ICES J. mar. Sci., 51: 315–324. 1994

# The potential of sampling with partial replacement for fisheries surveys

William G. Warren

Warren, W. G. 1994. The potential of sampling with partial replacement for fisheries surveys. – ICES J. mar. Sci., 51: 315–324.

The methodology of Nicholson *et al.* (ICES CM 1991 D: 11, 9pp.) for comparing sampling with fixed and random stations is extended to the case of sampling with partial replacement. The conditions under which estimates of abundance and, more particularly, change in abundance based on fixed samples, will be more accurate than those based on sampling with partial replacement and under which the latter will be more accurate than those based on random sampling are derived. Research trawl data from NAFO Divisions 2J, 3K, and 3L for the years 1985–1991 (1985–1992 for 3L) are used to gain some idea of how the accuracy of estimates of changes based on the three strategies (fixed, random, and partial replacement) would compare under the assumption that the stations fixed between years are selected at random. In practice, purposive selection of the fixed stations is, of course, possible.

Key words: fixed samples, random samples, accuracy, trawl surveys.

Received 30 June 1993; accepted 7 March 1994.

*William G. Warren: Department of Fisheries and Oceans, PO Box 5667, St John's, Newfoundland, A1C 5X1, Canada.*

## Introduction

In recent years there has been some debate on the relative merit of using fixed or random stations in fisheries surveys. For example, based on an analysis of North Sea 0-group haddock data from the English groundfish survey by Hunton (1986), Nicholson *et al.* (1991) reiterated her conclusion that, had the survey been rerandomized each year, the variation due to primary station within roundfish area would be included in the residual variation, which would be approximately doubled, and trend estimation thus made less precise. On the other hand, failure to randomize can result in biased estimation on not only estimates made within a year but also trends measured across years (Nicholson *et al.*, 1991). Sampling with partial replacement can be viewed as an attempt to improve accuracy in trend estimation by combining some reduction in variation from the fixed-station component, with, because of the random component, the introduction of little or no bias.

The genesis of sampling with partial replacement (SPR) appears to lie with Jessen (1942), with further development by Patterson (1950) and Yates (1960), these three being cited by Cochran (1977) in his brief treatment of the topic. The method is closely related to rotation sampling (Eckler, 1955; Rao and Graham 1964). Various aspects of the approach have been studied by Kulldorff (1963), Singh (1968), Manoussakis (1977), and Fong (1990). Much of this work was undertaken with agricultural applications in mind, although Sen (1977) applied the methodology to a survey of waterfowl harvest. In forestry, considerable interest in SPR in relation to forest inventories followed the publication of a monograph of Ware and Cunia (1962), with several illustrations of application being cited by Cunia (1974). As far as is known, no applications to fishery surveys have been published.

In this paper expressions are derived for comparing the accuracy, as measured by the mean squared error, of estimates based on fixed stations, random stations, and sampling with partial replacement. Data drawn from research trawl surveys of NAFO Division 2J, 3K, and 3L for the years 1985–1992 are used for illustration.

## Review of basic methodology

Following Nicholson *et al.* (1991) let $\mu_{iy}$ denote the (true) index of abundance at the ith station in year y, and suppose that the total number of (possible) stations in the area is N. Then, the (true) mean index of abundance in the area in year y is given by:

$$\bar{\mu}_y = \sum_{i=1}^{N} \mu_{iy}/N.$$

Let $x_{iy}$, i=1,2, . . . n, be the observations made at n (<N) sample stations in year y. Further, it is supposed that the

1054-3139/94/030315+10 $08.00/0

sampling variance of $x_{iy}$ at the ith station (in year y) is $\sigma^2_{iy}$. This variation comes about from replicate observations at the same location not yielding exactly the same value and the station being an area, rather than a point, so that repeated measurements within the same station are not necessarily made at exactly the same point. (This is akin to the nugget effect in kriging.)

First, suppose that the n sample stations are selected at random from the N available. Then, as is well known, the sample mean, $\bar{x}_y = \Sigma_{i=1}^{n} x_{iy}/n$, is an unbiased estimator of $\bar{\mu}_y$. It follows that the variance of $\bar{x}_y$ is the sum of two components, one stemming from the usual sampling variance, i.e.:

$$\sigma^2_{\mu} = \sum_{i=1}^{N} (\mu_{iy} - \bar{\mu}_y)^2/N$$

and the other from the "measurement" error, i.e.:

$$\sum_{i=1}^{N} \sigma^2_{iy}/nN.$$

An unbiased estimator of $\mathrm{Var}(\bar{x}_y)$ is:

$$\sum_{i=1}^{n} (x_{iy} - \bar{x}_y)^2/n\,(n-1) = s^2/n, \text{ say.}$$

If the sample stations are fixed, the data refer specifically to those stations and:

$$E(\bar{x}_y) = \sum_{i=1}^{n} \mu_{iy}/n \neq \bar{\mu}_y, \text{ in general.}$$

Further:

$$\mathrm{Var}(\bar{x}_y) = \sum_{i=1}^{n} \sigma^2_{iy}/n^2.$$

Commonly, but not necessarily:

$$\sum_{i=1}^{n} \sigma^2_{iy}/n^2 < \sigma^2_{\mu}/n + \sum_{i=1}^{N} \sigma^2_{iy}/Nn,$$

i.e. the variance of the fixed-station mean will be less than the variance of the random-station mean. In a sense this is not a fair comparison, since the former is a biased estimator of $\bar{\mu}_y$. Accuracy is, perhaps, best measured by the mean squared error which, for fixed stations, is:

$$(\bar{\mu}_y - \sum_{i=1}^{n} \mu_{iy}/n)^2 + \sum_{i=1}^{n} \sigma^2_{iy}/n^2.$$

It is difficult to generalize how the magnitude of this quantity compares with the variance of the random-station mean, which is also its mean squared error. A fortuitous set of fixed stations may yield a highly accurate mean; unfortunately there is no way of determining which circumstance exists from the sample data *per se*.

