In the case of proportional sampling, the same proportion used to determine the marginal sample sizes can be applied to the populations of individual cells to allocate the sample uniformly to the cells. In the case of disproportional or optimum sampling, cell allocation is not as straightforward. For example, suppose a 20% sample is required from a population, but with target marginal sample sizes as indicated in Table 2.

| A \ B | 1 | 2 | 3 | Total | |
|---|---|---|---|---|---|
| 1 | 0 / 0 | 20 / 20 | 130 / (20) | 150 / 40 | |
| 2 | 50 / 25 | 100 / (10) | 50 / 0 | 200 / 30 | ↑35 |
| Total | 50 / 25 | 120 / 30 | 180 / 15 | 350 / 70 | ↑75 |
| | | | ↑20 | | |

TABLE 2

Because of population constraints in individual cells, it is impossible to meet the minimum requirements for every marginal sample size without oversampling in some of the rows and columns. Cell (1,2), with a population of only 20, forces excessive units to be allocated to both row 2 and column 3. Cell (2,3) is excluded entirely from the sample to minimize oversampling.

In such a small matrix it is relatively simple to determine an allocation satisfying the marginal totals with mini-

mum oversampling, which is desirable when survey resources are limited. But as the matrix expands, the allocation process becomes more complicated, especially if the number of cells to be included in the sample is also of concern.

A general method for obtaining a feasible allocation when sampling disproportionally in a two-dimensional stratified sample design was developed using a network analysis formulation solved by the out-of-kilter algorithm, Ford and Fulkerson (4). Given marginal sample sizes, and population limits and minimum samples sizes for each cell, it guarantees that the resulting allocation will achieve the desired marginal totals with a minimum of oversampling. To apply the algorithm, the sample allocation problem must be formulated as a network model.

## NETWORK FORMULATION

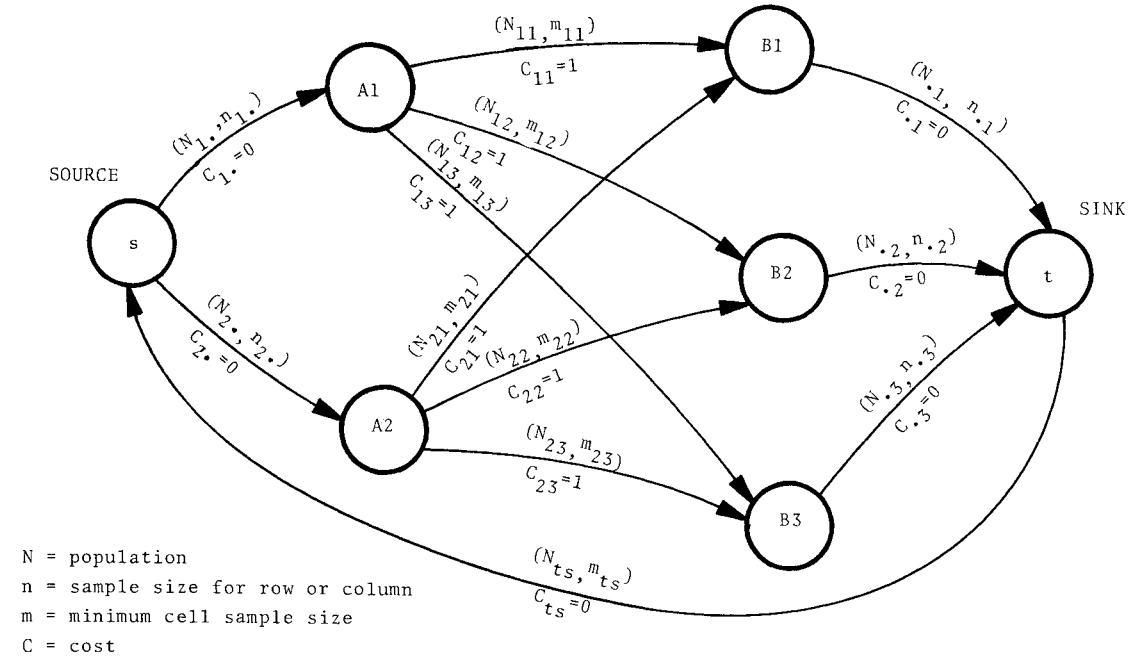
The minimum total sample size is treated as a particular flow which circulates in a closed system between two nodes: s, the source, and t, the sink. The intervening network structure consists of a set of arcs from s connected to respective row marginal nodes, connected in turn by cell arcs to respective column marginal nodes, leading finally through a set of arcs into t. One additional arc recirculates the flow from t back to s. When appropriate unit costs and capacity limitations are placed on each arc, the out-of-kilter algorithm finds the minimum cost flow, corresponding to the minimum total sample size. Flows on cell arcs then represent the allocation of the sample to the cells.

