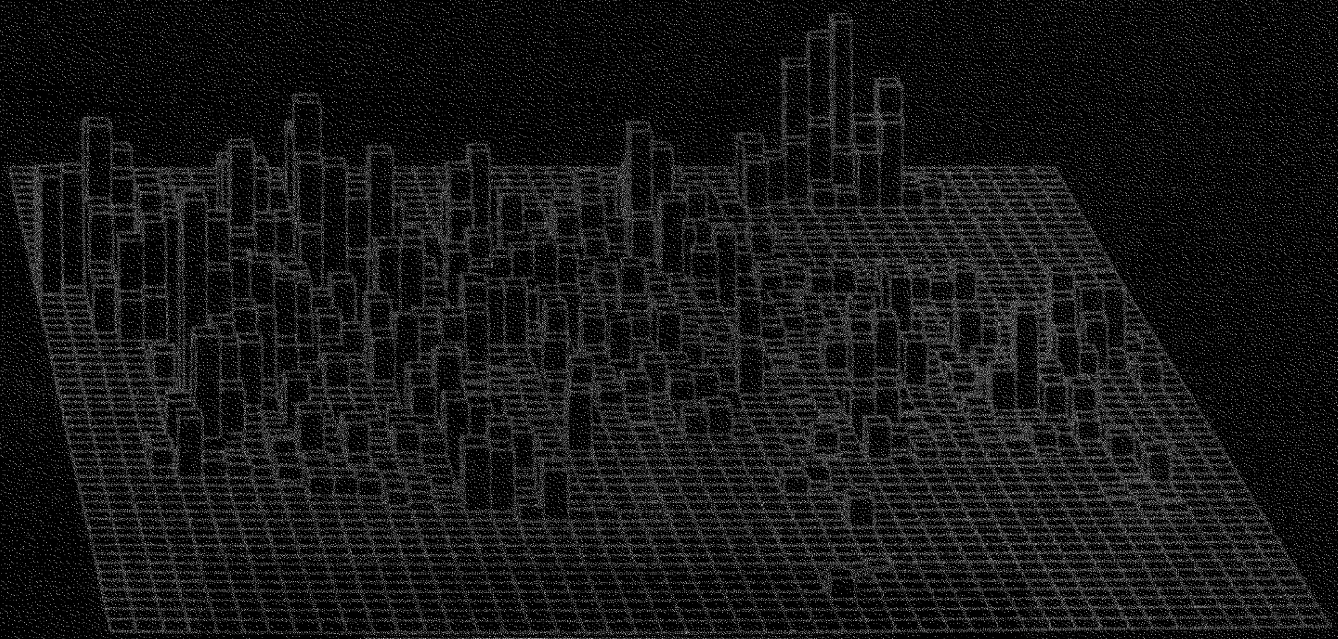
CRU

Storage of the grid-square based 1971 G.B. Census data; checking procedures

by M. Visvalingam and B.J. Perry



Census Research Unit
Department of Geography
University of Durham

Working paper 7

The Census Research Unit, Department of Geography, University of Durham, is a small group of research workers investigating aspects of the theory and use of census data. It is currently funded as a research project by the Social Science Research Council.

The diagram on the cover represents total population per 1 km grid square in the northern part of County Durham: the height of each column is proportional to the population in that square. The county is viewed from the west, Gateshead being at the extreme left margin, West Hartlepool at the far right and Bishop Auckland at the centre-right. The original surface was calculated and drawn by computer.

UNIVERSITY OF DURHAM

DEPARTMENT OF GEOGRAPHY

CENSUS RESEARCH UNIT

WORKING PAPER No. 7

OCTOBER 1976

STORAGE OF THE GRID-SQUARE BASED 1971 GREAT BRITAIN CENSUS DATA : CHECKING PROCEDURES

M. VISVALINGAM

and

B.J. PERRY

CONTENTS

STORAGE OF THE GRID-SQUARE BASED 1971 GREAT BRITAIN CENSUS DATA : CHECKING PROCEDURES

1. INTRODUCTION

The 1971 population census Small Area (Ward Library) Statistics for Great Britain were made available to the Census Research Unit (CRU) on a grid-square basis by the Office of Population Censuses and Surveys (OPCS). The structure and format of this data set are described in an OPCS handbook[1].

The 100% population and household data were received by CRU on nineteen ICL-phase-encoded nine-track magnetic tapes; later, the 10% data arrived on eleven tapes. As the Northumbrian Universities Multiple Access (NUMAC) system functions with IBM 360 and 370 computers, it was necessary to convert these 30 tapes from OPCS, ICL 1900 form to the IBM 360/370 form. Each ICL computer word corresponds to three IBM bytes, whereas an IBM word consists of four bytes; thus converting the data on a word-to-word basis increased the requirements to 40 tapes. The details of this conversion process have been discussed elsewhere[2].

Even with a virtual memory computer system such as NUMAC, it was impossible to process this volume of data as a single data set. To overcome this problem, the body of data was partitioned into subsets, each consisting of data for a strip of 100-kilometre blocks extending from west to east across Great Britain. These subsets were processed and checked independently, and were subsequently accumulated in a master data file.

To minimise storage requirements and to facilitate subsequent processing, the data were stored in a highly compacted[3] and geographically sorted form. However, the fundamental concern during the storage process was preservation of the data content throughout the series of computer-based manipulations. This paper outlines the limited logical and arithmetic checks used to ensure that the data received conformed to the definitions[4], and the subsequent manual and program checks which were employed to verify that the data remained as provided by OPCS.

2. SOURCES OF ERROR AND CHECKING PROCEDURES

Errors could enter the storage system at any stage of the manipulations. The magnetic tapes from OPCS were checked on receipt to verify that the information on the tape tallied with specifications in the explanatory notes supplied by OPCS. A few of the tapes were obviously in error and had to be returned. Normally, this type of error was very quickly identified in the pre-compression stages (see 2.1 below). Once the tapes had been converted from ICL to IBM form, further checks were undertaken to ensure that no other errors had been introduced by incorrect algorithms or carelessness during our own production runs. Procedures were devised whereby several algorithms checked each other and the control data. The program user in turn carefully checked the output from the programs at each manipulation stage.