The primary interest here is, however, the change in abundance over time. Again, following Nicholson *et al.* (1991), let:

$$\mu_{iy} = \mu + \varphi_i + \psi_y + \xi_{iy},$$

where the $\varphi_i$ are station effects and the $\psi_y$ are year effects and, for convenience, it is assumed that:

$$\sum_{i=1}^{N} \varphi_i = \sum_{i=1}^{Y} \psi_y = \sum_{i=1}^{N} \xi_{iy} = \sum_{y=1}^{Y} \xi_{iy} = 0.$$

(No loss of generality is introduced by these constraints, although they can be an inconvenience if the data set is augmented by an additional year.) This will be recognized as the usual two-way analysis-of-variance model with $\xi_{iy}$ denoting the interaction betwen year and station. Only when the $\xi_{iy} = 0$ for all i and y will the difference between years at any specific station be the same as the difference between the overall means; this property has been described by Houghton (1987) as persistence.

If the sample stations are selected at random in each year then:

$$E(\bar{x}_2 - \bar{x}_1) = \left[\sum_{i=1}^{N} (\mu + \varphi_i + \psi_2 + \xi_{i2}) - \sum_{i=1}^{N} (\mu + \varphi_i + \psi_1 + \xi_{i1})\right]/N$$
$$= \psi_2 - \psi_1,$$

i.e. $\bar{x}_2 - \bar{x}_1$ is an unbiased estimator of the difference between the year effects.

If, however, the same set of fixed stations is used in each year, then:

$$E(\bar{x}_2 - \bar{x}_1) = \left[\sum_{i=1}^{n} (\mu + \varphi_i + \psi_2 + \xi_{i2}) - \sum_{i=1}^{n} (\mu + \varphi_i + \psi_1 + \xi_{i1})\right]/n$$
$$= \psi_2 - \psi_2 + \sum_{i=1}^{n} (\xi_{i2} - \xi_{i1})/n.$$

Thus, in general, unless one has the property of persistence, the difference between the means of fixed stations is a biased estimator of the difference between year effects.

Consider now $\mathrm{Var}(\bar{x}_2 - \bar{x}_1)$ under independent random selection of stations in each year.

First, write $\bar{x}_2 - \bar{x}_1$ as:

$$[n(\psi_2 - \psi_1) + (\Sigma_2\varphi_i - \Sigma_1\varphi_i) + (\Sigma_2\xi_{i2} - \Sigma_1\xi_{i1}) + (\Sigma_2\varepsilon_{i2} - \Sigma_1\varepsilon_{i1})]/n,$$

where $\Sigma_y$ denotes summation over the stations in year y and the $\varepsilon_{iy}$ denote the "measurement" errors. For convenience, it is assumed that, in each year, there is one observation at each of the sampled stations. $\mathrm{Var}(\bar{x}_2 - \bar{x}_1)$ can be obtained as follows.

1. Clearly, $\mathrm{Var}(\psi_2 - \psi_1) = 0$.
2. $\mathrm{Var}(\Sigma_2\varphi_i - \Sigma_1\varphi_i) = \mathrm{Var}(\Sigma_2\varphi_i) + \mathrm{Var}(\Sigma_1\varphi_i) - 2\mathrm{Cov}(\Sigma_2\varphi_i, \Sigma_1\varphi_i)$.

   Now:

$$\mathrm{Var}(\Sigma_y\varphi_i) = n\,\frac{N-n}{N}\,\frac{\Sigma_{i=1}^{N}\varphi_i^2}{N-1}$$

   and, as shown in Appendix 1:

$\text{Cov}(\Sigma_2\varphi_i, \Sigma_1\varphi_i)=0.$

Define:

$$\sigma_\varphi^2=\sum_{i=1}^{N}\varphi_i^2/N.$$

Then:

$$\text{Var}(\Sigma_2\varphi_i-\Sigma_1\varphi_i)=2n\,\frac{N-n}{N}\,\frac{N}{N-1}\,\sigma_\varphi^2.$$

3. Likewise

$$\text{Var}(\Sigma_y\xi_{iy})=n\,\frac{N-n}{N}\,\frac{\Sigma_{i=1}^{N}\xi_{iy}^2}{N-1}$$

and, clearly:

$\text{Cov}(\Sigma_1\xi_{i1},\Sigma_2\xi_{i2})=0.$

Define:

$$\sigma_\xi^2=\sum_{i=1}^{N}\sum_{y=1}^{2}\xi_{iy}^2/2N,$$

i.e. for the time being, we take $Y=2$. Then:

$$\text{Var}(\Sigma_2\xi_{i2}-\Sigma_1\xi_{i1})=2n\,\frac{N-n}{N}\,\frac{N}{N-1}\,\sigma_\xi^2.$$

4. Next $\text{Var}(\Sigma_2\varepsilon_{i2}-\Sigma_1\varepsilon_{i1})=\Sigma_{i=1}^{n}\Sigma_{y=1}^{2}\sigma_{iy}^2$. It may not be unreasonable to suppose $\sigma_{iy}^2=\sigma_\varepsilon^2$, for all i, y, in which case:

$$\text{Var}(\Sigma_2\varepsilon_{i2}-\Sigma_1\varepsilon_{i1})=2n\sigma_\varepsilon^2.$$

5. Finally, all the covariances such as $\text{Cov}(\Sigma_2\varphi_i-\Sigma_1\varphi_i, \Sigma_2\xi_{i2}-\Sigma_1\xi_{i1})$ are zero.

Bringing the above together yields:

$$\text{Var}(\bar{x}_2-\bar{x}_1)=\frac{2}{n}\left[\sigma_\varepsilon^2+\frac{N-n}{N}\,\frac{N}{N-1}\,(\sigma_\varphi^2+\sigma_\xi^2)\right].$$

If, as is usually the case, $n \ll N$ then:

$$\text{Var}(\bar{x}_2-\bar{x}_1)\approx 2[\sigma_\varphi^2+\sigma_\xi^2+\sigma_\varepsilon^2]/n.$$

Next, consider $\text{Var}(\bar{x}_2-\bar{x}_1)$, under the assumption of fixed stations; it is assumed that these stations are purposively selected by some criterion. Then:

$$\bar{x}_2-\bar{x}_1=[(\psi_2-\psi_1)+\sum_{i=1}^{n}(\xi_{i2}-\xi_{i1})+\sum_{i=1}^{n}(\varepsilon_{i2}-\varepsilon_{i1})]/n.$$

The only random components here are the $\varepsilon_{iy}$. Thus:

$$\text{Var}(\bar{x}_2-\bar{x}_1)=2\sigma_\varepsilon^2/n.$$

It is clear that the variance of the difference is less with fixed stations than with random sampling, but what about the mean squared error? With fixed stations this is:

$$[2\sigma_\varepsilon^2+(\sum_{i=1}^{n}(\xi_{i2}-\xi_{i1}))^2/n]/n.$$

Thus, fixed stations will be more accurate in estimating change if, for $n \ll N$:

$$[\sum_{i=1}^{n}(\xi_{i2}-\xi_{i1})]^2/n<2(\sigma_\varphi^2+\sigma_\xi^2).$$

This will depend on what happens at the subset of fixed stations. If $\xi_{iy}=0$ for all i and y then, with fixed stations, the estimator will always be the more accurate (as well as unbiased). Otherwise, the differences $\xi_{i2}-\xi_{i1}$ might well be large, yielding appreciable bias and, hence, a mean squared error greater than under independent random sampling.

