United States
Department
of Agriculture

Forest Service

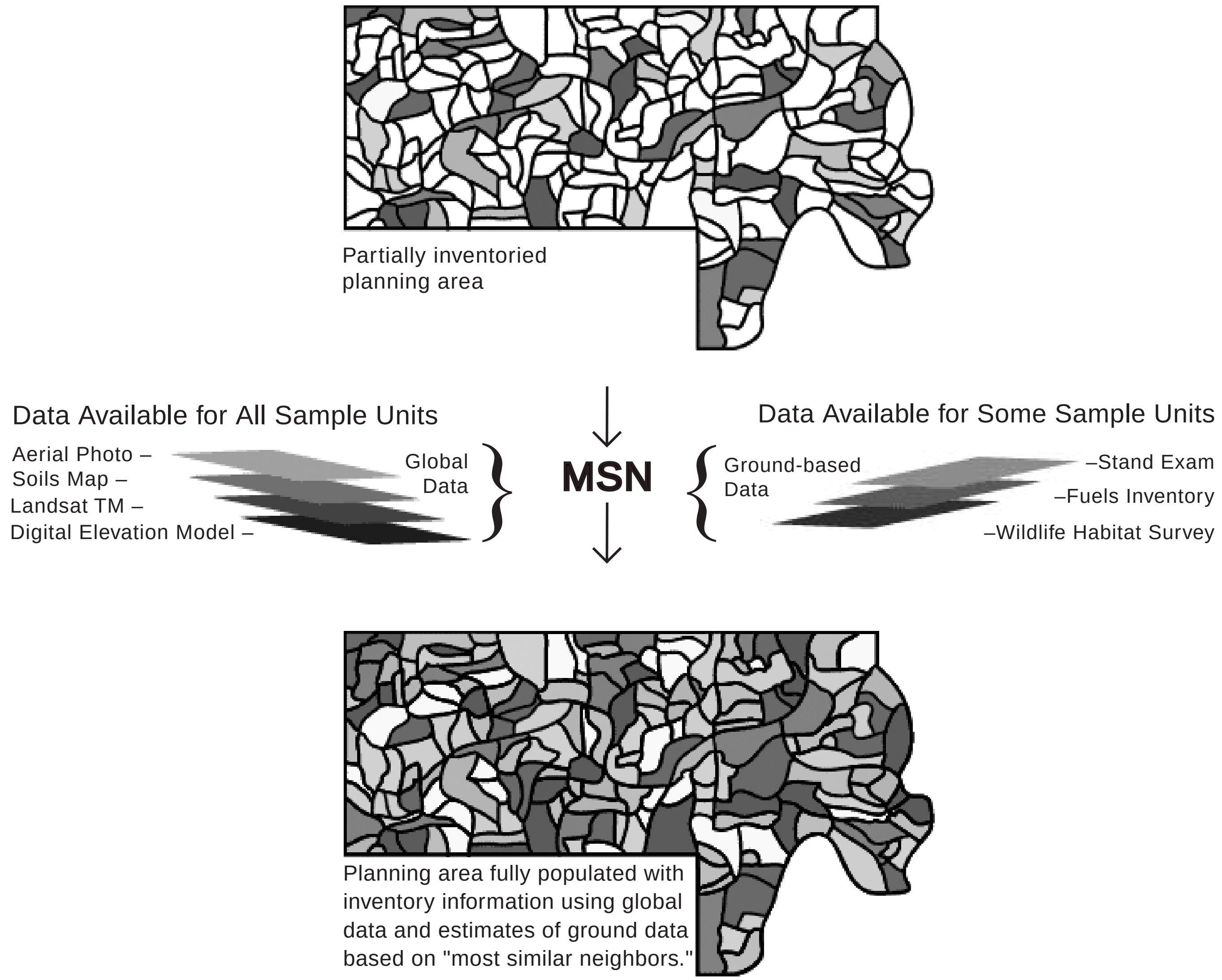
**Rocky Mountain
Research Station**

General Technical
Report RMRS-GTR-96

September 2002

Users Guide to the Most Similar Neighbor Imputation Program Version 2

Nicholas L. Crookston
Melinda Moeur
David Renner


Partially inventoried
planning area
Data Available for All Sample Units
Aerial Photo –
Soils Map –
Landsat TM –
Digital Elevation Model –
Global
Data
MSN
Data Available for Some Sample Units
Ground-based
Data
–Stand Exam
–Fuels Inventory
–Wildlife Habitat Survey
Planning area fully populated with
inventory information using global
data and estimates of ground data
based on "most similar neighbors."

Abstract

Crookston, Nicholas L.; Moeur, Melinda; Renner, David. 2002. Users guide to the Most Similar Neighbor Imputation Program Version 2. Gen. Tech. Rep. RMRS-GTR-96. Ogden, UT: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 35 p.

The Most Similar Neighbor (MSN, Moeur and Stage 1995) program is used to impute attributes measured on some sample units to sample units where they are not measured. In forestry applications, forest stands or vegetation polygons are examples of sample units. Attributes from detailed vegetation inventories are imputed to sample units where that information is not measured. MSN performs a canonical correlation analysis between information measured on all units and the detailed inventory data to guide the selection of measurements to impute. This report presents an introductory discussion of Most Similar Neighbor imputation and shows how to run the program. An example taken from a forest inventory application is presented with notes on other applications and experiences using MSN. Technical details of the way MSN works are included. Information on how to get and install the program and on computer system requirements is appended. The MSN Web address is: http://forest.moscowfsl.wsu.edu/gems/msn.html.

Keywords: canonical correlation, imputation, forest inventory, forest planning, landscape analysis

The Authors

Nicholas L. Crookston is an Operations Research Analyst at the Moscow Forestry Sciences Laboratory. His contributions have included developing extensions to the Forest Vegetation Simulator.

Melinda Moeur is the Vegetation Module Leader at the Interagency Monitoring Program, Northwest Forest Plan, Strategic Planning, Pacific Northwest Region, USDA Forest Service, Portland, OR. While she worked on this project she was a Research Forester at the Rocky Mountain Research Station's Forestry Sciences Laboratory in Moscow, ID. Her principal research interests include forest vegetation modeling and sampling inference.

David Renner is a freelance Computer Programmer in Moscow, ID. He has contributed to several projects at the USDA Forest Service, Rocky Mountain Research Station's Moscow Laboratory related to FVS and its application. He also works at the University of Idaho and as a private forestry consultant.

You may order additional copies of this publication by sending your mailing information in label form through one of the following media. Please specify the publication title and number.

Telephone	(970) 498-1392
FAX	(970) 498-1396
E-mail	rschneider@fs.fed.us
Web site	http://www.fs.fed.us/rm
Mailing Address	Publications Distribution Rocky Mountain Research Station 240 West Prospect Road Fort Collins, CO 80526

Contents

Page

Users Guide to the Most Similar Neighbor Imputation Program Version 2

Nicholas L. Crookston
Melinda Moeur
David Renner

Introduction to Most Similar Neighbor Imputation

Purpose

The Most Similar Neighbor (MSN, Moeur and Stage 1995) program is used to impute attributes measured on some sample units to sample units where they are not measured. A list of all sample units in a problem is constructed, and some attributes are measured for every unit. Additional attributes, typically those that are much more expensive to measure, are recorded for some of the units. The MSN program picks the most similar unit where the additional attributes are measured to impute to a unit where the additional attributes are not measured. Information about the relationship between the attributes measured for all units and the additional attributes measured on a sample of the units is compiled by the program using canonical correlation analysis and used to compute the similarity among units.

An application in forestry serves to illustrate the idea. A list of all the forest stands in a large watershed is made. In this case each forest stand is a sample unit. Then, easy-to-measure attributes are recorded for all the forest stands in the watershed. These attributes can come from any source; aerial photographs and topographic maps are good examples. For a sample of the stands, attributes are measured during on-the-ground visits. The program picks the most similar stand that was visited on the ground to characterize a stand that was not visited.

The technique was developed to support detailed, landscape-level analyses, required in today's forest planning environment. Geographic Information systems (GIS) linked with databases that store inventory information about landscape elements provide the data tools needed to create these plans. Despite the availability of sophisticated display and database tools, a comprehensive landscape wide plan is still difficult to create because the inventory is rarely complete. In other words, even though a GIS can easily display all the planning units (such as forest stands) in an analysis area, rarely do all planning units have a ground-based inventory. For planning purposes, it would be convenient to be able to operate as if detailed inventory information were available for all units in the planning area. For units in a landscape that lack ground-based attributes, MSN can be used to find the most similar ground-based unit and impute its inventory attributes to the unit that lacks those data (Moeur and others 1995, Ek and others 1997, Van Deusen 1997).

Important Terms

In the MSN system, all the units in a problem have some measured information and are therefore called *observations*. In MSN, as in databases or spreadsheets, an observation is a row in a table. The columns of the table are known as *variables*. In a forestry application, the observations might be stands, and an example variable is the number of trees per acre. A *value* is a cell in the table.

Variables measured for all the observations in the MSN analysis are known as *X-variables*. The additional variables measured only for a sample of the observations are known as *Y-variables*.

Common X-variables in forestry include those obtained from aerial photo interpretation such as crown closure, average tree height, species group, and stocking class. Variables derived from remotely sensed satellite spectral data and digital elevation models are also used.

Common Y-variables include those obtained from stand examinations or similar vegetation inventories such as Forest Inventory and Analysis grid plots. Examples include the basal area, percent canopy cover, trees per acre, volume, species composition, and size class.

Observations that have measured Y-variables and X-variables are called *reference* observations. Observations that lack measured Y-variables are called *target* observations. The objective of MSN is to pick a reference observation as a source of Y-variables to impute to a target observation. This is done by computing the weighted distances between each target observation and every reference observation. The reference observation with the shortest weighted distance between itself and a given target observation is the target observation's *most similar neighbor* (Moeur and Stage 1995).

Moeur and Stage (1995) defined an additional class of variables not used directly in the MSN process but that are imputed to the targets. These variables are measured on the sample of plots along with the Y-variables. For example, the measurements made on individual trees in a detailed inventory may be too numerous to be useful as Y-variables but may be aggregated into Y-variables. The basal area (a Y-variable) is computed using a function that sums the basal area of each sample tree weighted by the number of trees each represents in the sample. Frequently, it is the detailed inventory itself, as represented by all the detailed variables, that is imputed.

Elements in an MSN Run

MSN program follows these steps:

- Data for all observations are read and prepared for processing according to data input instructions you provide. The program classifies observations with both Y- and X-variables as reference observations, and those with only X-variables are classified as target observations.
- A canonical correlation analysis using the Y- and X-variables from the reference observations is performed. Outputs are produced that report on the strength of the canonical correlations.
- The most similar reference observation is selected for each target observation and output to a For-Use file (*for* a given target observation, *use* a specific reference observation's attributes). The canonical correlation analysis computed in the previous step is used in determining which reference observation is most similar to a target. When calculating how

similar a reference is to a target, MSN gives more weight to X-variables that are strongly correlated with the Y-variables than those that are weakly correlated. More information on this topic is provided in the following pages.

- Validation statistics display how well a run of MSN worked. These statistics compare the imputed values of Y- and X-variables to observed values. For target observations, only X-variables have observed values so comparisons are not possible for Y-variables. For reference observations, Y- and X-variables have observed values. We define the most similar neighbor to a reference observation to be the observation itself, and therefore the observed and imputed values are identical. Reference observations are therefore compared to their second most similar neighbor when the validation statistics are computed.

A Simplistic Example

The preceding explanation provides you with an overview of the process and its application. The following example illustrates several points that will help you understand the terms and concepts introduced so far.

This example contains three observations. The elevation and location, measured by the UTM coordinates of the plot centers, are known for each of the observations. For two of the observations, the basal area and volume are measured with a detailed inventory. The Y-variables are *basal area* and *volume*, and the X-variables are *easting, northing,* and *elevation*. There are two reference observations and one target observation.

Which one of the two reference observations is most like the target observation? Given the UTM coordinates and the elevations of all observations, it is a simple matter to compute the Euclidean distance between the target and each of the reference observations. The reference with the shortest distance is the nearest observation.

What if you have some evidence that *basal area* and *volume* of the observations do not change much along the gradient from east to west, but do change greatly along the *northing* and *elevation* gradients? Given this information you could simply leave out *easting*, recompute the distances in two dimensions, and conclude the analysis.

In real problems, leaving a variable out like *easting* out might not be a good idea as it could contribute important information in picking a similar neighbor. A solution is to give *easting* less weight in computing the distance as given *northing* and *elevation*.

How much weight should be given to each of the X-variables? Canonical correlation analysis finds linear combinations of Y-variables that have maximum simple correlations to linear combinations of X-variables. The MSN program computes the weights from the coefficients of these linear combinations of the X-variables and their respective canonical R-squares. The exact formula and technical details are presented in the section titled “Technical Details.” The reference observation with the shortest weighted distance between itself and a given target observation is the target observation’s *most similar neighbor* (Moeur and Stage 1995).