2.1. Pre-Compression Stages

2.1.1. Listing of OPCS tapes

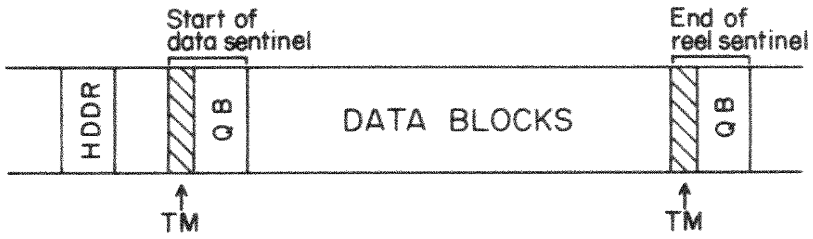
In general, each data file supplied by OPCS extended over one or more tapes (multi-reel format). Each tape was accompanied by a line-printer character and numeric listing of the first and last ten data blocks on the tape (Listing A), which proved useful for certain checks and verification procedures on the OPCS tapes and also for investigating the constancy of OPCS tape formats.

On receipt, these listings were visually scanned for obvious errors, such as repetition of the data record for a one-kilometre square and gaps in the page sequence[1] (the ICL nomenclature for the record number within a 100-kilometre block). One OPCS file, covering three tapes, had to be returned at this initial stage because data for a one- kilometre record were repeated several times under different page sequence numbers.

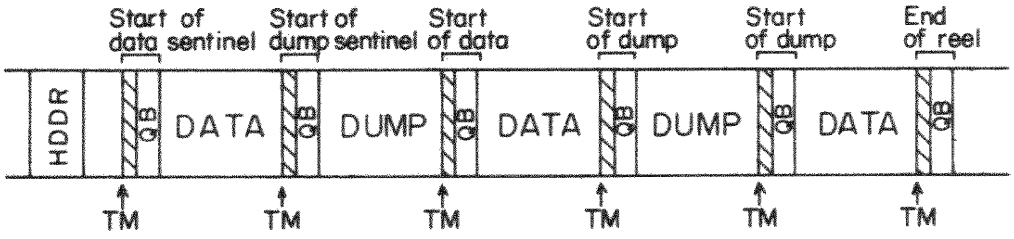
As the IBM magnetic tape system does not check the labels of "foreign" tapes, Listing A was used to verify that the correct tape had been mounted. Each time an OPCS tape was mounted at NUMAC, the ICL header was converted in terms of characters and visually checked against Listing A, which also gave the header in character form. This proved a useful check in some instances when tape identifications