Figure 1 shows the network representation of the matrix. Each row i is allocated at least $n_{i.}$ but no more than $N_{i.}$; these are then the bounds $(N_{i.}, n_{i.})$ on the flow along the arc from the source to the ith row marginal node. Similarly for each column j. Each unit of flow on these source and sink arcs is given a cost $C_{i.}$ or $C_{.j}$ of zero because the row and column totals are free to vary within the established bounds. Each cell (i,j) is allocated at least $m_{ij}$ (zero in the Table 2 example) but no more than $N_{ij}$. Each unit of flow on the cell arcs is given equal cost $C_{ij}$ (assumed arbitrarily to be one), to indicate indifference as to which cells are used as long as a minimal sample is found. The arc from the sink to the source has to recirculate the entire flow with no penalty; thus it has bounds of N above and n below, with $C_{ts}$ equal to zero.

Every linear program (known as a primal problem) has a corresponding problem known as its dual. Three conditions must be met for a solution to be optimal: (1) primal feasibility (it satisfies the constraints above), (2) dual feasibility (it satisfies the constraints in the dual of the above), and (3) complementary slackness (a set of conditions relating the primal and dual problems which states that a positive variable in one problem implies equality in the corresponding constraint of the other).

When a linear program has a minimal cost flow network formulation, the out-of-kilter method at each iteration classifies each arc either as in-kilter if its flow satisfies all three conditions,

FIGURE 1




N = population
n = sample size for row or column
m = minimum cell sample size
C = cost

or out-of-kilter if not. When every arc is in-kilter, an optimal solution has been found.

The algorithm commences with zero flow on all arcs. The arc most out-of-kilter is found, a circulation of flow (i.e., a directed loop containing the arc) detected, and the flow in the circulation increased or decreased in order to bring the arc into kilter. These iterations continue so that the number of arcs in-kilter is always increasing, hence convergence is assured. The advantage of this algorithm over the simplex method for solving linear programs which have no network structure is that the procedure is entirely additive, leading to a more efficient computation process.

Formulated as a linear program with integer-valued variables, the minimal cost flow problem is then to determine the set of $n_{ij}$ which minimizes $\sum_{ij} C_{ij}n_{ij}$ subject to:

$$\left.\begin{aligned} n_{i.} \leq n_{si} \leq N_{i.}, \ \forall\, i \\ m_{ij} \leq n_{ij} \leq N_{ij}, \ \forall\, ij \\ n_{.j} \leq n_{jt} \leq N_{.j}, \ \forall\, j \end{aligned}\right\} \text{Flow bounds on each arc}$$

$$\left.\begin{aligned} n_{si} &= \sum_j n_{ij}, \ \forall\, i \\ n_{jt} &= \sum_i n_{ij}, \ \forall\, j \\ n_{ts} &= \sum_i n_{si} = \sum_j n_{jt} \end{aligned}\right\} \text{Flow conservation at each node}$$

OPERATING CHARACTERISTICS

The algorithm was applied to a variety of sample allocation situations and subjected to a series of sensitivity tests to changes in parameters. With regard to the amount of oversampling necessary to meet the minimum marginal sample sizes, results indicated the optimum overall sample size depends on five factors, both individually and in combination. These factors are discussed in detail below.

1. Sample size relative to population.

FIGURE 2

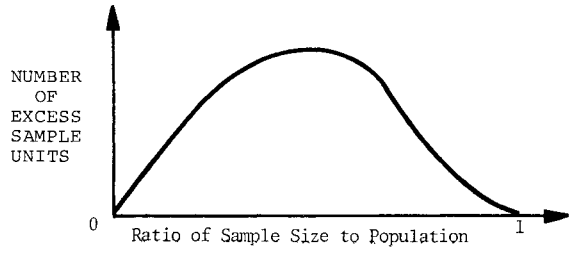



Figure 2 illustrates the general relationship of oversampling to the sample size factor given that the other four factors remain constant. The shape of the graph for any one population-sample-size case will, of course, depend on the particular characteristics of the matrices. For example, a proportional sample, no matter how large the proportion, will never require oversampling, and the corresponding graph would be flat. On the other hand, some samples may be so disproportional to the population that oversampling may be necessary even at the overall sample proportion of only 1%, and the corresponding graph would rise sharply from zero to its peak, then taper off slowly as the sample size increased.