The last point regarding this example is that if there is no relationship between the Y- and X-variables, MSN provides no justification for imputing either reference observation to the target. But even with little or no relationship and therefore with little or no justification, the program will assign the best neighbors it can find. It is left for you to decide if those neighbors are satisfactory for your purposes. The program does provide some statistical summaries that can be used to help you evaluate the utility of the results.

New in Version 2

This version of MSN program, including this user's guide, is based largely on a prepublication release. New features include the ability to read comma and tab delimited input files, the addition of Chi Square and Kappa statistics to the contingency tables, several new variations of the distance function, and a limited implementation of K-MSN providing more than one possible selection in the neighborhood of similar observations. In addition, the method used to compute the canonical correlations was improved, the output was remodeled, greatly improved diagnostic messages are generated, and the installation and operating instructions for Windows and Unix systems were standardized.

What Follows

Instructions on how to run the program, followed by descriptions of the output, are presented in the next two sections. An example taken from a real forest inventory application illustrates these sections. Applications of the technique in forestry are presented as a source of guidance on how to apply MSN to your situation. Some factors that control how the MSN process works are included. A technical presentation of the mathematics behind the calculations follows. Information about how to get and install the program and computer system requirements is presented in appendix A. As some MSN runs can take hours, even on fast computers, we have included some run time information in appendix B with some advice on how to manage large runs.

How to Run the Program

Commands

The MSN program runs from the command line on Unix and Windows systems. It can also be started from an icon, a file explorer, or a run dialog.

MSN reads a command file and follows the instructions found therein. The command file is a standard ASCII text file that can be created with a text editor. The format of the data inside the file is highly structured and must be carefully coded. You specify the name of the command file on the same line as you use to start MSN. For example, to run the command file called *example.msn*, you would enter this command:

```
MSN example.msn
```

Command files contain commands that start in the first column of each input line. Some commands require that additional information be entered on the same line, following the command, starting in column 14. Any line in the command file that contains an asterisk (*) in the first column is considered a comment and is ignored by the program. Several of the commands are optional. If they are left out, the program will perform according to a preset default. All the commands are described below grouped into functional sets.

Several commands are used to specify the names of data files used to enter data into the program and files to which reports are written. In all cases, those names may be up to 256 characters in length and start after column 14, following the command. Note that the MSN program automatically changes the working directory to the directory that contains the command file. If the file being named is not in the directory that contains the MSN command file, then full path names should be used. Remember that file names are case-sensitive on Unix systems.

Commands that control entering data and request output

INFILE Specifies the name of an input data file and signals that *data format lines* follow in the command file. Data format lines are described in detail in the following section titled "Input Data Formats and Identifying Y- and X-Variables." The line following the last data format line is the command ENDFILE coded starting in the first column, just like other commands. You can have as many input data files as you need. Each is made known to the MSN program with separate sets of INFILE/ENDFILE commands.

REPORT Specifies the name of the standard report file to which MSN summary information is written. The default is the command file name with the file suffix replaced with .rpt. For example, if the command file is called *example.msn* the default report file name is *example.rpt*.

FORUSE Requests that the *For-Use* file be output, and provides a way to specify the file name. This is the file to which MSN assignments are written. The default is to not write this file. The default file name is to replace the command file name suffix with .fus. For example, if the command file is called *example.msn* the default *For-Use* file name is *example.fus*.

OBSIMPU Requests that the *Observed-Imputed* file be output, and provides a way to specify the file name. This is the file to which MSN accuracy assessment information for individual observations is written; the default is to not write this file. The default file name is to replace the command file name suffix with .obi. For example, if the command file is called *example.msn* the default *Observed-Imputed* file name is *example.obi*.

Commands that control program execution

PROCESS Signals the end of commands related to a single MSN analysis and instructs the MSN program to process the analysis. An unlimited number of individual MSN analyses may be "stacked" within the command file.

STOP Signals the program to stop.

Commands that control optional report generation

PRTCORXX Signals that simple correlations for X-variables be written to the report file; the default is to not write this information.

PRTCORYY Signals that simple correlations for Y-variables be written to the report file; the default is to not write this information.

PRTCORYX Signals that simple cross-correlations between X-variables and Y-variables be written to the report file; the default is to not write this information.

PRTWGHTS Signals that coefficients for each Y- and X- variable used in the weight matrix in the MSN distance function be written to the report file; the default is to not write this information.

RUNTITLE Provides for entering a run title that is written to the report file; the default is no title. The title can be up to 256 characters long.

Commands that control the MSN selection process

DISTMETH Specifies the method used to compute the distances. There are five ways. The default, coded method zero, follows the original Moeur and Stage (1995) formulation. All of the other methods are presented in the section on “Technical Details.” Briefly, those methods are (1) use an alternative to the original formula, (2) compute a Mahalanobis distance on normalized X-variables and thereby ignore the canonical correlation results, (3) use a weight matrix input on data lines that follow this command, and (4) compute Euclidean distances on normalized X’s.

PROPVAR Specifies the proportion of total variance used in the distance calculations, entered as a real number following the command; the default is 0.9. This number is used to calculate the number of sets, or vectors, of canonical correlation coefficients used. See the section titled “Technical Details” for more information. This command cannot be used if the NVECTORS command is used (described next).

NVECTORS Specifies the number of vectors of canonical correlation coefficients used in the distance calculation; the default is to defer to the PROPVAR command, described above. This command cannot be used if the PROPVAR command is used. See the section titled “Technical Details” for more information.

RANNSEED Reseeds the random number generator that is invoked in the case of ties between most similar neighbors. Ties will be rare when the input global data consist of continuously valued Y-variables but may be common when using categorical Y-variables. This command also affects the results derived using the RANDOMIZE command described below.

KMSN Specifies the number of additional neighbors, besides the most similar one, that are desired. Using this command does not change the validation statistics or the contents of the For-Use file compared to not using it. However, using it can change the contents of the Observed-Imputed file. See the description of the Observed-Imputed file in the section entitled “Output” for more information on this limited implementation of KMSN and an explanation of how the contents of the file changes when the command is used.

NOREFS Suppresses the inclusion of the reference observations in the For-Use and Observed-Imputed files and saves the computer time used to find the second most similar neighbors for reference observations. The feature is useful in cases where you have many thousands of reference observations and have no need for the additional information and validation statistics that depend on knowing the second most similar neighbors for reference observations.

NOTARGS Suppresses the inclusion of the target observations in the For-Use and Observed-Imputed files and saves the computer time used to find the most similar neighbors for target observations. The feature is useful in cases where you have many thousands of target units in your data but do not want to wait for your computer to find the most similar neighbors. It is used when making runs where only the validation statistics are desired.

MOSTUSED Specifies the number of reference observations to report as those most frequently used to represent target observations; the default is 20. The same value is used to control the number of observations listed in the report of the largest distances between reference and target observations.

RANDOMIZE Randomizes the observed X-variables with respect to the observed Y-variables among the reference observations. It leaves intact the relationships within the Y’s and within the X’s while destroying the relationship between them. If using this feature provides results as good as those you get when you don’t use it, then there is no value in applying MSN in your case. This topic is revisited in the section entitled “Output.” The RANNSEED command can be used to create unique randomizations between runs.

Input Data Formats and Identifying Y- and X-Variables

Data format lines

Data format lines are entered between the INFILE and ENDFILE commands. Each line defines a field of data on the input data records. A field may be an input variable or an observation identification code. For each variable, the lines define

- if it is a Y- or an X-variable
- if you desire most similar neighbor accuracy to be assessed for this variable
- its name
- if it is a categorical or continuously valued variable
- where on the input data records the value is located

Table 1—Elements and function of the *data format lines*.

Name	Columns	Value	Meaning or description
Comment	1	*	Signals the line is a comment line.
		blank	Signals the line be processed as data (except that if the entire line is blank, it is skipped).
Data field type	2	L or l	The line describes observation identification (or identification *label*). The maximum length is 26 characters.
		Y or y	The line describes a Y-variable.
		X or x	The line describes an X-variable.
		blank	The variable is none of the above, but still may be entered for validation purposes.
Validation code	3	V or v	Signals that *validation* statistics be produced for the variable. This is meaningless for identification labels.
		blank	No validation data is to be produced with for the variable.
Data field name	5-16	*name*	The name of the continuously valued variable or identification code; note that there is no period following the name.
		name.	The name of the variable categorical variable if you want MSN to construct dummy variables for you (see the text on naming conventions); note that a period '.' follows the name.
		name.level	The name of the variable categorical variable, followed by its *level* or *value* when you have provided the dummy variable coding (see the text on naming conventions).
Data field number when free-form format specification is used (input data are not in specific columns).	20-34	*n*	Commas, spaces, tabs, or any combinations of spaces, commas, and tabs separate the data fields in the input records. Missing values are indicated by a period that is preceded and followed by a space, comma, or tab. Missing values are also indicated when two commas or two tabs are found with a blank string or nothing in between them. Identification codes and categorical data are processed as character strings (single or double quote marks may be present and are stripped by the program) and continuous data are processed as numbers. Free-form formats and fixed format specifications cannot be mixed in the same file. Values over 12 characters long are truncated. *n* is the field number.
Data field location when fixed format specification is used (input data are in specific columns).	20-34	*n,m*	The values are in specific columns on the input records. Commas and spaces are not used to separate fields. Tabs are illegal in the input file. Fixed format free-form format specifications cannot be mixed in the same file. *n* is the column location on an input record where the values for the variable start *m* is the column location on an input record where the values for the variable end. The length (m-n+1) must not exceed 12 characters.
		/	Signals that MSN should move to the next physical input line to continue reading information for the current observation. When / is used, the rest of the data format line is left blank.
Comment	35+	*text*	You can place descriptive information here if you wish.

The placement of the information on data format lines follows strict rules as outlined in table 1.

Observation identifications are used to merge data from several input files. Therefore, the identification codes (or *labels*) are required for each logical input line. A logical input line may contain more than one physical data line. The program reads a logical input line in a single input operation.

MSN supports two styles of input formats, free form and fixed format. Free-form files can have exactly one physical record for each input operation. Commas, tabs, spaces, in any combination, can separate the data (table 2 illustrates this). Missing values are those that are left blank or those that contain only a period as the input value. You specify the field number corresponding to the variable using free-form format specifications.

Fixed format files require that you enter the beginning and ending position on the data records corresponding to a variable. The fixed format files can have more than one physical input line for each input operation.

You can enter data from different input data files in the same run. Each file may contain data for some or all of the observations, some or all of the Y-, X-, or both kinds of variables, in any combination. MSN merges information that belongs to the same observations using the identification codes as keys. It merges the values corresponding to each variable using the variable names as keys.

The program checks for missing observations and classifies each observation as a reference or a target using these steps, applied in the order listed:

1. Observations that have missing values for all Y-variables are classified target observations. The others are classified as reference observations.
2. If over 80 percent of the values for a given variable are missing, then the variable is dropped. Y-variables are checked among reference observations, and X-variables are checked among all observations.
3. If a reference observation has any missing values among its Y-variables, it is converted to a target observation.
4. Observations that have missing values among the X-variables are dropped.

Messages are output journaling the actions taken. In general, the solution to having missing data is to ensure there are none. The output generated by MSN when missing values are detected will help you find the problems or conclude that you should accept the results gained with the data you have.

Continuous and categorical variable naming convention

Variables may be continuous or categorical. Continuous variables are real-valued variables measuring the magnitude of a particular attribute. Examples

Table 2—Examples showing the interpretation of free-form input data.

Example input line	Data field values		
1,2,3	1	2	3
1 . 3	1	.	3
, ,.	Blank	Blank	.
’a quote” ’,, .	a quote”	Blank	.
LP,PP,AF	LP	PP	AF
LP , PP , ”AF”	LP	PP	AF

of continuous variables are stand basal area, percent crown cover for a given species, and average elevation. Continuous variables may also be rankings or ratings, such as disease hazard rating (example, 1=low, 2=moderate, 3=severe) where the ranked values can be interpreted as equally spaced and relative in magnitude.

Categorical variables contain data with class levels that do not lend themselves to numerical ordering. Examples are species codes and cover type codes.

Within the MSN program, categorical variable data are processed using *dummy variables*. There are as many dummy variables as unique categories; each takes on a value of 1 if the observation is from the category, or 0 if the observation is from another category. You can set up the dummy variables yourself or let MSN set them up for you.

To signal that a variable is categorical, you use a period in the variable's name (table 1). When the period is the last character in the name, MSN will set up the dummy variables for you. In that case, MSN will create a variable corresponding to each unique category found in the data and name the variable in two parts. The first part is the name you provide, including the period, followed by the value of the category. For example, say you have named a variable `Species`. If MSN finds two species, PP and DF, for example, MSN will build two variables, `Species.PP` and `Species.DF`. For the variable called `Species.PP` each observation where PP was found will have the value of 1 and the value of zero when PP is not found.

You can set up the dummy variables yourself using exactly the same approach. If you do, name the variables using the same two-part naming convention. MSN computes different validation statistics for continuous versus categorical data, and the variable names are used to distinguish between the types.