Fig I

A) Simple Format



B) Tape with Dumps



TM tape mark
HDDR header label
QB qualifier block

Figure 1 : OPCS tape formats

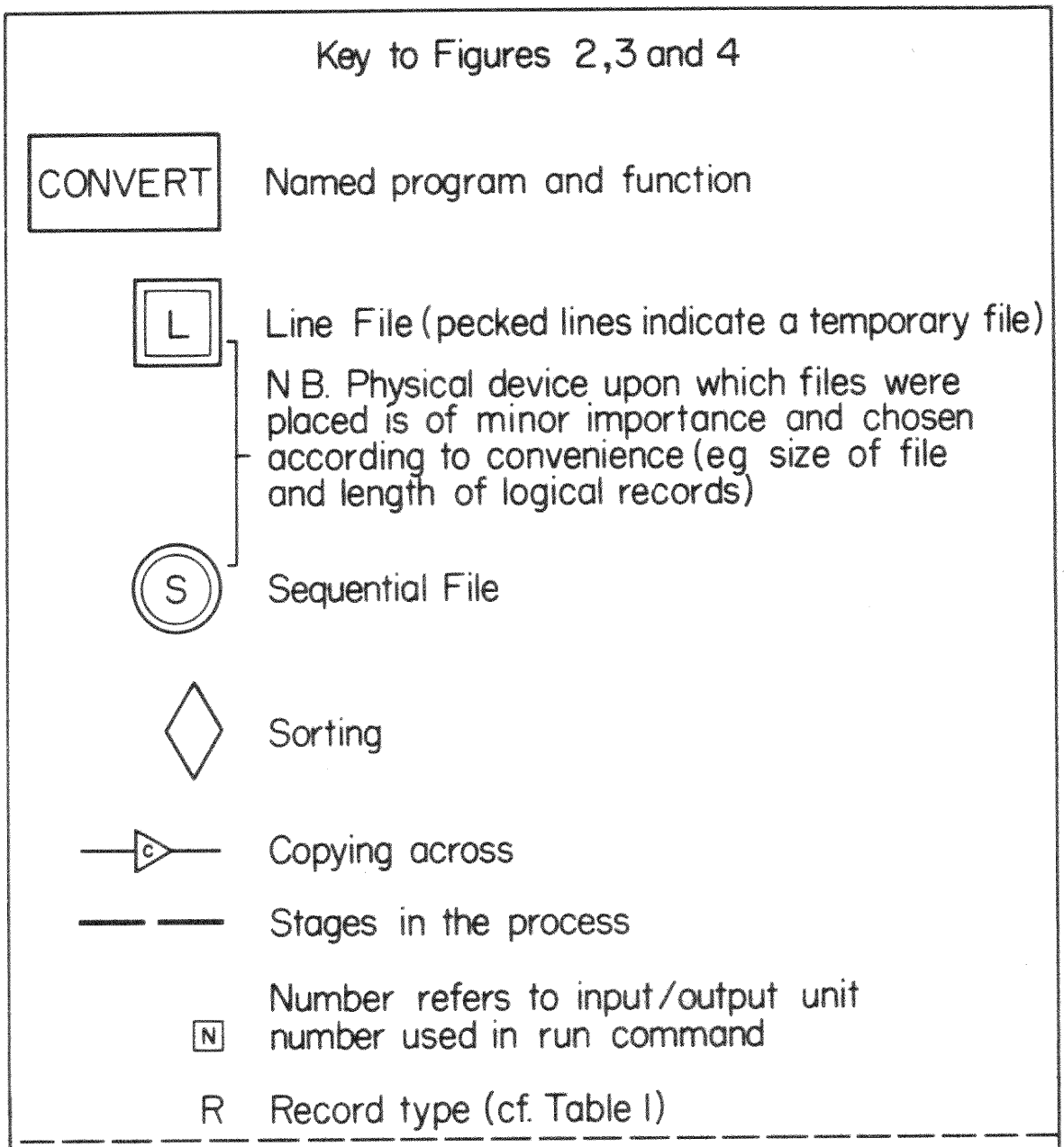


Figure 2 : Conversion from ICL 1900 to IBM 360/370 - readable form.

were quite similar.

Whilst visually scanning these listings, format differences between sets of 100% tapes were recognised. The first OPCS file received and two others (extending over eight tapes) had a very simple format (Fig. 1A) and consisted of uninterrupted pairs of records for each kilometre square (population followed by household), with reel boundaries (end of tape) occurring after such a pair. However, in one OPCS file covering four tapes, a slightly different pattern occurred. Here, the reel boundary came after a population record and the matching household record was the first data block on the succeeding tape. This arrangement was quite awkward as the CRU conversion process (section 2.1.2. below), which processed one tape at a time, separated population and household records into separate files. Both the resulting files accessed a single copy with common grid references in a third file. To ensure that the population and household records in each file corresponded to the single copy with common grid references, it was convenient to tag the individual household record from one tape to the corresponding population record on the preceding tape.

Of greater importance was the discovery that, in one OPCS file extending over seven tapes, the data sections on the tapes were separated by a variable number of interrupts containing dump information (Fig. 1B), which were used by OPCS during their internal data processing. A preliminary test conversion of the data blocks immediately preceding and succeeding each dump showed that no data records had been lost as a result of these interrupts; a population record preceded each dump and the matching household record followed the dump. To facilitate ICL to IBM conversion, the fragmented data blocks from these tapes were copied over to another magnetic tape. As the tape marks (TM) heralding the ICL sentinel are interpreted by the IBM system as end-of-file markers, the first (header label) and every alternate file (the dumps) were skipped in this copying-over process. The first record of each data section, being the qualifier block (QB) of the sentinel, was also irrelevant and was therefore ignored. This was an error-prone procedure and a large number of records (in 100-kilometre square SJ) at the end of one tape were accidentally omitted

during initial processing of the data. Consequently, CRU requested OPCS to supply the 10% data in the original, simple format (Fig. 1A).

2.1.2. Conversion from ICL 1900 to IBM 360/370 - readable form

The conversion from ICL to IBM format was effected by a specially-written utility program, CONVERT, discussed in reference 2. The OPCS 100% tapes, containing alternating population and household records, were separated into two files during conversion for reasons outlined in reference 3. The duplicate storage of grid reference in a separate file was convenient for some of the subsequent processes (Fig. 2). The same procedure was adopted for the alternating records in the 10% file.

TABLE 1 : RECORD TYPE CHARACTERISTICS

Record Type	Block Length	CRU Code	Number of OPCS Variables (elements)		Total Number of Records excluding seamen		
			Full Records	Suppressed Records	1 km	10 km	Pseudo
100% Population	2012	1	471	2	152440	2721	-
100% Household	1924	2	449	1	152440	2721	-
10% Type 1	1600	3	368	0	87975	2569	2549
10% Type 2	1260	4	283	0	87975	2569	2549

The program user checked that the number of records reported as converted by the routine was as specified by refering to Listing A. In addition, the MTS utility program *LABELSNIFF was used to determine the minimum and maximum block lengths on the IBM output tapes; these should both have been as in Table 1. The total number of records reported by the conversion routine was double-checked with the total number of blocks on the tape as reported by *LABELSNIFF. As data for a 100-kilometre block often spanned across two tapes and as the 100-kilometre blocks were not necessarily stored in any consistent order, the record number and the relevant parts of the ICL record headers on the IBM output tape were listed (Listing B) and filed for each converted tape. This made subsequent manual separation of records by 100-kilometre blocks relatively easy.

