Data input and data structure

R. van den Driessche, Office de la Recherche Scientifique et Technique
Outre-Mer, 70 route d'Aulnay, Bondy, France

Introduction

Before signing a contract with a hardware supplier, a user must choose among manual input, optical reading and automated capture of the information intended for the computer. Most people use a keyboard to enter data on paper tape, punched cards, magnetic tape, or disks. Some users have their alphanumeric data typewritten and loaded by an optical reader. The same readers can be fed with handwritten digits and a few letters. In the field, there is perhaps an opportunity for a few specific environmental items to be received by a satellite from data collection platforms, at regular intervals. In the laboratory, Technicons or other analysers can record the 'results' in a variety of ways (tape, cards, etc.) for input.

But the question is not only HOW but also WHAT shall we put in. This brief presentation of the problems of data will take the form of 4 case studies.

Case Study 1: class texture
More often than not a textural class is recorded in the field. Most manuals recommend a specific triangle of reference; some glossaries postpone such a choice between triangles and let the surveyor make a choice. The latter increases cost when it comes to data processing, unless processing is a mere editing of profiles. Statistical treatment of the profiles has gained wide interest in the last decade. Thus the question is raised: Which of the five following manipulations should texture undergo?

In the field, the item appears as a nominal variable. Usually this is the lot of a synthetic item. No relation other than that of equivalence being applicable to let's say 11 or 25 names of an unordered list. One of the names is recorded in the field, as such, or as a symbol.

e.g. sandy clay loam scl 8 16

Statistical treatment of nominal variables can take the form of a contingency table analysis (Dixon, 1964), of a similarity index based on probabilities (Goodall, 1966), or others. My involvement with rank-order statistics has led me to put the item in twice:

.sandy clay loam.scl.

is punched anywhere in the horizon record. The system is then instructed by a dictionary to treat the first item as a sand class number and the second as a clay class number. Both are ordinal variables. Here the relation 'greater than' holds. Upon retrieval, tape output is in the I-type format:

9 6

A very simple modification of the dictionary would transform the duplicated item into two interval variables (or ratio variables) sand percentage and clay percentage:

60 27

These percentages, by the way,

are the coordinates from the center point of the SCL polygon on the sand and clay percentage scales in the FAO triangle; they can be used -- with other items -- namely to compute a multi-level dissimilarity index between profiles.

A similar modification would transform the 'sandy clay loam. scl' input into class limits on the sand and clay percentage scales:

    4480   2035

these, in turn, are input for other programs by format card:

    44-80% sand and 20-35% clay.

From a cost point of view, the most appealing input seems to be symbol

    .scl.   punched

anywhere and transformed by the program into percentage sand and clay.

Case Study 2: particle size analysis

Let us consider 8 classes used in routine work by Hubert for INEAC in the Mahagi in the 1950'ies: LT 2, 2-20, 20-50, 50-100, 100-250, 250-500, 500-1000, 1000-2000 µm. The percentages were delivered by the laboratory to one decimal place:

36.8 7.0 5.1 4.2 10.6 10.4 20.8 5.1%

Is this not unwanted accuracy? Would rounded-off percentages not suffice? If so, one could allocate 8 specified 2-column fields to the percentages or one could keypunch

```
    .37C1.7C2.5C3.4C4.11C5.10C6
    .21C7.5C8.
```

anywhere in the record of v variables.

Case Study 3: field pH

The next case study on field pH is presented as a block diagram with all the alternatives open to the planner of a soil information system. (see p.19)

Case Study 4: profile descriptions

We recently created a reference file for profile descriptions from 230 sites in central Africa (Sys 1972), using techniques developed by Van den Driessche *et al.* (1974a, b; 1975a, b). 1. The master cards were in fixed format: one card for the profile number, ten cards for the ten recorded depths (1cm, 5cm, 12cm, 25cm, 41cm, 61cm, 85cm, 113cm, 145cm, 181cm).

```
    PROFIL/NO 92/
    - HORIZON/HRZ 5/41CM//
```

