MET O 11 TECHNICAL NOTE NO 93

Some experimental evidence on principal components

R DIXON

1. Introduction

Six years ago when the present operational analysis system based on orthogonal polynomials was being developed the idea was entertained that it might be possible to construct a set of eigenpolys which would be sufficiently similar to a set of principal components to share their renowned efficiency in the matter of extracting variance from a set of data. The intention was to use eigenpolys in place of standard orthogonal polynomials to speed up the analysis. Some experiments were carried out to establish just how much more efficient principal components are as a data fitting tool than orthogonal polynomials.

The experimental results were discouraging and the idea was abandoned. However the results are interesting and valuable in their own right and should be available to anyone having an interest in the use of principal components. The results raise a number of obvious theoretical questions and some of these questions I hope to deal with in a later note.

2. The experimental results

From daily values of 500 mb heights during 1965 to 1967 for the 130 points shown in Figure 1, 130 eigenvectors were computed. This data set will be referred to as the dependent data set. The eigenvectors were then ordered according to the amount of variance they each extracted from this dependent data set. The first 40 ordered eigenvectors were then taken, the rest were discarded and played no further part.

In carrying out this experiment we had in mind the well-known mathematical statistical theorem which states that such a set of ordered eigenvectors will be more efficient at extracting variance from the data set than any other set of linear functions of the data. This is certainly true, in a particular sense, of the dependent data set. What interested us was the extent to which this was likely to be true if the eigenvectors were used to extract variance from a similar but independent data set. We were especially interested in the case of single independent data vectors.

The single independent data vectors used were the 500 mb heights for 00Z on 1st January 1970, 2nd July 1970, and 29th July 1970. Bivariate orthogonal polynomials of the 14th power (120 terms) were fitted to each data vector and in each case the 40 most efficient terms with respect to variance extraction were selected. Then the three independent cases were fitted using the optimal ordered set of eigenvectors extracted from the dependent data set.

Figures 2, 3, and 4 show the cumulative % second-moment-about-zero accounted for by the eigenvectors (crosses) and the orthogonal polynomials (dots) in each case. The diagrams show that although the eigenvectors establish an early lead, this lead is lost by the time 10 terms have been used, although in the 2nd July case the eigenvectors regain the lead later. Of course it can be said that this comparison is unfair to the eigenvectors since the orthogonal polynomials have been variance-ordered with respect to each independent data vector. True, but the main point here is to test whether the theorem quoted above is likely to be true with respect to independent data. Is it to be expected that a set of eigenvectors will do better than any other set of linear functions on an independent set of data? There is no theoretical reason why they should as far as I know. The ordered sets of orthogonal polynomial terms can be regarded simply as other sets of linear functions. It is thus clearly unsafe to rely on the theorem holding for independent data. The most that can be said is that the first few eigenvectors will probably do better than some other functional terms but that as more terms are admitted the outcome is uncertain.

Another way of presenting evidence on this point is to number the optimal eigenvectors of the dependent data set from 1 to 40 and then list the corresponding variance ordering for the same eigenvectors used on the independent data vectors. In Table 1 below the top line numbers the optimal set and the second line shows the order in which they extracted variance from the independent data cases.

TABLE 1

1st Jan 1970

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	3	6	21	13	8	15	17	30	22	38	36	4	24	11	27	29	20	23

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
32	12	31	10	35	19	18	28	25	7	40	34	26	33	37	5	9	14	39	16

2nd July 1970

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	22	27	12	17	19	20	15	23	14	29	8	18	7	16	13	30	10	37	38

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
2	4	25	26	40	33	28	32	6	39	31	24	34	36	5	3	9	11	35	21

29th July 1970

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5	4	25	27	9	26	10	3	6	31	33	8	16	29	15	18	11	14	20

21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
24	19	7	12	38	2	34	32	40	39	21	30	17	23	35	37	28	13	22	36

It will be noticed that the dependent data set optimal ordering is not preserved on these independent cases. Indeed the departures from the original order are quite striking. As a rough guide to the correspondence between the original ordering and the order in which they extract variance from the independent data cases I have calculated the correlation coefficients. They are: 1st Jan 1970 $r = .35$, 2nd July 1970 $r = .19$, 29th July 1970 $r = .48$. The $r = .19$ for the 2nd July 1970 case is not even significant at the $p = .05$ level. I think this simply confirms what is apparent to the eye from Table 1.

It is also instructive to plot the cumulative percentages accounted for by the principal components when taken in their order of efficiency for each individual independent data vector, i.e. the order given by the second rows of Table 1. The results are shown in Figures 5, 6 and 7. It is seen that although in one case, 2nd July 1970 figure 6, the principal components now stay ahead of the orthogonal polynomials all the way, in the other two cases the orthogonal polynomials still catch up with and overtake the principal components. Thus in these two cases if a detailed representation of the data is required it might still be better to use specially ordered orthogonal polynomials than specially ordered principal components.