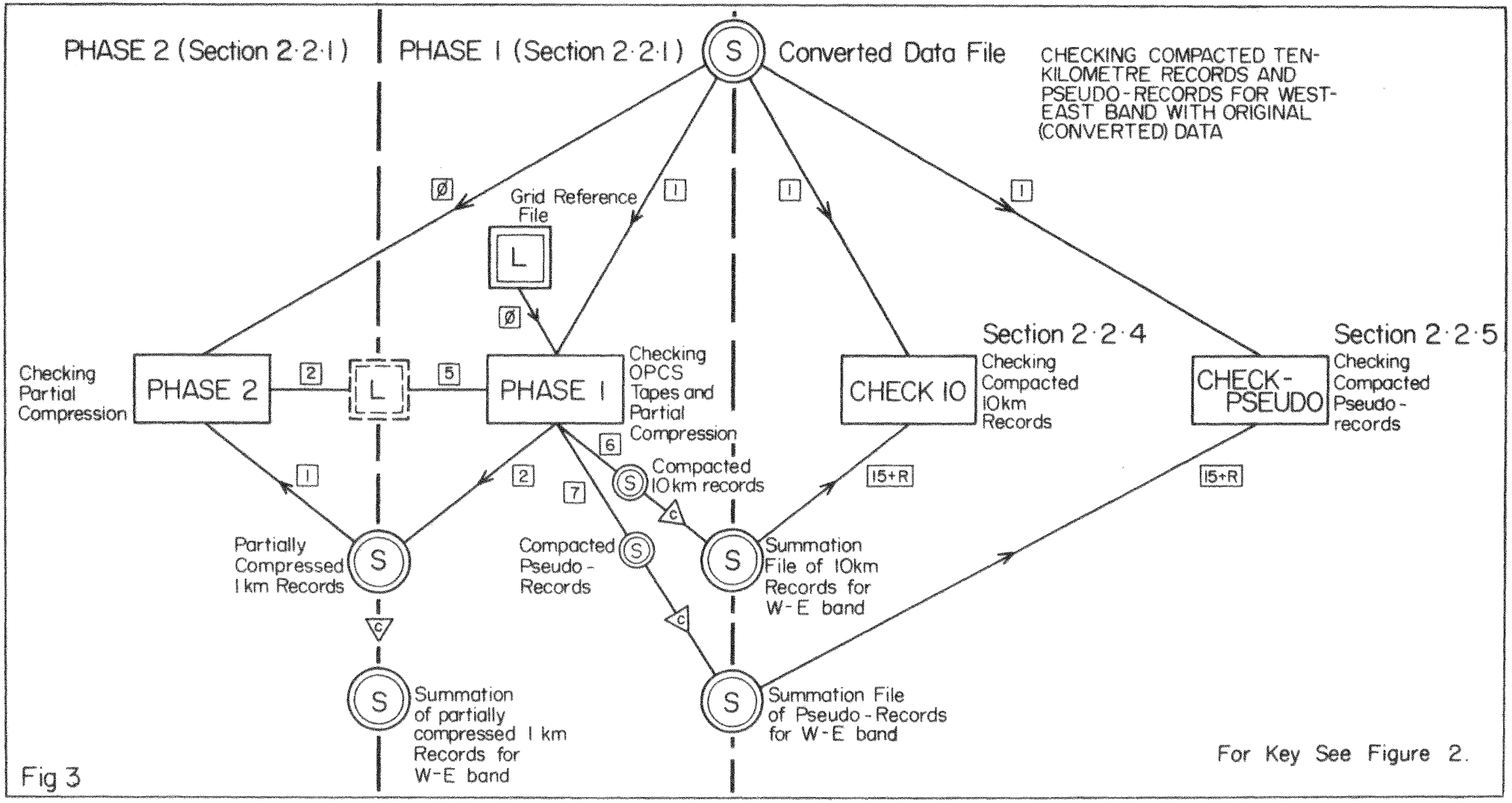


Figure 3 : Checking of OPCS tapes and partial compression.

2.2. Partial Compression

As far as possible, the following processes were organised to operate on subsets of the data, namely G.B.-wide west to east strips of one or more 100-kilometre blocks. In general, this meant the manual extraction of data from several IBM tapes.(At a later stage, while 10% data were being processed, NUMAC made available an additional 3330 disc pack on a temporary basis. This allowed the automatic streaming of data for each 100-kilometre strip into different files). The processes were similar for all record types : these were identified by codes 1, 2, 3 and 4 (see Table 1) to cope with any detailed differences such as record lengths, suppression thresholds,etc.

2.2.1. Checking of OPCS tapes and partial compression

This stage consisted of a two-phase operation (Fig. 3). Phase 1 among other functions (see below), checked the OPCS tapes and produced a partially compressed data file. Phase 2 checked the compressions performed in Phase 1 by expanding the compressed records and checking each data value with those in the original file. The detailed functions of each phase and the attendant manual checks are described below :

Phase 1 : The Phase 1 program for partial compression had several functions :

(i) It checked that the converted OPCS tape conformed to the specifications in the OPCS handbook[1]. Both the converted data file and the common grid reference file (see section 2.1.1. above) were input to the program to ensure that no incorrect data file had been assigned as a result of user carelessness. The page sequence numbers of records within each 100-kilometre block were checked to make certain that there were no gaps in the data (see below for procedure). The possibility of erroneous suppression of records for one-kilometre squares with more than 25 people or 8 households was investigated. The program also reported unsuppressed full records for which, as a result of OPCS adjustment procedures[4], the published totals in the 100% files fell below these suppression threshold levels.

(ii) Once the OPCS record had been checked for consistency, only the grid references were extracted from the OPCS record headers and attached to the CRU four-byte record header. The latter contained information on record type, on whether OPCS suppression had occurred, and on the storage requirement in bytes for each data element (based on the maximum data value in that record).

(iii) A count of the total number of males and females was added to the end of unsuppressed 100% population records, since these totals were provided by OPCS only in the case of suppressed records.

(iv) The suppressed records were truncated to contain only the required number of elements (Table 1).

(v) The resulting partially-compressed records were separated into a one-kilometre data file and a ten-kilometre summary file.

(vi) For the 10% data, a separate file was also required for pseudo-records[1].

(vii) For each one-kilometre record, some additional checking information was stored in a temporary file. This consisted of the grid reference, the length of record (in terms of the total number of data elements stored) and the number of bytes required for the storage of each element. The MTS editing program *ED was subsequently used to scan this file and to note if any of the one-kilometre records required more than two bytes of store per element. This storage strategy was based on the expectation that no value (except the count for seamen) at this spatial resolution level would require more than two bytes. The count of seamen was an exception because OPCS had allocated all seamen to one grid location (6000, 4000), providing identical one-kilometre and ten-kilometre records. As the total number of seamen (37,451) required more than two bytes of storage and because of their arbitrary spatial allocation, no further compaction of the one-kilometre seamen record was attempted and a separate file was created to hold these data.

Manual checks involved the visual scanning of the line-printer output of the Phase 1 program (Listing C). This was used initially to confirm that the expected number of records had been processed.

