7 UPDATING REFERENCE SITE DATA

7.1 Introduction

7.1.1 Why update reference site data?

Users may wish to add new variables to the reference site data, delete variables from the reference site data, modify variables in the reference site data or update the descriptive information for each variable in the reference site data. **The reference site data will also need to be updated whenever there is a change in composition of the reference sites in an AUSRIVAS model or a change in site membership of any classification group in an AUSRIVAS model.** This is because physical and chemical data are reported for each AUSRIVAS model and by each classification group contained within a model (see Part 5).

7.1.2 How are reference site data updated?

There are two parts to the reference site data. The first is the actual **reference data set**, which contains the values of each physical and chemical variable at each site (Section 7.2). The second is the **reference variables information sheet,** which provides descriptive information on the variables included in the reference data (Section 7.3). This sheet details the variable acronym, description of the variable, unit of measurement, the number of decimal places and whether the variable is continuous or categorical. The reference variables information sheet also sets the valid range of expected values, which is a function for identifying gross errors in the entry of reference and test data. The procedures for updating each of these components are detailed in the following sections.

7.2 How to set up or modify a reference data set

7.2.1 Step 1
Set up the reference data set in a spreadsheet

7.2.1.1 General instructions

It is easiest to set up a reference data set in Excel (or a similar spreadsheet package). Reference data have the following general properties:

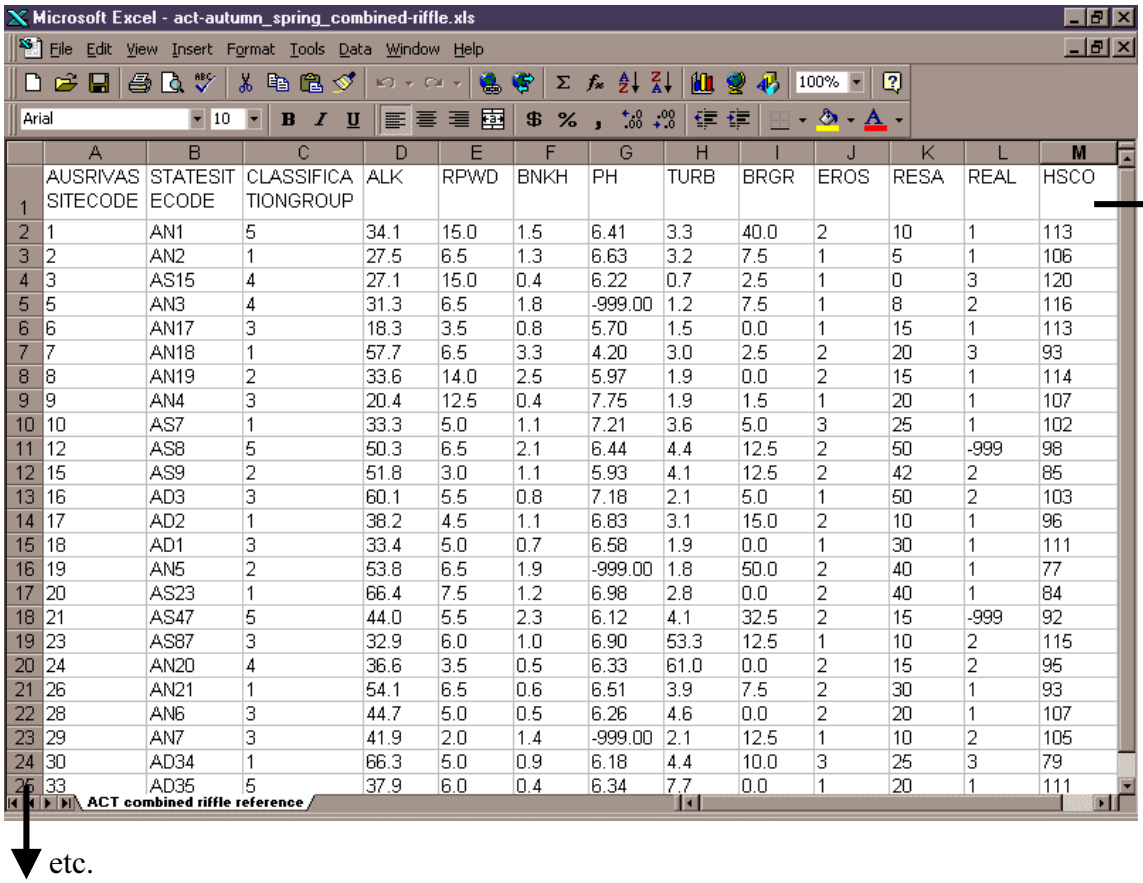
- Data are arranged with sites down the side (i.e. as rows) and variables across the top (i.e. as columns).
- The sites included in the reference data set must match the sites that comprise the corresponding AUSRIVAS model.
- Variable acronyms listed across the top must be in CAPITALS and CAN NOT CONTAIN SPACES. There is no character limit for the length of a variable acronym, but shorter names will be easier to display in the outputs.
- Variable acronyms must match those listed in the reference variables information sheet (see Section 7.3.1.1).
- Three columns are standard to ALL reference data sets and must be labelled exactly with the following acronyms:

Column	Exact acronym	What the column means
A	AUSRIVASSITECODE	This column is the site code used in the AUSRIVAS model building process.
B	STATESITECODE	Some states have site codes (e.g. alpha-numeric format) that differ from those used in the AUSRIVAS models. This column identifies the State site code and enables matching with the AUSRIVAS site codes. Columns 1 and 2 must BOTH be included in the data set but where the AUSRIVAS and State site codes are the same, both columns will contain identical information.
C	CLASSIFICATIONGROUP	This column is the classification group to which each site belongs. It is derived from construction of the AUSRIVAS models.

There are also several other requirements for the main body of the data. These are:

- Cells with missing data are denoted with the value -999 The number of sites that contain missing data for any particular variable is given in the output (see Section 5.1.3). However, the usefulness of any variable is reduced when there are large chunks of missing data and it is best to delete the entire variable from the reference data set if roughly more than one third of sites in any model have missing data for any one variable.
- Cells can not contain formatting such as subscript, superscript, bold or italicised characters.
- Continuous variables can not contain characters such as < or >. Where chemical variables are listed as detection limits (e.g. <0.001) the < symbol must be removed (e.g. <0.001 becomes 0.001). Notes on detection limits can be set in the reference variables information sheet (see Section 7.3).
- Categorical variables can contain text (e.g. high, low, medium or bedrock, boulder, cobble, pebble, gravel, sand, silt, clay etc.).
- Decimal places must be set for each variable. In Excel, this is easily done using the "Format Cells" function. Continuous variables must be set to a specific number of decimal places (e.g. velocity measurements that are metres per second usually have 3 decimal places (0.043), stream width in metres usually has one decimal place (12.4) etc.). The number of decimal places for a missing value will correspond to the number of decimal places set for that variable (e.g. -999 for 0 decimal places, -999.000 for three decimal places). The number of decimal places set for each variable is a required entry on the reference variables information sheet (see Section 7.3), because test data must be set up with a matching number of decimal places (see Section 3.2).

The reference data set will resemble the following spreadsheet:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	AUSRIVAS SITECODE	STATESITE CODE	CLASSIFICATIONGROUP	ALK	RPWD	BNKH	PH	TURB	BRGR	EROS	RESA	REAL	HSCO
2	1	AN1	5	34.1	15.0	1.5	6.41	3.3	40.0	2	10	1	113
3	2	AN2	1	27.5	6.5	1.3	6.63	3.2	7.5	1	5	1	106
4	3	AS15	4	27.1	15.0	0.4	6.22	0.7	2.5	1	0	3	120
5	5	AN3	4	31.3	6.5	1.8	-999.00	1.2	7.5	1	8	2	116
6	6	AN17	3	18.3	3.5	0.8	5.70	1.5	0.0	1	15	1	113
7	7	AN18	1	57.7	6.5	3.3	4.20	3.0	2.5	2	20	3	93
8	8	AN19	2	33.6	14.0	2.5	5.97	1.9	0.0	2	15	1	114
9	9	AN4	3	20.4	12.5	0.4	7.75	1.9	1.5	1	20	1	107
10	10	AS7	1	33.3	5.0	1.1	7.21	3.6	5.0	3	25	1	102
11	12	AS8	5	50.3	6.5	2.1	6.44	4.4	12.5	2	50	-999	98
12	15	AS9	2	51.8	3.0	1.1	5.93	4.1	12.5	2	42	2	85
13	16	AD3	3	60.1	5.5	0.8	7.18	2.1	5.0	1	50	2	103
14	17	AD2	1	38.2	4.5	1.1	6.83	3.1	15.0	2	10	1	96
15	18	AD1	3	33.4	5.0	0.7	6.58	1.9	0.0	1	30	1	111
16	19	AN5	2	53.8	6.5	1.9	-999.00	1.8	50.0	2	40	1	77
17	20	AS23	1	66.4	7.5	1.2	6.98	2.8	0.0	2	40	1	84
18	21	AS47	5	44.0	5.5	2.3	6.12	4.1	32.5	2	15	-999	92
19	23	AS87	3	32.9	6.0	1.0	6.90	53.3	12.5	1	10	2	115
20	24	AN20	4	36.6	3.5	0.5	6.33	61.0	0.0	2	15	2	95
21	26	AN21	1	54.1	6.5	0.6	6.51	3.9	7.5	2	30	1	93
22	28	AN6	3	44.7	5.0	0.5	6.26	4.6	0.0	2	20	1	107
23	29	AN7	3	41.9	2.0	1.4	-999.00	2.1	12.5	1	10	2	105
24	30	AD34	1	66.3	5.0	0.9	6.18	4.4	10.0	3	25	3	79
25	33	AD35	5	37.9	6.0	0.4	6.34	7.7	0.0	1	20	1	111