Finally some evidence supplied by M Colgate. On a grid of 106 points a set of surface pressures taken every ten days through the period 1949-1964 was used as a dependent data set. The eigenvectors were computed and ranked according to their extraction of variance efficiency with respect to this dependent data set. As an independent data set surface pressures every ten days from the period 1965 to 1971 were taken. The eigenvectors from the dependent set were used to extract variance from each of the data vectors making up the independent set. A frequency table was then compiled showing, for each eigenvector, how often it was the most efficient extractor of variance, how often it was the second most efficient, etc, etc. This table is shown as Figure 8, for the first 20 eigenvectors. To make the interpretation of this table quite clear, row one shows that eigenvector No.1. was the most efficient extractor of variance in 114 cases, it was the second most efficient extractor in 36 cases, the third most efficient extractor in 12 cases and so on; row 9 shows that eigenvector No.9 was the most efficient extractor of variance in 3 cases, the second most efficient in 9 cases, the third most efficient in 13 cases and so on. The table speaks for itself.

3. A few closing comments

Principal components are an appealing tool, but anyone intending to use them needs to consider this evidence. Such hard evidence is quite difficult to come by in the voluminous literature of the subject. It is not suggested that workers in this field are unaware that eigenvectors which are dominant with respect to the dependent data set may not be dominant with respect to any independent data, but they have not published much evidence showing just how badly this may be the case. Also it is not suggested that they are unaware that the optimality of the principal components applies only on average over the dependent data set and that eigenvectors which are dominant with respect to the dependent data set as a whole may not be dominant with respect to any individual data vector taken from that set, but again hard evidence on this point is very scarce. The point is briefly acknowledged on p.588 of W D Seller's paper in Vol.96, No.9, MWR Sept.1968, but it is a pity that he did not display the evidence.

There seems to be two further experiments which cry out to be done. First, a frequency table corresponding to Figure 8 should be compiled for the dependent data set. This would show to what extent one can rely upon the eigenvectors which are dominant for the dependent set as a whole being dominant for individual data, vectors drawn from that set. Second, a set of orthogonal polynomial terms optimal with respect to the dependent data set as a whole should be constructed and frequency tables corresponding to Figure 8 for the dependent and independent data sets compiled. We should then investigate on what percentage of occasions does the optimal orthogonal polynomial set do better than the optimal principal components when used on individual data vectors drawn from both the dependent set and the independent set. Finally, we can pose the question as to whether it is possible to construct a set of orthogonal polynomials which will do better on the dependent data set as a whole than the optimal set of principal components? Now this last question has a strong whiff of heresy about it, because it challenges the validity of the main theorem, and this theorem undoubtedly has a firm place in mathematical statistics. But all the evidence presented in this note and the further work suggested really bears on the problem of determining the practical validity and scope of this theorem, so we may as well ask the ultimate awful question. To paraphrase the opening sentence of Craddock's main Note on the topic, this theorem is one which is better known in the quoting than in the proving.

None of all this should be interpreted as meaning that I am against principal components. They appeal to ~~be~~ me too.