TABLE 2 : EFFECTS OF COMPRESSION

Record Type	Size of 1 km files (pages)		Compaction Factor	Read time* (CPU secs)	Transfer time (CPU sec)	Size of file (pages)*	
	Converted OPCS data (a)	Fully compacted CRU file (b)	(a/b)			Pseudo data	Summary data
1	74880	7482	10.01	132	41	-	687
2	71605	7699	9.30	133	34	-	674
3	34365	2390	14.38	105	18	439	456
4	27603	1513	18.24	87	17	339	349
TOTAL	208453	19084	10.92	457	110	778	2166

	778
	19084
Total storage requirements	22028

* can be reduced further

The output also reported the number of one-kilometre and ten-kilometre records within each 100-kilometre block. The number of one-kilometre and ten-kilometre pseudo-records was also reported for the 10% data tapes. These were checked with Listing B (section 2.1.2.) For each 100-kilometre block there should have been only one set of reports in Listing C, any repeats indicating either a duplication or a gap in page sequence. Where data for a 100-kilometre block spread over two tapes, Listings C for both tapes were compared to ensure that the page sequence numbers were consecutive. The parially-compressed one-kilometre and ten-kilometre record files were checked (using the MTS editor program *ED) to ensure that the number of records in each tallied with the totals indicated in Listing C. This picked out instances where the MTS sequential files used for intermediate storage had not been emptied for that particular job. As each write to a sequential file is always appended to the end of the file, this trap enabled the user to detect and recover from situations which resulted from carelessness. Previously, duplicates resulting from carelessness had not been detected until the final compression stage.

Phase 2 : The phase 2 program re-checked the content of the partially-compressed file by expanding each record and comparing data values with those of the comparable record in the original file. It also gave an indication of the proportion of zeroes in the precompressed data and estimated the size of the file necessary to store the data after full compression.

2.2.2. Comparison of original (after conversion) and compressed 100% and 10% summary records

The ten-kilometre summary files (section 2.2.1. above) were not compressed any further, owing to their small size and anticipated less frequent use. A routine (CHECK1Ø in Fig. 3) compared the original (converted) data with the expanded ten-kilometre records to ensure that all the relevant data had been processed and were free from corruption errors. After this check, the summary data for the west-east band being handled were sorted and appended to the ten-kilometre multivariate format master file for Great Britain : this file finally contained 2721 records (excluding the summary record for seamen); Table 2 gives details of the size of this file.

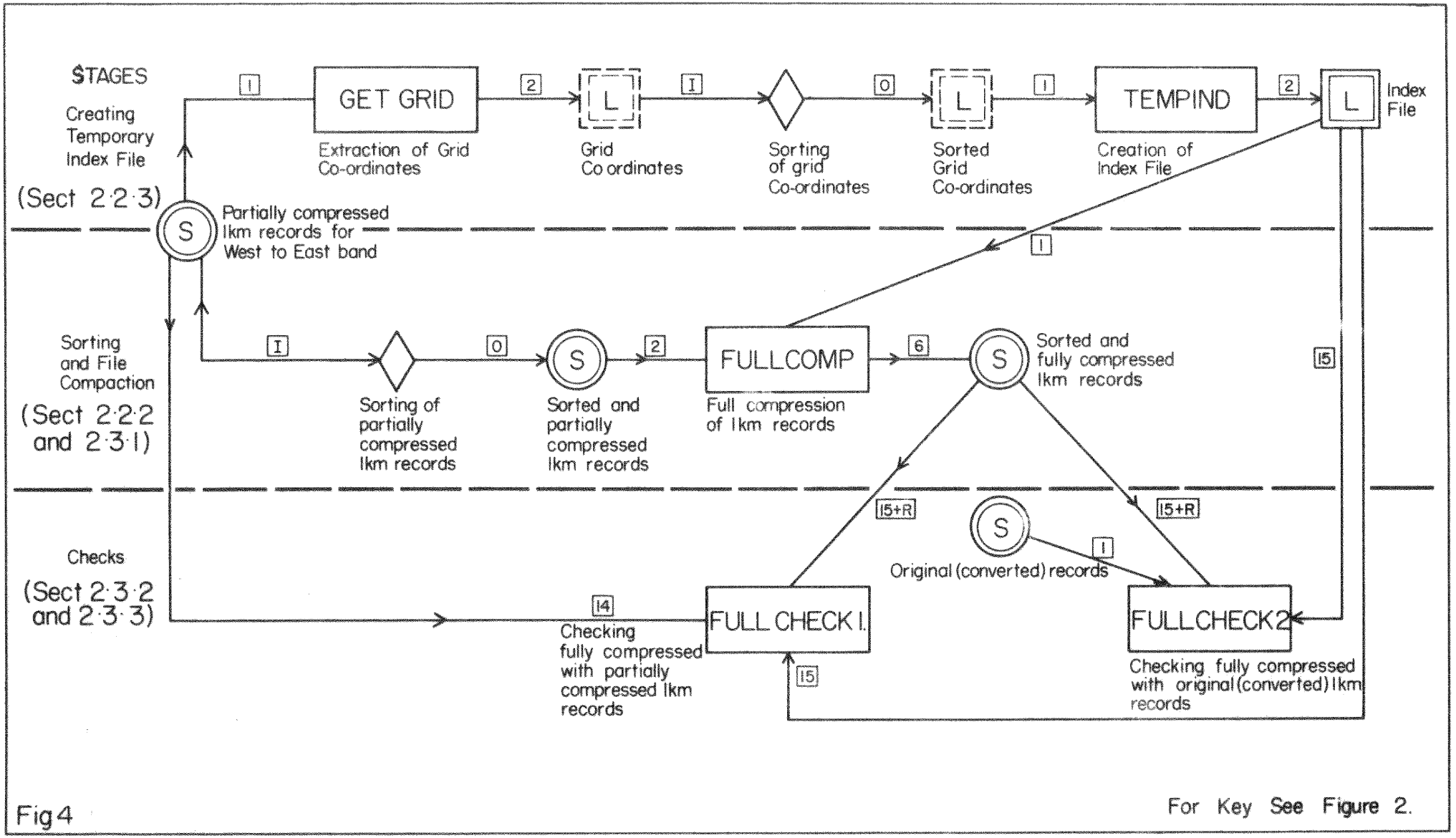


Figure 4 : Sorting, creating an index file, final compaction and checking of one west to east band of one-kilometre records.

