**Data File Format for AMMISOFT**

**Introduction to AMMI Analysis and AMMISOFT Software.** The Additive Main effects and Multiplicative Interaction (AMMI) model uses analysis of variance for the additive or main effects followed by principal components analysis for the multiplicative or interaction effects. The data structure that AMMI addresses is a two-way factorial design, with or without replication. In agricultural research, the principal application is yield trials which have a number of genotypes tested in a number of environments, often with replication. Although yield is the foremost trait, others are also common, including plant height, days to flowering or maturity, disease resistance, and nutritional value. AMMI analyses multiple traits one at a time, not all together. Also, when environmental factors such as rainfall or soil nitrogen have also been measured (beyond the yield trial itself), this additional information can help to interpret AMMI results, although this environmental data is not directly incorporated in AMMI analysis. When main and interaction effects are statistically significant and agriculturally important, as is most often the case, AMMI is extremely useful. The purpose of AMMISOFT is to make it easy to implement AMMI analysis with best practices as described in: Hugh G. Gauch, Jr., 2013, A Simple Protocol for AMMI Analysis of Yield Trials, Crop Science 53:1860–1869.

AMMISOFT uses the same format for datasets as did its predecessor MATMODEL (except for deletion of the optional format for single-entry lines of data). For all of the following items, enter full lines completing everything called for, even if this requires blanks or zeros at the end of a line. If the data were already entered into the computer for some other purpose, hopefully little work will be required to prepare the file for AMMISOFT. Data files must be created as plain text files and hence their filenames must end with .txt for the extension. Two or three lines precede the data, and they are supplied in the following order.

**Dataset Name.** The first line of the input data file identifies the dataset.  It has two entries.

(1) Columns 1–72, an alphanumeric title for the dataset of up to 72 characters.  This title is reproduced once at the beginning of the AMMISOFT output in order to identify or describe your dataset.

(2) Columns 77–80, an alphanumeric brief name for the dataset. This name is worth choosing carefully because it appears repeatedly in the AMMISOFT output file as a concise way to identify your dataset.  A useful convention is to use this brief name also in the dataset’s filename. For example, the supplied dataset with the brief name SOYB has been given the corresponding filename SOYB.txt. AMMISOFT automatically generates a name for its output file from the name for the input data file by adding the suffix “ammi” before the extension, so the result in this case is SOYBammi.txt. An output file is created in the same subdirectory as its corresponding input file so that these input and output files will appear together in your browser.

**Data Information.** The second line of the data file contains thirteen entries.

(1) Columns 1–20, an alphanumeric name for the data’s measurement and units.  For example, “Yield  kg/ha” is the measurement and units for dataset SOYB.

(2) Columns 21–25 right justified, a scaling integer.  This entry may be a positive or negative integer, or zero.  Ten is raised to the power of this integer, and then the data are multiplied by the result.  If this entry is 0 (or blank), then no action is taken.  Scaling can be useful if the data have been entered already as rather small or large numbers, or to avoid having to enter decimal points.  The measurement units given in the previous entry should reflect the result after multiplication by the indicated scaling factor.

*In order to produce results which print nicely in the print formats used by AMMISOFT, it is required that typical data entries be in the range of 0.001 to 1000000, or better yet within 0.01 to 100000, or ideally within 1 to 10000.* If necessary, the units of measurement should be changed in order to result in values in this range.

(3) Columns 26–30 right justified, number of genotypes.  For data file SOYB, there are 7 genotypes.

(4) Columns 32–35 left justified with 1 to 4 characters, brief name for genotypes.  As explained soon, three such names are given:  one for genotypes, environments, and replications.  The first letter of each of these three names must be unique (in columns 32, 42, and 52).  For SOYB, the name “GEN” is given for the genotypes.

(5) Columns 36–40 right justified, number of environments. For SOYB, there are 10 environments.

(6) Columns 42–45 left justified with 1 to 4 characters, brief name for environments.  Remember that the first letter of this name in column 42 must be unique (not identical to the entry in column 32 or 52).  For SOYB, the name “ENV” is given for the environments.

(7) Columns 46–50 right justified, number of replications.  If the data are unreplicated, then this entry should be 1. If the data are unbalanced, with various treatments having various numbers of replicates, then this entry should be the largest number of replicates encountered. For example, if various treatments have 2 to 4 replicates, then this entry should be 4.