R Dixon

R DIXON
Met O 11

August 1977

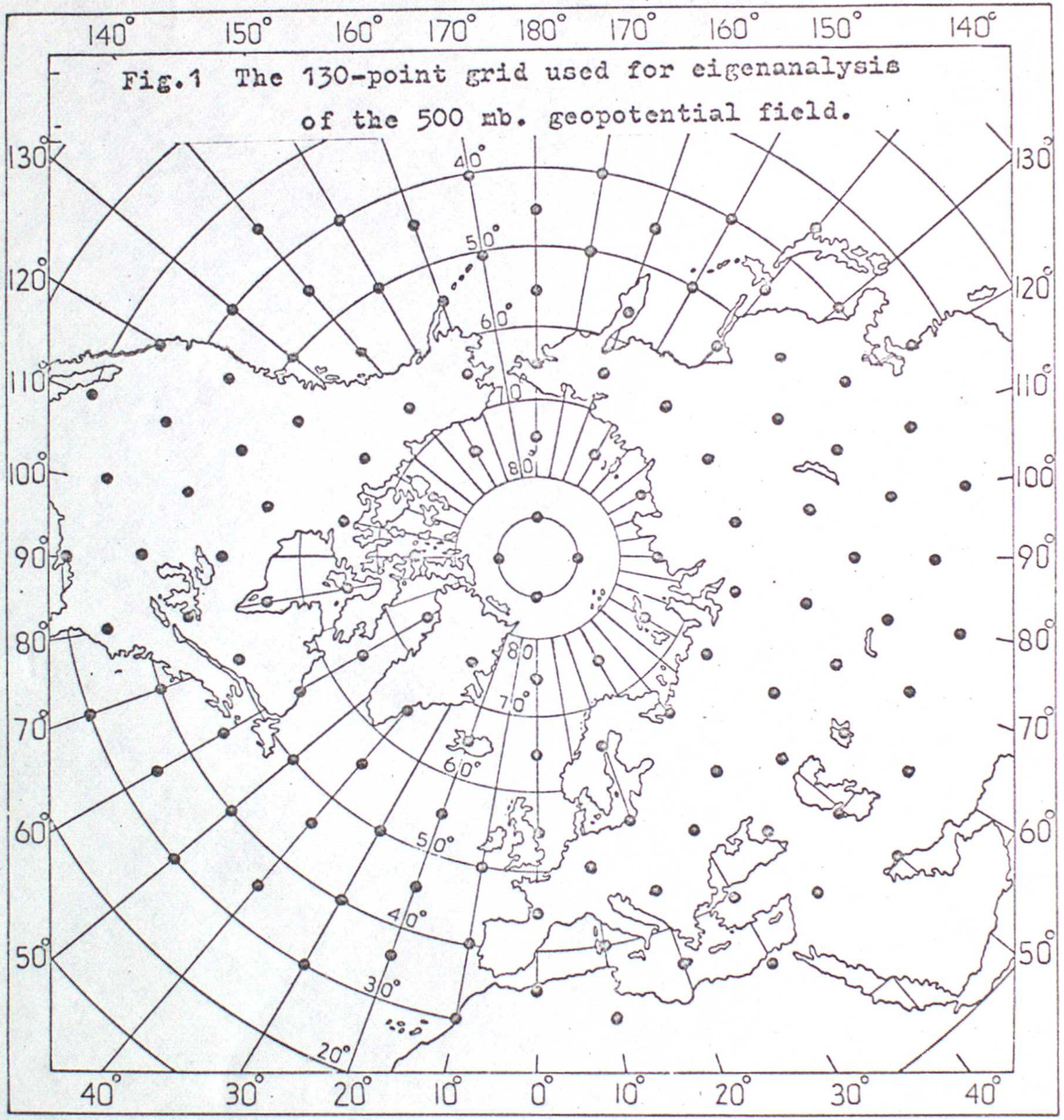


Fig.1 The 130-point grid used for eigenanalysis of the 500 mb. geopotential field.