2.2.3. Comparison of original (after conversion) records and compressed 10% pseudo-records

For the same reasons, the pseudo-records were also only partially compressed and were checked by a routine (CHECKPSEUDO) similar to that used for the summary records. The master file of pseudo-records contained 2549 records (see Table 2 for size details). Since the 10% files refer to households, there are no pseudo-records for seamen.

After checking, both the ten-kilometre summary files and the pseudo-record files were dumped onto magnetic tapes.

2.2.4. Accumulation of one-kilometre square data and sorting for each west-east band

As mentioned above, the data for one west-east band of 100-kilometre blocks generally spread over several tapes. This, together with the large size of both the partially compressed one-kilometre data and the check files, meant that only one or two tapes could be processed on each run. Each batch was subjected to the checks already described and then accumulated in separate one-kilometre and ten-kilometre "summation" files (see Fig.3). Each time a file containing part of the west-east band was appended to the summation file, the number of records in the latter was double-checked to ensure that the correct file(s) had been appended. When all the data for a band of one or more 100-kilometre blocks had been partially compressed and accumulated, the records in the one-kilometre summation file were sorted (using the MTS utility program *SORT) into descending northing then ascending easting grid co-ordinate order; i.e. to occur in west-east strips from north to south across the country (see Fig.4). The grid co-ordinates of the first and last few records of the sorted file were then examined to ensure that they were within the anticipated Y-bounds.

2.2.5. Pre-formatting a temporary index file

In order that, during subsequent checking, each fully-compressed record could be accessed randomly by reference to its absolute location on the disc pack, temporary index files were formed by program TEMPIND, using sorted grid co-ordinates (see Fig.4). The grid co-ordinates were extracted from the one-kilometre summation files and sorted into increasing easting then increasing northing sequence, so

that index records were written into the file in increasing (optimal for MTS) key sequence. For each grid square, an index record was stored, giving the length of each record of each record type for that particular grid square. These index files were intended only for use during the storage and checking processes. A much more efficient scheme has since been devised for indexing the multivariate files.

2.3. Final Compression of One-Kilometre Records

2.3.1. Full compression

The program FULLCOMP, wherever possible, further compacted the partially-compressed files by converting two-byte storage of data elements to single-byte storage and by zero suppression using the bitmap and coded deltas schemes.[3,6] The address and length of fully-compressed records were stored in index files (section 2.2.3 above). The data records were checked to ensure that they corresponded to the record type specified. The program user was responsible for checking that the correct files were attached to the program and that the correct number of records (i.e. the total number in the particular west-east band) had been processed.

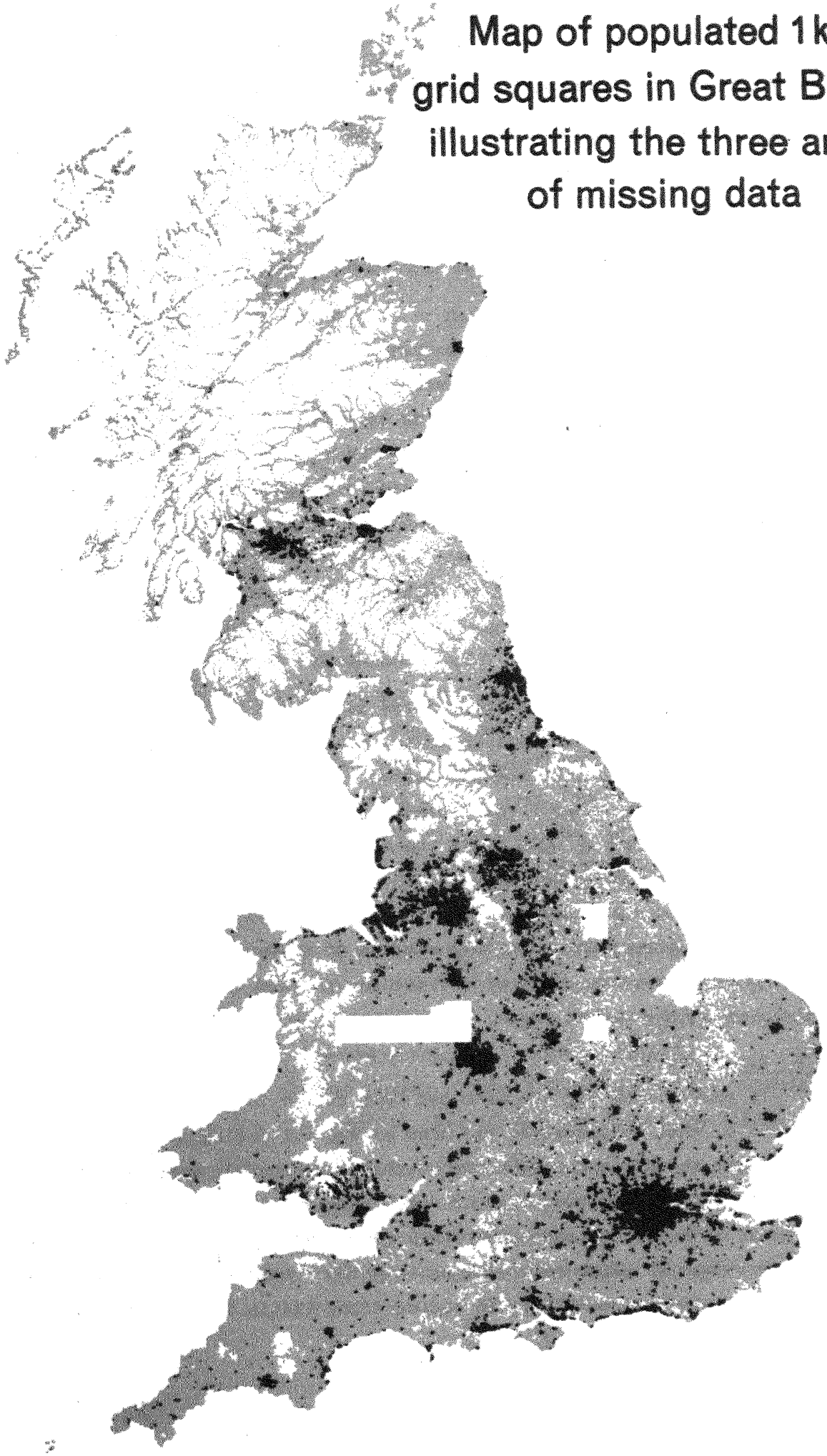
2.3.2. Comparison of the fully-compressed and partially-compressed files