Even if the fixed-station estimator is the more accurate, independent random samples may still be preferred, especially if there is a cost differential in favour of the latter. For example, in forestry, permanent plots are considerably more expensive to establish and maintain than temporary plots. The plot boundaries have to be carefully marked to ensure that, in future surveys, exactly the same trees, apart from ingrowth and mortality, are measured. Such cost differential is unlikely to occur in fishery surveys, although there be some added navigational cost in returning to a relatively precise location if the station area is small. On the other hand, foresters are dealing with units that are fixed in space and have to suffer the environmental conditions imposed on them. Fish, however, are mobile and can react spatially to environmental changes. Accordingly, lack of persistence seems likely to be a more serious problem with fishery than with forestry surveys.

## Extension to sampling with partial replacement

As noted above, sampling with partial replacement, i.e. keeping some sampled units fixed but selecting the others at random on each occasion, has been used with some success in forestry and may, perhaps, produce some gains in fishery surveys, albeit for a different reason (lack of persistence as opposed to a cost differential between fixed and random stations). The methodology is, therefore, here extended to cover sampling with partial replacement.

Let the sample be as before except that $n_2$ of the observations are from fixed stations and $n_1=n-n_2$ are independently selected at random from the remaining $N-n_2$ stations in each year. We now let $\Sigma_y$ denote the sum over the random stations in year y, and $\Sigma_0$ the sum over the fixed stations. Then:

$$\bar{x}_2-\bar{x}_1=[\Sigma_2\varphi_i-\Sigma_1\varphi_i+n(\psi_2-\psi_1)+\Sigma_2\xi_{i2}+\Sigma_0\xi_{i2}-\Sigma_1\xi_{i1}-\Sigma_0\xi_{i1}+\sum_{i=1}^{n}\varepsilon_{i2}-\sum_{i=1}^{n}\varepsilon_{i1}]/n.$$

Now:

$$E(\Sigma_2\varphi_i)=E(\Sigma_1\varphi_i)=-n_1\Sigma_0\varphi_i/(N-n_2).$$

By the same token:

$$E(\Sigma_2\xi_{iy})=-n_1\Sigma_0\xi_{iy}/(N-n_2),$$

whence

$$E(\bar{x}_2-\bar{x}_1)=\psi_2-\psi_1+\frac{N-n}{n\,(N-n_2)}\Sigma_0(\xi_{i2}-\xi_{i1}).$$

The result for fixed stations ($n_2=n$, $n_1=0$) and independent random samples ($n_2=0$, $n_1=n$) can, of course, be obtained from this general case.

Consider, finally, $\mathrm{Var}(\bar{x}_2-\bar{x}_1)$. Since, in effect, one is dealing with independent random samples of $n_1$ stations out of a possible $N-n_2$:

$$\mathrm{Var}(\Sigma_2\varphi_i-\Sigma_1\varphi_i)=n_1\frac{N-n}{N-n_2}\frac{\Sigma^{N-n_2}\sigma_i^2}{N-n_2-1},$$

where $\Sigma^{N-n_2}$ denotes summation over the $N-n_2$ non-fixed stations. Now:

$$\sum^{N-n_2}\varphi_i^2=\sum_{i=1}^{N}\varphi_i^2-\Sigma_0\varphi_i^2$$
$$=N\sigma_\varphi^2-\Sigma_0\varphi_i^2.$$

Define:

$$\sigma_{\varphi,0}^2=\Sigma_0(\varphi_i-\Sigma_0\varphi_i/n_2)^2/n_2$$
$$=(\Sigma_0\varphi_i^2-n_2\bar{\varphi}_0^2)/n_2,\text{ say,}$$

whence:

$$\mathrm{Var}(\Sigma_2\varphi_i-\Sigma_1\varphi_i)=n_1\frac{N-n}{N-n_2}\frac{N(\sigma_\varphi^2-n_2[\sigma_{\varphi,0}^2+\bar{\varphi}_0^2]/N)}{N-n_2-1}.$$

Likewise, define:

$$\bar{\xi}_0=\Sigma_0\sum_{y=1}^{2}\xi_{iy}/2n_2$$

and

$$\sigma_{\xi,0}^2=\Sigma_0\sum_{y=1}^{2}\xi_{iy}^2/2n_2.$$

Then:

$$\mathrm{Var}(\Sigma_2\xi_{i2}-\Sigma_1\xi_{i1})=2n_1\frac{N-n}{N-n_2}\frac{N(\sigma_\xi^2-n_2[\sigma_{\xi,0}^2+\bar{\xi}_0^2]/N)}{N-n_2-1}.$$

Thus:

$$\mathrm{Var}(\bar{x}_2-\bar{x}_1)=\frac{2\sigma_\varepsilon^2}{n}+\frac{2n_1}{n^2}\frac{N-n}{N-n_2}\frac{N}{N-n_2-1}$$
$$\left[\sigma_\varphi^2+\sigma_\xi^2-\frac{n_2}{N}(\sigma_{\varphi,0}^2+\sigma_{\xi,0}^2+\bar{\sigma}_0^2+\bar{\xi}_0^2)\right].$$

## An index of persistence: estimation and behaviour

Let $x_{iy}$, $y=1,2$ denote the observed catches in two, not necessarily successive, years. From the above, for a given year:

$$\mathrm{Var}(x_{iy})=\sigma_\varphi^2+\sigma_\xi^2+\sigma_\varepsilon^2$$

for which a pooled estimate may be obtained in the usual way as:

$$s_x^2=[\sum_{y=1}^{2}\sum_{y=1}^{n_y}(x_{iy}-\bar{x}_y)^2]/(m_1+m_2-2),$$

where $m_1$ and $m_2$ denote the number of stations in the two years. Further:

$$d_i=x_{i2}-x_{i1}=\psi_2-\psi_1+\xi_{i2}-\xi_{i1}+\varepsilon_{i2}-\varepsilon_{i1}$$
$$=\psi_2-\psi_1+2\xi_{i2}+\varepsilon_{i2}-\varepsilon_{i1}$$

since $\xi_{i1}+\xi_{i2}=0$. Thus:

$$\mathrm{Var}(d_i)=4\sigma_\xi^2+2\sigma_\varepsilon^2$$

which can be estimated as:

$$s_d^2=\sum_{i=1}^{n}(d_i-\bar{d})^2/(m-1),$$

where, here, m is the number of "fixed" stations.