2. Difference in marginal distributions between population and sample size.

The more disproportional the sample, the greater the probability that the solution will require excess sample units, given the other four factors remain constant. The precise relationship would again depend on the particular sampling situation being examined. For instance, for a very small total sample size, oversampling may never occur no matter how disproportional the distributions between the population and the sample. In contrast, if the sample were large relative to the population, the number of excess sampling units would increase dramatically as disproportionality increased.

3. Number of cells with low or zero populations.

The number and location of cells with very low or zero populations within the two-dimensional matrix affects the amount of oversampling necessary to achieve marginal sample sizes. The example in Table 3 illustrates this point in the extreme. In Table 3A six of the nine cells are empty, but their location within the matrix makes it possible nevertheless to allocate the sample without exceeding any marginal totals. In Table 3B the non-empty cells were shifted to different positions within the matrix, yielding an allocation which necessitates oversampling by 10 units.

3A.

| A \ B | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 1 | 20 / 5 | 0 / 0 | 0 / 0 | 20 / 5 |
| 2 | 0 / 0 | 50 / 10 | 0 / 0 | 50 / 10 |
| 3 | 0 / 0 | 0 / 0 | 30 / 15 | 30 / 15 |
| Total | 20 / 5 | 50 / 10 | 30 / 15 | 100 / 30 |

TABLE 3

3B.

| A \ B | 1 | 2 | 3 | Total | |
|---|---|---|---|---|---|
| 1 | 0 / 0 | 0 / 0 | 20 / 15 | 20 / 5 | ↑15 |
| 2 | 50 / 10 | 0 / 0 | 0 / 0 | 50 / 10 | |
| 3 | 0 / 0 | 30 / 15 | 0 / 0 | 30 / 15 | |
| Total | 50 / 5 | 30 / 10 | 20 / 15 | 100 / 30 | ↑40 |
| | ↑10 | ↑15 | | | |

TABLE 3 (continued)

4. Magnitude of cell minima.

Manipulation of the cell minima for a matrix determines the spread of the resulting allocation across cells of the matrix and can also affect the amount of oversampling required. If the cell minima are set at zero, the algorithm produces a solution which not only minimizes the excess sample units required to satisfy marginal totals, but also reduces the number of cells to which the sample is allocated. This may be a desirable feature of the algorithm if there are reasons, such as cost or geographical location of cells, for limiting the number of cells included in a sample. On the other hand it may be a requirement, as when the desired sample size is less than the number of cells with non-zero population.

Setting the minima of cells with non-zero population to one or greater assures that every cell will be included in the sample with a probability of one. This may be desirable to eliminate the possibility of any biases in the survey results caused by exclusion of certain segments of the population from the sample. However, in any given matrix, minima of zero will yield the smallest sample size of all possible minima.

5. Relative costs of cells.

Since the main objective of the algorithm is to minimize the overall cost of flow through the network, it will tend to oversample from cells with lower unit costs. The two matrices in Table 4 illustrate how costs influence the resulting allocation when certain costs are lowered to give those cells a greater probability of being included in the sample or of being more heavily sampled than others. The costs associated with the cells are shown to the right of the population in each cell. The first matrix shows that, ignoring the costs, it is possible to find an allocation of the marginal sample sizes that does not require oversampling. Matrix 4B shows the allocation that the algorithm produced, which has a lower cost function, but which also has five extra sample units.

4A.

| A \ B | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| 1 | 40 / 5 / 0 | 40 / 1 / 5 | 40 / 5 / 0 | 120 / 5 |
| 2 | 40 / 5 / 0 | 40 / 10 / 15 | 40 / 5 / 5 | 120 / 20 |
| 3 | 40 / 5 / 10 | 40 / 10 / 0 | 40 / 5 / 10 | 120 / 0 |
| Total | 120 / 10 | 120 / 20 | 120 / 15 | 360 / 45 |

Total Cost = 280

4B.

| A \ B | 1 | 2 | 3 | Total | |
|---|---|---|---|---|---|
| 1 | 40 / 5 / 0 | 40 / 1 / 5 | 40 / 5 / 0 | 120 / 5 | ↑10 |
| 2 | 40 / 5 / 0 | 40 / 10 / 10 | 40 / 5 / 10 | 120 / 20 | |
| 3 | 40 / 5 / 10 | 40 / 10 / 0 | 40 / 5 / 10 | 120 / 20 | |
| Total | 120 / 10 | 120 / 20 | 120 / 15 | 360 / 45 | ↑50 |
| | | | ↑20 | | |

Total Cost = 260

TABLE 4

These five matrix characteristics thus affect the final sample allocation. There remains an interesting associated question: In the case where there is more than one allocation minimizing the cost function, how does the algorithm pick among them? Experimentation with numerous matrices has led to these ad hoc observations:

1. If all lower capacities are set to zero, the algorithm limits the allocation to a subset of cells of the matrix, but does not necessarily choose the minimum number of cells for which the cost function is minimized.
2. If all costs are uniform over the

network, the algorithm as programmed distributes extra required sample units to cells with excess capacity beginning in the upper righthand corner of the matrix, ending in the lower left.
3. If costs differ from cell to cell, the algorithm distributes extra required sample units to cells with the lowest costs and with excess capacity.