(8) Columns 52–55 left justified with 1 to 4 columns, brief name for replications.  Remember that the first letter of this name in column 52 must be unique (not identical to the entry in column 32 or 42).  For SOYB, this and the previous entry indicate 4 REP for replications.

(9) Columns 58–60, order of data entries. Subscripts for genotypes, environments, and replicates are identified by the first letters of their brief names, which are located in columns 32, 42, and 52.  The order of data entries is specified by listing these subscripts according to their rate of change, *the first changing fastest and the last changing slowest.* If only one replicate is read, replication changes most slowly (in fact, never), so it is listed last (in the third position; it is *not* omitted).  For the SOYB data, the first data entry 2729 is for genotype 1, environment 1, and replicate 1.  The second entry, 2662, is for genotype 2, environment 1, and replicate 1.  Hence, genotype changes fastest, so it is listed first.  There are 7 genotypes.  The eighth data entry, 2747, is for genotype 1, environment 1, and replicate 2, so replicates change second fastest.  There are 7x4 = 28 entries for the first environment, and the twenty‑ninth entry, 983, begins the data for environment 2, the subscript for environment thus changing slowest.  Consequently, the data order specification for this dataset is GRE. Note that the subscript code letters given earlier in columns 32, 42, and 52, and the letters given here in columns 58–60 must match exactly, including case.

(10) Columns 62–65 left justified with 1 to 4 characters, brief name for treatments. A treatment is defined as a genotype-and-environment combination. The first letter of this name is *not* required to be unique relative to the first letters of the names for genotypes, environments, and replications (in columns 32, 42, and 52). For SOYB, this brief name is “TRT.”

(11) Columns 68–70 left justified with 1 to 3 characters, experimental design. For SOYB the experimental design is RCB for the randomized complete block design. Likewise, LAT could indicate a lattice design, or RAN a completely randomized design. The one and only experimental design that AMMISOFT is designed to recognize and analyze is the randomized complete block design, denoted by RCB, and this design makes possible calculation of an error mean square to be used for significance tests.

(12) Columns 71–75 right justified or else including decimal point, missing value indicator. This value must be numeric. Data entries with this value will be marked by AMMISOFT as missing data. Provided that no actual observations happen to have a value of zero, the missing value indicator may be set to zero in order to mark entries of zero as missing observations. A zero may be entered by blanks in columns 71–75, but it is better to enter an explicit “0” or “0.0” in order to make this entry visible. Note that a missing value indicator of zero causes data entries of zero to be treated as missing values, not as actual observations of zero. If there are no missing data, and observations are always positive, then a specification of zero will have the intended effect of marking no data as missing. Likewise, if there are no missing data, but some observations are zero although none are negative, then a missing value indicator of –1 will have the intended effect of marking no data as missing.

(13) Column 80, format supplied or not. If column 80 contains “N” (in upper or lower case), then no format is supplied and AMMISOFT reads the data in free format. Free format is suitable if no extraneous information is interspersed with the data, if a line break or at least one space always separates entries, and if entries of zero are given as “0” or equivalent rather than as a blank. If these conditions are not met, enter “F” (in upper or lower case) in column 80 and also supply the format, as explained momentarily.

If no format is supplied, that is if the second line contains “N” in column 80, then the following data lines are to be read in free format. Given G genotypes, E environments, and R replications, exactly GER entries must follow. Missing values, if any, are indicated by the missing value indicator (specified in columns 71–75 of line 2). Free format allows data entries to appear anywhere on a line, but all entries must be separated by a line break or at least one space, and there must be no extraneous information, including no commas.

**Optional Data Format.** Another line of the input data file is optional, included only if the format is supplied, that is, if column 80 of the second input line contains “F” rather than “N.” AMMISOFT uses extremely fast code written in FORTRAN to read the data file and to perform AMMI analysis (and additional code written in Python to read the output file and to make graphs). This format line contains one item, the supplied FORTRAN format in columns 1–72. Be sure to include the enclosing left and right parentheses. Field specifications for data entries must be real, not integer, because these entries will be read as single‑precision real numbers. Even if the data entries are typed as integers that are right justified in their fields, they must still be read as real numbers.