There are, thus, two equations from which to estimate the three unknowns, $\sigma_\varphi^2$, $\sigma_\xi^2$, and $\sigma_\varepsilon^2$. Suppose that the measurement error is negligible, then, if $\sigma_\xi^2=\bar{\omega}\sigma_\varphi^2$, $\bar{\omega}$ can be estimated as:

$$\bar{\omega}=\frac{s_d^2/4}{s_x^2-s_d^2/4}.$$

The quantity $\bar{\omega}$ can be regarded as a measure of the degree of persistence, with the case of $\bar{\omega}=0$, referred to above as the property of persistence, described as being persistent. The smaller the value of $\bar{\omega}$, the greater the degree of persistence.

No doubt some degree of measurement error will exist, in which case, provided $\sigma_\varphi^2>\sigma_\xi^2$, which would normally be the case, $\bar{\omega}$ will be overestimated; i.e. the persistence will be stronger than indicated. The bias is approximately $(1-\sigma_\varphi^2/\sigma_\xi^2)\sigma_\varepsilon^2$.

It has been shown above that fixed stations will be more accurate in estimating change if:

$$[\sum_{i=1}^{n}(\xi_{i2}-\xi_{i1})]^2/n<2(\sigma_\varphi^2+\sigma_\xi^2).$$

Since it is assumed $\xi_{i1}+\xi_{i2}=0$, the left-hand side can be written:

$$4[\sum_{i=1}^{n}\xi_{i1}]^2/n.$$

With N large the $\xi_{i1}$ will approximate a sample from a continuous distribution. Suppose, therefore, that the $\xi_{i1}$ are normally distributed and that the "fixed" stations

Table 1. Probability that fixed stations will be more accurate than random stations in estimating change, for selected values of the persistence index $\bar{\omega}$.

| $\bar{\omega}$ | Prob($\chi_1^2<(1+\bar{\omega})/2\bar{\omega}$) (%) |
|---|---|
| 0.1 | 98.1 |
| 0.2 | 91.7 |
| 0.3 | 85.9 |
| 0.4 | 81.4 |
| 0.5 | 77.9 |
| 0.6 | 75.2 |
| 0.7 | 73.0 |
| 0.8 | 71.1 |
| 0.9 | 69.6 |
| 1.0 | 68.3 |

Table 2. Probability that fixed stations will be more accurate than random stations by a specified amount, $\alpha$, in estimating change, for selected values of the persistence index $\bar{\omega}$.

| | Prob($\chi_1^2<(1+\bar{\omega})\alpha^2/2\bar{\omega}$) (%) | | | | | |
|---|---|---|---|---|---|---|
| | $\alpha$ | | | | | |
| $\bar{\omega}$ | 0.95 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 |
| 0.1 | 97.4 | 96.5 | 93.9 | 89.9 | 84.1 | 75.9 |
| 0.2 | 90.0 | 88.1 | 83.4 | 77.5 | 70.1 | 61.4 |
| 0.3 | 83.8 | 81.5 | 76.1 | 69.7 | 62.3 | 53.8 |
| 0.4 | 79.1 | 76.6 | 71.0 | 64.6 | 57.3 | 49.2 |
| 0.5 | 75.5 | 73.0 | 67.3 | 60.9 | 53.8 | 46.0 |
| 0.6 | 72.7 | 70.1 | 64.4 | 58.1 | 51.2 | 43.6 |
| 0.7 | 70.5 | 67.9 | 62.2 | 56.0 | 49.1 | 41.8 |
| 0.8 | 68.6 | 66.0 | 60.4 | 54.2 | 47.5 | 40.4 |
| 0.9 | 67.1 | 64.5 | 58.9 | 52.8 | 46.2 | 39.3 |
| 1.0 | 65.8 | 63.2 | 57.6 | 51.6 | 45.1 | 38.3 |

have been chosen at random. Then $[\Sigma_{i=1}^{n}\ \xi_{i1}]^2$ will be distributed as $n\sigma_\xi^2$ times a chi-squared random variate on 1 degree of freedom. The condition thus becomes:

$\chi_1^2<(1+\bar{\omega})/2\bar{\omega}$.

In this sense, the probability that fixed samples yield a more accurate estimate of change than random independent random samples is given in Table 1 for selected values of $\bar{\omega}$. Clearly, the probability increases with the persistence and tends to unity as $\sigma_\xi^2$ (or $\bar{\omega}$) tends to zero. The smaller the persistence (the greater $\sigma_\xi^2$ relative to $\sigma_\varphi^2$) the smaller this probability, with, for the worst possible scenario, a theoretical limit of:

$\text{Prob}(\chi_1^2<0.5)=52.0\%$

which would require, unrealistically, $\sigma_\varphi^2=0$.

It is here assumed that the costs for fixed and random stations are equal.

It is, perhaps, more germane to determine the chance of fixed samples yielding a substantially more accurate estimate of change than independent random samples. Recall that to halve a confidence interval one must quarter the variance. Hence, in Table 2, values of $\text{Prob}(\chi_1^2<(1+\bar{\omega})\alpha^2/2\bar{\omega})$ for selected $\bar{\omega}$ and $\alpha$ are presented.

