ICES 1993

ICES C.M. 1993/D:56
Statistics Cttee
Ref. B,G,H,J,K,N



# More on Persistence and the Potential of Sampling with Partial Replacement

William G. Warren
Dept. of Fisheries and Oceans
P.O. Box 5667, St. John's, NF
Canada A1C 5X1

## Summary

The calculation of the values of the index of persistence for cod given in Warren (1992) is extended to cover the years 1985-1991 for NAFO Divisions 2J and 3K and the years 1985-1992 for Division 3L. Persistence remains strong throughout the period in 3L but degrades over time in 2J and 3K. Cluster analyses, using the index of persistence and an estimated associated probability value as measures of association, suggest that, for Divisions 2J and 3K, years 1985-1988 form one group and years 1989-91 another group. Examination of the centres of gravity of trawl biomass confirms a sudden and substantial shift in the spatial distribution between 1988 and 1989 for these two divisions; a similar phenomenon does not appear with Division 3L. Accordingly, for the years and regions studied, sampling with partial replacement would appear to have been a viable procedure, particularly if the subset of "fixed" stations were progressively changed. The instances of a sudden and substantial shift in the spatial distribution would, however, caused a loss of persistence and, thus, a reduction in the expected precision in the estimate of the change in abundance between these years.

## Introduction

Warren (1992) extended the methodology of Nicholson et al. (1991) for comparing sampling with fixed and random stations to the case of sampling with partial repacement. The gain in the precision of the estimate of change in abundance between two years that could be expected from the use of sampling with partial replacement (with fixed stations as a special case) was shown to be a function of a measure, $\varpi$, of the degree of "persistence" between the two years, with $\varpi = 0$ implying that the spatial pattern was "persistent", i.e. the change was the same at all points in the survey area. Based on research trawl data from NAFO Division 2J, estimates of $\varpi$ were presented for all pairwise combinations of years from 1985 to 1990. These suggested that persistence may be relatively strong between successive years but, in general, falls off over time. This would imply that little would be gained by keeping the same set of fixed stations over long time periods. On the other hand, sampling with partial replacement with the subset of "fixed" stations varying over time may well be a reasonable compromise.

It was of interest to determine if these inferences would hold over a larger data base, and if the measure of persistence, $\varpi$, were reflected in the degree of similarity between spatial distributions as determined by, say, kriging. Accordingly, estimates of $\varpi$ are presented here for pairwise combinations of the years 1985-91 for NAFO Divisions 2J and 3K and of the years 1985-92 for NAFO Division 3L. Some preliminary results of relating $\varpi$ to the change in spatial distribution are included.

## Methodology

For the full methodological development reference is made to Warren (1992). Let $x_{iy}$ denote the observation made at the $i^{th}$ station in the $y^{th}$ year. For any two years $Var(x_{iy})$ is estimated as

$$s_x^2 = [\sum_{y=1}^{2}\sum_{i=1}^{n_y}(x_{iy} - \bar{x}_y)^2]/(n_1 + n_2 - 2)$$

where $n_1$ and $n_2$ denote the numbers of stations in the two years, and

$$\bar{x}_y = \sum_{i=1}^{n_y} x_{iy}/n_y.$$

Further, let $d_i = x_{i1} - x_{i2}$. Then $Var(d_i)$ is estimated as

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

where the summation is over the $n$ "fixed" stations, i.e. those stations common to the two years, and $\bar{d} = \sum_{i=1}^{n} d_i/n$.

The persistence measure is then estimated as

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

## Implementation

The research trawl surveys in NAFO Divisions 2J, 2K and 3L are stratified random surveys with the number of succesful stations ranging from 107 (3K, 1986) to 232 (3L, 1985). In Warren (1991) stations in any two years within 2.5 nm of each other were regarded as being at the same location. The 2.5 nm was somewhat arbitrary; it was chosen as the smallest distance from which a minimally reasonable number of "fixed" stations could be generated from these data. This procedure was followed for the present study but, to obtain some idea of how critical was the choice of distance, estimates with the distance extended to 4.0 nm were also obtained.

