COMPUTERIZED DATA VALIDATION, MAPPING, AND TALLY SHEETS FOR FOREST GENETICS STUDIES1/

James M. Kucera and Samuel B. Land, Jr.2/

ABSTRACT

Computer techniques have been developed at Mississippi State University to increase data processing efficiency and accuracy of research documentation for large, long-term forest genetics field plantings. Advantages of using these techniques include: (a) early diagnosis of errors in documentation of tree identity in field plantings; (b) computer validation to check for errors in measurement records; (c) flexibility and accuracy in methods for recording field data; (d) more efficient data-punching procedures when intermediate analyses of data are required during a long--term study; and (e) opportunities for examination of non-random environmental gradients within a site on performance of trees in the study. Examples of computer-printed maps of field studies and computer-printed tally sheets are given.

INTRODUCTION

Forest genetics field studies often have been reduced in effectiveness or even lost, due to poor mapping and data-handling procedures. The long-term nature and large numbers of observations involved in such studies contribute to the seriousness of the problem. Poorly planned, non-legible field notes placed in filing cabinets without adequate identification are not acceptable when one considers the large investments in time and money expended in establishing and measuring these studies.

This paper presents: (a) advantages of using computers in data validation, mapping, and printing tally sheets in forest genetics research plantations and (b) examples of such uses of the computer for an American sycamore (Platanus occidentalis L.) regional provenance test at Mississippi State University.

[1] Contribution No. 3401 of the Mississippi Agricultural and Forestry Experiment Station, Mississippi State.

[2] Research Associate and Associate Professor, respectively, in the Department of Forestry, Mississippi Agricultural and Forestry Experiment Station, Mississippi State University.

DATA VALIDATION

Large numbers of observations in forest genetics studies inevitably result in errors, both at the stage of recording data and during the transfer of data to a computer processable medium. Data validation and data verification are two separate steps in data processing used to correct these errors. Verification is concerned with correct key punching of data, while validation is a check of internal consistency of the recorded data.

Validation and verification of data can be accomplished by careful visual examination of records. But with large numbers of entries visual examination is expensive, time-consuming, and usually ineffective after the first hour of scrutiny. Often, such checks of forest genetics data have been foregone for these reasons. The argument that a few incorrect records will not significantly affect the overall results of the analyses cannot be accepted when there are other validation and verification procedures available.

A machine varifier that involves "double punching" can be used to correct errors in copying data from tally sheets to cards or tape. For smaller studies verification can also be accomplished visually by cross-checking a computer listing of cards against the actual tally sheets.

Validation of data can be accomplished with a computer, using a program that scans measurement data and identifier codes for reasonableness. One method is "range checking", where error messages are printed for any records falling outside a specified range of values. Included in the print-out is the necessary identification to allow location of the faulty record in the data file. One useful "range checking" program sets the limits on record values equal to two standard deviations above and below the mean for the trait in the data file.

MAPPING

Most researchers have heard the rule to always label the trees or plots in the study with permanent labels and to map the study at time of planting. But practice often lags behind preaching. Labeling trees in itself is not sufficient documentation, nor is mapping alone. Labels can be lost, such as has happened at Mississippi State during discing of hardwood progeny tests. Without maps, identities of the plots would have been lost. Conversely, maps without some remaining field labels on trees can never be verified for accurancy.

Even with tree labeling and mapping at time of planting, the hand-written map from the field has often been the only map made of the study. Some disadvantages of this procedure are:

(a) Mapping errors at the time of planting might not be rectifiable three to five years later, when the maps are next examined, since flagging and some labels will be missing;

(b) Changes in project leaders may occur, and insufficient information on the field map may not allow the new leader to decipher the study layout or code system.

One method of avoiding problems of poor mapping is to use computerized mapping of studies. Not only are the standard identifiers of replication, seed source, family, tree, treatment, etc. provided for each tree, but also row and column positions of trees in the planting are assigned. Computer programs can then be written to provide a computer-printed map of the study, with each tree identified on a row and column grid.