7.2.2 Step 2
Convert the spreadsheet to .csv file format

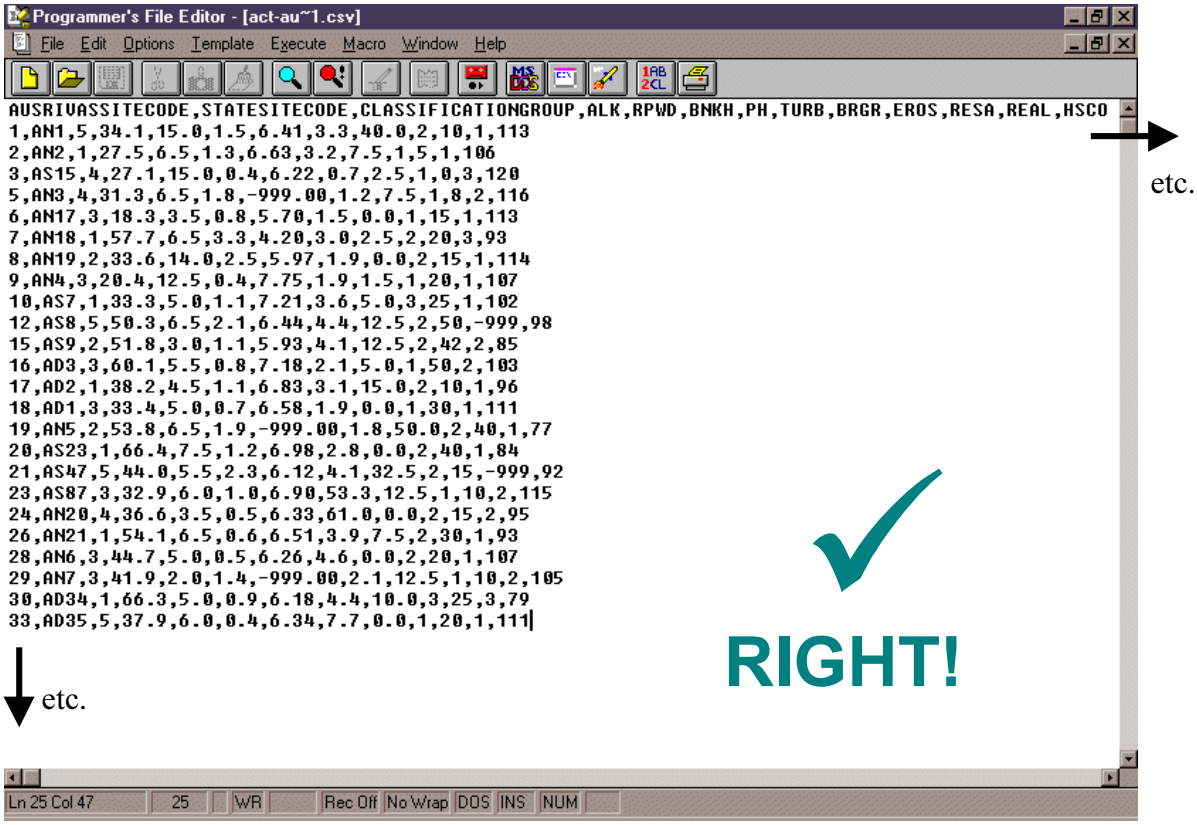
7.2.2.1 General instructions

- For the program to be able to read the reference data, the spreadsheet must be converted to a comma delimited .csv file. This is easily done using the "Save As" function in Excel – choose the file type called "CSV (comma delimited) (*.csv)".
- File names must follow the following format and be in lower case:

  state-model-habitat.csv

  Spaces within any portion must be underscored. For example, the ACT combined season riffle model reference data file name is act-autumn_spring_combined-riffle.csv

When converted to .csv format the reference site data will resemble the following:



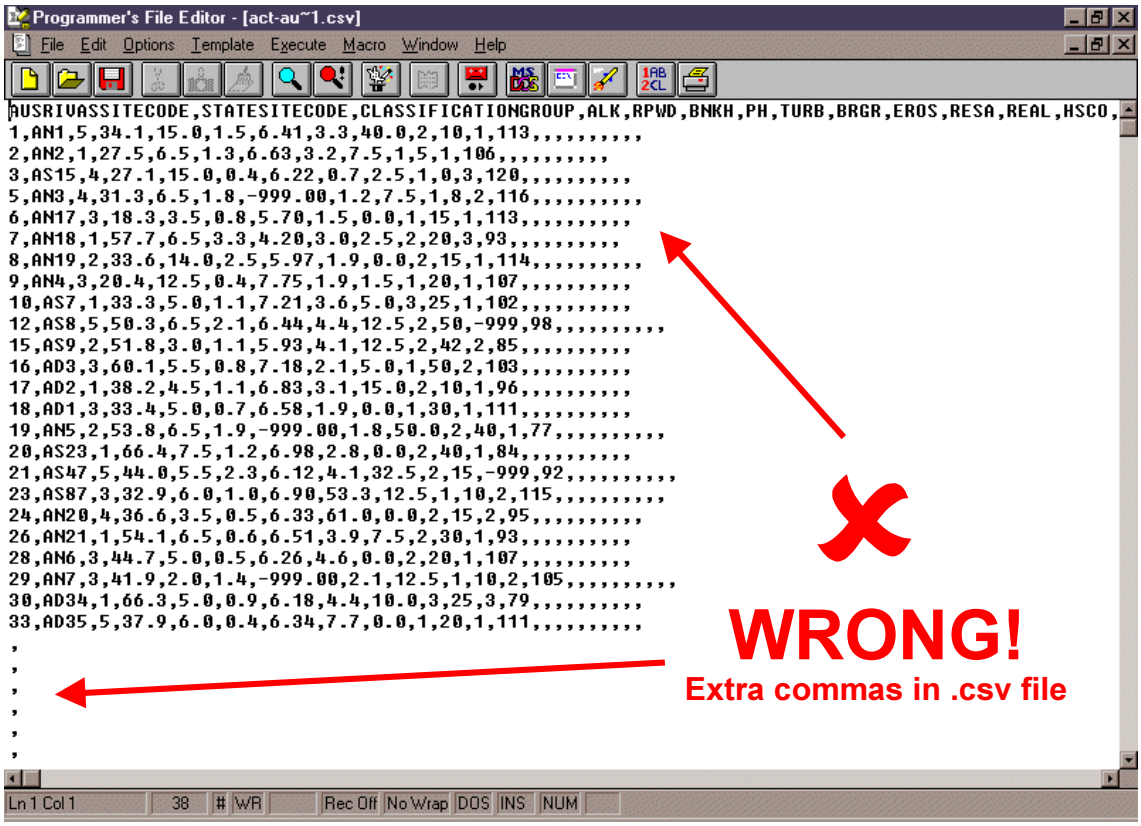
All .csv files **must be checked after conversion.** This can be done by viewing in a text editor such as Word Pad or similar.

7.2.2.2 Problems encountered when converting spreadsheets to .csv files

Some common problems encountered when converting spreadsheets to .csv files are outlined in the following sections

Extra commas in .csv file

As far as we can tell, this problem occurs where certain formatting (such as bold) is switched on in empty rows and columns in the Excel spreadsheet. The resulting .csv file will look like this:

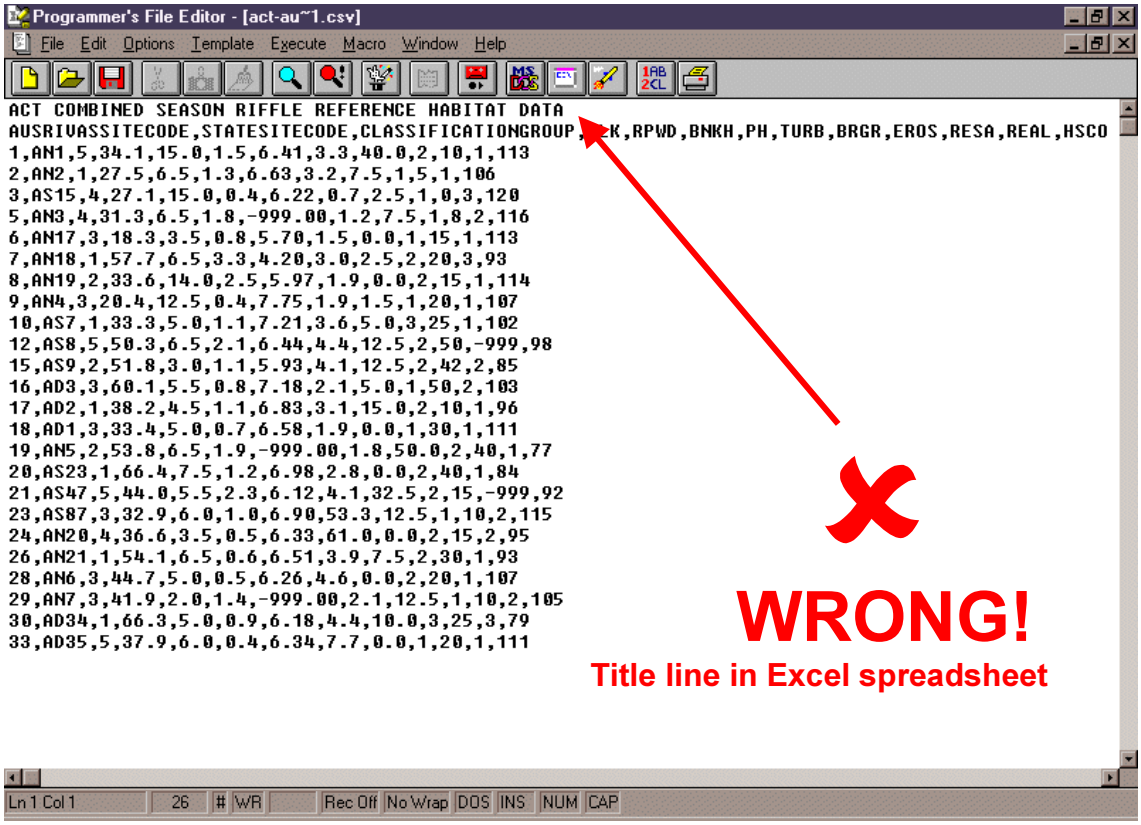


The **solution** is to either:

1. Remove any formatting from empty rows and columns in excel before saving as .csv format. Check the .csv file again after this is done though, as it is not 100% certain that this is the cause of the problem.
2. Remove the commas manually from the .csv file using a text editor such as Word Pad.

Title line in Excel spreadsheet

This problem occurs where the user has inserted a title line as Row 1 of the Excel spreadsheet. The resulting .csv file will resemble the following:

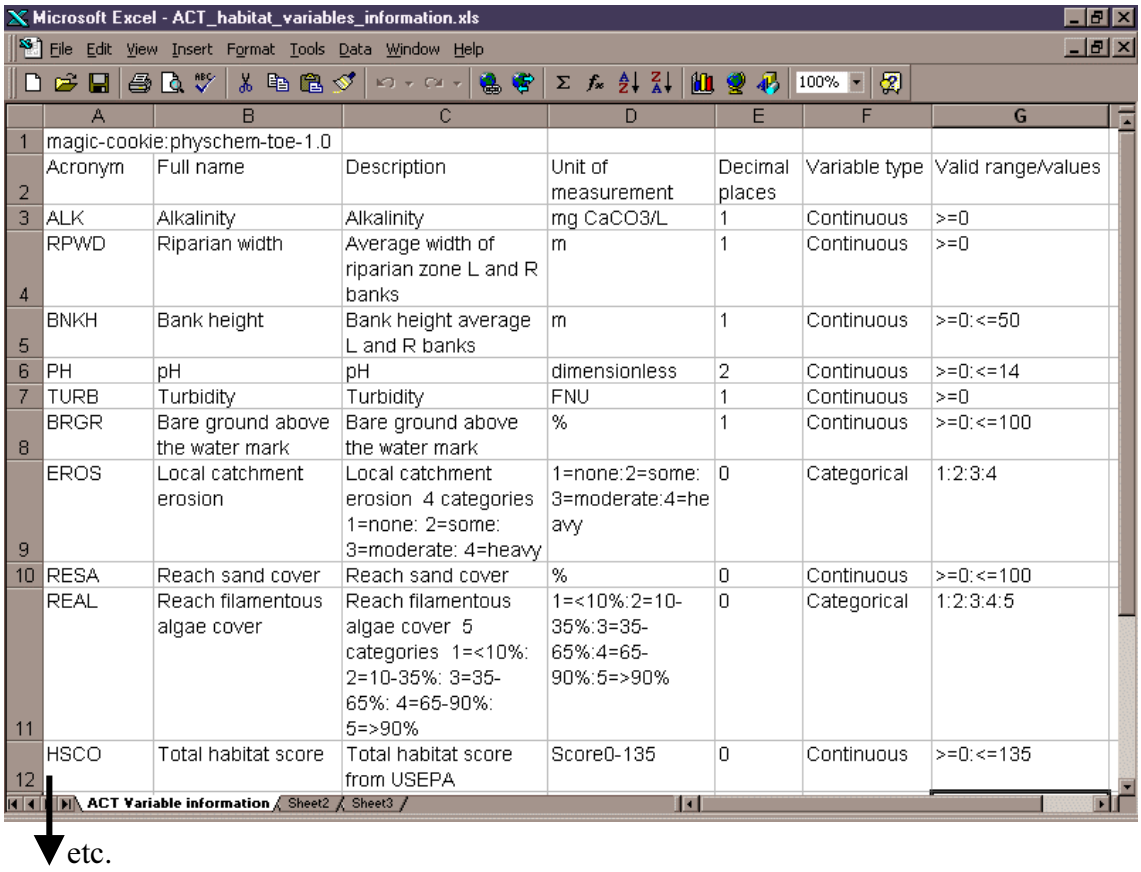


The **solution** is to remove the title line from the Excel spreadsheet before converting to .csv format. Row 1 should be the column title acronyms.

7.3 How to set up or modify a reference variables information sheet

7.3.1 Step 1 Set up the reference variables information sheet in a spreadsheet

As with a reference data set, it is easiest to set up the reference variables information sheet in Excel, or a similar spreadsheet package. There is only one reference variables information sheet for each State, but this contains details of all the variables included in constituent models. The reference variables information sheet resembles the following spreadsheet:



	A	B	C	D	E	F	G
1	magic-cookie:physchem-toe-1.0						
2	Acronym	Full name	Description	Unit of measurement	Decimal places	Variable type	Valid range/values
3	ALK	Alkalinity	Alkalinity	mg CaCO3/L	1	Continuous	>=0
4	RPWD	Riparian width	Average width of riparian zone L and R banks	m	1	Continuous	>=0
5	BNKH	Bank height	Bank height average L and R banks	m	1	Continuous	>=0:<=50
6	PH	pH	pH	dimensionless	2	Continuous	>=0:<=14
7	TURB	Turbidity	Turbidity	FNU	1	Continuous	>=0
8	BRGR	Bare ground above the water mark	Bare ground above the water mark	%	1	Continuous	>=0:<=100
9	EROS	Local catchment erosion	Local catchment erosion 4 categories 1=none: 2=some: 3=moderate: 4=heavy	1=none:2=some: 3=moderate:4=heavy	0	Categorical	1:2:3:4
10	RESA	Reach sand cover	Reach sand cover	%	0	Continuous	>=0:<=100
11	REAL	Reach filamentous algae cover	Reach filamentous algae cover 5 categories 1=<10%: 2=10-35%: 3=35-65%: 4=65-90%: 5=>90%	1=<10%:2=10-35%:3=35-65%:4=65-90%:5=>90%	0	Categorical	1:2:3:4:5
12	HSCO	Total habitat score	Total habitat score from USEPA	Score0-135	0	Continuous	>=0:<=135

7.3.1.1 General properties of the reference variables information spreadsheet

- The first row of the reference variables information sheet must contain the words

  `magic-cookie:physchem-toe-1.0`

  This string is called a "magic cookie" and its purpose is to allow the Physical & Chemical Reporting software to check that the file it has obtained from the server is actually the reference variables information sheet, rather than, say, an error message sent back to the software from a proxy server.

  The string "magic-cookie:physchem-toe-1.0" must be reproduced exactly in the first entry of the first row of the reference variables information sheet as follows:

  - It is all in lower case;
  - It doesn't contain any spaces;
  - It it plain text, just like the rest of the file (i.e. no formatting or non-keyboard characters)

- All variables in all models used in the physical and chemical assessment software must be included in the one reference variables information sheet (refer to Section 3.2.1 for instructions on how to find out which models are included in the package).

However, when using the software, only those variables included in the selected model will be displayed (see Section 3.2.2).

- It is critical that the acronyms used in the reference data sets MATCH EXACTLY with the acronym given in the reference variables information sheet. This is because the Physical & Chemical Reporting software uses these variable acronyms to match up the information in different files. For example, if you have a variable in your reference data called "ALK", the program will look for this code in the reference variables information sheet to find this variable's descriptive information. If you've called this variable "ALKA" rather than "ALK" in the reference variables information sheet, the program won't be able to recognise that "ALK" and "ALKA" are the same variable.