EXAMPLES

The work which motivated this research was the development of the sample design for the U.S. Department of Transportation's General Aviation Activity and Avionics Survey. This survey, implemented for the first time in January 1978, is an annual national survey of approximately 30,000 of the 215,000 registered general aviation aircraft in the United States. The sample design for the survey is a two-dimensional stratified sample of manufacturer/model of aircraft by state of registration. There are approximately 350 manufacturer/models and 54 states and territories for which estimates are required, yielding more than 18,900 cells, about half of which are empty in the design matrix.

Determination of cell sample sizes is a two-step process. First, allocation of the target sample size of 30,000 across manufacturer/model and state is optimally determined. Then the marginal sample sizes are allocated across the cells. In a test matrix involving 10 manufacturer/models of aircraft and all 54 states and territories, the target total sample size was 740, to be distributed over the 540 cells, 214 of which were empty. The out-of-kilter algorithm was run with all cell costs and all non-zero cell lower capacities set to one to assure that all segments of the population would appear in the sample. It produced a solution calling for 755 aircraft, only 15 more than initially desired. The 15 extra aircraft were allocated to five of the states and territories, and only one of the manufacturer/models.

In W.G. Cochran (3) there is a two-way stratification example for small samples, where the sample size n is less than the number of cells and it is desired to have the sample give proportional representation to each stratification criterion. The matrix is shown in Table 5 beside the allocation of the sample obtained by using the method developed by Bryant, Hartley and Jessen (1). The out-of-kilter algorithm was applied to this matrix using

$C_{ij} = \frac{N}{N_{ij}}$, $C_{i.} = 1$, $C_{.j} = 1$, $\forall i,j$. This cell cost scheme was designed to imitate as closely as possible the proportional probabilities used by the Bryant, Hartley and Jessen method. The sample allocation determined by the network analysis appears in Table 5, too. Note the similarities between the two outcomes.

| A \ B | 1 | | 2 | | 3 | | 4 | | TOTAL | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15 | 11 | 21 | 7 | 17 | 9 | 9 | 18 | 62 | |
| | (2) | [2] | (1) | [1] | (1) | [1] | (0) | [0] | | 4 |
| 2 | 10 | 16 | 8 | 20 | 13 | 12 | 7 | 23 | 38 | |
| | (0) | [0] | (0) | [0] | (2) | [2] | (0) | [0] | | 2 |
| 3 | 6 | 27 | 9 | 18 | 5 | 33 | 8 | 20 | 28 | |
| | (0) | [0] | (1) | [2] | (0) | [0] | (1) | [0] | | 2 |
| 4 | 4 | 41 | 3 | 55 | 6 | 27 | 6 | 27 | 19 | |
| | (0) | [0] | (1) | [0] | (0) | [0] | (0) | [1] | | 1 |
| 5 | 3 | 55 | 2 | 82 | 5 | 33 | 8 | 20 | 18 | |
| | (0) | [0] | (0) | [0] | (0) | [0] | (1) | [1] | | 1 |
| TOTAL | 38 | | 43 | | 46 | | 38 | | 165 | |
| | | 2 | | 3 | | 3 | | 2 | | 10 |

Solution in Parentheses from Cochran

Solution in brackets from network analysis

TABLE 5

SUMMARY

The main features of the out-of-kilter algorithm as applied to a two-way stratified sample design are:
1. It allocates disproportional samples.
2. It minimizes the amount of oversampling.
3. It will not assign sample units to empty cells.
4. It works efficiently on both small and large matrices.
5. The spread of sample units across cells can be controlled.
6. It can incorporate the costs associated with specific cells into the allocation process to minimize survey costs.

REFERENCES

1. Bryant, E.C., H.O. Hartley, and R.J. Jessen. "Design and Estimation in Two-Way Stratification." Journal of the American Statistical Association, 55(1960), pp.105-124.
2. Cochran, R.S. "Sampling in Two or More Dimensions." Contributions to Survey Sampling and Applied Statistics. Edited by H.A. David. New York: Academic Press, Inc., 1978.
3. Cochran, W.G. Sampling Techniques. New York: John Wiley & Sons, Inc., 1963.
4. Ford, L.R. Jr., and D.R. Fulkerson. Flows in Networks. Princeton: Princeton University Press, 1962.