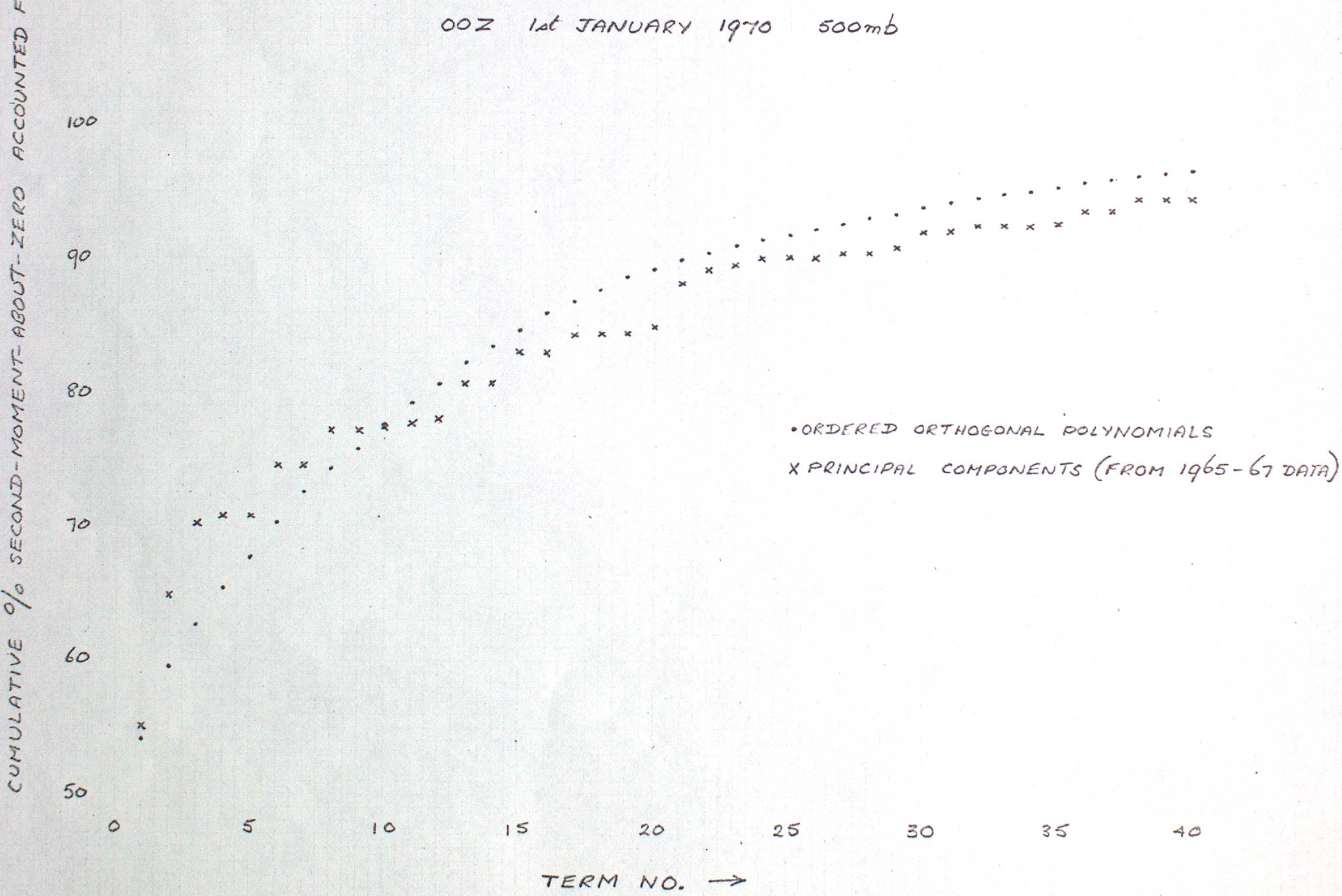
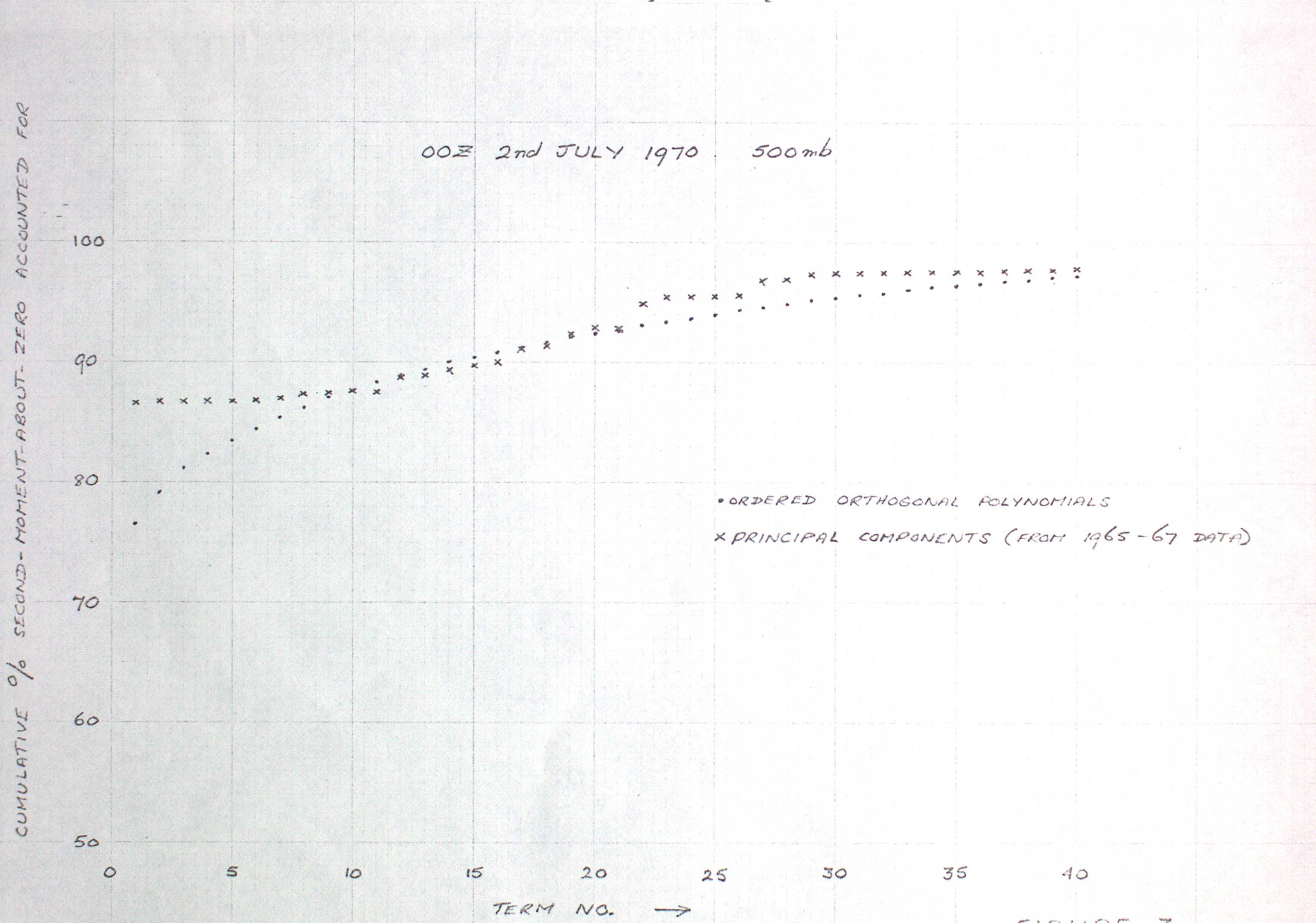

00Z 1st JANUARY 1970 500mb
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
100
90
80
70
60
50
0
5
10
15
20
25
30
35
40
TERM NO. →
• ORDERED ORTHOGONAL POLYNOMIALS
X PRINCIPAL COMPONENTS (FROM 1965-67 DATA)

FIGURE 2



00Z 2nd JULY 1970 500mb
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
100
90
80
70
60
50
0
5
10
15
20
25
30
35
40
TERM NO. →
• ORDERED ORTHOGONAL POLYNOMIALS
× PRINCIPAL COMPONENTS (FROM 1965-67 DATA)