Example Run Input

An example illustrates most MSN features and concepts. It is taken from the Deschutes National Forest in central Oregon (Moeur 2000). In this case, 197 stands with ground-based inventory data (reference observations) are used to impute attributes to 399 additional stands (target observations) that have no ground-based inventory data.

The contents of the command file are presented in figure 1. Line 1 is a comment as it starts with an asterisk ('*'), and line 2 is ignored because it is blank. We have used most of the commands and asked for all the reports so we can discuss the outputs. Inspecting lines 3 through 11 displayed in figure 1 reveals the name MSN will give to the report and for-use files and displays other options.

There are two input files, one contains the Y-variables and the other contains the X-variables. As we said above, this is not a requirement because we could have several files where each contains both kinds of data.

Line 15 of figure 1 names the first input data file, from which the Y-variables are read. Figure 2 displays a few example records for these data. The data format lines related to these data start on line 17 and end with the ENDFILE command at line 54. Each logical data line in each file must have an identification code. The data in line 17 specifies that its name is ExamNo and it is coded in the input data starting in column 1 and ends in column 8. Note that both the beginning and ending columns are specified for this file

```
  Line |                             Column ruler
Number | ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----
     1 | * example_command.txt  Command file for User's Guide example
     2 |
     3 | REPORT       example_report.txt
     4 | RUNTITL      User's Guide example run
     5 | OBSIMPU      example_obsimpu.txt
     6 | FORUSE       example_foruse.txt
     7 | PROPVAR       .90
     8 | PRTCORXX
     9 | PRTCORYY
    10 | PRTCORYX
    11 | PRTWGHTS
    12 |
    13 | *Training Observations, Inventory data
    14 |
    15 | INFILE       inventory_data.txt
    16 | *Record 1--ExamNo, Constants, Major Spp Code, Plant Assoc Group
    17 |  L  ExamNo          1,8              Stand number
    18 |  YV MSC.            22,23            Tree species code of plurality
    19 |   V PAG.            26,28            Plant Association Group
    20 |                     /                Line feed to next record
    21 | *Record 2--Percent Canopy Cover by dbh class
    22 |
    23 | *   COV_SML         11,16            Cover in dbh class 0- 4.9
    24 |  Y  COV_MED         18,24            Cover in dbh class 5-19.9
    25 |  YV COV_LRG         25,31            Cover in dbh class   20+
    26 |                     /                Line feed to next record
    27 | *Record 3--Stand-level totals from FVS summary
    28 |
    29 |  Y  Tot_TPA         11,17            Trees per acre
    30 |  YV Tot_BA          18,24            Basal area per acre
    31 |  YV TopHt           25,31            Avg. height of 40 largest-dbh trees/acre
    32 |  YV Tot_Vol         32,38            Total cubic foot volume
    33 |   V Tot_Mort        39,45            Mortality (CuFt/yr)
    34 |                     /                Line feed to next record
    35 |
    36 | *Record 4--Basal area per acre by species group
    37 |
    38 |  YV BA_LP           11,17            Lodgepole pine
    39 | *   BA_DF           18,24            Douglas-fir
    40 |  YV BA_IINE         25,31            Ponderosa + Sugar + Western White pines
    41 |  YV BA_FIR          32,38            True Firs
    42 | *   BA_OTHR         39,45            Other Species (MH,IC,ES,HWD)
    43 |                     /                Line feed to next record
    44 | *Record 5--Basal area per acre by size class
    45 |
    46 |  YV BA_IOLE         11,17            5.0-15.9" dbh (small & large poles)
    47 |  YV BA_SAW          18,24            16.0+" dbh (small, med & large sawtimber)
    48 |                     /                Line feed to next record
    49 |
    50 | *Record 6--Trees per acre by size class
    51 |
    52 |  YV TPA_SML         11,17            0.0-4.9" dbh (seedlings & saplings)
    53 |  YV TPA_LRG         18,24            16.0+" dbh (large trees)
    54 | ENDFILE
    55 |
    56 | *Satellite and DEM data
    57 |
    58 | INFILE       satellite_data.txt
    59 | * ExamNo, UTM Coordinates, Elevation, LandSat TM, Tasselled Cap and NDVI.
    60 |  L  ExamNo            1              Stand number
    61 | *X  UTMX              2              E UTM coords of stand polygon centroid
```

Figure 1 (Con.)

```
     62 | *X  UTMY            3             N UTM coords of stand polygon centroid
     63 |  XV ELEV            4             Elevation (m)
     64 | *   BAND1           5             Band 1 450-520 nm (blue)
     65 |  XV BAND2           6             Band 2 520-600 nm (green)
     66 |  XV BAND3           7             Band 3 630-690 nm (red)
     67 |  X  BAND4           8             Band 4 760-900 nm (near-IR)
     68 |  X  BAND5           9             Band 5 1550-1750 nm (mid-IR)
     69 | *   BAND7          10             Band 7 2080-2350 nm (mid-IR)
     70 | *   BRT            11             Tasselled Cap Component 1
     71 |  X  GRN            12             Tasselled Cap Component 2
     72 | *   WET            13             Tasselled Cap Component 3
     73 |  XV NDVI           14             Normalized Difference Vegetation Index
     74 | ENDFILE
     75 |
     76 | PROCESS
     77 | STOP
```

Figure 1—Command file used in the example.

```
 Line  |                            Column ruler
 Number| ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----
     1 | 29610287  601  1997  PP  PPD
     2 |      Cover    0.9   24.8   17.2
     3 |    StandSum     97    147     90   4753     33
     4 |       BAxSp      0      0    147      0      0
     5 |     BAxSize     39    107
     6 |    TPAxSize     10     39

       | [observations omitted]

  1177 | 29651979  601  1996  LP  LPD
  1178 |      Cover    0.8    5.9    1.8
  1179 |    StandSum    821     23     32    702      0
  1180 |       BAxSp      2      0     21      0      0
  1181 |     BAxSize     11     11
  1182 |    TPAxSize    806      4
```

Figure 2—First and last observation from the file inventory_data.txt.

using the fixed format specification (table 1). Line 18 identifies a categorical variable named MSC. Its values start in column 22 and end in column 23. A comment identifies this variable as a species code.

All the variables in the first input file are Y-variables and so signified by a Y coded in the second column of the data format lines. An exception is found in line 33 that shows the variable Tot_Mort is entered into the program for validation but neither as a Y- nor an X-variable in the canonical correlation analysis. Variables are included in the validation process when a V is found in the third column.

Not all the data for each observation are on the same physical record, so MSN is instructed to move on to the next record using the "/" symbol as shown in line 20. In this example, there are six physical data lines for each logical input operation.

The X-variables are read on a separate file and are composed of data derived from LandSat TM data and the physical location of the observations (fig. 3 contains some example lines). The INFILE command for this file is on

```
 Line                              Column ruler
Number   ----+----1----+----2----+----3----+----4----+----5----+----6----+----7----+----8----+----
     1   29610001,594744.44,4793027.50,1583,68.8,26.5,28.1,60.5,66.4,28.7,114.50,2.91,14.91,0.3657

         [Observations omitted]

   596   29651979,601605.06,4799502.5,1424,88.1,38.4,50.1,61.6,114.5,59.9,165.0,-18.22,50.54,0.103
```

Figure 3—First and last observation from the file satellite_data.txt

comma-delimited, and a free-form format specification (table 1) is used. In this case, the field numbers are coded on the data format lines rather than the beginning and ending columns. Line 60 illustrates that the label, called ExamNo, is the first value on the input records, and line 67 shows that BAND4, is the eighth value on the input data record.

MSN processes the run upon reading the PROCESS command, shown in line 76.

Output

The output from MSN contains a report and two optional files, the For-Use file and the Observed-Imputed file. Output is also written to the console as the program progresses. The informational messages in this file are self-explanatory.

Report File

The report file provides a summary of the input data, the results of the canonical correlation analysis used to create the weighting matrix for the MSN distance model, the neighbor selection procedure, and overall run validation including accuracy assessment statistics for individual variables. Notable sections of the report are described below.

Input data preparation

The input data preparation section summarizes the data format lines read from the command file for each of the files (fig. 4). The formats are contained between the INFILE and ENDFILE commands for L-, Y-, and X-variables read from each of the input files.

Output file names and data checking operations

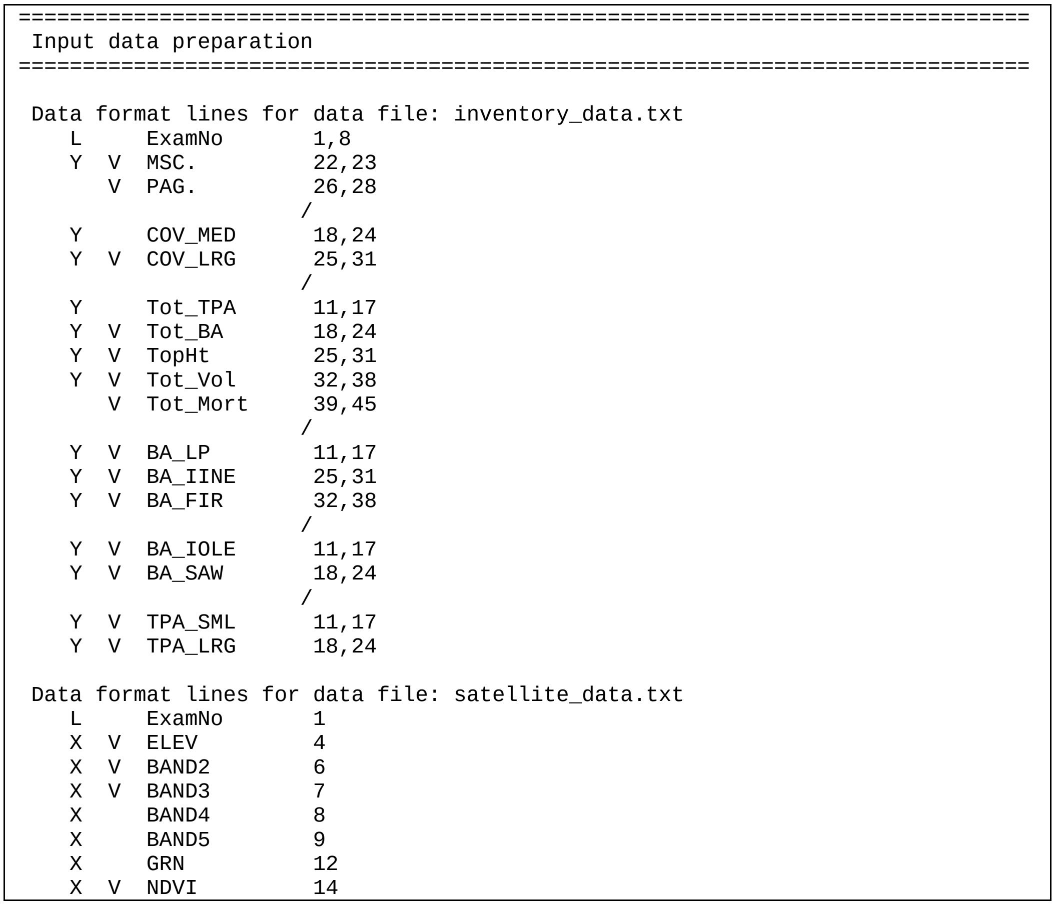
The output file names and information generated as the program merges your data and prepares it for further analysis are printed next. If there are no errors or problems to report, this section will contain only the file names.

Reference and target data reports

Figure 5 illustrates the target data report. The number of observations is reported with descriptive statistics for the X-variables in the target observation set. A similar report is produced for the reference data.

Correlations

Figure 6 illustrates a simple correlation table between the X- and Y-variables for the reference observations. Correlation tables are also generated between the X-variables and the Y-variables for the reference observations and the X-variables for the target observations.