Advantages of computerized mapping include:

(a) A clearly legible map with study identification and location information on each page;

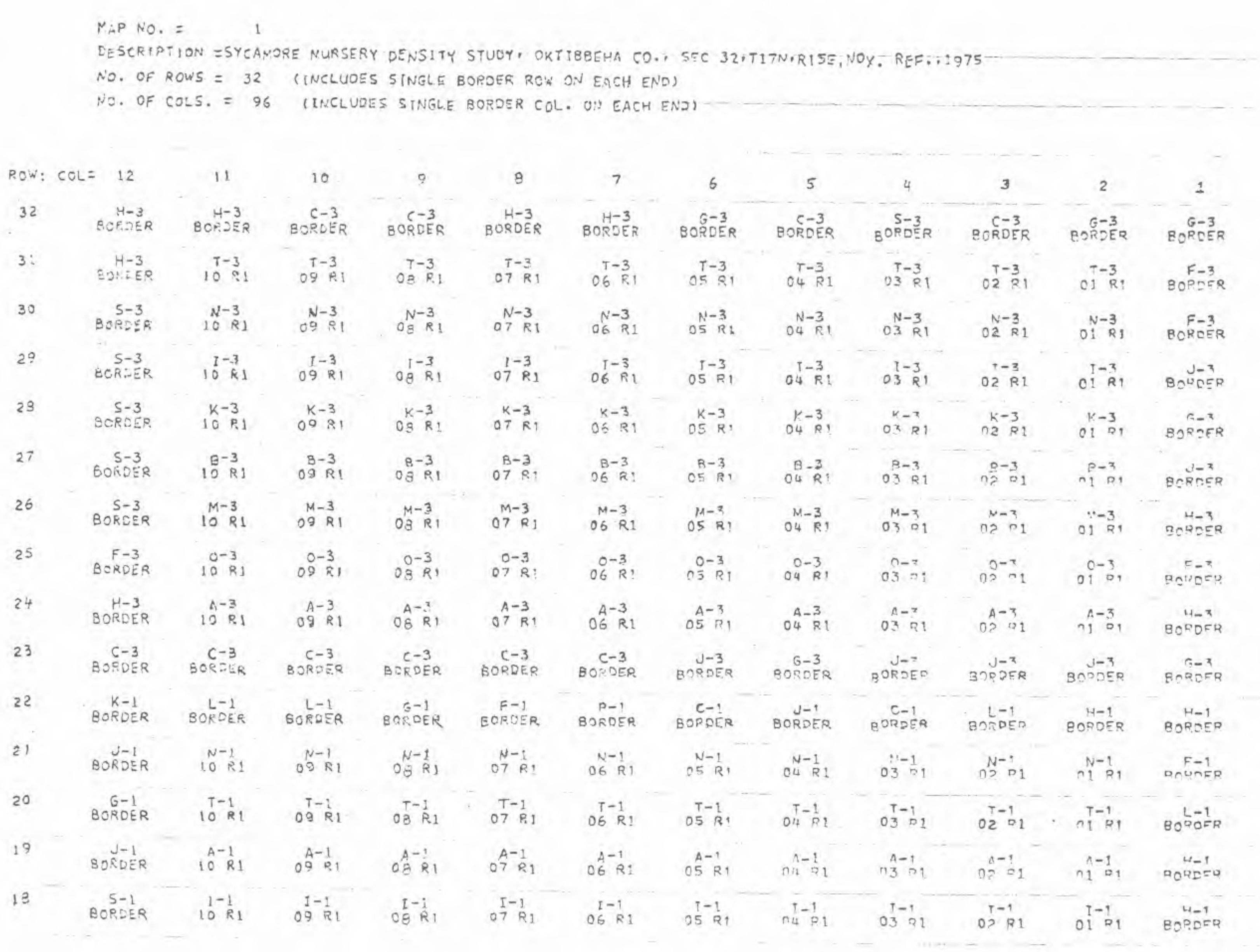
(b) An opportunity to check for errors in row and column identifiers, and to some extent for other identifiers in the individual tree's record; and

(c) Increased capabilities for use of the computer in other aspects of the study, such as analyses of site variation within the study plantation.

Use of well-designed, clearly legible maps can alleviate the problem associated with changes in project leaders and can allow multiple copies for cooperators. Computerized mapping can also serve as a validation program for detecting errors in row and column coordinates of trees. If a computer-printed map is planned as an early requirement of the study, early diagnosis of errors in documentation of tree identities can be made and the errors corrected while all labels are still present. Since the trees will be identified by grid coordinates over the field planting, capabilities for examining patterns of environmental variation across the planting are available as a third advantage of the mapping procedure. This could lend itself to adjustments to remove some environmental variation through analyses of multiple covariance on row and column positions.