Warren (1991) presented values of $\varpi_{max}$. This is not the maximum value that $\varpi$ could take, but the value obtained by, in effect, treating all stations, no matter how far apart, as being at the same location. Since this varied between pairs of years, it was felt to provide a yardstick by which the calculated values of $\varpi$ could be compared. While sometimes useful, $\varpi$ was subsequently found to be potentially misleading and, in the present study, is replaced by the following. For any pair of years the number, $n$, of fixed stations was determined. Random samples of size $n$ were drawn from the data for each of these years. These were matched in the sense that the first sample drawn from year 1 was matched with the first drawn from year 2, although, in reality, these are independent. These pairs were used to calculate a $s_d^2$ and, hence, a $\varpi$. The procedure was repeated 1000 times for each pair of years, and the number of times the value so calculated fell below the actual $\varpi$ recorded. The number being less than 50 (out of 1000) is then equivalant to $\varpi$ being "significantly small" at the 5% level.

It should be noted that the $x_{iy}$ used were the logarithms of the station biomass of cod (plus 1.0 to allow for zero catches). Thus $\varpi = 0$ implies a constant percentage, rather than absolute, change at each location.

## Results

The values of $\varpi$ are given in Tables 1, 2 and 3 for Divisions 2J, 3K and 3L, respectively. The values above the diagonal are for same-location distances of 2.5 nm those below the diagonal for 4.0 nm distances.

Table 1.
Values of $\varpi$ for 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 | - |

[Note: the above-diagonal values are the same as in Warren (1992) with the exception of the 1987-1990 combination; the 0.58 given therein is the result of a transcription error].

Table 2.
Values of $\varpi$ for 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 | - |

Table 3.
Values of $\varpi$ for 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 | - |

The probabilities of obtaining values of $\varpi$ less than those observed, as estimated by the above repeated-sampling approach, are given in Tables 4, 5 and 6 for Divisions 2J, 3K and 3L, respectively. As with Tables 1-3, the above- and below-diagonal values refer to the 2.5 nm and 4.0 nm distances, respectively.

Table 4.
Estimated probabilities (%) of $\varpi$ for 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 | - |

Table 5.
Estimated probabilities (%) of $\varpi$ for 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 | - |

Table 6.
Estimated probabilities (%) of $\varpi$ for 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 | - |

It will be observed that there are more than a few instances where the observed value of $\varpi$ was less than *all* of the 1000 values generated for the null distribution, and that these instances are more common when the 4.0 nm distance is used. There are likewise numerous instances where the observed value of $\varpi$ 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 for Division 2J that, as observed in Warren (1992), the persistence between adjacent years is relatively strong but gradually degrades over time; this appears to be very much the case when the results for the 4.0 mn distance are considered. On the other hand, for Division 3L, persistence appears relatively strong throughout the whole period 1985-92. The picture for Division 3K is less clear.

Both $\varpi$ and its associated estimated probability can be regarded as measures of the degree of association between pairs of years. Accordingly, they may be used to construct clusters of similar years. Various clustering algorithms are available. We here use a very simple method described by Scott (1972) and attributed by him to Jolliffe (1970). The association between two groups is defined as the arithmetic mean of the measures of association between the members of one group and the members of the other. At each state of the clustering process, the two groups (which may

contain only one member, here a year) having the largest measure of association (here the smallest numerical value, since zero implies fully persistent) are combined. If one wishes, the procedure may be terminanted when a predetermined number of groups is attained, or the measure of association reaches some prescribed value.

Dendrograms generated by the above method are presented in Figs. 1-6. Figs 1,3,5 use $\varpi$ and Figs. 2,4,6 use the associated probability as measure of association. Each figure has two components; that labelled (a) corresponds to the 2.5 nm distance, that labelled (b) to the 4.0 nm distance.

For Division 2J these dendrograms suggest that two distinct clusters can be formed from the years, one containing the years 1985 through 1988, and the other 1989 through 1991. The picture is the same for each combination of distance and measure of association with the exception that, for the probabilty measure and 4.0 nm, 1989 is grouped with 1985-88 rather than 1990-91.

For Divison 3L no clear-cut grouping emerges.

For Division 3K the pattern is similar to that for 2J in that, for the probability measure, 1985-1987 form one group and 1989-91 another, with 1988 grouped with 1989-91 when the 4.0 nm distance is used and with 1985-87 when the 2.5 nm distance is used. When $\varpi$ is used, with 2.5 nm, 1986-87 and 1989 appear to form one group and 1985, 1988 and 1990-91 another, although 1985 and 1988 could, perhaps, be regarded as a group separate from 1990-91. With 4.0 nm, 1985 and 1988, perhaps along with 1986, form one group and 1987 and 1989-1991 another. With both distances, 1990 and 1991 occur in the same group. In this sense, Division 3K appears to be somewhere between Divisions 2J and 3L, although perhaps more akin to 2J.