FIGURE 3

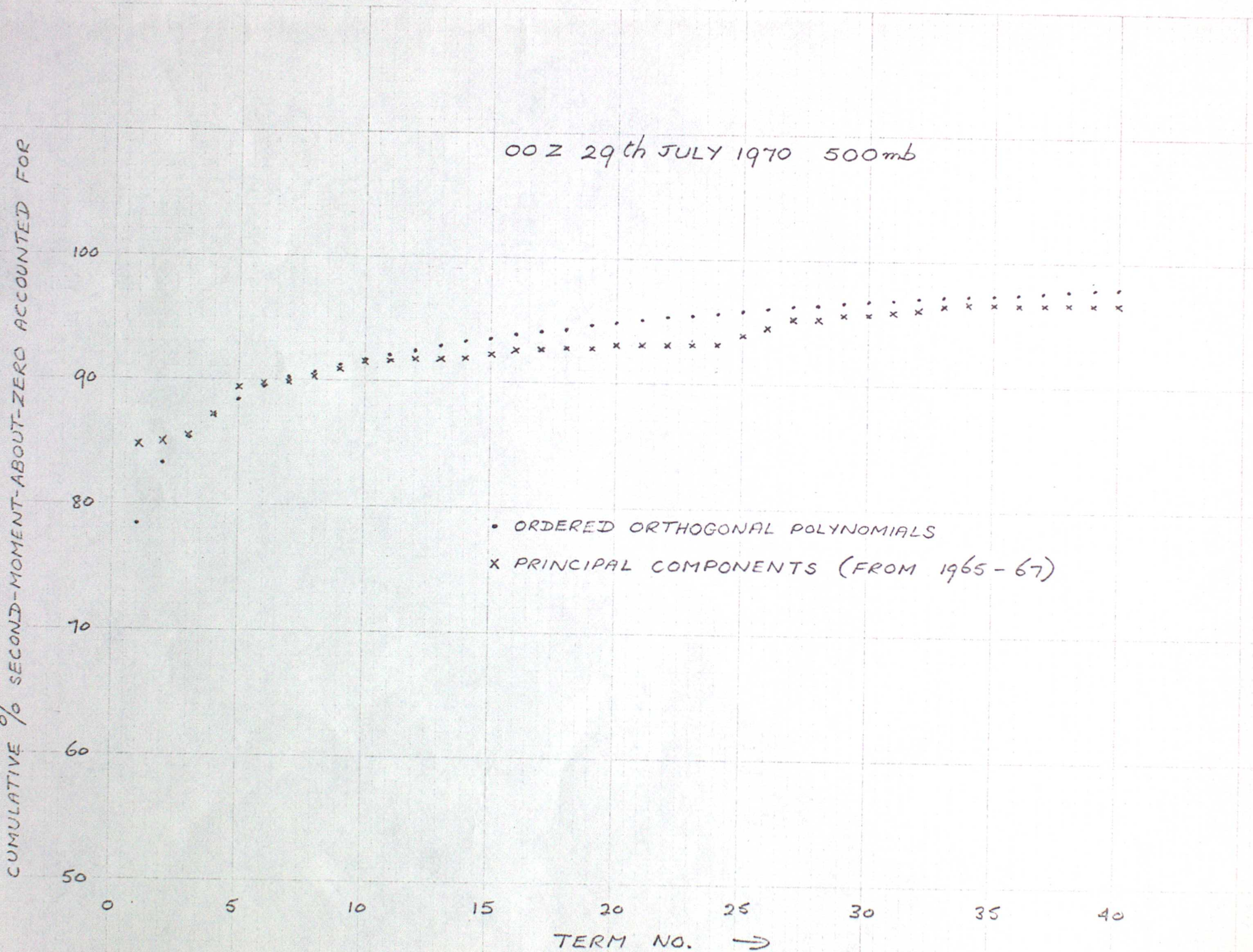