7.3.1.2 Explanation of individual columns in the reference variables information spreadsheet

The reference variables information sheet has seven columns, all of which must be included. The exact heading and purpose of each column is as follows:

Column No.	Exact heading	Purpose
A	Acronym	Is the acronym for each variable used in the reference data sets
B	Full name	Gives the full name of the variable
C	Description	Provides additional information about the variable. This can include notes about detection limits, categories required, source of the variable, collection method etc.
D	Unit of measurement	Gives the unit of measurement for each variable. Different format for continuous or categorical variables.
E	Decimal places	Gives the number of decimal places for each variable
F	Variable type	Identifies the variable as either continuous or categorical
G	Valid range/values	Sets the valid range of values that can be expected for each variable in both the reference and test data.

Instructions for the format of each column are detailed in the following sections. These instructions must be followed strictly, otherwise the program can't accurately match test data with reference site data. Rogue commas, spaces or formatted characters will generate errors that disable the program!

Column A Acronym

- The acronyms in the reference variables information sheet must exactly match the acronyms used in the reference data sets
- Acronyms must be in CAPITALS with no spaces

Column B Full name

- The full name of each variable can be in upper or lower case, but don't use any formatted characters such as subscripts or superscripts. Use only the characters that can be seen on the keyboard.
- The full name of the variable CAN NOT contain any commas, because this will distort the file when saved as a comma delimited .csv file

Column C Description

- The description of each variable can be in upper or lower case, but don't use any formatted characters such as subscripts or superscripts. Use only the characters that can be seen on the keyboard.
- The description of the variable CAN NOT contain any commas, because this will distort the file when saved as a comma delimited .csv file

Column D Unit of measurement

- The unit of measurement for each variable can be in upper or lower case, but don't use any formatted characters such as subscripts or superscripts. For example, mg $CaCO_3$ l^{-1} becomes mg CaCO3/L, or something similar that retains meaning. Use only the characters that can be seen on the keyboard.
- The unit of measurement for each variable CAN NOT contain any commas, because this will distort the file when saved as a comma delimited .csv file
- For **continuous** variables, the unit of measurement corresponds to the unit of measurement used for each variable.
- For **categorical** variables, the unit of measurement corresponds to the boundaries of each category. The unit of measurement for a categorical variable must have colons separating each category and CAN NOT contain spaces. For example, the categories and corresponding values for a variable such as algal cover should be entered as:

1=<10%:2=10-35%:3=35-65%:4=65-90%:5=>90% ✓ **Right!**

Items separated by colons
No spaces between colons

1=<10%, 2=10-35%, 3=35-65%, 4=65-90%, 5=>90% ✗ **Wrong!**

Items separated by commas
Spaces between items

- For text-based categorical variables lettering case must match between the reference data set and the unit of measurement column. For example, a variable that was called DOMSUB (dominant substrate type) may have categories of bedrock:boulder:cobble:pebble:gravel:sand:silt:clay. These words should be entered in lower case in the reference data set.

Column E Decimal places

- This is a numerical entry of the number of decimal places set for each variable (see Section 7.2.1.1).

Column F Variable type

- Two types of variables are possible:

  Continuous variables are those for which data are arranged along a continuum. These variables include items like chemical parameters measured in mg l^{-1}, percent cover of substrate, habitat assessment scores, stream width etc. To denote a continuous variable, column F should contain the string:

  `continuous`

  Categorical variables are those for which data are arranged in finite categories. These variables include items like site observations, percent cover categories and dominant substrate category. To denote a categorical variable, column F should contain the string:

  `categorical`

- It is important to set the type of variable carefully, as continuous and categorical variables are calculated and reported differently (see Part 5).

Column G Valid range/values

- The upper and lower limits for a variable are included to guard against gross errors in data entry in both the reference data set and the test data set.
- Upper and lower limits for **continuous** variables are set by using the greater than, equals and less than characters. For example:

>=0 means that the variable is expecting values that are greater than or equal to zero

`>=0:<135` means that the variable is expecting values that are greater than or equal to zero and less than one hundred and thirty five. Note here that the "**:**" symbol is used to represent "and". This variable should be greater than or equal to 0 **and** less than 135.

- **Categorical** variables do not have upper and lower limits but have values corresponding to each category. For example a variable with percent occurrence categories of 1-5 would have a valid range/values entry of 1:2:3:4:5 This means that the reference or test data can not be entered as any value other than these individual numbers. Note here that the ":" symbol is used to represent "or", rather than "and" as for continuous variables. The example variable given here should have a value of 1 **or** 2 **or** 3 **or** 4 **or** 5.
- All cells in the valid range/values column must contain information.
- The valid range/values entries must be separated by colons and CAN NOT contain spaces. For example:

>=0:<=100 1:2:3:4:5	✓ **Right!** Items separated by colons No spaces between colons
>= 0, <= 100 1, 2, 3, 4, 5	✗ **Wrong!** Items separated by commas Spaces between items

***** IMPORTANT NOTE *****

Do not set an upper limit to a variable on the basis of the reference site information only, because the value for a test site can fall outside these limits. For example, the range of pH values within a reference data set may be 5.87-8.13 but an acidified test site may have a pH of 2.14. However, the pH scale will always run from 0-14 and these values can be set as upper or lower limits. Thus, a pH value that is mis-entered as 81.3 instead of 8.13 will be flagged as an error.