Table 7.
Centres of Gravity of Stations and Trawl Biomass

| | | Trawl Biomass | | Stations | |
|---|---|---|---|---|---|
| Division | Year | lat. | long. | lat. | long. |
| 2J | 1985 | 53.98 | 54.93 | 53.78 | 54.34 |
| | 1986 | 54.23 | 54.79 | 53.77 | 54.19 |
| | 1987 | 54.26 | 54.58 | 53.81 | 54.31 |
| | 1988 | 54.50 | 54.76 | 53.97 | 54 38 |
| | 1989 | 53.49 | 53.47 | 53.78 | 54.16 |
| | 1990 | 53.88 | 53.42 | 53.84 | 54.13 |
| | 1991 | 52.89 | 53.46 | 53.89 | 54.15 |
| 3K | 1985 | 51.08 | 52.08 | 50.92 | 52.39 |
| | 1986 | 51.33 | 51.86 | 50.93 | 52.87 |
| | 1987 | 51.29 | 51.82 | 50.88 | 52.51 |
| | 1988 | 50.91 | 51.93 | 50.86 | 52.73 |
| | 1989 | 50.94 | 50.89 | 51.00 | 52.57 |
| | 1990 | 49.81 | 50.39 | 50.69 | 52.05 |
| | 1991 | 50.71 | 50.66 | 50.97 | 51.97 |
| 3L | 1985 | 47.37 | 50.35 | 47.41 | 50.25 |
| | 1986 | 47.82 | 50.29 | 47.28 | 50.45 |
| | 1987 | 47.49 | 49.82 | 47.23 | 50.24 |
| | 1988 | 48.06 | 50.54 | 47.32 | 50.36 |
| | 1989 | 47.66 | 50.30 | 47.34 | 50.42 |
| | 1990 | 47.01 | 49.52 | 47.25 | 50.04 |
| | 1991 | 48.48 | 50.13 | 47.40 | 49.91 |
| | 1992 | 48.77 | 50.23 | 47.45 | 49.83 |

As yet, the measures of persistence have not been related to changes in the spatial distribution as estimated by kriging, in part because of possible instability of the variograms caused by movement of the fish during the period of a survey. However, the centre of gravity of the trawled biomass for each combination of year and division can be calculated from the study data. The latitude and longitude of the centres of gravity are given in Table 7, along with those for the centre of gravity of the trawl stations. The latter are presented to determine whether any shift in the centre of gravity of the biomass was due to a change in the centre of gravity of the stations caused by a change in the distribution of sampling effort.

The locationsaof the biomass centres of gravity are plotted in Fig. 7. While the centres of gravity of the stations remain fairly compact, especially in Division 2J, the centres of gravity of the trawl biomass in Divisions 2J and 3K take a sudden and pronounced shift to the east in 1989. In 2J this also coincides with a shift to the south. In 3K there is a shift to the south in 1990 followed by a return to the north in 1991. The picture in Divisin 3L is quite different. There is somewhat of a shift to the southest in 1990 followed by a return and additional move northwards in 1991-92. The behaviour is consistent with the pattern observed in the measures of persistence, particularly in the identification of the two groups of years, namely 1985-88 and 1989-91 in Divisions 2J and 3K.

## Discussion

Warren (1992) shows, quantitatively, the gains in precision in the estimation of the change in abundance that would be obtained under sampling with partial replacement (or, as a special case, fixed stations) as function of $\varpi$. In Division 3L, sampling with partial replacement or, even, fixed stations, would appear to have been a viable option for the period 1985-92. 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, etc. The advantages of such a strategy would have been lost, however, for the change that occurred between 1988 and 1989.

Thus, over the period and regions studied, persistence between adjacent years is commonly strong, justifying the use of, at least, sampling with partial replacement, albeit preferably with the "fixed" stations progressively changed. It appears, however, that sudden and substantial changes in spatial distribution can, and do, occur causing a loss in persistence and, thus, a reduction in the expected precision of the estimated change in abundance. The reasons for these such changes in spatial distribution are outside the scope of the present paper.

## References

Jolliffe, I.T. 1970. Redundant variables in multivariate analysis. Unpublished D. Phil. Thesis, University of Sussex.