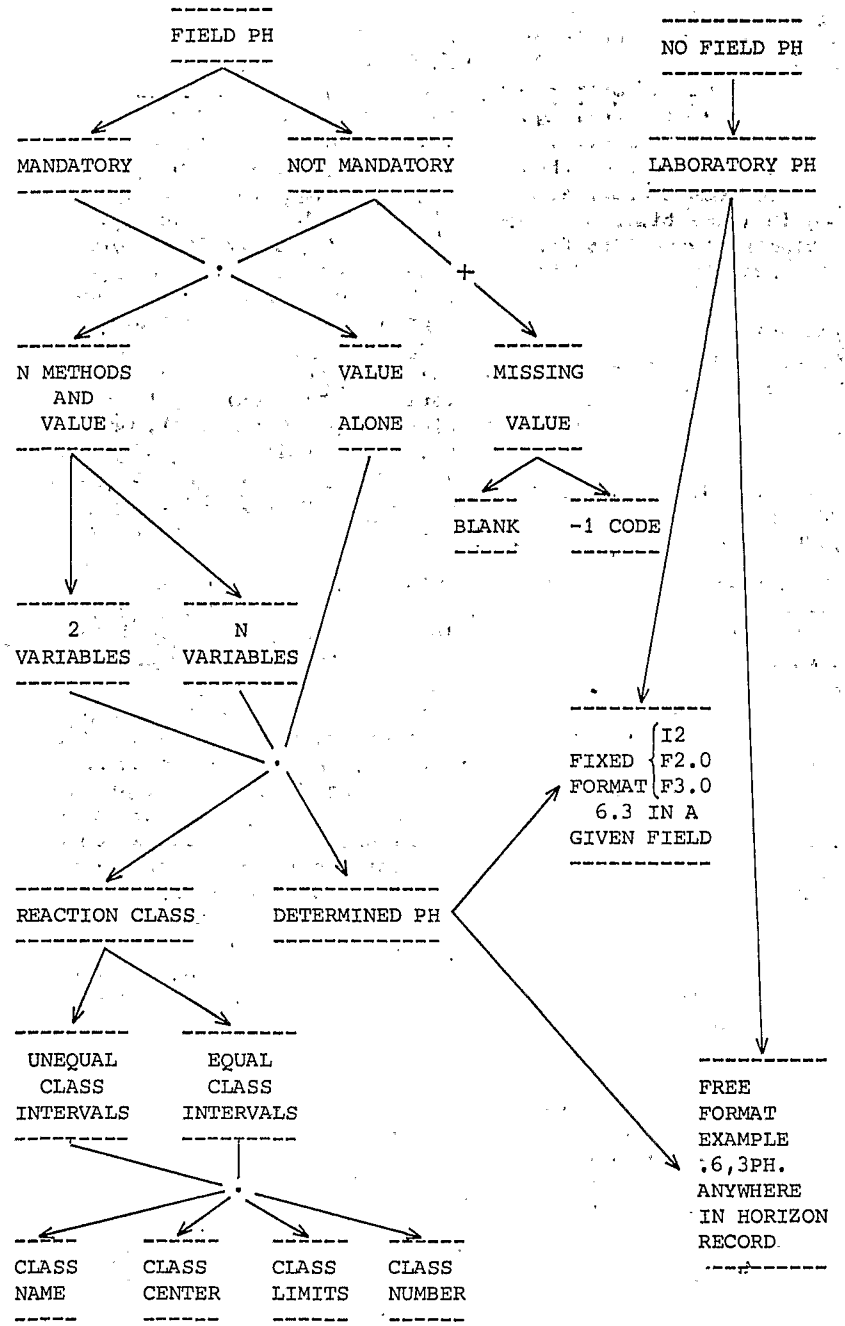
2. The data cards were in free format: English names of the observed and analysed items were punched, continuously, disregarding any order within the record, with just a period as separator. No blank space nor code for missing data:

```
    SANDY CLAY.SC.7,5YR.5/.6.
    MEDIUM.BLOCKY STRUCTURE.
    .MODERATE.FIRM.ROOTS.35F1.
    3F2.5F3.10F4.21F5.12F6.14F7.
    0F8.0RE.0,4C.0,07N.5,0PH
    1CA.0,3K.6T.POINT 92.DEPTH 5.
```

The same horizon was punched twice (or duplicated) when coverage extended over two depths of the above list. When maximum depth was not attained in a profile, the master card remained so to retrieve a record of missing data codes. Conversely, for the deepest profiles, a horizon may have been overlooked.