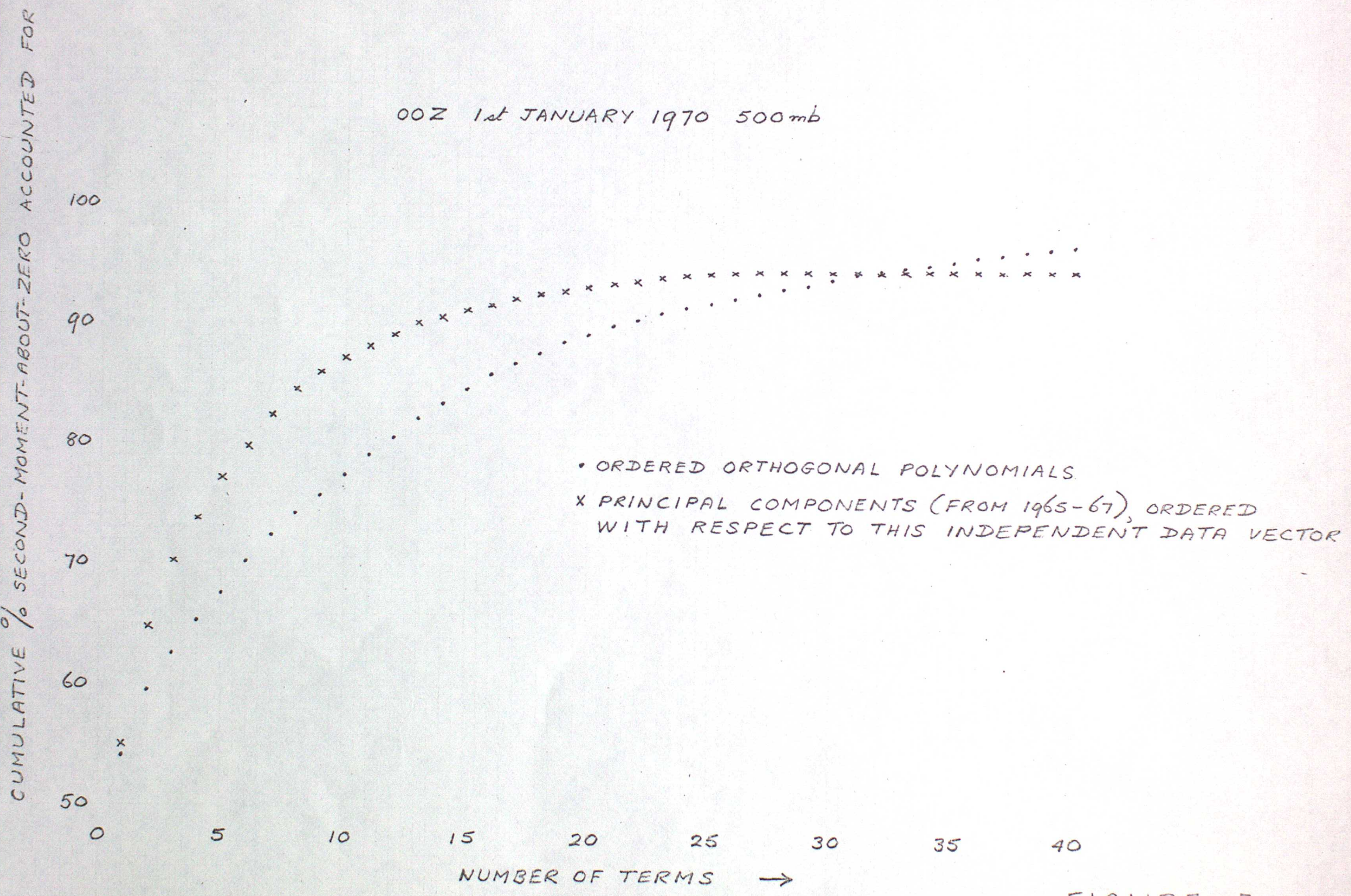
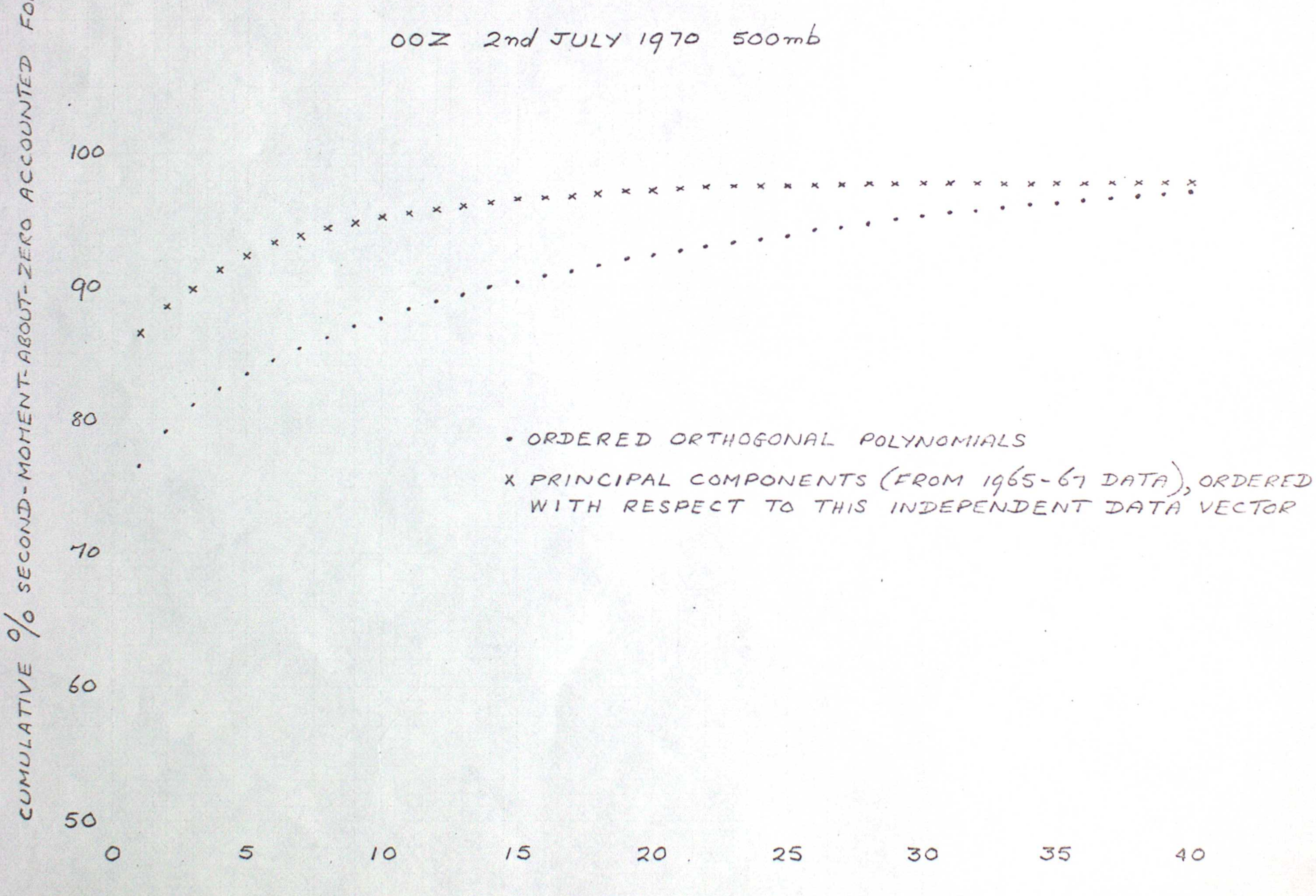