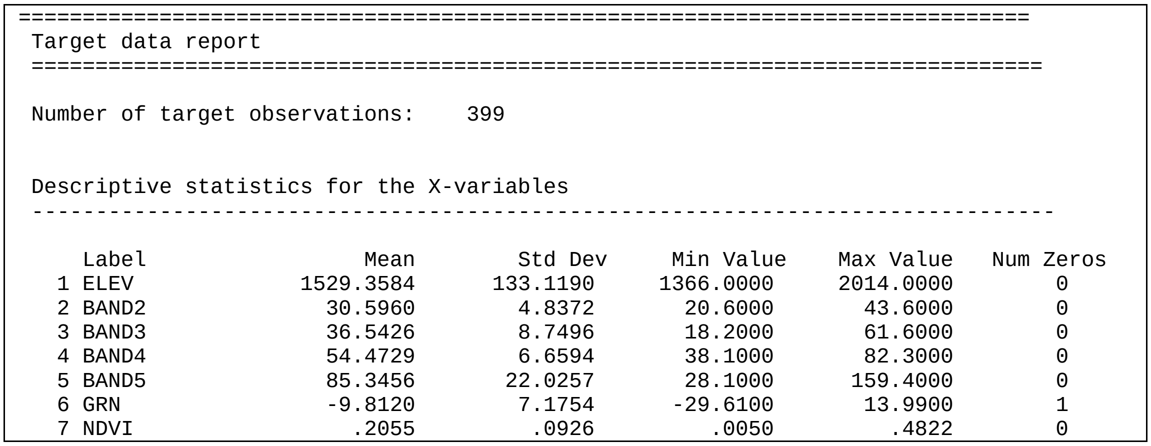


```
===============================================================================
 Input data preparation
===============================================================================

 Data format lines for data file: inventory_data.txt
    L     ExamNo       1,8
    Y  V  MSC.         22,23
       V  PAG.         26,28
                      /
    Y     COV_MED      18,24
    Y  V  COV_LRG      25,31
                      /
    Y     Tot_TPA      11,17
    Y  V  Tot_BA       18,24
    Y  V  TopHt        25,31
    Y  V  Tot_Vol      32,38
       V  Tot_Mort     39,45
                      /
    Y  V  BA_LP        11,17
    Y  V  BA_IINE      25,31
    Y  V  BA_FIR       32,38
                      /
    Y  V  BA_IOLE      11,17
    Y  V  BA_SAW       18,24
                      /
    Y  V  TPA_SML      11,17
    Y  V  TPA_LRG      18,24

 Data format lines for data file: satellite_data.txt
    L     ExamNo       1
    X  V  ELEV         4
    X  V  BAND2        6
    X  V  BAND3        7
    X     BAND4        8
    X     BAND5        9
    X     GRN          12
    X  V  NDVI         14
```

Figure 4—Input data preparation part of the report file.



```
===============================================================================
 Target data report
 ===============================================================================

 Number of target observations:    399


 Descriptive statistics for the X-variables
 -------------------------------------------------------------------------------

     Label                 Mean        Std Dev     Min Value    Max Value   Num Zeros
   1 ELEV             1529.3584       133.1190     1366.0000    2014.0000       0
   2 BAND2              30.5960         4.8372       20.6000      43.6000       0
   3 BAND3              36.5426         8.7496       18.2000      61.6000       0
   4 BAND4              54.4729         6.6594       38.1000      82.3000       0
   5 BAND5              85.3456        22.0257       28.1000     159.4000       0
   6 GRN                -9.8120         7.1754      -29.6100      13.9900       1
   7 NDVI                 .2055          .0926         .0050        .4822       0
```

Figure 5—Target data report. A similar report is produced for the reference data.

```
==============================================================================
Correlations in the Reference data
==============================================================================

Correlations between Y and X-variables
------------------------------------------------------------------------------

               ELEV         BAND2         BAND3         BAND4         BAND5          GRN          NDVI
COV_MED      .32102       -.63056       -.62762       -.43637       -.60393       .45860        .49444
COV_LRG      .33550       -.45466       -.45582       -.23436       -.48866       .40561        .44986
Tot_TPA      .01826       -.10430       -.11902        .02737       -.05647       .16210        .14342
Tot_BA       .39637       -.73297       -.73825       -.43171       -.72932       .60327        .64402
TopHt        .30856       -.65944       -.65090       -.47951       -.62917       .45786        .49219
Tot_Vol      .40693       -.72588       -.72785       -.44697       -.73116       .58084        .63625
BA_LP       -.07464       -.22165       -.21202       -.31052       -.14111       .01884        .00816
BA_IINE      .31009       -.44539       -.44694       -.21116       -.48617       .41400        .45301
BA_FIR       .37305       -.30585       -.32112       -.03159       -.34802       .38429        .40889
BA_IOLE      .24717       -.51967       -.51211       -.42244       -.46564       .32063        .34199
BA_SAW       .34256       -.52564       -.53073       -.26398       -.57096       .47624        .53424
TPA_SML      .01041       -.09107       -.10571        .03547       -.04349       .15148        .13191
TPA_LRG      .29857       -.51789       -.52403       -.25905       -.56377       .47096        .53050
MSC.PP       .04411       -.11719       -.11967       -.01367       -.14833       .15158        .14486
MSC.LP      -.24360        .26108        .27121        .00908        .31472      -.34911       -.35274
MSC.WF       .27483       -.27877       -.29636       -.06999       -.32759       .31637        .36013
MSC.RF       .24855       -.10000       -.10155        .09112       -.11162       .21202        .18964
```

Figure 6—Simple correlations between Y- and X-variables. Similar reports are available among the Y variables and among the X-variables. This example illustrates the correlations for the reference observations; tables are also output for the X-variables in target observation set.

line 58, and the data format lines end at line 74. The data for this file are

Canonical correlation report

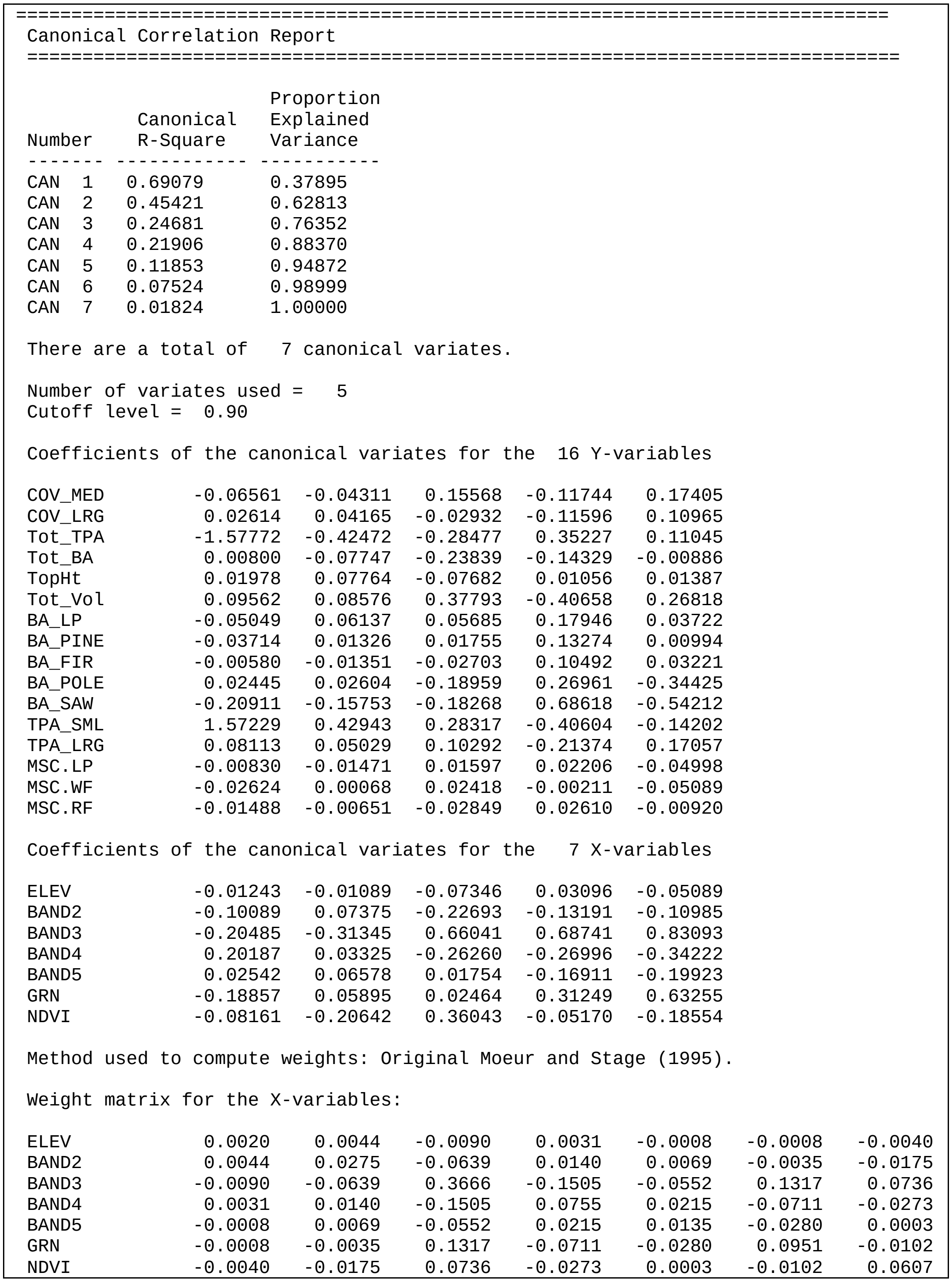
Figure 7 shows the results of the canonical correlation phase of the MSN program. For each canonical vector, the squared canonical correlation and the proportion of total explained variance are reported. The total number of canonical vectors computed from the data is reported next, followed by the number of vectors selected for use in computing the distances. The PROPVAR or NVECTORS commands control the number of vectors used in the distance calculations. There are seven canonical vectors in the example, of which five are used (because it takes five to account for 90 percent of the variation explained by the canonical correlation as set on line 7 of figure 1).

When the PRTWGHTS command is used, the coefficients in the weight matrix for each variable are printed, first for the Y-variables and then for the X-variables. Variables are in rows and canonical vectors are in columns in the tables. The entries in these tables can be interpreted as loadings that indicate the relative importance of variables in the MSN weighting function.

The method used to compute the weighting function used in the distance calculation is reported next, followed by a listing of the weight matrix.

Validation report

The beginning of the validation report is illustrated in figure 8. It summarizes the results of the MSN run, including validation statistics for selected variables and overall MSN distance results. A summary of the number of observations is followed by a list of the variables picked for validation. The number of ties broken by random numbers indicates how frequently the tiebreaker logic is used.



```
=============================================================================
 Canonical Correlation Report
 =============================================================================

                      Proportion
          Canonical   Explained
 Number    R-Square   Variance
 ------- ------------ -----------
 CAN  1   0.69079      0.37895
 CAN  2   0.45421      0.62813
 CAN  3   0.24681      0.76352
 CAN  4   0.21906      0.88370
 CAN  5   0.11853      0.94872
 CAN  6   0.07524      0.98999
 CAN  7   0.01824      1.00000

 There are a total of   7 canonical variates.

 Number of variates used =   5
 Cutoff level =  0.90

 Coefficients of the canonical variates for the  16 Y-variables

 COV_MED        -0.06561  -0.04311   0.15568  -0.11744   0.17405
 COV_LRG         0.02614   0.04165  -0.02932  -0.11596   0.10965
 Tot_TPA        -1.57772  -0.42472  -0.28477   0.35227   0.11045
 Tot_BA          0.00800  -0.07747  -0.23839  -0.14329  -0.00886
 TopHt           0.01978   0.07764  -0.07682   0.01056   0.01387
 Tot_Vol         0.09562   0.08576   0.37793  -0.40658   0.26818
 BA_LP          -0.05049   0.06137   0.05685   0.17946   0.03722
 BA_PINE        -0.03714   0.01326   0.01755   0.13274   0.00994
 BA_FIR         -0.00580  -0.01351  -0.02703   0.10492   0.03221
 BA_POLE         0.02445   0.02604  -0.18959   0.26961  -0.34425
 BA_SAW         -0.20911  -0.15753  -0.18268   0.68618  -0.54212
 TPA_SML         1.57229   0.42943   0.28317  -0.40604  -0.14202
 TPA_LRG         0.08113   0.05029   0.10292  -0.21374   0.17057
 MSC.LP         -0.00830  -0.01471   0.01597   0.02206  -0.04998
 MSC.WF         -0.02624   0.00068   0.02418  -0.00211  -0.05089
 MSC.RF         -0.01488  -0.00651  -0.02849   0.02610  -0.00920

 Coefficients of the canonical variates for the   7 X-variables

 ELEV           -0.01243  -0.01089  -0.07346   0.03096  -0.05089
 BAND2          -0.10089   0.07375  -0.22693  -0.13191  -0.10985
 BAND3          -0.20485  -0.31345   0.66041   0.68741   0.83093
 BAND4           0.20187   0.03325  -0.26260  -0.26996  -0.34222
 BAND5           0.02542   0.06578   0.01754  -0.16911  -0.19923
 GRN            -0.18857   0.05895   0.02464   0.31249   0.63255
 NDVI           -0.08161  -0.20642   0.36043  -0.05170  -0.18554

 Method used to compute weights: Original Moeur and Stage (1995).

 Weight matrix for the X-variables:

 ELEV            0.0020    0.0044   -0.0090    0.0031   -0.0008   -0.0008   -0.0040
 BAND2           0.0044    0.0275   -0.0639    0.0140    0.0069   -0.0035   -0.0175
 BAND3          -0.0090   -0.0639    0.3666   -0.1505   -0.0552    0.1317    0.0736
 BAND4           0.0031    0.0140   -0.1505    0.0755    0.0215   -0.0711   -0.0273
 BAND5          -0.0008    0.0069   -0.0552    0.0215    0.0135   -0.0280    0.0003
 GRN            -0.0008   -0.0035    0.1317   -0.0711   -0.0280    0.0951   -0.0102
 NDVI           -0.0040   -0.0175    0.0736   -0.0273    0.0003   -0.0102    0.0607
```

Figure 7—The canonical correlation report includes the canonical coefficients for the Y- and X-variables. Note that the number of variates used is limited. You can specify the cut off proportion or the number of variates using the PROPVAR or NVECTORS commands.

```
===============================================================================
Validation Report
===============================================================================

Number of reference obsrevations                  =    197
Number of target observations                     =    399

**Number of ties broken with random numbers       =      0
  (**Random number seed =           55329.)

25 variables selected for validation:

 COV_LRG       Tot_BA        TopHt         Tot_Vol       Tot_Mort      BA_LP         BA_IINE       BA_FIR
 BA_IOLE       BA_SAW        TPA_SML       TPA_LRG       ELEV          BAND2         BAND3         NDVI
 MSC.PP        PAG.PPD       MSC.LP        PAG.LPD       PAG.LPW       PAG.MCD       MSC.WF        PAG.PPW
 MSC.RF
```

Figure 8—The beginning of the validation report.