A separate routine, FULLCHECK 1, expanded the fully-compressed records and checked the data values against those in the partially-compressed files. The primary aim of this process was to check the validity of the compression algorithms and to obviate any trivial errors, for example the presence of duplicate records for spatial units. This is a much faster process than that described below (section 2.3.3.) and was used online before continuing with the next phase (section 2.3.4. below).

2.3.3. Comparison of fully-compressed records with the original (converted) records

Another routine, FULLCHECK 2, compared the original one-kilometre data for the west-east band with the expanded final one-kilometre records. The aim in this case was to ensure that all the relevant data had been processed and that no records or values had been lost or corrupted. This somewhat slower process (compared with the process described in 2.3.2. above) was conducted in batch mode and was useful in detecting one instance where the wrong file had been

Map of populated 1 km grid squares in Great Britain, illustrating the three areas of missing data



subjected to the processes described in 2.3.1. and 2.3.2.

2.3.4. Final storage in master files

The data for each 100 kilometre-deep west-east band of 100-kilometre blocks were then appended to the appropriate one-kilometre multivariate format master file for that record type for Great Britain.

2.4. Checks on Record Duplication

As the data were processed in batches, it was necessary to check that there was no duplication of records between batches as a result of carelessness. The grid references of each record were extracted and sorted into increasing key order. The routine to preformat the primary index (see below) was also used to check that no spatial key was duplicated. This isolated an instance where the wrong file had been copied into the master file, resulting in the omission of one 100 kilometre-deep band of records and the duplication of parts of another.

2.5. Graphic Checks on Gaps

Since none of the above procedures is capable of discovering geographical gaps in the data, a distribution map of the populated areas (i.e. of all one-kilometre records received) was produced by D. Rhind. This revealed three "holes" (see Map 1), where data had been omitted either by OPCS or in the course of CRU processing. All known omissions have since been corrected and the previously described checks repeated where necessary. It must be remembered, however, that even graphic checks are only capable of detecting the omission of a geographic block of data; the absence of data for a single kilometre square cannot easily be detected.

2.6. Final Checks

To ensure that all the data had been stored in a recoverable form (i.e. that no record(s) had been missed out during selective manual extraction as described in section 2.2. above), each of the original IBM tapes was mounted and each record on the tape was compared with its corresponding fully-compacted record from the one-kilometre master files held on disc. As there were a large number of records in the one-kilometre master files (Table 1), random

access using unsorted keys was made possible via the primary index file[5]. Since the entire file could be recovered in this fashion, the fully-compressed data files were obviously intact and as received from OPCS. This procedure also checked the reliability of the index files and retrieval routines. As a security measure, copies of all index and compacted data files were dumped onto tapes.

3. CONCLUSION

The extremely large volume of grid-based data for Great Britain necessitates the use of procedures for minimising data storage and input/output time requirements. The procedures adopted completely changed the form of the data while maintaining the data content. The record headers and data values were represented in a different manner and the record structure was no longer that of a simple array of values. Moreover, the various record types were segregated and the sequence of data records altered. As the change in format and structure was a complex process, it was essential to check and double-check that no errors had been introduced by CRU. Some of the checks were obviously needed, while others were introduced, as a result of previous unfortunate occurrences, to trap user-carelessness as quickly as possible. Rectification often involved back-tracking only one stage in the storage process, but errors of the sort described in sections 2.4 and 2.5. involved re-conversions, a major shuffle of the appropriate master files and a repetition of the whole series of checks.

Table 2 summarises some characteristics of the master data files in the CRU system. On average, the data size was reduced by a factor of ten. This enabled all the grid-based census data to reside on one double density 333Ø disc pack and had the added benefit of reducing both CPU and elapsed times. Table 2 also gives data transfer time and CPU time for a sequential read of the one-kilometre files.

Tables 3A and 3B summarise the distribution by 100- kilometre squares of the one-kilometre and ten-kilometre records respectively in the 100% and 10% files. These are intended to assist users in ascertaining the availability of data within sub-areas of interest to them.