In the case of sampling with partial replacement, if $N \gg n$, the bias is approximately $\Sigma_0(\xi_{i2}-\xi_{i1})/n$ and $\text{Var}(\bar{x}_2-\bar{x}_1)$ approximately:

$$\frac{2\sigma_\varepsilon^2}{n}+\frac{2n_1}{n^2}[\sigma_\varphi^2+\sigma_\xi^2].$$

Thus, the mean squared error is, appproximately:

$$\frac{[\Sigma_0(\xi_{i2}-\xi_{i1})]^2}{n^2}+\frac{2\sigma_\varepsilon^2}{n}+\frac{2n_1}{n^2}[\sigma_\varphi^2+\sigma_\xi^2].$$

How does this relate to the special cases of all samples fixed ($n_1=0$, $n_2=n$) and independent random samples ($n_1=n$, $n_2=0$)? If the common term, $2\sigma_\varepsilon^2/n$ is omitted, this means comparing:

$$\frac{[\Sigma_{i=1}^{n}(\xi_{i2}-\xi_{i1})]^2}{n^2},$$

$$\frac{[\Sigma_0(\xi_{i2}-\xi_{i1})]^2}{n^2}+\frac{2n_1}{n^2}[\sigma_\varphi^2+\sigma_\xi^2]$$

and

$2[\sigma_\varphi^2+\sigma_\xi^2]/n$.

Consider the probability that sampling with partial replacement will have a smaller mean squared error than independent random sampling under the assumption that the "fixed" stations are selected at random. Then:

$$\frac{[\Sigma_0(\xi_{i2}-\xi_{i1})]^2}{n^2}+\frac{2n_1}{n^2}[\sigma_\varphi^2+\sigma_\xi^2]<2[\sigma_\varphi^2+\sigma_\xi^2]/n$$

is equivalent to:

$$\frac{4n_2\sigma_\xi^2\chi_1^2}{n}<2(\sigma_\varphi^2+\sigma_\xi^2)[1-n_1/n]$$

which reduces to:

$\chi_1^2<\frac{1}{2}[1+\sigma_\varphi^2/\sigma_\xi^2]$,

that is, the same condition, for fixed samples to be more accurate in estimating change than independent random samples. The difference is that, under partial replacement ($n_1$ strictly less than n), the variance component is reduced; this may or may not be counterbalanced by the introduction of bias.

The condition for fixed samples to yield smaller mean squared error than sampling with partial replacement is somewhat more complicated. Consider:

Table 3. Probability that fixed stations will be more accurate than sampling with partial replacement for estimating change for selected values of the persistence index, $\bar{\omega}$, and the proportion, p, of stations held fixed in SPR.

| | $\text{Prob}(\chi^2_{1(F)} - p\chi^2_{1(P)} < \frac{1-p}{2}[1+\sigma^2_\varphi/\sigma^2_\xi])$ (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\bar{\omega}$ | | | | | | | | | |
| p | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| 0 | 98.1 | 91.7 | 85.9 | 81.4 | 77.9 | 75.2 | 73.0 | 71.1 | 69.6 | 68.3 |
| 0.1 | 97.5 | 90.9 | 84.7 | 80.6 | 77.0 | 74.4 | 72.3 | 70.4 | 68.7 | 68.1 |
| 0.2 | 96.4 | 89.2 | 83.0 | 78.7 | 76.2 | 73.5 | 72.0 | 70.4 | 68.8 | 67.2 |
| 0.3 | 95.6 | 87.4 | 81.0 | 77.2 | 74.2 | 72.1 | 70.3 | 69.3 | 68.3 | 66.8 |
| 0.4 | 94.4 | 85.5 | 80.0 | 74.6 | 72.4 | 70.5 | 69.3 | 68.0 | 66.2 | 65.4 |
| 0.5 | 92.0 | 82.5 | 77.6 | 73.2 | 71.0 | 68.8 | 67.8 | 65.2 | 65.5 | 64.2 |
| 0.6 | 89.5 | 80.0 | 74.5 | 70.6 | 68.3 | 67.2 | 65.4 | 64.2 | 63.6 | 62.6 |
| 0.7 | 85.7 | 76.0 | 70.4 | 68.0 | 65.9 | 63.6 | 62.9 | 62.1 | 61.4 | 60.7 |
| 0.8 | 79.4 | 70.3 | 66.2 | 62.9 | 62.4 | 61.1 | 60.6 | 59.5 | 59.2 | 58.0 |
| 0.9 | 70.6 | 63.1 | 60.2 | 58.9 | 58.0 | 55.8 | 56.0 | 55.3 | 55.1 | 54.4 |

Table 4. Probablity that the bias will be less under sampling with partial replacement than with fixed stations for selected values of the proportion, p, of stations held fixed in SPR.

| | p | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| Probability (%) | 80.5 | 73.2 | 68.1 | 64.1 | 60.8 | 58.0 | 55.6 | 53.5 | 51.7 |

$$\frac{[\Sigma_{i=1}^{n}(\xi_{i2}-\xi_{i1})]^2}{n^2} < \frac{[\Sigma_0(\xi_{i2}-\xi_{i1})]^2}{n^2} + \frac{2n_1}{n^2}[\sigma^2_\varphi+\sigma^2_\xi].$$

This is equivalent to:

$$4n_2\sigma^2_\xi\chi^2_{1(P)}+2n_1(\sigma^2_\varphi+\sigma^2_\xi)>4n\sigma^2_\xi\chi^2_{1(F)},$$

where $\chi^2_{1(P)}$ and $\chi^2_{1(F)}$ denote two separate chi-squared variables. If it can be assumed that the fixed stations in these two situations are independently chosen [at random] then these chi-squared variables can be regarded as independent. Let p denote the proportion of fixed stations ($=n_2/n$). The condition can then be written:

$$\chi^2_{1(F)} - p\chi^2_{1(P)} < \frac{1-p}{2}[1+\sigma^2_\varphi/\sigma^2_\xi].$$

Since, in general, a weighted sum of chi-squared variables is *not* itself a chi-squared variable (the exception is when the weights are equal) the probability:

$$\text{Prob}(\chi^2_{1(F)} - p\chi^2_{1(P)} < \frac{1-p}{2}[1+\sigma^2_\varphi/\sigma^2_\xi])$$

would have to be evaluated by double integration. This probability can, however, be approximated as follows. For given p and $\sigma^2_\varphi/\sigma^2_\xi$, $(1/2)(1-p)[1+\sigma^2_\varphi/\sigma^2_\xi]=X(p,\bar{\omega})$ can be evaluated. Two random chi-squared variables can be generated by, for example, squaring random Normal(0,1) random variables, and $\chi^2_{1(F)} - p\chi^2_{1(P)}=\Xi$, say, calculated. The latter can be repeated a large number of times, $N_t$ and the proportion of times that $\Xi<X(p,\bar{\omega})$ taken as an estimate of the probability. Table 3 was so generated with $N_t=20\,000$. (The error should be less than 0.7%, 19 times out of 20.) For completeness, the values for p=0 (from Table 1) are included.