00 Z 29th JULY 1970 500mb
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
100
90
80
70
60
50
0
5
10
15
20
25
30
35
40
TERM NO.
• ORDERED ORTHOGONAL POLYNOMIALS
x PRINCIPAL COMPONENTS (FROM 1965 - 67)

FIGURE 4



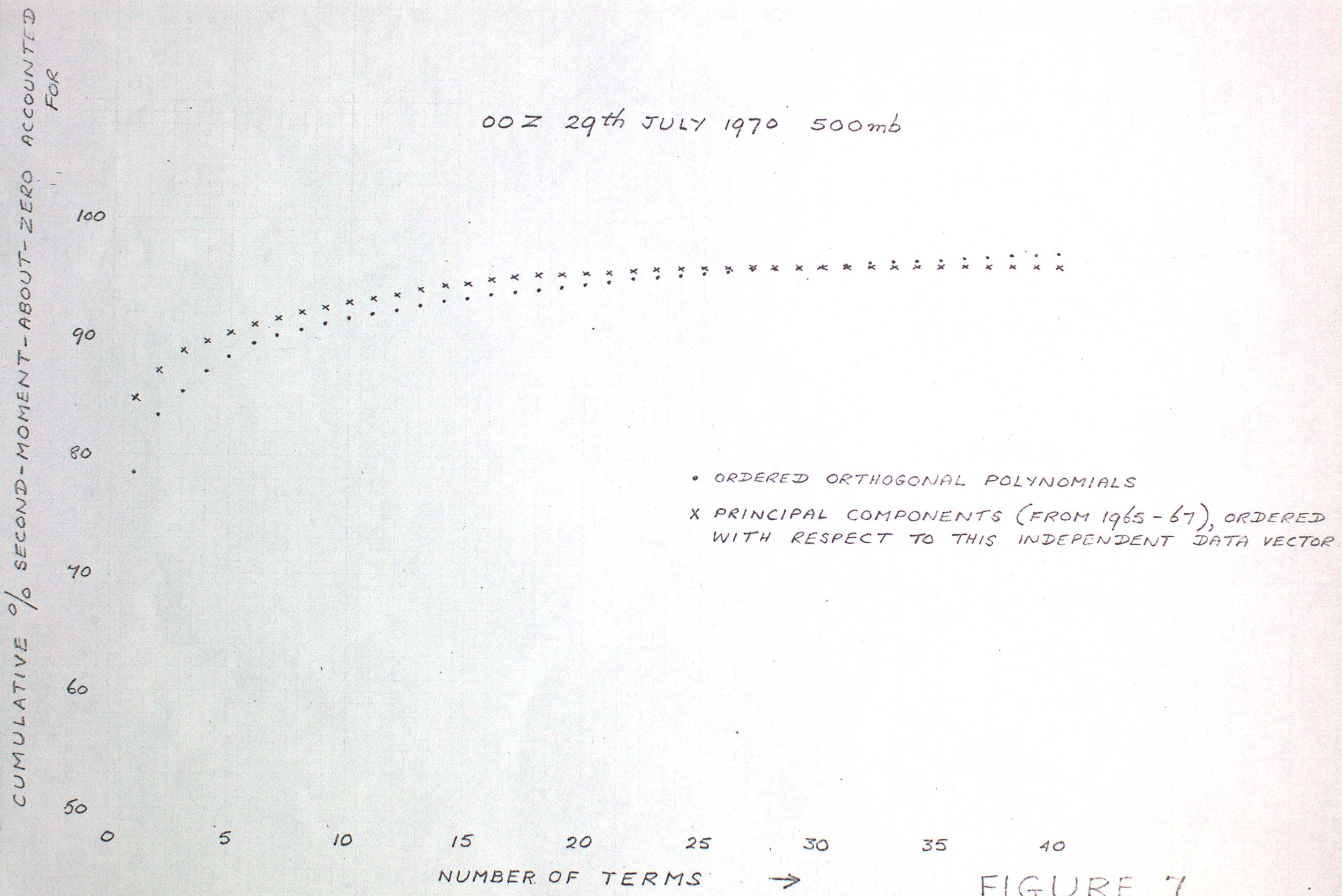
00Z 1st JANUARY 1970 500mb
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
100
90
80
70
60
50
0
5
10
15
20
25
30
35
40
NUMBER OF TERMS →
• ORDERED ORTHOGONAL POLYNOMIALS
x PRINCIPAL COMPONENTS (FROM 1965-67), ORDERED WITH RESPECT TO THIS INDEPENDENT DATA VECTOR