For most state and federal research projects, and many company projects, a programmer and computer are available. Computer time can be expensive, but very little time is actually used in a mapping program when a "table look-up" programming technique can be employed. Except for machine-planted studies, most forest genetics studies are carefully laid out in a systematic spacing amenable to computer documentation. One of the most difficult designs, the Nelder's spacing study in a wheel design, can be assigned spoke and tree-within-spoke positions and printed in rectangular coordinates on the map. Thus, use of the computer in mapping forest genetics studies is feasible and should be considered.

A Midsouth American sycamore provenance test at Mississippi State University provides an example of the use of the computerized mapping of a forest genetics study. Twenty seed sources with two stands per seed source and five trees per stand are represented by open-pollinated progenies planted at each of four locations in each of two years. The total number of individuals in the study is 28,800. An example of one page of the resulting map for one of the plantings is given in Figure 1.



```
    MAP NO. =        1
    DESCRIPTION =SYCAMORE NURSERY DENSITY STUDY, OKTIBBEHA CO., SEC 32,T17N,R15E,NOV. REF.,1975
    NO. OF ROWS =  32   (INCLUDES SINGLE BORDER ROW ON EACH END)
    NO. OF COLS. =  96   (INCLUDES SINGLE BORDER COL. ON EACH END)

ROW: COL=  12       11       10        9        8        7        6        5        4        3        2        1

 32       H-3      H-3      C-3      C-3      H-3      H-3      G-3      C-3      S-3      C-3      G-3      G-3
         BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER

 31       H-3      T-3      T-3      T-3      T-3      T-3      T-3      T-3      T-3      T-3      T-3      F-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 30       S-3      N-3      N-3      N-3      N-3      N-3      N-3      N-3      N-3      N-3      N-3      F-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 29       S-3      I-3      I-3      I-3      I-3      I-3      I-3      I-3      I-3      I-3      I-3      J-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 28       S-3      K-3      K-3      K-3      K-3      K-3      K-3      K-3      K-3      K-3      K-3      G-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 27       S-3      B-3      B-3      B-3      B-3      B-3      B-3      B-3      B-3      B-3      B-3      J-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 26       S-3      M-3      M-3      M-3      M-3      M-3      M-3      M-3      M-3      M-3      M-3      H-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 25       F-3      O-3      O-3      O-3      O-3      O-3      O-3      O-3      O-3      O-3      O-3      F-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 24       H-3      A-3      A-3      A-3      A-3      A-3      A-3      A-3      A-3      A-3      A-3      H-3
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 23       C-3      C-3      C-3      C-3      C-3      C-3      J-3      G-3      J-3      J-3      J-3      G-3
         BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER

 22       K-1      L-1      L-1      G-1      F-1      P-1      C-1      J-1      C-1      L-1      H-1      H-1
         BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER   BORDER

 21       J-1      N-1      N-1      N-1      N-1      N-1      N-1      N-1      N-1      N-1      N-1      F-1
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 20       G-1      T-1      T-1      T-1      T-1      T-1      T-1      T-1      T-1      T-1      T-1      L-1
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 19       J-1      A-1      A-1      A-1      A-1      A-1      A-1      A-1      A-1      A-1      A-1      H-1
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER

 18       S-1      I-1      I-1      I-1      I-1      I-1      I-1      I-1      I-1      I-1      I-1      H-1
         BORDER   10 R1    09 R1    08 R1    07 R1    06 R1    05 R1    04 R1    03 R1    02 R1    01 R1    BORDER
```

Figure 1. - Example of a computer-printed map of a forest genetics study of American sycamore at Mississippi State University

North is always at the top of the page, and each page contains a description of the study, specific location, date of planting, and total number of rows and columns in the study. The various pages of the map can be attached together, if desired, to give a mosaic of the total planting. The mapping procedure included a validation check that printed error messages when two trees had the same row and column coordinates or where the coordinates of a tree exceeded the designated number of rows and columns in the study. Ease of visual examination of the map for tree identifications also led to detection of errors. A xerox copy of all pages in a plantation map, along with a less-detailed map of the study layout and a description of the individual tree-identification codes on the computer-printed map, was sent to each cooperator in the study.