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

Scott, J.F. 1972. Redundent variables in multivariate analysis. Procedings of the Third Conference of the Advisory Group of Forest Statisticians, International Union of Forest Research Organisations. Institut National de la Recherche Agronomique, Publ. 72-3, 153-160.

Warren. W.G. 1992. The potential of sampling with partial replacement for fisheries surveys. ICES C.M. 1992/D:21.

Fig. 1
Division 2J
Dendrogram based on $\varpi$

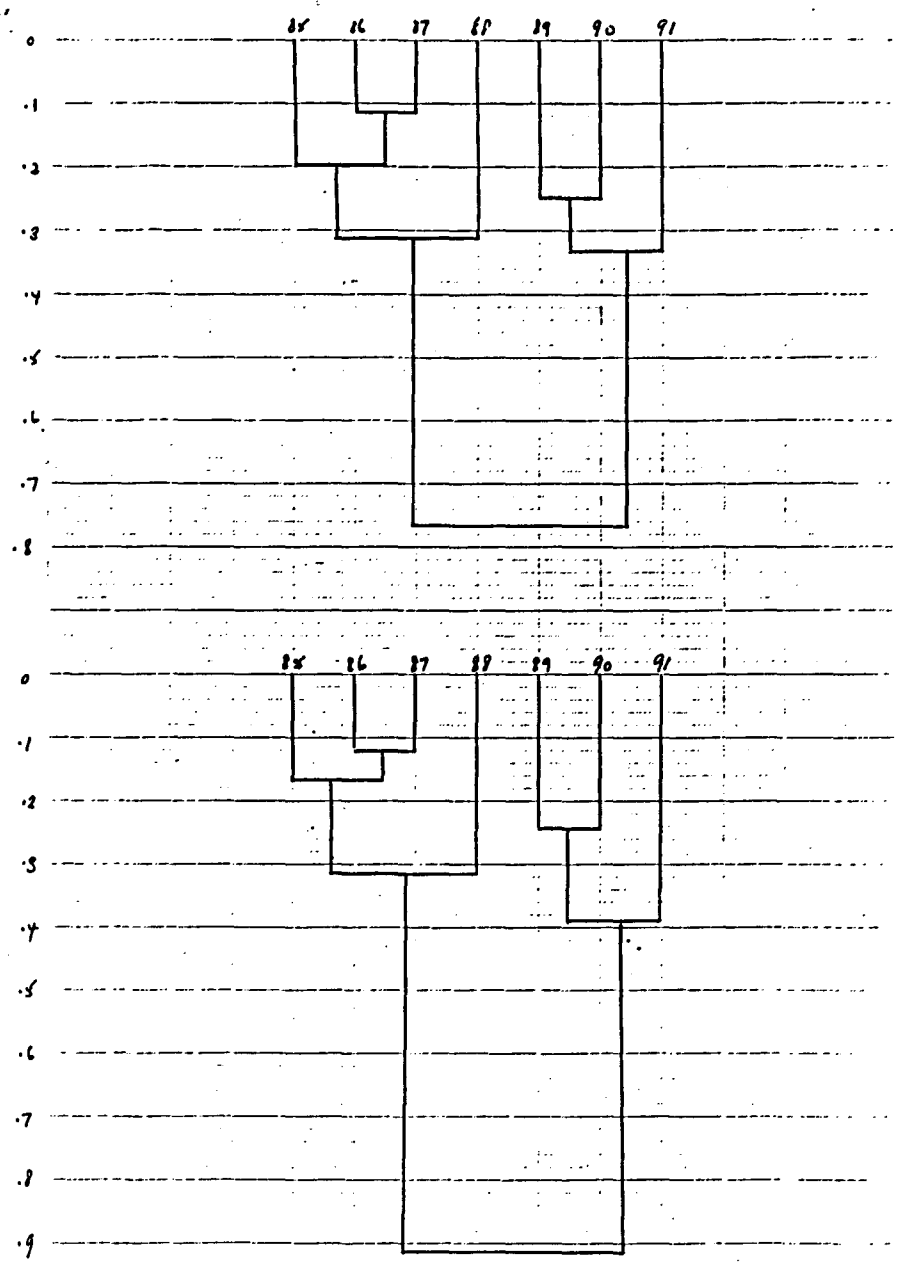


Fig. 2
Division 2J
Dendrogram based on estimated probability

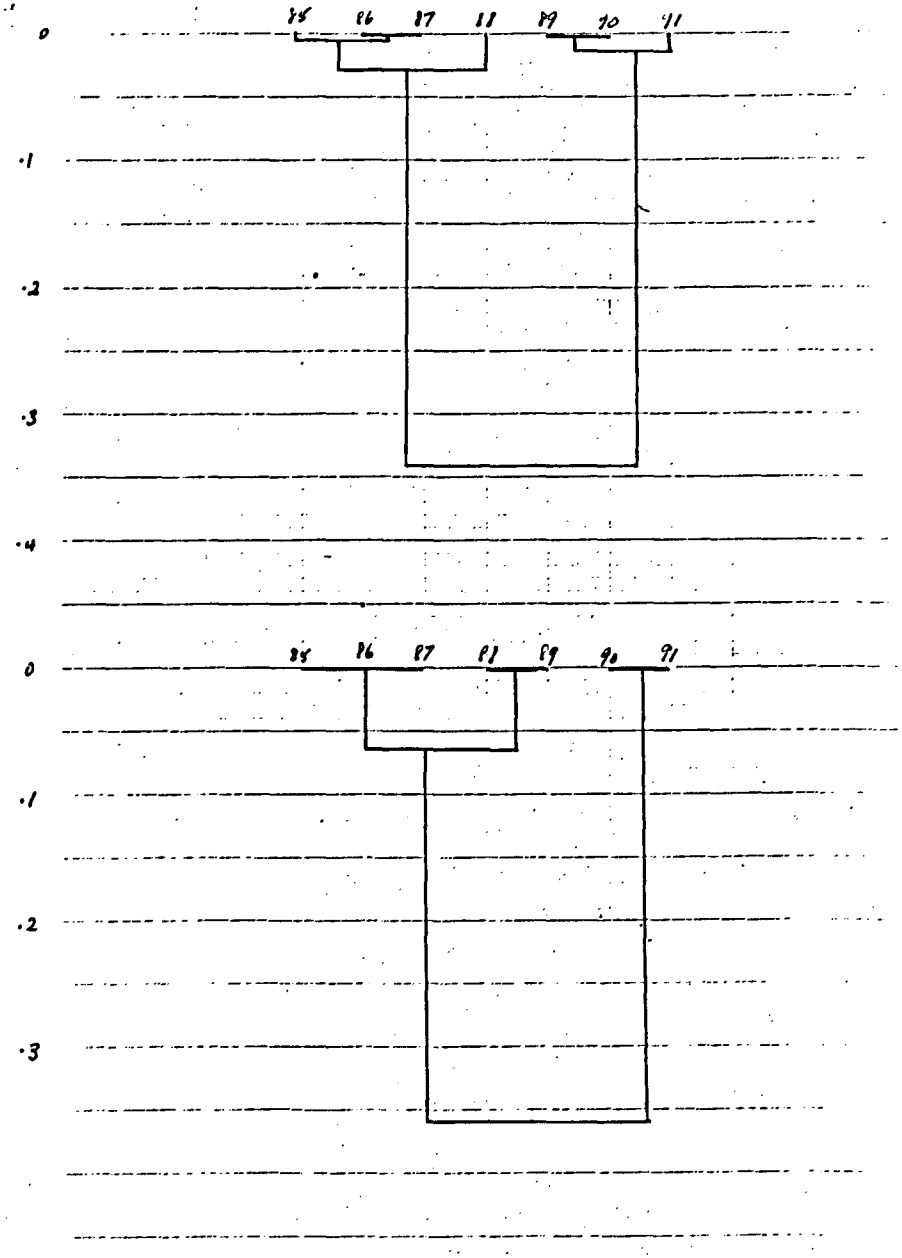

Fig. 3