Validation statistics are computed by comparing the observed value for a variable to an imputed value. For reference observations, the imputed values are from the second-most similar neighbors because the most similar neighbor is the observation itself. If it were the basis of comparison, the differences would all be zero and the classifications would be perfect.

Figure 9 shows the validation statistics for continuous variables among the reference observations. The number of observations is listed in the table title, the mean and standard deviation of the observed and imputed values are followed by the mean and standard deviation of the *standardized difference values*. Standardized differences are computed as the absolute difference between observed and imputed values, divided by the truncated range of the variable in the data. The standardization permits the differences to be directly compared between variables (the section on “Technical Details” shows the formulas for this calculation). The mean difference between the observed and imputed values is followed by a t-Ratio computed for a paired comparison under the null hypothesis that the mean of the residuals is zero. In the case of MSN, we hope that the mean residual is essentially zero and that we will not be forced to reject the null hypothesis. A large t-Ratio and a

```
-------------------------------------------------------------------------------------------------------------------------------------
                               MSN Validation Results for    197 Reference Observations
                               OBS=Observed,  IMPU=MSN Imputed,  RESID=OBS-IMPU
-------------------------------------------------------------------------------------------------------------------------------------
                        -------- OBS --------  --------- IMPU --------  ---- STD DIFF ----  --------------- RESIDUALS ---------------
LABEL       TYP    N        MEAN      STD.DEV         MEAN     STD.DEV      MEAN   STD.DEV       MEAN t-Ratio   P>|t|       RMSE    R
COV_LRG      Y   197       4.164        5.479        3.801       4.848     0.017    0.284      0.363   0.862  0.390      5.920  0.350
Tot_BA       Y   197      88.467       53.124       87.406      45.775     0.005    0.208      1.061   0.309  0.758     48.250  0.533
TopHt        Y   197      58.959       17.116       59.492      15.748     0.007    0.233     -0.533  -0.446  0.656     16.775  0.482
Tot_Vol      Y   197    2068.421     1443.447     2018.213    1265.839     0.008    0.207     50.208   0.555  0.579   1270.163  0.568
Tot_Mort         197      11.924        9.091       12.315       8.662     0.011    0.260     -0.391  -0.603  0.547      9.101  0.476
BA_LP        Y   197      41.594       42.193       41.462      38.988     0.001    0.264      0.132   0.037  0.971     50.626  0.224
BA_PINE      Y   197      37.284       43.924       39.142      41.789     0.011    0.296     -1.858  -0.522  0.602     50.008  0.321
BA_FIR       Y   197       6.477       21.848        5.061      17.662     0.010    0.176      1.416   0.813  0.417     24.488  0.248
BA_POLE      Y   197      46.538       35.880       47.350      32.103     0.005    0.224     -0.812  -0.304  0.761     37.446  0.398
BA_SAW       Y   197      27.284       32.644       26.325      29.682     0.007    0.249      0.959   0.423  0.673     31.870  0.481
TPA_SML      Y   197    2129.975     1706.194     2181.391    1606.926     0.006    0.256    -51.416  -0.350  0.726   2060.745  0.228
TPA_LRG      Y   197       9.797       11.903        9.848      11.244     0.001    0.249     -0.051  -0.060  0.953     11.954  0.468
ELEV         X   197    1503.218      126.859     1497.787     122.353     0.009    0.088      5.431   1.400  0.163     54.714  0.905
BAND2        X   197      30.740        4.691       30.676       4.516     0.003    0.050      0.064   0.802  0.424      1.130  0.971
BAND3        X   197      36.714        8.335       36.647       7.963     0.002    0.049      0.066   0.476  0.635      1.962  0.972
NDVI         X   197       0.204        0.086        0.203       0.080     0.004    0.061      0.001   0.855  0.394      0.023  0.964
```

Figure 9—Validation statistics for continuous variables among reference observations.

low probability of getting a larger one indicate that the imputed values are significantly different than the observed. The column headed RMSE is the *root mean square error;* see the “Technical Details” section for the formula. The last column, labeled R, is the *correlation coefficient* between the imputed and observed values. Values near 1 suggest the imputations are of good quality, and those approaching zero and those less than zero suggest the imputations are of little utility. Note that strict interpretation of these statistics is limited by the conservative nature of comparisons to second-nearest neighbor.

For categorical variables, the program produces contingency tables that report the frequency of classification and related statistics (fig. 10). The rows in the table show the number of observations at each imputed class, the percent of the row total, the expected count under the assumption that the classification is done completely at random, and the cell Chi Square statistic. The bordering column and last row show the totals for rows and columns, respectively with the number and percent correct classification. A total Chi Square and the probability of a larger value are reported when the degrees of freedom permits. Chi Square statistics that are near zero indicate little predictive power, implying that the classification is no different from imputation using a random process. Finally, the Kappa statistic, labeled Khat,

```
Categorical variable group: PAG.

                         Imputed
 Observed

             PAG.PPD       PAG.LPD       PAG.LPW       PAG.MCD       PAG.PPW        Total          %CC
 PAG.PPD          18            25             0             3             2           48
      Row%     37.50         52.08          0.00          6.25          4.17        24.37
  Expected     13.64         21.93          0.24          9.75          2.44
    Chi-Sq      1.39          0.43          0.24          4.67          0.08

 PAG.LPD          30            52             1             7             5           95
      Row%     31.58         54.74          1.05          7.37          5.26        48.22
  Expected     27.01         43.40          0.48         19.29          4.82
    Chi-Sq      0.33          1.70          0.56          7.83          0.01

 PAG.LPW           2             1             0             0             0            3
      Row%     66.67         33.33          0.00          0.00          0.00         1.52
  Expected      0.85          1.37          0.02          0.61          0.15
    Chi-Sq      1.54          0.10          0.02          0.61          0.15

 PAG.MCD           3             7             0            28             3           41
      Row%      7.32         17.07          0.00         68.29          7.32        20.81
  Expected     11.65         18.73          0.21          8.32          2.08
    Chi-Sq      6.43          7.35          0.21         46.50          0.41

 PAG.PPW           3             5             0             2             0           10
      Row%     30.00         50.00          0.00         20.00          0.00         5.08
  Expected      2.84          4.57          0.05          2.03          0.51
    Chi-Sq      0.01          0.04          0.05          0.00          0.51

 Total            56            90             1            40            10          197           98
 Row%          28.43         45.69          0.51         20.30          5.08       100.00        49.75

 Total Chi-Sq:     81.2  DF:   16  Prob>Chi-Sq:  <0.0001

 Kappa statistics: Khat:0.24489  Std Err:0.05269
 Interpretation of Khat: <=0 Poor, 0-.2 Slight, .2-.4 Fair, .4-.6 Moderate, .6-.8 Substantial, .8-1 Almost perfect.
```

Figure 10—The two-way contingency table for a categorical variable in the reference observation set. One-way tables are used to show the predictions for categorical Y-variables in the target observation set as there are no observations for these variables among the target observations.

and its standard error are reported. Like Chi Square, this statistic measures how well the observed and imputed values match. A guide to the interpretation of these values is also output.

For target observations, the imputed values are compared to their most similar neighbors, rather than the second neighbor, as done for the reference observations. Figure 11 shows the validation report for continuous variables among target observations. As observed values do not exist for Y-variables, many of the columns are left blank. Observed values do exist for the X-variables, permitting the calculation of all the statistics. Note that one-way contingency tables are output for categorical Y-variables and two-way tables are produced categorical X-variables. Examples of these tables are not illustrated.

Distance report

Figure 12 shows an example of the distribution of MSN distances for the reference observations, where the second most similar neighbor is used to represent the observation. Similar tables are produced for the target observations, and overall observations are combined. For each set, the mean, standard deviation, minimum, and maximum values of MSN distances are given, followed by a table of distance values summarized within 10 percent classes. Columns in the percentile table are the upper and lower boundaries of the class, the frequency of observations in the class, relative frequency (frequency/N), cumulative frequency (sum of frequency), and relative cumulative frequency (sum of relative frequency). Large MSN distances may signal serious mismatches between observed values and most similar neighbor imputations. Figure 13 lists some of the largest distances between reference and target observations. The first column is the identification of a target, called the *ForID*. The second is the identification of the imputed reference observation, called the *UseID*, and the third is the distance between them. The number of items listed is controlled using the MOSTUSED command.

Figure 14 is a summary of the most-used reference observations. References that are used to represent many targets can indicate that your

```
------------------------------------------------------------------------------------------------------------------------------------
                               MSN Validation Results for    399 Target Observations
                               OBS=Observed,  IMPU=MSN Imputed,  RESID=OBS-IMPU
-------------------------------------------------------------------------------------------------------------------------------------
                        -------- OBS --------  --------- IMPU -------- ---- STD DIFF ----   --------------- RESIDUALS ---------------
LABEL       TYP    N        MEAN      STD.DEV         MEAN      STD.DEV     MEAN    STD.DEV     MEAN  t-Ratio   P>|t|      RMSE      R
COV_LRG      Y   399                                 4.960       5.853
Tot_BA       Y   399                                91.083      53.260
TopHt        Y   399                                60.271      17.149
Tot_Vol      Y   399                              2195.807    1497.571
Tot_Mort         399                                11.238       8.775
BA_LP        Y   399                                38.110      38.525
BA_PINE      Y   399                                42.905      48.094
BA_FIR       Y   399                                 6.744      20.738
BA_POLE      Y   399                                46.331      35.105
BA_SAW       Y   399                                31.133      33.565
TPA_SML      Y   399                              2089.882    1718.090
TPA_LRG      Y   399                                10.749      11.974
ELEV         X   399    1529.358      133.119     1524.095     140.036     0.009    0.093      5.263   2.017  0.044     52.377  0.928
BAND2        X   399      30.596        4.837       30.595       4.741     0.000    0.057      0.001   0.021  0.983      1.183  0.970
BAND3        X   399      36.543        8.750       36.487       8.427     0.001    0.054      0.055   0.550  0.582      2.011  0.973
NDVI         X   399       0.205        0.093        0.205       0.088     0.001    0.057      0.000   0.241  0.810      0.023  0.968
```

Figure 11—Validation statistics for continuous variables among target observations. As the value of Y-variables is not know for target observations, only mean and standard deviation of the imputed values can be computed. For the X-variables, the observed values are known making possible full statistics.

```
===============================================================================
 Distance Report
 ===============================================================================

 MSN Distance distributional statistics for Reference Observations
--------------------------------------------------------------------------------

     MEAN      STD.DEV.    MINIMUM     MAXIMUM
   ---------  ---------   ---------   ---------
    0.026184   0.013865   0.006307     0.091998

      10TH PCTILE
         CLASS             CLASS   REL.    CUM.    REL.CUM.
      BOUNDARIES           FREQ.   FREQ.   FREQ.   FREQ.
 -----------------------   -----   -----   -----   -----
     0.006307    0.014876      35    0.18      35    0.18
     0.014876    0.023445      61    0.31      96    0.49
     0.023445    0.032014      52    0.26     148    0.75
     0.032014    0.040584      25    0.13     173    0.88
     0.040584    0.049153      13    0.07     186    0.94
     0.049153    0.057722       5    0.03     191    0.97
     0.057722    0.066291       3    0.02     194    0.98
     0.066291    0.074860       0    0.00     194    0.98
     0.074860    0.083429       1    0.01     195    0.99
     0.083430    0.091999       2    0.01     197    1.00
```