TALLY SHEETS

Hand-written tally sheets used for repeated collection of field data in long-term forest genetics studies are inefficient and risk loss of data. Usually the time required to write column identifications and study descriptions at the top of each page results in only a few of the tally sheets having such information. Unlabeled sheets are confusing and can result in mistakes at both the field recording and key punching stages. Furthermore, hand-written tally sheets are often arranged in a sequence of tree records that is awkward to follow in the field and inefficient in time of movement between trees. Finally, when tally sheets are used for successive years of measurement (and this is often done to avoid rewriting tree or plot identifications), there is the danger of obscuring or destroying hand-written data from previous years. Anyone working with paper in field studies on rainy days is familiar with this problem.

Alternative methods of data collection include (a) punching data directly onto magnetic tape in the field and sending the taped information through phone hookups to the computer terminal or (b) using separate computer-printed tally sheets for recording each year's field measurements, followed by normal key punching of data to cards or tape. The first alternative is fast, since it eliminates the key-punching phase, but the equipment is expensive for small projects. Also, confusion and errors could arise in the field recording phase because of recorder's mix-up in tree or plot identities. This disorientation is avoided with copies of previous records. The second alternative has the advantages of (a) being less expensive than the procedure of field punching data on tape and (b) providing a clear, permanent record of field data that can be understood by recorder, key puncher, and future project leaders. It has the disadvantages of (a) being more expensive than hand-written tally sheets and (b) requiring recopy of data to cards or tape. But it provides extra flexibility, when individual trees are located by row and column coordinates, in arranging record entries in any design deemed most efficient for movement and measurement in the field. New tally sheets can be printed for each successive year's measurements, so that record-entry design can be varied with year or study. Furthermore, previous measurements and a field map can be printed on the tally sheet to reduce chance of error from disorientation during recording.

An example of the computer-printed tally sheets being used for forest genetics studies at Mississippi State is shown in Figure 2. Features include:

(a) Description of the study, including location;

(b) Page number and spaces for date of measurement and name of recorder;

(c) Three sections of data -- the first being that which is recorded in the field and is to be punched on cards, the second being the full tree identification and previous measurements, and the third being a field map of that portion of trees to be measured on that page;

(d) Complete column headings;

(e) A predetermined sequence of record entries (in this example a sweep of six columns of trees in one pass across the field was used); and

(f) Use of location, year of establishment, and tree row and column as sufficient identification for each record.