Division 3K
Dendrogram based on $\varpi$

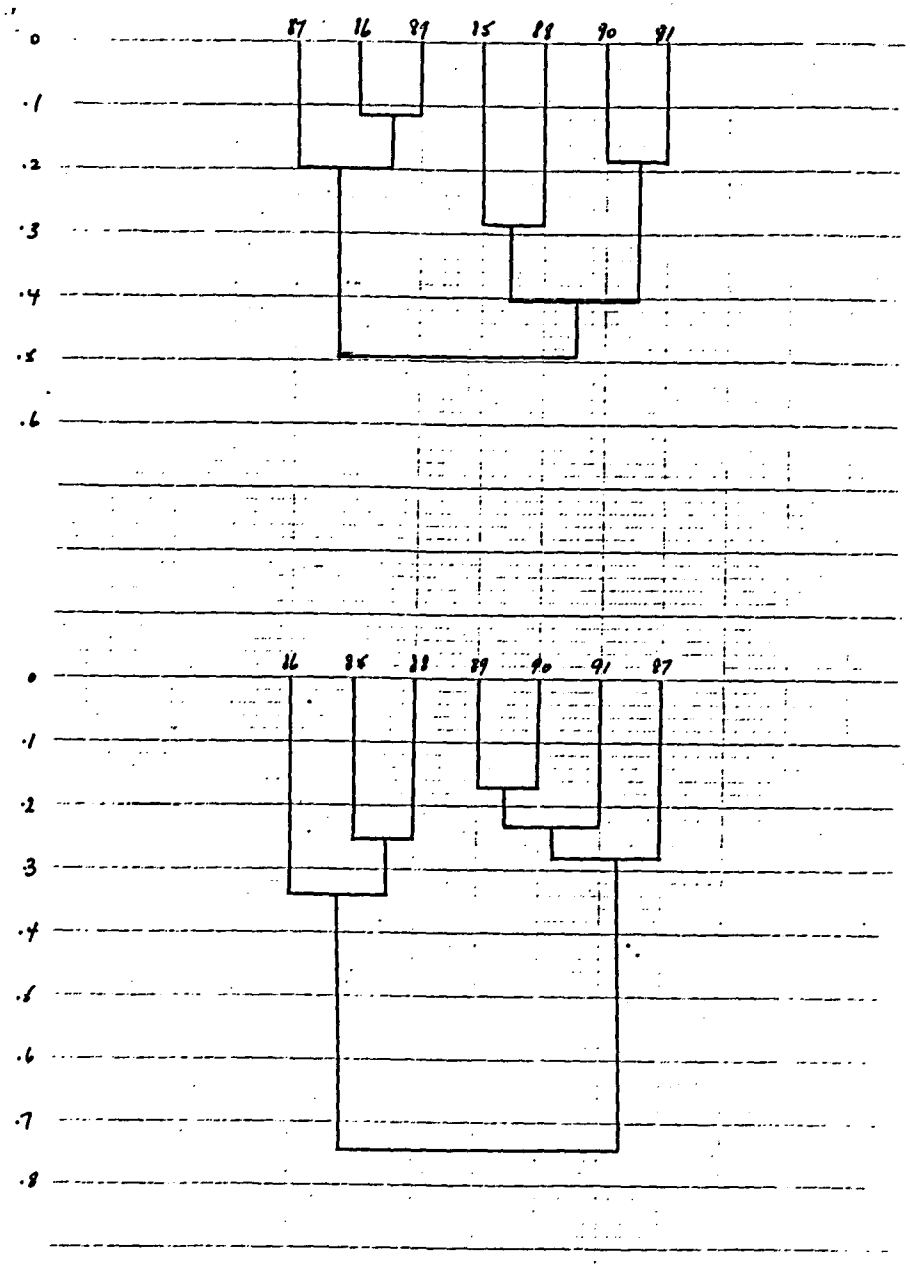


Fig. 4
Division 3K
Dendrogram based on estimated probability

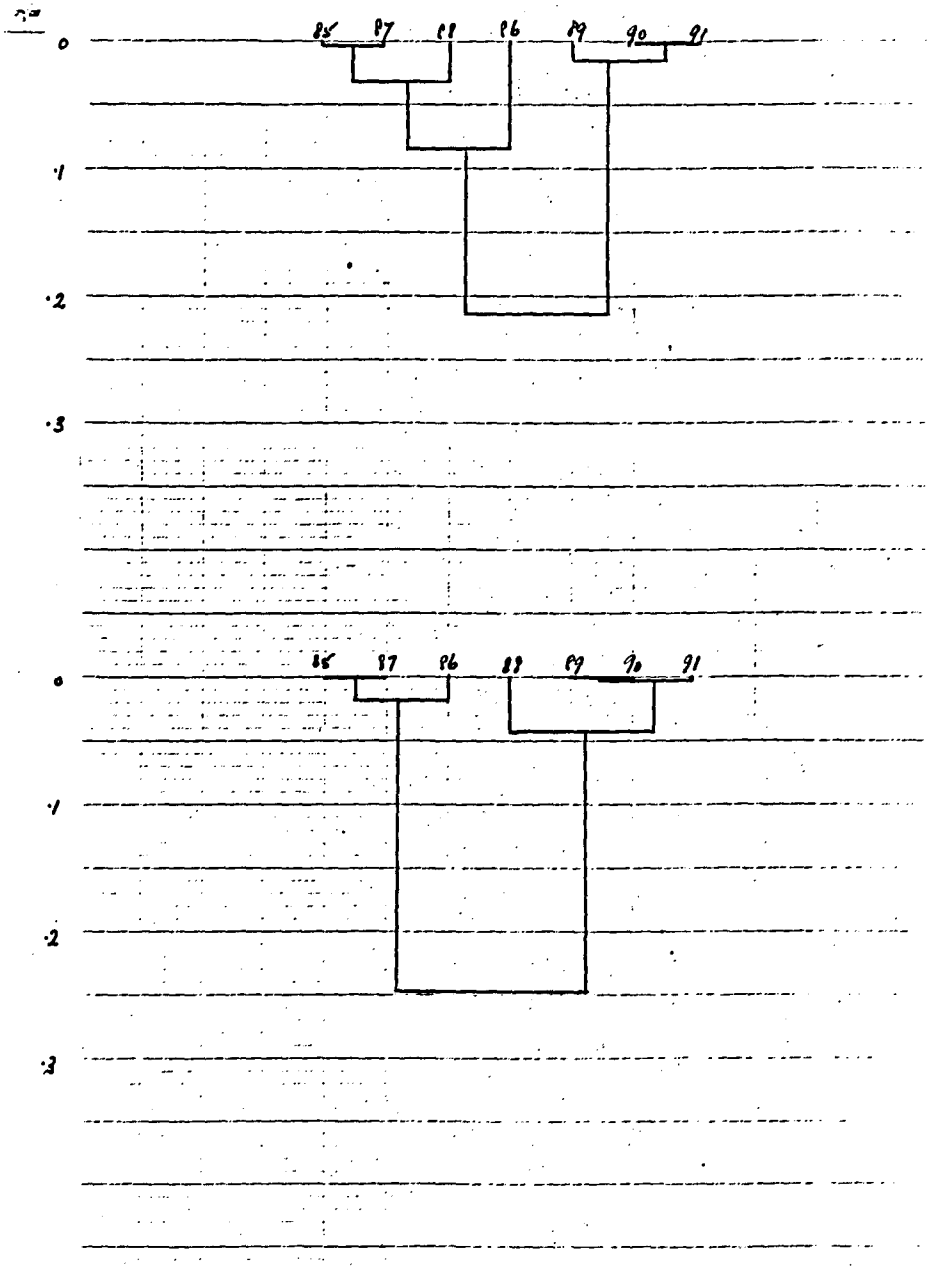

Fig. 5
Division 3L