Note that with p=1 the condition reduces to $\chi^2_{1(F)}<\chi^2_{1(P)}=50\%$. In other words, except for the pathological case of equality, the probability that the mean square error under fixed samples will be less than under sampling with partial replacement will always exceed 50%. It must be remembered, however, that a probability of, say, 55% that fixed samples will more accurate than sampling with partial replacement implies a 45% probability of sampling with partial replacement being more accurate than fixed samples. Thus, the advantage of fixed samples over sampling with partial replacement could then be said to be 55:45=1.22.

On the other hand, the bias component will be reduced under partial replacement if:

$$p\chi^2_{1(P)}<\chi^2_{1(F)}$$

or:

$$F_{1,1}<1/p,$$

where $F_{1,1}$ is distributed as Snedecor's F on 1 and 1 degrees of freedom. This leads to the values in Table 4

Table 5. Values of the persistence index, $\bar{\omega}$, calculated, by NAFO Division, for the pairs of years indicated.

| Division 2J | Year | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 |
| 1985 | — | 0.19 | 0.21 | 0.32 | 0.62 | 0.38 | 0.51 |
| 1986 | 0.12 | — | 0.11 | 0.22 | 0.17 | 0.75 | 1.31 |
| 1987 | 0.21 | 0.12 | — | 0.39 | 0.91 | 0.83 | 1.10 |
| 1988 | 0.45 | 0.23 | 0.27 | — | 0.68 | 0.55 | 1.41 |
| 1989 | 0.65 | 0.39 | 0.76 | 0.36 | — | 0.25 | 0.39 |
| 1990 | 0.82 | 0.79 | 0.80 | 0.56 | 0.24 | — | 0.28 |
| 1991 | 1.00 | 2.65 | 1.36 | 0.85 | 0.53 | 0.25 | — |

| Division 3K | Year | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 |
| 1985 | — | 0.32 | 0.25 | 0.28 | 0.90 | 0.26 | 0.47 |
| 1986 | 0.33 | — | 0.25 | 0.54 | 0.12 | 0.84 | 1.42 |
| 1987 | 0.26 | 0.41 | — | 0.28 | 0.15 | 0.22 | 0.42 |
| 1988 | 0.25 | 0.35 | 0.54 | — | 0.33 | 0.37 | 0.52 |
| 1989 | 1.53 | 0.83 | 0.21 | 0.32 | — | 0.26 | 0.15 |
| 1990 | 0.51 | 2.13 | 0.27 | 0.52 | 0.17 | — | 0.18 |
| 1991 | 0.53 | 1.10 | 0.33 | 0.28 | 0.21 | 0.25 | — |

| Division 3L | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 |
| 1985 | — | 0.20 | 0.14 | 0.23 | 0.26 | 0.37 | 0.30 | 0.27 |
| 1986 | 0.24 | — | 0.67 | 0.32 | 0.33 | 0.38 | 0.25 | 0.12 |
| 1987 | 0.27 | 0.68 | — | 0.27 | 0.21 | 0.18 | 0.46 | 0.35 |
| 1988 | 0.19 | 0.44 | 0.21 | — | 0.12 | 0.32 | 0.17 | 0.18 |
| 1989 | 0.16 | 0.27 | 0.24 | 0.15 | — | 0.07 | 0.36 | 0.82 |
| 1990 | 0.34 | 0.61 | 0.17 | 0.28 | 0.12 | — | 0.43 | 0.26 |
| 1991 | 0.59 | 0.18 | 0.49 | 0.25 | 0.37 | 0.49 | — | 0.26 |
| 1992 | 0.28 | 0.18 | 0.49 | 0.25 | 0.42 | 0.55 | 0.32 | — |

Note: the values above and below the diagonal are for same-location distances of 2.5 nm and 4.0 nm, respectively.

where the limits as p tends to 0 and 1.0 are 100% and 50%, respectively. In other words, partial replacement is likely to reduce the (absolute) bias although, in terms of accuracy, this is not counterbalanced by the introduced variance.

At the other extreme it could be supposed that the fixed stations of the partial replacement scheme are a subset of the fixed stations, i.e. $\Sigma_0(\xi_{i2} - \xi_{i1})$ is contained within $\Sigma_{i=1}^{n}(\xi_{2i} - \xi_{1i})$. The condition then reduces to:

$$\chi^2_{1<F,} < \frac{\sigma_\varphi^2 + \sigma_\xi^2}{2\sigma_\xi^2} = (1+\bar{\omega})/2\bar{\omega}.$$

## Example

To obtain some idea of what might happen in practice, persistence, i.e. $\bar{\omega}$, was estimated from research trawl survey data for cod in NAFO Divisions 2J and 3K for seven successive years (1985–1991) and in Division 3L for eight years (1985–1992). These are stratified random surveys with the number of successful stations ranging from 107 to 232. For the present purpose, the stratification was ignored and stations for any two years within a specified (short) distance of each other were regarded as being at the same location for those two years. To obtain some idea of how critical the choice of this distance is, two values, namely 2.5 and 4 nm, were used. These choices are somewhat arbitrary; the 2.5 nm appeared to be about the smallest distance from which a reasonable number of "fixed" stations could be generated from these data, this number ranging from 12 and to 25 depending on the Division and pair of years chosen.