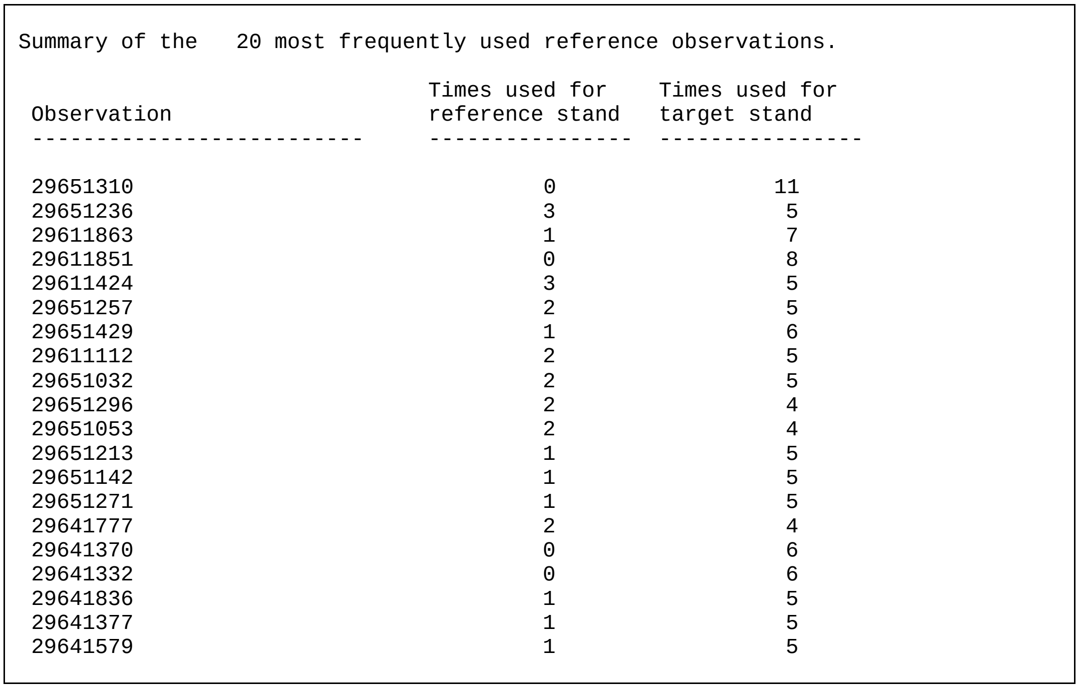
Figure 12—The distance report shows the distribution of distances among observations like this one that highlights the differences among reference observations, where the second most similar neighbor is imputed. Two similar tables show distributions of distances among target observations (using the most similar neighbor) and all observations.

```
 Summary of the   20 largest distances between reference and target observations.

 ForID                      UseID                      Distance
 -------------------------  -------------------------  ----------------

 29641543                   29651005                          0.08937
 29651099                   29651213                          0.07585
 29641060                   29641427                          0.07510
 29641321                   29641332                          0.07169
 29641307                   29651271                          0.06890
 29611551                   29651142                          0.06653
 29641966                   29651090                          0.06572
 29611521                   29641172                          0.06477
 29651119                   29651116                          0.06466
 29651254                   29651252                          0.06437
 29641288                   29641149                          0.06245
 29641168                   29641056                          0.06224
 29611467                   29611468                          0.06101
 29651127                   29651090                          0.05828
 29610309                   29651090                          0.05792
 29641268                   29641264                          0.05770
 29651117                   29651252                          0.05765
 29611500                   29641427                          0.05397
 29641110                   29641149                          0.05344
 29651438                   29641517                          0.05253
```

Figure 13—The matches with the highest distances between reference and target observations. Relatively high distances indicate relatively low similarity.



```
Summary of the   20 most frequently used reference observations.

                                Times used for    Times used for
 Observation                    reference stand   target stand
 --------------------------     ---------------  ----------------

 29651310                                0              11
 29651236                                3               5
 29611863                                1               7
 29611851                                0               8
 29611424                                3               5
 29651257                                2               5
 29651429                                1               6
 29611112                                2               5
 29651032                                2               5
 29651296                                2               4
 29651053                                2               4
 29651213                                1               5
 29651142                                1               5
 29651271                                1               5
 29641777                                2               4
 29641370                                0               6
 29641332                                0               6
 29641836                                1               5
 29641377                                1               5
 29641579                                1               5
```

Figure 14—The most frequently used reference observations.

reference data may not include enough observations for the condition being represented by an often-used reference. The number of entries is the same as the number in the largest distance report as it is controlled with the MOSTUSED command.

Interpreting the validation statistics

Recall that most of the validation statistics are based on comparing the observed values to imputed values where the imputations are the second most similar neighbor. Here are two examples where these validations statistics will lead you to incorrect conclusions about your analysis.

First, consider a case where several target observations form a subset of observations that are similar to each other and very different from all the others. Furthermore, there is exactly one reference observation that is like those in the subset and unlike all the other reference observations. In a reasonable application of MSN, this reference observation will be used to represent all the target observations in the subset, providing a perfectly good imputation. But the validation statistics for this reference observation will be computed by comparing it to its second most similar neighbor. This unique observation will be matched with another, very unlike observation, thereby inflating the error statistics and showing a lack of predictive power. Yet in practice, because the reference observation would be used to represent its self plus its similar target observations, the overall results would be satisfactory.

Second, consider the same data except that the one unique subset of observations is not represented by a reference observation. The validation statistics for the reference data will not show a lack of predictive power because they would not include the observation from the unique condition.

The error statistics would be blind to this situation showing that there is predictive power where none exists. You might detect this condition by reviewing the validation statistics for the target observations. Recall that these statistics do not report the performance of the imputed Y-variables, only the X-variables, and the lack of predictive power for the subset in question could easily be missed. You might detect this condition by reviewing some of the tables described below that characterize the distribution of distance calculations.

We suggest that you independently assess the results, perhaps by mapping some of the imputed variables. You should also independently ensure that you have representative reference observations that cover the full range of conditions in the target set (see the section on “Sampling,” below).

You could conclude that the predictive power illustrated in the user’s guide example is not very strong. However, the merits of the imputations must be weighed with respect to the application and the alternatives to collecting more data rather than using imputed data.

The validation statistics can generally be used to compare one MSN model formulation to another. Use the model and other parameters of MSN that provide the best imputations of the variables important to you.

Randomize your data

The RANDOMIZE command randomizes the observations of X-variables with respect to the observations of Y-variables for the reference data. It leaves intact the relationships within the Y’s and within the X’s yet destroys the relationship between the Y’s and X’s. The idea behind this command is to provide a way to ensure that relationships reported in the validation statistics are meaningful. You should see much better results when you leave this command out of the run. Note that the simple statistics for each variable and the correlation matrices among Y-variables and among the X-variables will not change when RANDOMIZE command is used, as compared to not using this command. The rest of the outputs should change, beginning with the correlations between Y- and X-variables.

In the example presented in this guide, adding the RANDOMIZE command reduced the largest canonical R-Squared from 0.691 to 0.164. The percent correctly classified for plant association group (*PAG*) dropped from 50 percent (fig. 10) to 27 percent, the corresponding Chi Square statistics were significant in both cases, and the Kappa statistic dropped from 0.25 (fair) to –0.09 (poor or none). However, the classification of tree species (*MSC*) only dropped from 58 to 53 percent, had significant Chi Square statistics in both cases, and the Kappa statistic dropped from 0.158 to 0.082. All of the continuous Y-variables in this problem suffered much lower R values when regressing observed on imputed values and higher RMSE values. For total cubic volume, the RMSE doubled. We conclude that MSN is providing real predictive power in making its selections. Yet despite significant Chi Square statistics, the classification of species composition is not better than would be made by chance alone.

Run the example distributed with the MSN program to see all the results. Add the RAMDOMIZE command and use the RANNSEED command to get different randomizations.

For-Use File

The FORUSE command is used to enter this output file name and to signal that the file be created. The file lists the MSN assignments and the distance between them. The file follows the same format as shown in the body of the data in figure 13. It is designed to be used to load databases and other programs with the imputations. When you leave out the NOTARGS and NOREFS commands, this file contains reference plus the target observations so that the file will be complete with respect to all the observations in the problem.

Observed-Imputed File

The OBSIMPU command signals that the Observed-Imputed file be created and specifies its name. The file contains observed and imputed values for variables flagged for validation on the data format lines. Figure 15 shows the partial file for the example analysis. The file is comma-separated and the first record is a header record with variable names.

The order of columns in the file is: ForID, UseID, ObsType, K, and Distance, followed by observed (*variable*_O), imputed (*variable*_I), for categorical variables, plus standardized difference (*variable*_D) continuous variables. ObsType is the observation type; R is output for reference observations and T for target observations. Categorical variables that have dummy variables created by MSN are output and those where you have created the dummy variables are not (this omission may be fixed in future versions).

The value of K is the ordinal value of the number of most similar neighbors, where 1 is the most similar neighbor, 2 is the second most similar neighbor, …, and K is the Kth most similar neighbor, output in order of increasing distances.

There are two styles of this file. The first style is output when the KMSN command is not used or when a zero is entered for the number of nearby neighbors. The second style is output when the KMSN command is used and the number of nearby neighbors is 1 or more.

The contents of the file are different depending on the style. In the first style, illustrated in figure 14, the file contains one entry for each observation. For reference observations, the entry is created where K=2, the second neighbor. For target observations, the entry is created for K=1, the

```
ForID,UseID,ObsType,K,Distance,MSC_O,MSC_I,PAG_O,PAG_I,COV_LRG_O,COV_LRG_I,COV_LRG_D,Tot_BA_O,Tot_BA_I,Tot_BA_D
29610287,29651043,R,2,0.03117732,PP,PP,PPD,MCD,17.2000,12.6000,0.2212,147,143,0.0172
29610599,29611725,R,2,0.01505923,PP,LP,PPD,LPD,4.5000,0.8000,0.1779,96,114,-0.0776
[Observations omitted]

29651461,29641674,T,1,0.02046493, ,PP, ,PPD, ,11.1000, , ,104
29651466,29641849,T,1,0.00968060, ,LP, ,LPD, ,0, , ,76
```

Figure 15—Partial listing of the Observed-Imputed file. Most observations and columns are omitted from this example. For the reference observations, the observed, predicted, and standardized difference values are known for each continuous variable. For the target observations only the observed values are known and the others are output as missing values.

first neighbor. This style provides information best suited for additional validation.

When the second style is used, the file contains as many entries for each observation as were requested using the KMSN command. When K is 1, the most similar neighbor is output, when K is 2, the second neighbor is output, and so on.

The second style is designed for two purposes. One is to provide a simple way to recover the best imputed value for all observations. Another purpose is to support additional processing of the file when you want to further analyze the set of K neighbors, perhaps with the goal of picking a second or third neighbor for reasons beyond the purview of the MSN program. For example, you might use an indicator of data reliability to override strict proximity, picking a reference observation that is very close, but not the closest to a target, but of better quality. Quality in this case might be an attribute of the data indicating that it was measured to higher standards. A third purpose is to support the calculation of K-based means for imputation. The MSN program does not compute those values, but the necessary data are available when this output style is used.

Factors that Influence MSN Imputations

In this section we discuss factors that influence how well the MSN methods work. The quality of MSN imputations is controlled by the extent to which the sample of reference observations covers the range of variation of the target observations, and how well the Y-variables are related to X-variables. In turn, the choice of variables and the use for transformations influence how well the variables are related. You can also influence the procedure by setting parameters in the MSN distance calculations. Those settings are discussed in the “Technical Details” section.

Sampling

We have pointed out that to use the MSN system, you start with a list of sample units, measure some attributes on all the sample units (the X-variables), and measure additional attributes (the Y-variables) on *some* of the sample of the units. Here, we begin to address the issue of how you chose the units on which the additional attributes are measured and why it matters.

When estimating a characteristic of a population, say total volume on a National Forest, for example, the way you sample dictates the statistical properties of the estimate. If you are interested in the statistical properties of averages and totals when MSN is used, we suggest you consult a statistician for advice, consult the statistical literature, or both. For example, Shao and Sitter (1996) have presented methods for computing variance estimates for imputed survey data.

Imputation does not *extrapolate* nor does it *interpolate* the way regression-based estimates do. The simple graph in figure 16 illustrates the point. Let’s say that you have a variable, *x* that is perfectly related to another variable *y* through a simple linear function, *f(x)*. In this case, suppose that the range of values of *x* is actually as shown in the figure, from zero to 100. Any sampling scheme that provides for two or more samples will be sufficient to characterize the regression line in this utopia. Furthermore, the use of *f(x)*

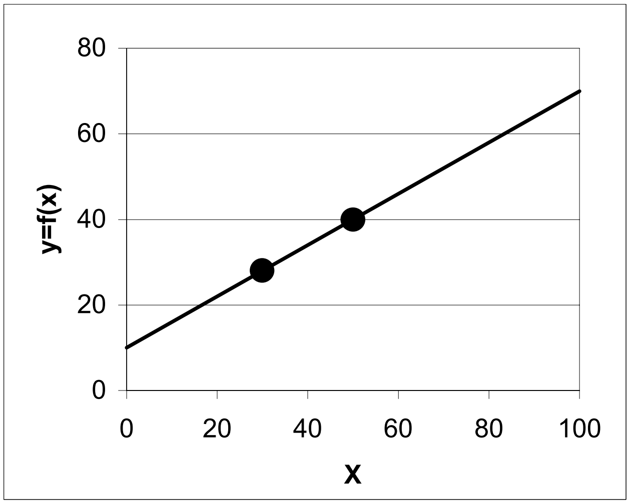


Figure 16—A regression line, *f(x),* allows for extrapolation beyond the two points and interpolation between them. Imputation limits the values of *y* to those observed at the data points. Your sampling strategy must insure that the entire range of values for Y- and X-variables is represented and at appropriate intervals along their axes.

will provide perfect estimates of *y* for every *x* because the function will interpolate between the measured values, represented by the points on the line, and it will extrapolate beyond the observed data.

