

Open Problem for the May 2009 SUA Ve meeting in Victoria.

The challenge here is to read in several separate text files and combine them into one SAS data set for further analysis. Each file has exactly the same format for the data, but important information is also located in the file name.

Each file contains simulated map data for tree locations, their height diameter and volume as well as a tree number. Each file can be read in using the following input statement:

```
input x y height diameter treeno volume;
```

The example files are on a ftp site at

<http://www.for.gov.bc.ca/ftp/HRE/external/outgoing/wbergerud/May2009Open/>. You will probably wish to start by downloading a copy of these files into a directory on your computer or LAN. The files are:

```
p180020015.dat  
p180020080.dat  
p220120015.dat  
p220120080.dat  
r240350015.dat  
r240350080.dat  
r260140015.dat  
r260140080.dat
```

New variables must be extracted from the file name so that they can be used in the analysis. The first letter of the file name refers to the spatial distribution with r = Random and p = Planted. The next two digits refer to what is known in forestry as the site index which is a measure of how productive a site is. The next five digits specify the beginning density while the last two digits are the age of the modelled stand.

While there are only eight files in this example, this problem becomes much more of a challenge when there are 480! Will your solution work well with a larger number of files?

Good luck and we look forward to seeing your suggested solutions!

Solution 1. Ming Guo (Management Information Branch, BC Ministry of Health Services)

```
*written by Ming Guo, Apr. 27, 2009;
%let inputdir=H:\document\SUAVE\;
OPTIONS LS=256 PS=30000 NOCENTER MPRINT OBS=MAX;

%LET fstart = 1; *file index to start;
%LET fend   = 8; *file index to end;

*get file name list;
*file name list, index.txt, can be obtained using "dir /b > index.txt" (MSDOS)
    or "ls > index.txt" (Unix);
data fileind ;
infile "&inputdir.index.txt";
input @1 filename $10.;
run;

%Macro mAddfile(fn_,nf_);*macro to load data from one file;

%let fname=&fn_;
data f&nf_;
infile "&inputdir.&fname..dat";
input x y height diameter treeno volume;
run;

data f&nf_;
set f&nf_;
sd=substr("&fname.",1,1);    *spatial distribution;
si=substr("&fname.",2,2)+0;  *site index;
bd=substr("&fname.",4,5)+0;  *beginning density;
ams=substr("&fname.",9,2)+0; *age of modelled stand;
run;
%mend;

%MACRO mMain(n1_,n2_); *macro including a loop to load all files;
%DO y = &n1_ %TO &n2_;
    %let mydataid=%sysfunc(open(fileind,i)); *open dataset containing file name list;
    %let rc=%sysfunc(fetchobs(&mydataid,&y+0)); *set cursor to a specified row;
    %let fname=%sysfunc(getvarc(&mydataid,%sysfunc(varnum(&mydataid,filename))));
*pass file name to the macro ;
    %let rc=%sysfunc(close(&mydataid));
    %mAddfile(&fname,&y)
%END;
%MEND;

%mMain(&fstart,&fend);

%MACRO mData(s1_, n1_, n2_);
    %DO i = &n1_ %TO &n2_;
        &s1_&i
    %END;
%MEND;

DATA dAll; *combine all data into one dataset;
SET %mData(f, &fstart, &fend);
RUN;
```

Solution 2. Aijun Yang (Management Information Branch, BC Ministry of Health Services), and Mike Atkinson (Acko Systems)

```
* Aijun Yang and Mike Atkinson - April 2009;
* Set name of directory containing the files;
%let dir_name = \\path\to\Multiple Files;

* DIR command to list filenames only, without including directories;
* Note: the following pipe works on Windows XP;
filename files pipe "dir "&dir_name" /a:-d /b" lrecl=1000;

* Under Unix, something like the following commands might be used;
* %let dir_name = /hlth/some/unix/directory;
* filename files pipe "ls "&dir_name"" lrecl=1000;

* Create a dataset with all of the tree map data;

data tree_map_data;

    * Read each file name from the pipe defined above;
    infile files truncover;
    input file_name $1000.;

    * Only continue for files ending with file type ".dat";
    if (prxmatch("/\.dat\s*$/", file_name));

    * Get the parts out of the file name;
    spatial_distribution = upcase(substr(file_name, 1, 1));
    site_index           = input(substr(file_name, 2, 2), 2.);
    beginning_density    = input(substr(file_name, 4, 5), 5.);
    age_of_modelled_stand = input(substr(file_name, 9, 2), 2.);

    * Read the contents of the file, and output each observation;
    file_name = "&dir_name" || '\' || file_name;
    infile dummy filevar=file_name end=eof;
    do while (not eof);
        input x y height diameter treeno volume;
        output;
    end;
run;
```

Solution 3. Jim Goudie (Research Branch, BC Ministry of Forests and Range)

```
*Jim Goudie, April 2007;
filename dat 'C:\temp\*.dat'; *path where files are located. Note use of wildcard to read-in several of
them;

DATA dat;
length fn $ 40 distribution $1;
infile dat missover filename=fn; *fn is a temporary variable;
input x y height diameter treeno volume;

tree_filename = trim(fn);
dot_position = INDEX(tree_filename, '.'); *find dot in the path name;
distribution = substr(tree_filename, dot_position-10, 1); *determine spatial distrib;
site_index = substr(tree_filename, dot_position-9, 2)+0; *determine SI and convert to numeric;
density = substr(tree_filename, dot_position-7, 5)+0; *determine dens and convert to numeric;
age = substr(tree_filename, dot_position-2, 2)+0; *determine age and convert to numeric;

run;

proc print;
by tree_filename;
var treeno x y distribution site_index density age height diameter volume;
run;

quit;
```