Because the data exhibited positive skewness, the $x_{iy}$ used were taken as the logarithms of the station biomass

Table 6. Estimated probability (%) of obtaining a more extreme (smaller) value of the persistence index, $\bar{\omega}$, than that observed if the spatial distributions in different years were independent.

| Division 2J | Year | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 |
| 1985 | — | 0.6 | 0.7 | 6.0 | 23.3 | 8.2 | 16.1 |
| 1986 | 0 | — | 0 | 0.1 | 0 | 32.5 | 64.0 |
| 1987 | 0 | 0 | — | 5.3 | 48.2 | 42.4 | 59.0 |
| 1988 | 2.7 | 0 | 0 | — | 22.6 | 20.0 | 73.1 |
| 1989 | 13.7 | 0.3 | 23.0 | 0 | — | 0.2 | 1.7 |
| 1990 | 33.8 | 22.8 | 32.9 | 8.2 | 0 | — | 1.1 |
| 1991 | 50.3 | 96.1 | 76.8 | 37.8 | 0.6 | 0 | — |

| Division 3K | Year | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 |
| 1985 | — | 5.9 | 0.4 | 3.2 | 50.5 | 1.9 | 11.5 |
| 1986 | 0.9 | — | 2.7 | 32.2 | 3.6 | 54.8 | 73.0 |
| 1987 | 0 | 2.7 | — | 3.2 | 0.5 | 0.6 | 6.7 |
| 1988 | 0.2 | 3.4 | 10.0 | — | 9.1 | 16.9 | 27.5 |
| 1989 | 87.2 | 39.3 | 0 | 0.7 | — | 3.0 | 0.2 |
| 1990 | 4.1 | 84.5 | 0 | 11.7 | 0 | — | 0.1 |
| 1991 | 4.8 | 62.2 | 0.5 | 0.1 | 0 | 0.1 | — |

| Division 3L | Year | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Year | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 |
| 1985 | — | 0.3 | 0.7 | 0 | 1.3 | 3.9 | 0.3 | 0.3 |
| 1986 | 0 | — | 29.5 | 6.5 | 3.6 | 8.6 | 3.0 | 0 |
| 1987 | 0 | 21.8 | — | 2.2 | 0.5 | 0.2 | 9.0 | 3.0 |
| 1988 | 0 | 3.3 | 0.1 | — | 0 | 5.1 | 0.5 | 0.5 |
| 1989 | 0 | 0.1 | 0 | 0 | — | 0 | 3.0 | 44.8 |
| 1990 | 0 | 13.8 | 0 | 0 | 0 | — | 12.7 | 0.3 |
| 1991 | 5.0 | 0 | 2.8 | 0 | 0.2 | 1.4 | — | 0.4 |
| 1992 | 0 | 0 | 3.2 | 0 | 0.3 | 3.8 | 0 | — |

Note: the values above and below the diagonal are for same-location distances of 2.5 nm and 4.0 nm, respectively.

of cod, to better satisfy the normality and homogeneity of variance assumptions. Thus, $\bar{\omega}=0$ implies a constant percentage, rather than an absolute, change at each location.

Table 5 gives the estimated values of $\bar{\omega}$, by NAFO Division, for each pair of years.

The following resampling procedure was undertaken to determine the "significance" of the calculated $\bar{\omega}$. For any pair of years the number, $n_2$, of fixed stations was determined. Random samples of size $n_2$ were drawn from the data for each of these years. The samples were matched in the sense that the first observation drawn from year 1 was matched with the first observation drawn from year 2, etc., although, in reality, these were independent. These pairs were used to calculate an $s_d^2$ and, hence, a $\bar{\omega}$. The procedure was repeated 1000 times for each pair of years to approximate the null distribution of $\bar{\omega}$, i.e. the distribution under the assumption that the spatial distribution in one year is independent of that in the other year. The number of times the value so calculated fell below the actual $\bar{\omega}$ was recorded. This number being less than 50 (out of 1000) is then equivalent to the observed $\bar{\omega}$ being "significantly small" at the 5% level. The so-estimated p-values are given in Table 6.

## Discussion

From Table 6 it will be observed that there were more than a few instances where the observed value of $\bar{\omega}$ was less than *all* of the 1000 values generated for the null distribution and that these instances are more common when the 4-nm distance is used. There are likewise numerous instances where the observed value of $\bar{\omega}$ would be judged as significant at the 5% level, with the

frequency of such instances increasing from Division 2J through 3K to 3L.

It seems clear that in the most northern Division, 2J, persistence between adjacent years is relatively strong but gradually degrades over time. On the other hand, in the most southern Division, 3L, persistence appears relatively strong throughout the whole period 1985–1992. The picture for Division 3K is less clear but, as with its geographical position, seems to lie between that of 2J and 3L.

Tables 1–3 show the probability of increased precision in the estimation of change that would be achieved by sampling with partial replacement (or, as a special case, fixed sample locations) over independent random sampling, as a function of $\varpi$. In division 3L, sampling with partial replacement, or even fixed stations, would appear to have been a viable option for Division 3L for the period 1985–1992. In Divisions 2J and 3K sampling with partial replacement would appear to have been a viable option, but with the subset of "fixed" stations varying over time, i.e. the stations common to 1985 and 1986, say, would not be the same as the stations common to 1986 and 1987. Notwithstanding, it appears that sudden and substantial changes in spatial distribution can, and do, sometimes occur between successive years, causing a loss in persistence and, thus, a reduction in the expected precision of the estimated change in abundance.

It must be emphasized that, in these illustrations, it has been assumed that the "fixed" stations have been randomly selected. Therefore, the inferences do not necessarily apply to purposive selection of the fixed stations. With sufficient prior knowledge it may be possible to select the fixed stations with small or negligible bias in the estimate of change. On the other hand, we have not here considered stratification to reduce the inter-station component of variation. Also, we have not considered specifically any cost differential in sampling fixed versus random stations which, again, could alter the balance. Finally, the focus here has been on the estimation of change. The balance could again be changed if, in addition to change, estimates of the actual abundance are required in one or both years, and how much emphasis is given to the latter.

## References


Cochran, W. G. 1977. Sampling techniques, 3rd edn. J. Wiley & Sons, New York. 428 pp.

Cunia, T. 1974. Independent versus dependent successive measurements. *In* Monitoring forest environment through successive sampling, pp. 1–18. Ed. by T. Cunia. College of Environmental Science and Forestry, State University of New York, Syracuse, New York. 390 pp.

Eckler, A. R. 1955. Rotation sampling. Annals of Mathematical Statistics, 26: 664–685.

Fong, W.-K. 1990. A Bayesian approach to successive sampling with partial replacement of units on two occasions. Biometrika, 77: 383–388.

Houghton, R. G. 1987. The consistency of spatial distribution of young gadoids with time. ICES CM 1987/D:15, 7 pp.

Hunton, J. K. 1986. Areal and temporal sources of variation in the English groundfish survey. ICES CM 1986/G:15, 7 pp.