Dendrogram based on $\varpi$

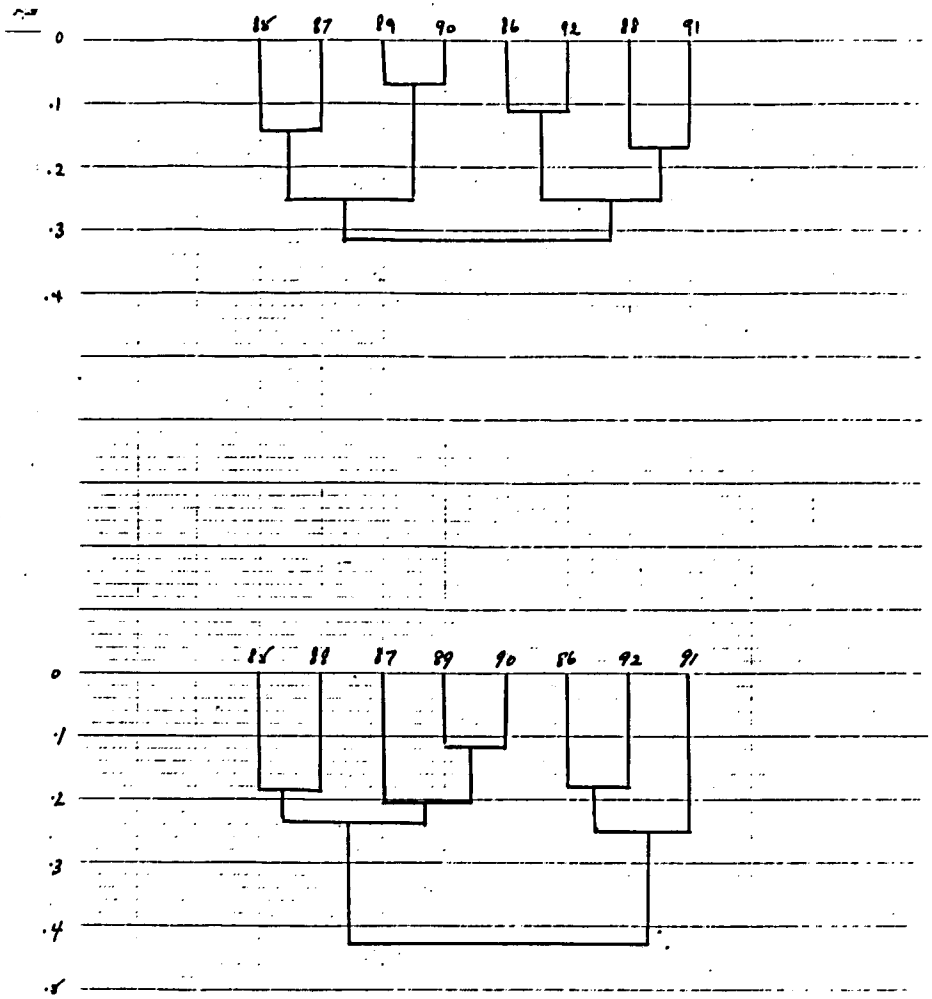


Fig. 6
Division 3L
Dendrogram based on estimated probability

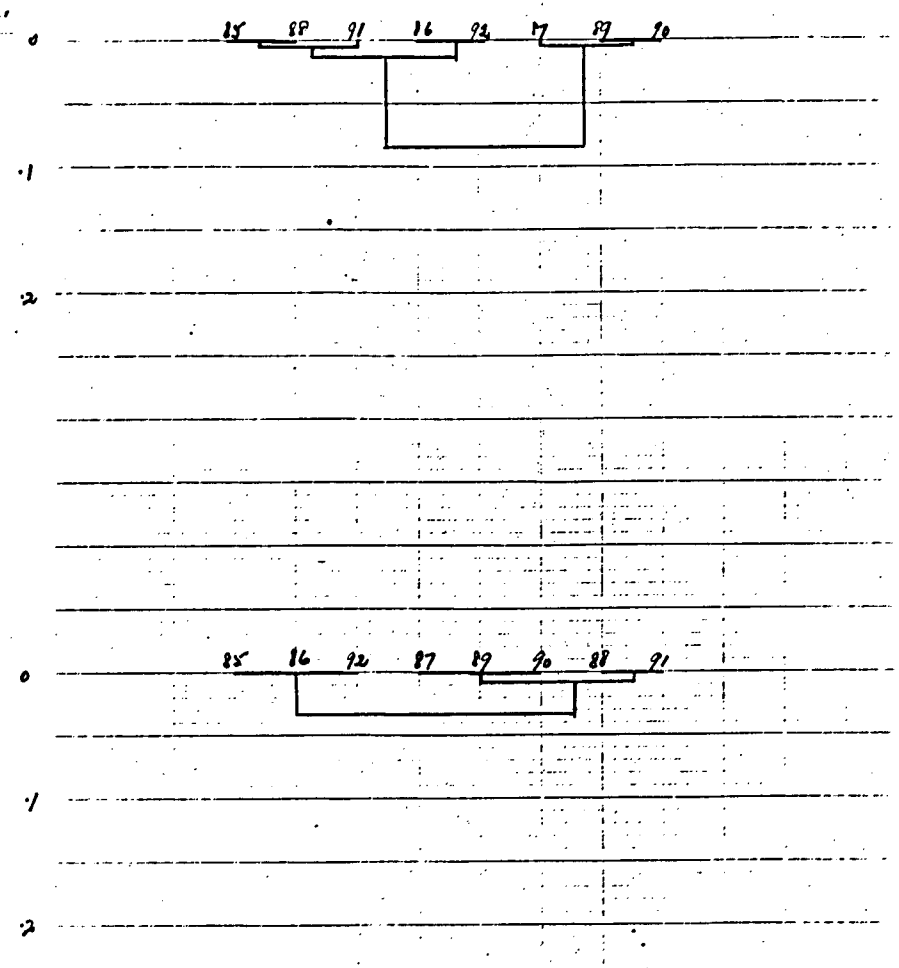

Fig. 7
Centres of Gravity of Trawl Biomass