Per profile: 23 field variables + 16 laboratory variables at 10 depths = 390 variables. Two or three cards were needed for each depth record, exactly 3530 cards, master cards included.

3. We ran the data with our program DGY (69K) on our UNIVAC 1108 (192K) and listed the profiles. A magnetic tape was loaded simultaneously with codes. This happened in flip-flop mode and



took 10 min (CPU) for a bundle of 100 profiles.

Unexpected, redundant, or unseparated items, or mis-spellings were annotated by the computer. The corresponding decks were then amended, and corrected profiles were run again (same time: 6 s/profile). Editing soil profiles took 1 s/profile. Example of an edited record:

```
- HORIZON/HRZ 5//
POINT 92.AT 41CM.SANDY CLAY.SC.
MODERATE.MEDIUM.BLOCKY
STRUCTURE.FIRM.
7,5YR.5/.6.
ROOTS.
35F1.3F2.5F3.10F4.21F5,12F6.
14F7.0F8.0RE.
0,4C.0,07N.5,0PH.1CA.0,3K.6T.
```

4. Our input procedure required previous loading of a dictionary, in which every item had an equivalent code number. Program RGY (68K) was used for that (28 s), but only once. A second, equivalent, dictionary afforded automated translation, simply by changing a control card.
5. Spatial and temporal referencing of the sites, plus environmental items (vegetation, climate, ..., center points of satellite imagery), plus soil classification data, were keypunched in the same way (24 variables) and loaded onto a second tape set. Large data files could be accommodated by the programs: 100 000 records of 500 variables.
6. Records were retrieved by input of a Boolean expression including less than 64 conditions written in plain language or in code (SGYT 80K and SGY 74K Fortran programs). Retrieval time is 1 s/profile.
7. For a geosearch input (Fortran program COOR 11K with worldwide coverage), 4 cards were needed inside an area rectangle which included the geographic latitude and longitude coordinates (accuracy ± 1'). The purpose was threefold: printout in English; data matrix on tape; spatial referencing of ground truth, satellite and Slar imagery was graphically displayed on the lineprinter (CARTO program 15K in Fortran), with a 24°/18° range and a 5'/3' accuracy for latitude and longitude, respectively.
8. All F-type formats were accepted by our Fortran statistical programs: one-level dissimilarity (DISSIM 47K), multi-level dissimilarity (+MERGE 8K), agglomerative clustering (AGGLOM 89K), multivariate identification of profiles (IDENT 12K), multivariate ordering of the sites, and the S' test (4K).

Conclusion

Through these 4 case studies I have attempted to show some alternatives in the input of punched items. The need for standardization across borders is felt. However, there is a danger in postponing the technical problems (hardware, software, statistics) until agreement is reached on an international list of priority items (point data + area data + ephemeral data + interpretive data) for exchange of data files. If we want to scrutinize various existing data base management systems, we must be prepared to reformat our data to achieve a standard test. The processing of foreign files, as in our central Africa reference file, is a stimulating experiment.

References

Dixon, W.J. ed. (1964). BMD Biomedical computer programs. 1st ed., UCLA, Los Angeles, 585 p.

Goodall, D. (1966). A new similarity index based on probabilities. Biometrics, 22, 882-907

Sys, C. (1972). Caractérisation morphologique et physico-chimique de profils types de l'Afrique centrale. INEAC, Brussels, 497 p.

Van den Driessche, R., Garcia Gomez, A., Giey, A., (1974a). Un système informatique en langage naturel: application didactique aux coloris ISCC-NBS. Init. doc. tech., ORSTOM, 25, Paris, 120 p.

Van den Driessche, R., Garcia Gomez, A., Giey, A. (1974b). SIDA Satellite imagery descriptors analysis. ORSTOM, Bondy, 165 p.

Van den Driessche, R., Garcia Gomez, A., Aubry, A.M. (1975a). A multi-level dissimilarity index between soil profiles. Cah. ORSTOM sér. pédol., 13, 2, in press.

Van den Driessche, R., Garcia Gomez, A., Giey, A., Aubry, A.M. (1975b). POSEIDON Procédures opérationelles en statistique et informatique pour données en langage naturel. ORSTOM, Bondy, in press