FIGURE 5



00Z 2nd JULY 1970 500mb
100
90
80
70
60
50
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
0
5
10
15
20
25
30
35
40
NUMBER OF TERMS →
· ORDERED ORTHOGONAL POLYNOMIALS
x PRINCIPAL COMPONENTS (FROM 1965-67 DATA), ORDERED WITH RESPECT TO THIS INDEPENDENT DATA VECTOR

FIGURE 6



00Z 29th JULY 1970 500mb
CUMULATIVE % SECOND-MOMENT-ABOUT-ZERO ACCOUNTED FOR
100
90
80
70
60
50
0
5
10
15
20
25
30
35
40
NUMBER OF TERMS →
• ORDERED ORTHOGONAL POLYNOMIALS
X PRINCIPAL COMPONENTS (FROM 1965-67), ORDERED WITH RESPECT TO THIS INDEPENDENT DATA VECTOR

FIGURE 7

FREQUENCY AS n-th MOST EFFICIENT EXTRACTOR OF VARIANCE FROM DATA VECTORS OF THE INDEPENDENT DATA SET

EIGENVECTOR No.																														
1	114	36	12	8	18	11	10	11	2	5	1	0	2	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	24	27	26	30	25	16	20	20	9	5	3	3	2	4	3	0	4	3	1	2	1	0	5	1	0	0	0	0	0	0
3	15	15	19	26	16	12	32	12	9	12	8	10	5	8	9	4	9	1	1	2	1	3	3	0	0	0	0	1	0	1
4	7	24	13	14	14	15	7	16	14	13	11	13	10	13	1	5	3	6	3	7	6	3	0	2	2	2	1	1	2	0
5	10	10	13	15	14	9	22	16	11	8	10	6	8	11	6	6	7	5	10	6	3	3	6	0	4	1	2	1	1	1
6	8	12	13	8	14	13	13	10	10	13	14	11	5	6	5	9	8	7	10	4	5	6	4	3	0	3	2	0	2	2
7	5	6	10	9	12	15	10	11	11	5	6	11	7	13	8	15	4	4	6	8	4	4	2	4	2	5	2	7	3	6
8	4	3	5	9	11	14	12	6	5	9	11	10	7	8	5	8	2	5	7	4	7	6	6	6	6	5	3	3	4	3
9	3	9	13	9	6	6	5	11	14	5	7	4	10	3	7	4	11	11	9	8	7	7	6	2	2	8	7	3	2	6
10	2	4	7	9	6	7	7	10	5	2	5	7	11	11	4	6	8	9	15	9	5	6	6	4	5	3	4	3	6	3
11	2	7	7	8	5	9	7	7	4	12	6	9	9	3	7	9	5	7	3	11	11	5	7	9	7	3	4	6	2	4
12	3	10	1	4	6	4	11	7	6	9	4	8	7	3	3	10	5	4	6	7	5	5	2	5	6	8	9	5	8	3
13	3	4	5	5	8	3	1	5	7	7	5	9	11	1	8	8	8	12	5	10	4	6	6	5	8	6	4	4	8	3
14	3	5	5	4	3	3	3	2	10	4	5	6	7	3	9	5	9	8	13	10	5	6	8	7	7	5	6	6	7	3
15	0	3	1	4	3	5	4	8	5	4	8	7	4	4	11	5	6	7	8	5	7	13	1	1	2	2	8	9	6	6
16	3	4	3	1	4	8	1	4	9	10	8	9	4	7	5	5	6	5	4	5	10	4	5	6	7	7	6	4	8	1
17	3	1	3	3	6	3	1	4	5	5	4	1	5	5	4	6	7	6	5	5	3	6	7	2	6	10	10	8	4	5
18	1	1	1	6	4	7	5	6	5	6	4	3	4	5	6	6	3	6	5	7	2	5	1	8	6	5	2	5	4	3
19	1	1	3	3	4	4	2	1	5	4	5	6	8	3	5	2	8	5	7	4	7	7	5	4	4	3	6	7	7	7
20	0	2	3	3	1	6	5	4	5	2	3	3	3	3	5	2	4	10	10	5	2	8	11	3	2	2	7	7	3	7

FIGURE 8

FREQUENCY TABLE SHOWING THE NUMBER OF TIMES THE RANKED EIGENVECTORS OF THE 1949 - 1964 DEPENDENT DATA SET CAME FIRST, SECOND, THIRD, ETC. IN TERMS OF VARIANCE ACCOUNTED FOR WHEN USED ON THE 1965 - 71 INDEPENDENT DATA SET. THE TABLE GIVES THE FIGURES FOR THE FIRST 20 RANKED EIGENVECTORS OF THE DEPENDENT DATA SET.