If you use imputation rather than regression, there are only two possible values for *y*, the one that corresponds to the first observation and the one corresponding to the second. The estimates of *y* will have some degree of error for all values of *x* except those that happen to exactly match either of the two observations. The only solution to this situation is to observe enough values along the *x*-axes so that the data used for the imputation covers the range of values and that there are no large gaps along the axes. Deficiencies in sample coverage can be located by studying the tables of large distances between target and reference observations and by comparing the imputations of the X-variable to their true values. Resolving the sampling deficiencies requires more data in the underrepresented parts of the population.

Choice of Variables

Some important points about the choice of variables are best made using an example. Our experience with MSN is that it is generally easier to predict stand volume than it is to predict species composition. If measures of both are included as Y-variables, and predictors of both are included as X-variables, then the volume measurement often gets the most weight, and the imputations favor matching references to targets where volume is similar and not when species composition is similar. But what if you would rather have MSN do its best matching the species composition and accept weaker matches on volume? The only way this version of MSN can reduce the strength of the volume prediction with respect to species composition is to leave volume out as a Y-variable or leave X-variables out that are strong predictors of volume. If the same X-variables are well related to volume as are related to species composition, your only choice is to leave out volume as a Y-variable. Note that the species composition imputations may not improve when you leave out volume. That will be the case when species composition is simply not well related to the X-variables or when the relationships between the X's and Y's that measure species are independent of, and weaker than, those that predict volume.

MSN is based on canonical correlation analysis that assumes linear relationships between variables. You need to provide variables that are related; see the section on “Applications” for some ideas you might find useful in your application.

Choice of Transformations

Many biological relations are essentially nonlinear. But with some ingenuity, transformations of the variables are available that can render the relations linear. Consider a situation in which species composition is important, and in which elevation, slope, and aspect are expected to be useful predictors. A common expectation in ecological analysis is that the probability of a species being present plots as a humped-back curve over elevation. At low elevations, a less-than-maximum value might be attributed to the species being limited to north-facing slopes. At high elevations, the reverse—the species is squeezed out by better adapted species on north slopes. In between, it is ubiquitous. For this example, a useful transformation of the species presence (p) is logarithm ((1-p)/p). The corresponding transformation of elevation is to add elevation squared as an X-variable. Aspect and slope effects can be represented by the trigonometric transformations described by Stage (1976). The next section, titled “Applications,” lists examples where other transformations were used.

Applications

Moeur and Stage (1995) published the MSN procedure and documented a test application. The X-variables included several derived using photo interpretation techniques including crown canopy coverage and photographic texture. Others represented topographic information, including slope and aspect. The Y-variables were derived from forestry traditional inventory measures such as board feet and density. Table 3 lists several Y- and X-variables used in MSN applications with citations to their use.

Lower RMSE using MSN imputations as compared to averages from stratified sampling and the Swedish inventory estimates were reported for the Moeur and Stage (1995) example. MSN produced higher RMSE compared to individual correlation estimates fit to individual Y-variables because it deliberately includes the variance about regression in the imputed values. MSN maintained the variance of the Y-variables and covariances among them as intended. The ability to maintain the covariances among Y-variables is one of the strong points of MSN. The imputations were done at the stand level.

Moeur and others (1995) tested the MSN application at a new location, the Clearwater National Forest in northern Idaho, using much larger sample sizes. They included Landsat TM data as X-variables. In this case, as in some others (for another example, see LeMay and Temesgen 2001), the entire tree list from the sample location was imputed.

Moeur (2000) reported an application on the Deschutes National Forest in central Oregon. Those data are used as the example in this guide. Landsat TM data were useful in imputing percent canopy cover and basal area when used in addition to geo- and topographical data. Cover type was correctly

Table 3—Example variables used in MSN analyses.

Variable	Citation
Examples of *Y-variables*	
Board feet per acre	Moeur and Stage (1995)
Cubic volume per acre, total and by species	Moeur and Stage (1995)
Basal area total and by species	Moeur and Stage (1995)
Trees per acre total and by species	Moeur (2000)
Average yield	Moeur and Stage (1995)
Percent stocking	Moeur and Stage (1995)
Stand density index	Moeur (2000)
Crown competition factor	Moeur (2000)
Site index	Moeur and Stage (1995)
Cover type	Moeur (2000)
Potential vegetation class	McDonald and others (in press)
Examples of *X-variables*	
Slope times the sine of aspect and slope times the cosine of aspect.	Stage (1976), Moeur and Stage (1995)
Average crown length	Moeur and Stage (1995)
Average crown area	Moeur and Stage (1995)
Percent crown canopy cover	Moeur and Stage (1995)
Elevation and its natural log	Moeur and Stage (1995)
Forest type	Moeur and Stage (1995)
Precipitation and its natural log	Moeur and Stage (1995)
Photographic texture	Moeur and Stage (1995)
Slope percent	Moeur and Stage (1995)
Topographic position code	Moeur and Stage (1995)
Average LandSat TM digital channels for bands 1-7 and their standard deviation as texture.	Moeur and others (1995), Moeur (2000)
Normalized difference vegetation index	Reed and others (1994), Moeur (2000)
3 Tasselled cap indices	Crist and Cicone (1984), Moeur (2000)
UTM easting and northing coordinates	Moeur (2000)
Slope curvature	McDonald and others (in press)
Tangent (planform) curvature	McDonald and others (in press)
Profile curvature	McDonald and others (in press)
Solar insolation	Fu and Rich (1999), McDonald and others (in press)
Light duration	Fu and Rich (1999), McDonald and others (in press)
Topographic wetness index	Wilson and Gallant (2000, p 108), McDonald and others (in press)
Variogram models	Moeur and Riemann Hershey (1999); Muinonen and others (2001)
Biogeoclimatic ecosystem classes	LeMay and Temesgen (2001)

classified in 76 percent of the cases. RMSE as a function of sampling intensity was reported showing how this information can be used to support sample intensity decisions.

Moeur and Hershey (1999) compared geostatistical simulation to MSN, and combined them, in a study of the Finger Lakes region of New York State. Their goal was to estimate forest species composition where it was not measured. The estimates were done over an entire Landsat scene at 7 by 7 pixel resolution (30m by 30m for each pixel). This was the first study done where the imputations were done for grid points rather than polygons representing stands. Like the studies reported above, the Y-variables represented a measure of current vegetation. Results suggested that the combined methods might result in a dramatic improvement in species modeling.

McDonald and others (in press) used MSN to impute potential vegetation class for 27,775 individual pixels (30m by 30m), based on 245 reference observations, in a study of the Priest River Experimental Forest in northern Idaho. This study is the first time where only one Y-variable was of interest, the first to impute data at the Landsat pixel level, and the first to impute potential rather than current vegetation. Some of the X-variables used are noteworthy, including planform, profile curvature, solar insolation, light duration, and wetness (table 3).

Muinonen and others (2001) used geostatistical methods and the method of *k*-MSN to estimate volume. They found that averaging three neighbors provided the lowest error. Sironen and others (2001) used *k*-MSN to estimate individual tree growth and compared it to traditional regression growth models. While these examples illustrate the use of MSN, it appears that neither was accomplished using this MSN program.

Technical Details

Computing the Canonical Correlations

Canonical correlation analysis (CCA) fits linear transformations of Y and X, such that the correlation between the two sets is maximized (see Gittins 1985 for an excellent discussion of CCA). The steps outlined by Gittins (1985, page 33) are used to compute the coefficients of the analysis. This procedure starts with a QR-decomposition of normalized X-variables and Y-variables. The numerical aspects of this approach are done using LAPACK subroutines (Anderson and others 1999). The answers we get using this approach match those from the R statistical program (Venables and others 2002). Prior to computing the CCA, extra columns of categorical variables are removed from the analysis. Recall that MSN creates a new dummy variable for each level of a categorical variable. This strategy works well when descriptive statistics and correlation tables are produced, but it introduces colinearity to the CCA. This colinearity is removed by deleting the dummy variable that corresponds to the first level of each categorical variable.

Most Similar Neighbor Selection

The most similar neighbor to target observation *i* is defined as the reference observation *j* (over all reference observations) that minimizes the weighted Euclidean distance on the set of X-variables (Moeur and Stage 1995).

$$\text{MSN}(i) = \text{reference observation } j \text{ with minimum } d^2_{ij} = (\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{j})\, \mathbf{W}\, (\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{j})', \text{ for all } j = 1, \ldots, n \tag{1}$$

where

MSN(i) is the most similar neighbor to the i^{th} target observation,

d^2_{ij} is the squared distance between the i^{th} target observation and the j^{th} observation, over n reference observations,

$\mathbf{X}_\mathrm{i}$ is the vector of normalized X-variables for the i^{th} target observation,

$\mathbf{X}_j$ is the vector of normalized X-variables for the j^{th} reference observation, and

$\mathbf{W}$ is a weight matrix.

The MSN program provides five ways for you to define the weight matrix using the DISTMETH command. The methods correspond to a numeric code entered on the command. The least likely to use, but simplest to understand is method 4, which sets the weight matrix, **W**, to the identity matrix. When that method is used, the distance formula (1) reduces to simple Euclidean distance on the normalized X-variables. In this simple case, it is fairly easy to conceptualize the MSN as the reference observation that is most similar in the combination of X-variables to the target observation. That is, it is the *closest* observation in the multivariate space of X-variables.

However, we can refine the MSN selection by using the data from the reference observations to quantify the relationships between Y- and X-variables, and replace **W** with a weighting matrix that maximizes the information content between the two variable sets. The resulting weight matrix is computed from the linear combination of the **X**'s, called the canonical vectors, $\mathbf{\Gamma}$, and the squared canonical correlations, $\mathbf{\Lambda}^2$. This weight matrix summarizes the best linear relationship between the sets of multivariate **Y** and multivariate **X** taken simultaneously, while incorporating the covariance between the elements of **X** and elements of **Y**.

Two formulas for **W** that use the results of CCA are offered. The first corresponds to the formula presented by Moeur and Stage (1995) and it is the default (used when DISTMETH is zero):

$$\mathbf{W} = \mathbf{\Gamma}\,\mathbf{\Lambda}^2\,\mathbf{\Gamma}' \qquad (2)$$

and the second (used when DISTMETH is 1) is an alternative recently proposed by Stage (in preparation):

$$\mathbf{W} = \mathbf{\Gamma\Lambda}(I - \mathbf{\Lambda}^2)^{-1}\,\mathbf{\Lambda\Gamma}'. \qquad (3)$$

Both formulas use the results of CCA. We have included formula 3 as an option to facilitate further study of its practical implications. Stage (in preparation) argues that it more closely represents some of the intended properties of MSN imputation compared to using formula 2. Using either formula provides exactly the same answers when only one set of canonical variates is used. Formula 3 provides more weight to the first canonical vectors than the subsequent vectors when more than one vector is used. Note that our experiences with MSN, and what current published literatures refer to as MSN, is based on formula 2.

There are other ways the program defines **W**. First, **W** can be the generalized inverse of the correlation matrix among the X-variables. This approach is used when DISTMETH is set to 2 and results in the distances being Mahalanobis distances on normalized X-variables. This approach ignores the canonical correlation results and may be the best approach when there is no relationship between X- and Y-variables. It is used automatically when there are fewer observations than variables. We use singular value decomposition programs from Dongarra and others (1979) as a step in finding the generalized inverse.

The next method (used when DISTMETH is set to 3) allows you to enter a weight matrix input on data lines that follow the command. You must enter a square, symmetric matrix. Each line of the matrix begins with the name of an X-variable. Following the name, you enter the coefficients of the matrix separated by blanks, commas, or tabs using the rules for free-form input.

Remember that you need to ensure that the columns of the matrix you enter correspond to the rows. After the last row, enter a line that contains only this string: “EndWeights”. The program matches the rows to the X-variables using the variable names you enter. Only the variables for which there are weights are used. The program assumes that you entered the matrix correctly. Lastly, when DISTMETH is set to 4, Euclidean distances on normalized X’s are computed.

The canonical vectors are illustrated at the bottom of figure 7. Because the numbers in these tables are normalized, they have comparable units, and the individual values in these tables can be interpreted as *loadings* that indicate the relative importance of each variable for each canonical vector in the MSN distance model.

In general, the more canonical vectors included in the MSN distance model, the more the most similar neighbor selections are guided toward the best all-around neighbor, taking into consideration the correlations between all the Y- and X- variables. On the other hand, restricting the MSN distance function to using only the first few canonical vectors will guide the neighbor selection toward the variables that are most predictable. You can select a cutoff for the number of canonical vectors to be used in the weighting matrix, using either the PROPVAR or NVECTORS command in the command file.

When two or more target units are exactly the same distance to the same reference unit, the program declares a tie. Ties occur when all the X-variables are categorical and are rare when at least one is continuous. When a tie is detected, the program picks one at random to be the most similar. In the case of KMSN, where several neighbors are output, the program adds a tiny random number to otherwise equal distances so that there are no numerical ties, just very close distances.