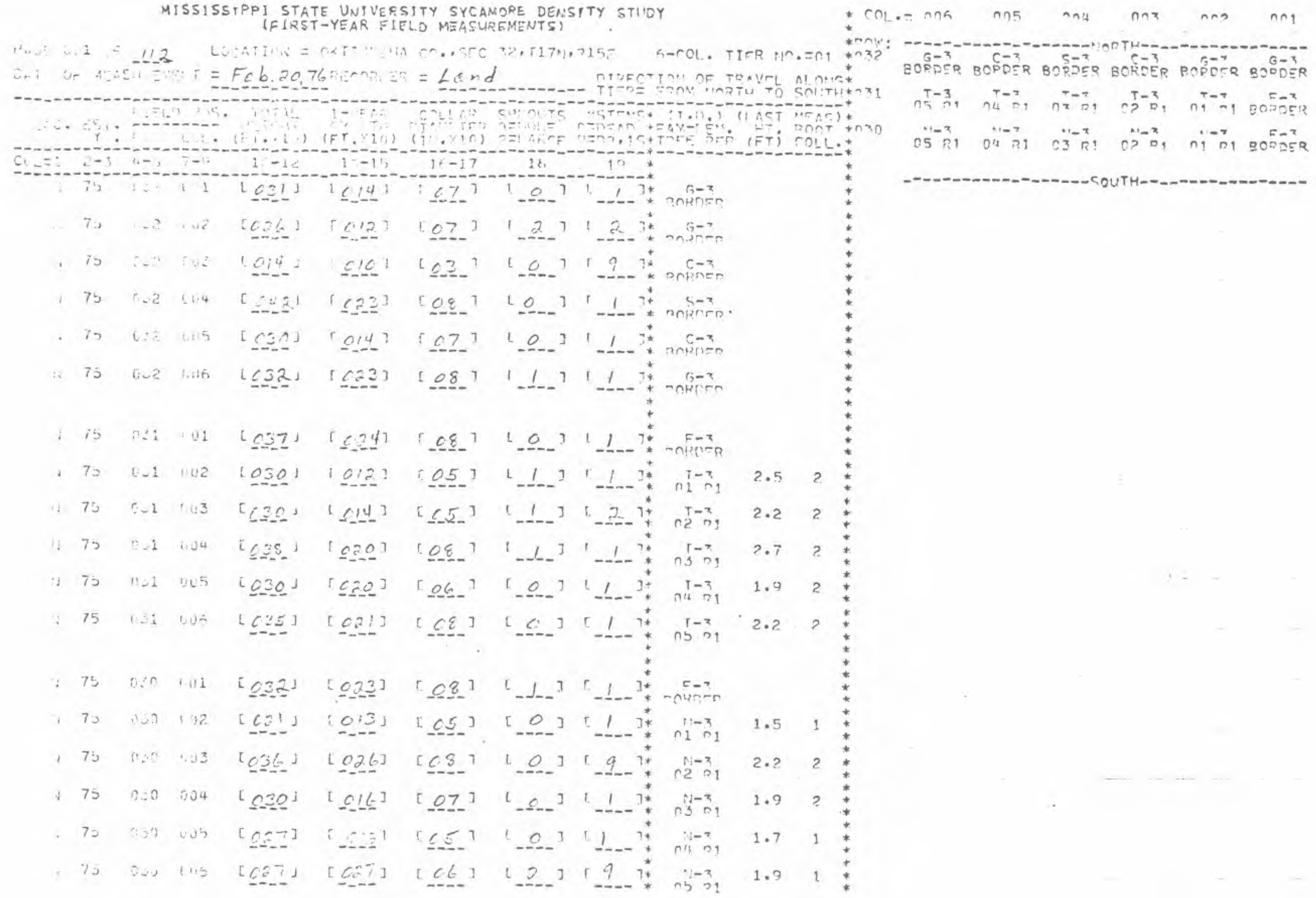
The last feature saves key--punch time, and preprinting that information eliminates one source of possible error in data collection. This feature is very important when many years of measurements and large numbers of individual trees are involved. Adequate space is also provided for notes, an important aspect of data collection and one which cannot be easily handled in the procedure where data is entered directly onto tape in the field.

SUMMARY AND RECOMMENDATIONS

Data validation and data verification are separate steps in correction of errors in data sets. Visual validation and verification of large sets can be inefficient. Data should be verified at time of key punching, and validation can most effectively be handled on the computer.

Poor documentation of long-term forest genetics field studies leads to irretrievable loss of data. Both tree labeling in the field and well-documented, easily-read maps of the planting are required. By assigning row and column coordinates to individual trees, the computer can be used to produce suitable maps. Validation of tree identities can be combined with the mapping program to provide an early check for errors in field mapping records. The assignment of row and column coordinates to trees adds capabilities for examination of site variation in studies and for computer printed tally sheets.