Jessen, R. J. 1942. Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Station Research Bulletin No. 304. 104 pp.

Kulldorf, G. 1963. Some problems in optimal allocation for sampling on two occasions. Review of the International Statistical Institute, 31: 24–57.

Manoussakis, E. 1977. Repeated sampling with partial replacement of units. Annals of Statistics, 4: 795–802.

Nicholson, M. D., Stokes, T. K., and Thompson, A. B. 1991. The interaction between fish distribution, survey design and analysis. ICES CM 1991/D:11, 9 pp.

Patterson, H. D. 1950. Sampling on successive occasions with partial replacement of units. Journal of the Royal Statistical Society, Series B, 12: 241–255.

Rao, J. N. K., and Graham, J. E. 1964. Rotation designs for sampling on repeated occasions. Journal of the American Statistical Association, 59: 492–509.

Sen, A. R. 1977. Sampling theory on repeated occasions with ecological applications. *In* Statistical ecology, Vol. 5: Sampling biological populations, pp. 315–328. Ed. by R. M. Cormack, G. P. Patil, and D. S. Robson. International Co-operative Publishing House, Fairland, Maryland. 392 pp.

Singh, D. 1968. Estimates in successive sampling using multistage designs. Journal of the American Statistical Association, 63: 99–112.

Ware, K., and Cunia, T. 1962. Continuous forest inventory with partial replacement of samples. Forest Science Monograph no. 3, 40 pp.

Yates, F. 1960. Sampling methods for censuses and surveys, 3rd edn. C. Griffin and Co., London. 458 pp.


## Appendix 1

We here show that $\mathrm{Cov}(\Sigma_1\varphi_i, \Sigma_2\varphi_i)=0$.

Since the sample stations in each year are selected independently and at random, there is a possibility that one or more stations will be common to both samples; indeed, the number of common stations can be 0, 1, 2, . . . n.

The total number of possible sample combinations is:

$$\binom{N}{n}^2,$$

and the number of these for which there will no be station in common is:

$$\binom{N}{n}\binom{N-n}{n}.$$

The number of sample combinations that will contain exactly one station in common is:

$$\binom{N}{1}\binom{N-1}{n-1}\binom{N-n}{n-1},$$

the number with exactly two stations in common is:

$$\binom{N}{2}\binom{N-2}{n-2}\binom{N-n}{n-2},$$

and, in general, the number with exactly j stations in common is:

$$\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j},$$

Accordingly:

$$\binom{N}{n}^2=\sum_{j=1}^{n}\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}.$$

For simplicity we assume $N>2n$ which, in practice, will always be the case.

*If there is no station in common then* $\Sigma_1\varphi_i\Sigma_2\varphi_i$ *will* contain terms $\varphi_i\varphi_j$, $i<j$, and there will be $n^2$ such terms. On taking expectations every possible $\varphi_i\varphi_j$ $(i<j)$ must occur an equal number of times. Thus, the number of times any particular $\varphi_i\varphi_j$, with $i\neq j$, will occur is:

$$n^2\binom{N}{0}\binom{N}{n}\binom{N-n}{n}\Big/\binom{N}{2}$$

Next, suppose that the samples have exactly one *station in common. Then,* $\Sigma_1\varphi_i\Sigma_2\varphi_i$ *will contain one term* of the form $\varphi_i^2$ and $n^2-1$ terms such as $\varphi_i\varphi_j$, $i<j$. Again, on taking expectations, each $\varphi_i^2$ must occur an equal number of times, as must each $\varphi_i\varphi_j$ with $i<j$. The number of occurrences of each $\varphi_i^2$ is, therefore:

$$1\binom{N}{1}\binom{N-1}{n-1}\binom{N-n}{n-1}\Big/\binom{N}{1}$$

and the number of occurrences of each $\varphi_i\varphi_j$, $i<j$:

$$(n^2-1)\binom{N}{1}\binom{N-1}{n-1}\binom{N-n}{n-1}\Big/\binom{N}{2}$$

More generally, suppose that the samples have exactly j stations in common. Then, $\Sigma_1\varphi_i\Sigma_2\varphi_i$ will contain j terms of the form $\varphi_i^2$ and $n^2-j$ terms like $\varphi_i\varphi_j$, $i<j$. Then, on taking expectations, the number of occurrences of any particular $\varphi_i^2$ will be:

$$j\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}\Big/\binom{N}{1}$$

and the number of occurrences of each $\varphi_i\varphi_j$, $i<j$:

$$(n^2-j)\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}\Big/\binom{N}{2}$$

Thus, overall, the number of occurrences of each $\varphi_i^2$ will be:

$$\sum_{j=0}^{n} j\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}\Big/\binom{N}{1}$$

and the number of occurrences of each $\varphi_i\varphi_j$, $i<j$:

$$\sum_{j=0}^{n}(n^2-j)\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}\Big/\binom{N}{2}$$

Now:

$$\sum_{j=0}^{n} n^2\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}\Big/\binom{N}{2}=n^2\binom{N}{n}^2\Big/\binom{N}{2}$$

We therefore focus on:

$$\sum_{j=0}^{n} j\binom{N}{j}\binom{N-j}{n-j}\binom{N-n}{n-j}$$

$$=\sum_{j=1}^{n} N\binom{N-1}{j-1}\binom{N-j}{j-1}\binom{N-n}{n-j}$$

$$=\sum_{j=0}^{n-1} N\binom{N-1-j}{j}\binom{N-1-j}{n-j-1}\binom{N-n}{n-j-1}=$$

$$N\binom{N-1}{n-1}^2$$

Thus, each $\varphi_i^2$ occurs:

$$N\binom{N-1}{n-1}^2\Big/\binom{N}{1}=\binom{N-1}{n-1}^2$$

times and each $\varphi_i\varphi_j$ occurs:

$$\left[n^2\binom{N}{n}^2-N\binom{N-1}{n-1}^2\right]\Big/\binom{N}{2}=2\binom{N-1}{n-1}^2$$

times. Hence:

$$\mathrm{Cov}(\Sigma_1\varphi_i\Sigma_2\varphi_i)=(n/N)\left[\sum_{i=1}^{N}\varphi_i^2+2\sum_{i=1}^{N}\sum_{j>1}^{N}\varphi_i\varphi_j\right]$$

$$=(n/N)\left[\sum_{j=1}^{N}\varphi_i\right]^2=0.$$