The option F is taken (in column 80 of line 2) when the data are to be read with a supplied format that uses the FORTRAN rules for a format statement. A supplied format is needed if the data lines contain extraneous information interspersed with the data entries, if line breaks or spaces do not always separate entries, or if entries of zero are indicated by blanks. Three examples follow of valid format specifications for this option F:

(7F9.2)

(5F5.0, 10X, F5.0, 15X, 2F5.0)

(3(10F8.3), /, 4F8.3)

The first and second examples are rather simple. The third applies to data organized in sets of four lines, with three lines similar and the fourth different.

Explicit decimal points in data entries override the format statement. For example, the format specifier F5.0 will correctly read all of the following fields of 5 characters where “^” is used to indicate the presence of a blank column: ^2.3^, ^1653, 1653., 1.2E3, and ^^^^5. However, 1653^ would *not* be read as 1653, but rather as 16530. Values entered like integers, without decimal points, must be right justified in their fields in order to be read correctly.

**Data Entries.** The data entries then follow in the input data file. Values may be positive, negative, or zero, but values equaling the missing value indicator will be marked as missing data. The data are stored in single precision, so up to about 7 significant digits can be retained. For a specified format, be sure to enter complete data lines satisfying the entire format, even if some entries are zero (regardless whether a zero represents an actual observation of zero or else a missing observation). For example, in option F with format (4F5.0), given a line containing only two actual observations of 2541 and 2619 and two missing observations indicated by values of zero, this line must still contain the full 20 characters called for by this format, although two of the fields (of 5 characters each) may be filled with blanks.

As noted earlier regarding scaling, in order to produce results which print nicely in the print formats used by AMMISOFT, it is required that typical data entries be in the range of 0.001 to 1000000, or better yet within 0.01 to 100000, or ideally within 1 to 10000.

Because the order of change of subscripts can be specified, if the data have already been entered into the computer for other purposes, the data may already be nearly in a suitable format for AMMISOFT.  But if the data file contains any extraneous information, such as line numbers or code names for anything, a specified format should be used in order to ignore this extraneous information.  In unusually complex cases, a small computer program may need to be written to transcribe the data into a format acceptable to AMMISOFT.

**Genotype and Environment Code Names.** Next the input data file contains 4‑character code or brief names for the genotypes and environments.  The format is 4 characters followed by a blank, with 15 entries per line.  The names use alphanumeric characters and may contain blanks.  Genotype names are given first, and then, beginning on a new line, environment names.  Should the genotypes or environments have no particular names or significance, it is still necessary to supply these lines, even if this is done by supplying the appropriate number of lines merely containing blanks.  For SOYB, the 7 genotype code names use one line, and the 10 environment code names (indicating location-and-year combinations) occur on the following line.

**Example.** The following simple example illustrates a correct format to be read by AMMISOFT. It is the supplied data file TOYD.txt that appears on pages 55 and 86 of my text on AMMI: Hugh G. Gauch, Jr., 1992, Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs, Elsevier.

A ruler has been placed above this example to facilitate determining column positions.

1 2 3 4 5 6 7 8

12345678901234567890123456789012345678901234567890123456789012345678901234567890

Toy data, Table 3.1 page 55 in Gauch 1992 Elsevier text on AMMI. TOYD

Yield kg/ha 0 5 GEN 4 ENV 3 REP RGE TRT RAN 0 N

352 298 331 249 295 251 266 226 219 231 254 196 182 199 216

268 271 232 253 270 218 243 215 205 199 205 217 234 184 221

201 170 169 165 214 206 154 206 153 157 176 144 171 159 195

125 121 102 180 133 158 159 146 124 170 180 139 121 138 164

GEN1 GEN2 GEN3 GEN4 GEN5

ENV1 ENV2 ENV3 ENV4

**Two Additional Examples**. Dataset SOYB.txt has data from a New York soybean trial that appears on page 56 of my text. Dataset ESW8.txt has an international wheat trial published in Crossa, Fox, Pfeiffer, Rajaram, and Gauch (1991), Theoretical and Applied Genetics 81:27–37. It was also used as the main example in my 2013 paper on AMMI protocol that was already cited.