Hand-written tally sheets for field data collection are inefficient and subject to loss of information when large, long-term studies are involved. Alternative methods include entering data directly on tape in



```
                 MISSISSIPPI STATE UNIVERSITY SYCAMORE DENSITY STUDY
                          (FIRST-YEAR FIELD MEASUREMENTS)

PAGE 001 OF 112    LOCATION = OKTIBBEHA CO.,SEC 32,T17N,R15E    6-COL. TIER NO.=01 *032
DATE OF MEASUREMENT = Feb.20,76 RECORDER = Land
-------------                     ---------------   DIRECTION OF TRAVEL ALONG*
                                  --------------- TIERS FROM NORTH TO SOUTH*031
-------------------------------------------------------------------------------*
       FIELD POS.  TOTAL    1-YEAR   COLLAR  SPROUTS  #STEMS* (I.D.) (LAST MEAS)*
REC. BLK. --------- HEIGHT  GROWTH  DIAMETER DOUBLE  SPREAD *FAM-GEN.  HT.  ROOT *030
   Y .  ROW  COL. (FT.X10) (FT.X10) (IN.X10) REL. LEADER  ROOT. *TREE REP (FT) COLL.*
-------------------------------------------------------------------------------*
COL=1 2-3  4-6  7-9   10-12    13-15    16-17    18      19   *                  *
-----------------------------------------------------------------*                  *
    75  032  001  [031]   [014]   [07 ]   [ 0 ]   [ 1 ]*   G-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  032  002  [036]   [012]   [07 ]   [ 2 ]   [ 2 ]*   G-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  032  003  [014]   [010]   [03 ]   [ 0 ]   [ 9 ]*   C-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  032  004  [022]   [023]   [08 ]   [ 0 ]   [ 1 ]*   S-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  032  005  [030]   [014]   [07 ]   [ 0 ]   [ 1 ]*   C-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  032  006  [032]   [023]   [08 ]   [ 1 ]   [ 1 ]*   G-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *

    75  031  001  [037]   [024]   [08 ]   [ 0 ]   [ 1 ]*   E-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  031  002  [030]   [012]   [05 ]   [ 1 ]   [ 1 ]*   T-3        2.5    2     *
                   ----    ----    ----    ----    ---- *  01 R1                   *
    75  031  003  [030]   [014]   [05 ]   [ 1 ]   [ 2 ]*   T-3        2.2    2     *
                   ----    ----    ----    ----    ---- *  02 R1                   *
    75  031  004  [038]   [020]   [08 ]   [ 1 ]   [ 1 ]*   T-3        2.7    2     *
                   ----    ----    ----    ----    ---- *  03 R1                   *
    75  031  005  [030]   [020]   [06 ]   [ 0 ]   [ 1 ]*   T-3        1.9    2     *
                   ----    ----    ----    ----    ---- *  04 R1                   *
    75  031  006  [025]   [021]   [08 ]   [ 0 ]   [ 1 ]*   T-3        2.2    2     *
                   ----    ----    ----    ----    ---- *  05 R1                   *

    75  030  001  [032]   [033]   [08 ]   [ 1 ]   [ 1 ]*   E-3                     *
                   ----    ----    ----    ----    ---- *  BORDER                  *
    75  030  002  [031]   [013]   [05 ]   [ 0 ]   [ 1 ]*   N-3        1.5    1     *
                   ----    ----    ----    ----    ---- *  01 R1                   *
    75  030  003  [036]   [026]   [08 ]   [ 0 ]   [ 9 ]*   N-3        2.2    2     *
                   ----    ----    ----    ----    ---- *  02 R1                   *
    75  030  004  [030]   [016]   [07 ]   [ 0 ]   [ 1 ]*   N-3        1.9    2     *
                   ----    ----    ----    ----    ---- *  03 R1                   *
    75  030  005  [027]   [[illegible]]   [05 ]   [ 0 ]   [ 1 ]*   N-3        1.7    1     *
                   ----    ----    ----    ----    ---- *  04 R1                   *
    75  030  006  [027]   [027]   [06 ]   [ 0 ]   [ 9 ]*   N-3        1.9    1     *
                   ----    ----    ----    ----    ---- *  05 R1                   *
```

```
* COL.= 006     005     004     003     002     001
*
*ROW:  ----------------------NORTH----------------------
*032    G-3     C-3     S-3     C-3     G-3     G-3
*     BORDER  BORDER  BORDER  BORDER  BORDER  BORDER
*
*031    T-3     T-3     T-3     T-3     T-3     E-3
*      05 R1   04 R1   03 R1   02 R1   01 R1  BORDER
*
*030    N-3     N-3     N-3     N-3     N-3     E-3
*      05 R1   04 R1   03 R1   02 R1   01 R1  BORDER
*
*
*      ----------------------SOUTH----------------------
```

Figure 2 - Example of a computer-printed tally sheet used for a forest genetics study of American sycamore at Mississippi State University

the field or entering data on computer-printed tally sheets. The latter alternative is intermediate in expense, reduces the chance for errors, and provides flexibility in sequence of record entries in the field.

At Mississippi State University data validation, computerized mapping, and computer-printed tally sheets are used to (a) increase data processing efficiency and (b) provide complete and accurate documentation of genetics research. It is recommended that careful planning, including consideration of the above-mentioned uses of the computer, be done before initiation of forest genetics studies to provide this increased accuracy and efficiency.