Validation Statistics

The output produced in the validation statistics is described in the preceding sections. An important attribute about the validation statistics is that they are mostly based on selecting the second most similar neighbor among the reference observations. This is analogous to so-called *leave one out*, or *Jackknife* technique. In full Jackknife, the entire model would be refit (n-1) times, holding out a different observation each time. Our MSN program does not refit the model.

The formulas used to compute some of the values are presented below.

Standardized difference values

Standardized difference values are reported in the Observed-Imputed file. The formula used to compute these values is: $D = \frac{\left|\left(V_{obs} - V_{impu}\right)\right|}{V_{.95} - V_{.05}}$, where $V_{.95}$ and $V_{.05}$ are the 95th- and 5th-percentile values of the variable. The same formula is used for Y- and X-variables. The mean and standard deviation for the standardized differences is computed in the traditional way.

Paired t-ratio

The mean residual for a variable is computed as $\bar{d} = \frac{\sum_i d_i}{n}$, where n is the number of observations and d_i is the difference between observed and imputed values. The standard deviation of differences is $s_d = \sqrt{\frac{\sum_i \left(d_i - \bar{d}\right)^2}{n-1}}$ and of the mean difference it is $s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$. The t-Ratio is $t = \frac{s_d}{s_{\bar{d}}}$ with n-1 degrees of freedom. The probability of a greater t-Ratio is computed using the routine BETAI of Press and others (1988) with an updated version of BETACF (Press and others 1992).

Root mean square error

The root-mean-squared error (RMSE) of the residual values is computed once for each continuous variable, V: $RMSE(V) = \sqrt{\frac{\sum_{i=1}^{N} (V_{obs} - V_{impu})^2}{N-1}}$.

Contingency table values

Cells in the two-way contingency tables contain three entries following the cell count. The row percentage is the percent of each cell count to its row total, the expected number is the expected frequency (E_{ij}, see the formula below) of each cell under the assumption that the imputations were made using a random process, and last entry is an approximate cell Chi Square statistic.

Let the quantity n_{ij} denote the cell count for row i and column j. Then $n = \sum_i \sum_j n_{ij}$ is total of all counts, $n_{\bullet j} = \sum_i n_{ij}$ is row total, and $n_{i\bullet} = \sum_j n_{ij}$ is the column total. The cell Chi Square statistic is: $\chi^2_{ij} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$, where $E_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{ij}}$. The total Chi Square is $\chi^2 = \sum_i \sum_j \chi^2_{ij}$. The degrees of freedom (df) associated with the total are computed as follows. Let r be the number of rows with non-zero totals and c the number of columns with non-zero totals. Then $df = (r - 1)(c - 1)$. When df is greater than zero, the probability of getting a larger Chi Square by chance alone is computed using the routine GAMMQ from Press and others (1988).

The Kappa statistic, $\hat{\kappa}$ (Khat), is computed as described by Cohen (1960), cited by SAS (1999, under the Statistical Computations for the FREQ procedure). The formula is $\hat{\kappa} = \frac{P_0 - P_e}{1 - P_e}$, where $P_0 = \sum_i p_{ii}$, $P_e = \sum_i p_{i\bullet} p_{\bullet i}$, $p_{\bullet i} = n_{\bullet i} / n$, and. The standard error is the square root of the asymptotic variance according to Fleiss and others (1969) also cited by SAS (1999) and computed as follows: $\text{var} = \frac{A + B - C}{\left(1 - P_e\right)^2 n}$, where $A = \sum_i p_{ii}\left[1 - \left(p_{i\bullet} + p_{\bullet i}\right)\left(1 - \hat{\kappa}\right)\right]^2$, $B = \left(1 - \hat{\kappa}\right)^2 \sum_{i \neq j}\sum p_{ij}\left(p_{\bullet i} + p_{j\bullet}\right)^2$, and $C = \left[\hat{\kappa} - P_e\left(1 - \hat{\kappa}\right)\right]^2$.

MSN output includes an interpretation of the range of Khat taken from Landis and Koch (1977), as follows: <=0 Poor, 0-0.2 Slight, 0.2-0.4 Fair, 0.4-0.6 Moderate, 0.6-.8 Substantial, and 0.8-1.0 Almost perfect.

Acknowledgments

We thank the numerous users of MSN for their suggestions for improvements. Their questions and comments lead to a greatly improved product. Comments and mentoring provided by Albert R. Stage were particularly valuable. We also acknowledge the contribution of David S. Watkins, Department of Mathematics, Washington State University, who coded the canonical correlation analysis.

References

Crist, E.P.; Cicone, R.C. 1984. A physically-based transformation of thematic mapper data—the TM tasseled cap. IEEE Transactions on Geoscience and Remote Sensing. Vol. GE-22(3):256-263.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20:37-46.

Dongarra, J.J.; Moler, C.B.; Bunch, J.R.; Stewart, G.W. 1979. LINPACK User's Guide. Philadelphia: Society for Industrial and Applied Mathematics. 11 chapters, 4 Appendices.

Ek, A. R.; Robinson, A.P.; Radtke, P.J.; Walters, D.W. 1997. Development and testing of regeneration imputation models for forests in Minnesota. Forest Ecology and Management. 94:129-140.

Fleiss, J.L.; Cohen, J.; Everitt, B.S. 1969. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin. 72:323-337.

Fu, P.; Rich, P.M. 1999. Design and implementation of the solar analyst: ArcView extension for modeling solar radiation at landscape scales. Proceedings of the 19th Annual ESRI User conference, San Diego, CA, USA.

Gittins, Robert. 1985. Canonical Analysis. A review with applications in ecology. Volume 12 in Biomathematics. New York: Springer-Verlag. 351 p.

Landis, J. Richard; Koch, Gary G. 1977. The measurement of observer agreement for categorical data. Biometrics. 33:159-174.

LeMay, V.M.; Temesgen, H. 2001. Nearest neighbour methods for generating tree-lists from aerial inventory data. Presented at the IUFRO 4.11 Forestry Biometry, Modelling, and Information Sciences Conference, The University of Greenwich, School of Computing and Mathematical Sciences, June 26-29, Greenwich, London. 24 p. [limited distribution on http://cms1.gre.ac.uk/conferneces/iufro/proceedings/index.htm]

McDonald, G.I.; Evans, J.; Moeur, M. Rice, T.M.; Strand, E. In press. Using digital terrain modeling and satellite imagery to map interactions among fire and forest microbes. In: Brennan, L.A.; and others (eds) National Congress on Fire Ecology, Prevention, and Management Proceedings, No. 1 Tall Timbers Research Station, Tallahassee, FL.

Moeur, Melinda. 2000. Extending stand exam data with most similar neighbor inference. In: Proceedings of the Society of American Foresters National Convention; 1999 September 11-15; Portland, OR. SAF Publication 00-1. Bethesda, MD: Society of American Foresters: 99-107.

Moeur, Melinda; Hershey, Rachel Riemann. 1999. Preserving spatial and attribute correlation in the interpolation of forest inventory data. Chapter 49 (pp. 419-430) in: K. Lowell, and A. Jaton (eds.), Spatial accuracy assessment: land information uncertainty in natural resources. Ann Arbor Press.

Moeur, Melinda; Stage, Albert R. 1995. Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science. 41:337-359.

Moeur, Melinda; Crookston, Nicholas L.; Stage, Albert R. 1995. Most similar neighbor analysis: a tool to support ecosystem management. Pages 31-44 in: Thompson, J. E. (comp.). 1995. Analysis in support of ecosystem management. Analysis Workshop III, April 10-13, 1995, Fort Collins, CO. Washington, D.C.: USDA, Forest Service, Ecosystem Management Analysis Center. 360 p.

Muinonen, Eero; Maltamo, Matti; Hyppänen, Harri; Vainikainen, Vesa. 2001. Forest stand characteristics estimation using a most similar neighbor approach and image spatial structure information. Remote Sensing of Environment. 78:223-228.

Press, William H.; Flannery, Brian P.; Teukolsky, Saul A.; Vetterling, William T. 1988. Numerical Recipes The Art of Scientific Computing. Cambridge: Cambridge University Press. 818 p.

Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P.1992. Numerical Recipes in Fortran 77 The Art of Scientific Computing. Second Edition. Cambridge: Cambridge University Press. 973 p. (web published version at http://lib-ww.lanl.gov/numerical/bookfpdf.html).

Reed, B.; Brown, J.F.; Vanderzee, D.; Loveland, T.R.; Merchant, J.W. 1994. Measuring phonological variability from satellite. Journal of Vegetation Science. 5:703-714.

SAS. 1999. SAS OnlineDoc$_{(TM)}$, Version 7-1. Cary, NC: SAS Institute Inc.

Shao, Jun; Sitter, Randy R. 1996. Bootstrap for imputed survey data. Journal of the American Statistical Association. 91(435):1278-1288.

Sironen, Susanna; Kangas, Annika; Maltamo, Matti; Kangas, Jyrki. 2001. Estimating individual tree growth with the k-nearest neighbour and k-Most Similar Neighbour methods. Silva Fennica. 35(4):453-467.

Stage, A.R. 1976. An expression for the effect of aspect, slope, and habitat type on tree growth. Forest Science. 2:457-460.

Van Deusen, P.C. 1997. Annual forest inventory statistical concepts with emphasis on multiple imputation. Canadian Journal of Forest Research. 27:379-384.

Venables, W.N.; Smith, D.M.; R Development Core Team. 2002. An introduction to R. Version 1.5.0. ISBN 3-901167-55-2. (Web published at http://www.r-project.org).

Wilson, John P.; Gallant, John C. 2000. Secondary Topographic Attributes. Chapter 4 in Terrain Analysis: Principles and Applications. Wilson, John P.; Gallant, John C. (Eds.). New York: John Wiley & Sons, Inc. 479 p.

Appendix A: Program Installation and Computer Requirements

MSN program is available from the Rocky Mountain Research Station, Forestry Sciences Laboratory, Moscow, ID: the Web site for MSN is at http://forest.moscowfsl.wsu.edu/4155/msn.html. The installation instructions are posted with the software. It is free and in the public domain. It is simple to install and is available for Microsoft Windows family and IBM AIX version 4. Standard PC's with Pentium Pro processors or better can be used. Contact the authors if you want the source code.

The current configuration of MSN accepts problems that have up to 40,000 observations total and up to 300 Y- and 300 X-variables. This number includes one column for each unique value of categorical data.

Appendix B: Run Time Information

Large MSN runs can take a lot of computer time, sometimes several hours, even on fast computers. It is the computation of distances that takes so much time. Reading the input data, merging observations, and the canonical correlation steps are not big issues.

In searching for the most similar neighbor, the program computes the distance between target and reference observations. This is a so-called N-by-N search, and the amount of time required increases exponentially with the value of N. The number of calculations required to compute one distance is a function of the number of X-variables and the number of canonical variates.

We explored the influence of these factors on processing time for our Unix system: an IBM/RS6000 F40 server, with 768 MB main memory and dual power 604e processors running under AIX 4.3 operating system. We measured the central processing unit (CPU) time expended for processing various combinations of numbers of observations, numbers of variables, and numbers of canonical variates.

The number of observations has the greatest effect on CPU time. A relatively small run consisting of about 550 reference and 550 target observations uses about 4.5 minutes CPU time. Medium-sized runs—10,000 to 15,000 observations—can use 6 to 7 hours of CPU time, and larger runs can use much more. Increasing the number of reference observations greatly increases run time more than increasing the target observations. Increasing the number of X-variables in the problem increases CPU time, while increasing the number of Y-variables and number of canonical variates has little effect.

For large problems, it may be necessary to break the problem into smaller segments. Here are some strategies that take advantage of some of the tools built into the MSN program. First, you can experiment with different canonical correlation models and study the most important validation statistics using only the reference observations. Use the NOTARGS option to suppress generating imputations for target observations until you are satisfied with the model formulation and results. Finally, replace the NOTARGS option with NOREFS and rerun program. You will get the imputations for target observations without repeating the calculations achieved with the first step. For large problems, you can break the target observations into separate files and run them one at a time, perhaps using separate computers at the same time. Each run would require identical reference data; only the target data would change.

RMRS
ROCKY MOUNTAIN RESEARCH STATION

The Rocky Mountain Research Station develops scientific information and technology to improve management, protection, and use of the forests and rangelands. Research is designed to meet the needs of National Forest managers, Federal and State agencies, public and private organizations, academic institutions, industry, and individuals.

Studies accelerate solutions to problems involving ecosystems, range, forests, water, recreation, fire, resource inventory, land reclamation, community sustainability, forest engineering technology, multiple use economics, wildlife and fish habitat, and forest insects and diseases. Studies are conducted cooperatively, and applications may be found worldwide.

Research Locations

Flagstaff, Arizona
Fort Collins, Colorado*
Boise, Idaho
Moscow, Idaho
Bozeman, Montana
Missoula, Montana
Lincoln, Nebraska
Reno, Nevada
Albuquerque, New Mexico
Rapid City, South Dakota
Logan, Utah
Ogden, Utah
Provo, Utah
Laramie, Wyoming

*Station Headquarters, Natural Resources Research Center,
2150 Centre Avenue, Building A, Fort Collins, CO 80526

Federal Recycling Program ♻ Printed on Recycled Paper