TABLE 3A : DISTRIBUTION OF 1 KM DATA RECORDS OVER GREAT BRITAIN

100 km Grid Square X	Y	All 100%	Unsuppressed Population	Unsuppressed Household	All 10%	Unsupressed 10%
4000	12000	62	15	16	24	9
3000	11000	8	0	0	1	0
4000	11000	583	110	117	230	66
2000	10000	1	0	0	0	0
3000	10000	701	88	105	254	53
4000	10000	5	1	1	1	1
0000	9000	9	1	1	3	0
1000	9000	373	167	167	219	120
2000	9000	603	60	59	173	55
3000	9000	769	103	114	244	77
0000	8000	314	77	69	130	36
1000	8000	842	143	163	317	90
2000	8000	2028	363	382	740	264
3000	8000	4380	842	835	1686	561
4000	8000	344	74	64	143	51
0000	7000	61	15	16	26	8
1000	7000	899	86	92	248	78
2000	7000	1698	271	261	570	223
3000	7000	3949	1050	1053	1772	689
1000	6000	900	115	126	270	96
2000	6000	5102	2077	2036	2737	1759
3000	6000	4084	1217	1244	1934	927
4000	6000	654	163	179	288	125
1000	5000	47	1	1	12	4
2000	5000	2372	375	374	814	295
3000	5000	5001	1251	1236	2102	907
4000	5000	4165	1920	1935	2430	1713
3000	4000	5225	2890	2862	3383	2430
4000	4000	7567	3410	3417	4387	2950
5000	4000	1607	689	689	889	594
6000	4000	1	1	0	0	0
2000	3000	3239	1002	1034	1524	787
3000	3000	8646	4891	4973	5793	3888
4000	3000	8221	4355	4386	5283	3698
5000	3000	5248	2171	2200	2898	1619
6000	3000	1734	953	1001	1190	782
1000	2000	652	207	206	312	160
2000	2000	5269	1300	1379	2279	977
3000	2000	8581	3687	3793	4904	2765
4000	2000	8074	4116	4155	5054	3514
5000	2000	7952	4204	4277	5130	3353
6000	2000	3291	1757	1825	2179	1299
1000	1000	42	11	13	19	8
2000	1000	3706	1226	1239	1872	988
3000	1000	7669	4204	4253	5096	3275
4000	1000	7934	4814	4874	5586	4009
5000	1000	8541	6511	6554	7015	5587
6000	1000	1260	767	779	890	615
0000	0000	20	15	15	15	9
1000	0000	1682	895	909	1107	678
2000	0000	4148	1625	1659	2305	1220
3000	0000	1295	638	646	801	488
4000	0000	839	597	610	663	539
5000	0000	44	26	28	33	25
Total		152441	67546	68422	87975	54464

TABLE 3B : DISTRIBUTION OF 10 KM DATA RECORDS OVER GREAT BRITAIN

100 km Grid Square X	Y	All 100%	Unsuppressed Population	Unsuppressed Household	All 10%	Unsuppressed 10%
4000	12000	4	3	3	4	3
3000	11000	2	1	1	1	0
4000	11000	37	31	31	32	28
2000	10000	1	0	0	0	0
3000	10000	26	22	23	24	23
4000	10000	2	1	1	1	1
0000	9000	3	2	2	2	1
1000	9000	29	26	26	26	24
2000	9000	60	43	40	46	32
3000	9000	32	26	26	26	22
0000	8000	21	18	17	18	18
1000	8000	67	53	54	53	47
2000	8000	92	68	66	70	60
3000	8000	71	69	69	69	66
4000	8000	8	8	8	8	8
0000	7000	6	3	4	4	2
1000	7000	66	51	51	54	41
2000	7000	93	69	68	72	59
3000	7000	80	70	70	74	68
1000	6000	55	45	45	45	42
2000	6000	94	93	93	93	89
3000	6000	91	89	89	88	84
4000	6000	15	15	15	15	15
1000	5000	3	3	3	3	2
2000	5000	67	61	61	61	55
3000	5000	100	97	97	96	93
4000	5000	60	59	59	59	59
3000	4000	80	79	79	80	79
4000	4000	100	100	100	100	100
5000	4000	32	31	31	31	31
6000	4000	1	1	0	0	0
2000	3000	67	66	66	66	66
3000	3000	98	98	98	98	98
4000	3000	100	100	100	100	100
5000	3000	77	77	77	77	77
6000	3000	24	24	24	24	24
1000	2000	15	12	11	12	10
2000	2000	81	81	80	80	79
3000	2000	100	100	100	100	100
4000	2000	100	100	100	100	100
5000	2000	100	100	100	100	100
6000	2000	44	44	44	44	44
1000	1000	2	1	1	1	1
2000	1000	57	55	55	55	53
3000	1000	97	97	97	97	97
4000	1000	100	100	100	100	100
5000	1000	99	99	99	99	99
6000	1000	24	23	23	23	23
0000	0000	3	3	3	3	3
1000	0000	30	30	30	29	29
2000	0000	59	59	59	59	59
3000	0000	25	25	25	25	13
4000	0000	19	18	18	19	19
5000	0000	3	3	3	3	3
Total		2722	2552	2545	2569	2459

Not all users of census data may need to modify the OPCS data organisation and thus may not be concerned with the complex procedures for checking their own processing. However, when the data being used are part of a gigantic set, the end product of an involved series of operations conducted by different departments, we accept that the onus is ultimately on the user to ensure that the data conform to definitions and broad expectations. Hence, some limited logical and arithmetic tests were performed during the storage process. Further quality checks have to be undertaken for specific studies. For example, it is advisable to double-check values when the same entity appears in one or more files as different census variables or is also quoted in a disaggregate form, given that census data have not been subjected to verification procedures before dissemination.

ACKNOWLEDGEMENTS

We are extremely grateful to the other members of the CRU, especially John Dewdney, Ian Evans and David Rhind, for their comments on drafts of this paper. We would also like to thank Mrs. Joan Dresser and Paul Wilson for preparing the manuscript and diagrams respectively and John Normile and Derek Hudspeth for the reprographics.

REFERENCES

1. OPCS — 1971 Census small area statistics (Ward Library). Magnetic Tape Layout Explanatory Notes, 1975.

2. VISVALINGAM, M., M.J. NORMAN, and R. SHEEHAN, Data Interchange on Industry-Compatible Tapes, Working Paper No.6 Census Research Unit, Department of Geography, University of Durham 1976.

3. VISVALINGAM, M., — Storage of the 1971 UK Census data, some technical considerations, Working Paper No.4, Department of Geography, University of Durham 1975.

4. RHIND, D.W., — Geographical analysis and mapping of the 1971 UK Census Data, Working Paper No.3, Census Research Unit, Department of Geography, University of Durham 1975.

5. VISVALINGAM, M., — A locational index for the 1971 kilometre-square population census for Great Britain (forthcoming Working Paper, Census Research Unit, Department of Geography, University of Durham).

6. VISVALINGAM, M., — Indexing with coded deltas - A data compaction technique, Software - practice and experience, Vol. 6 p. 397-403, 1976.