7.3.2 Step 2
Convert the spreadsheet to .csv file format

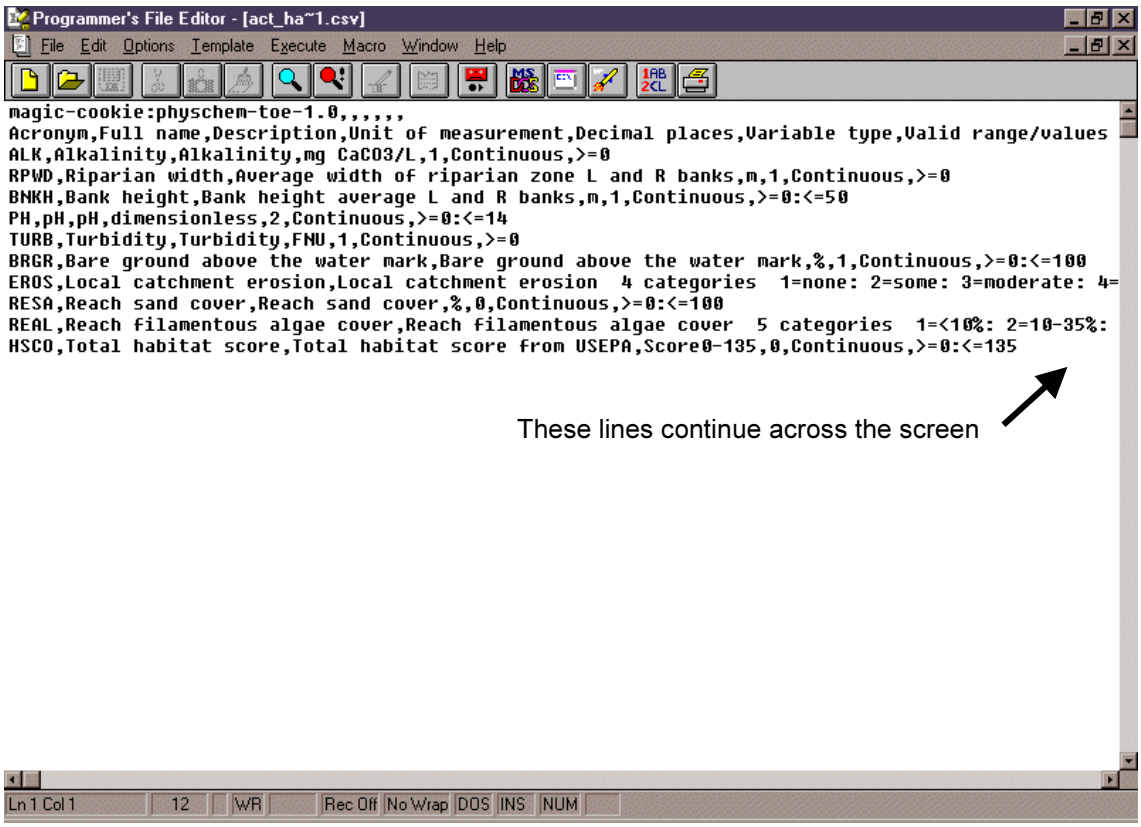
7.3.2.1 General instructions

- For the program to be able to read the reference variables information, the spreadsheet must be converted to a comma delimited .csv file. This is easily done using the "Save As" function in Excel – choose the file type called "CSV (comma delimited) (*.csv)".
- File names must follow the following format and be in lower case:

state_habitat_variables_information.csv

Spaces within any portion must be underscored. For example, the ACT reference variables information sheet file name is act_habitat_variables_information.csv

When converted to .csv format the reference variables information sheet will resemble the following format:



Note that the first line contains six additional commas. THESE SHOULD BE LEFT IN THE REFERENCE VARIABLES INFORMATION SHEET .CSV FILE. This is

because the "magic cookie" sequence is entered in the first cell of Column A in the Excel spreadsheet, and the remaining six columns are blank. Leaving these commas (in the reference variables information file only!) will tell the program to move to the next row, but to expect sequential rows to contain items in columns B-G.

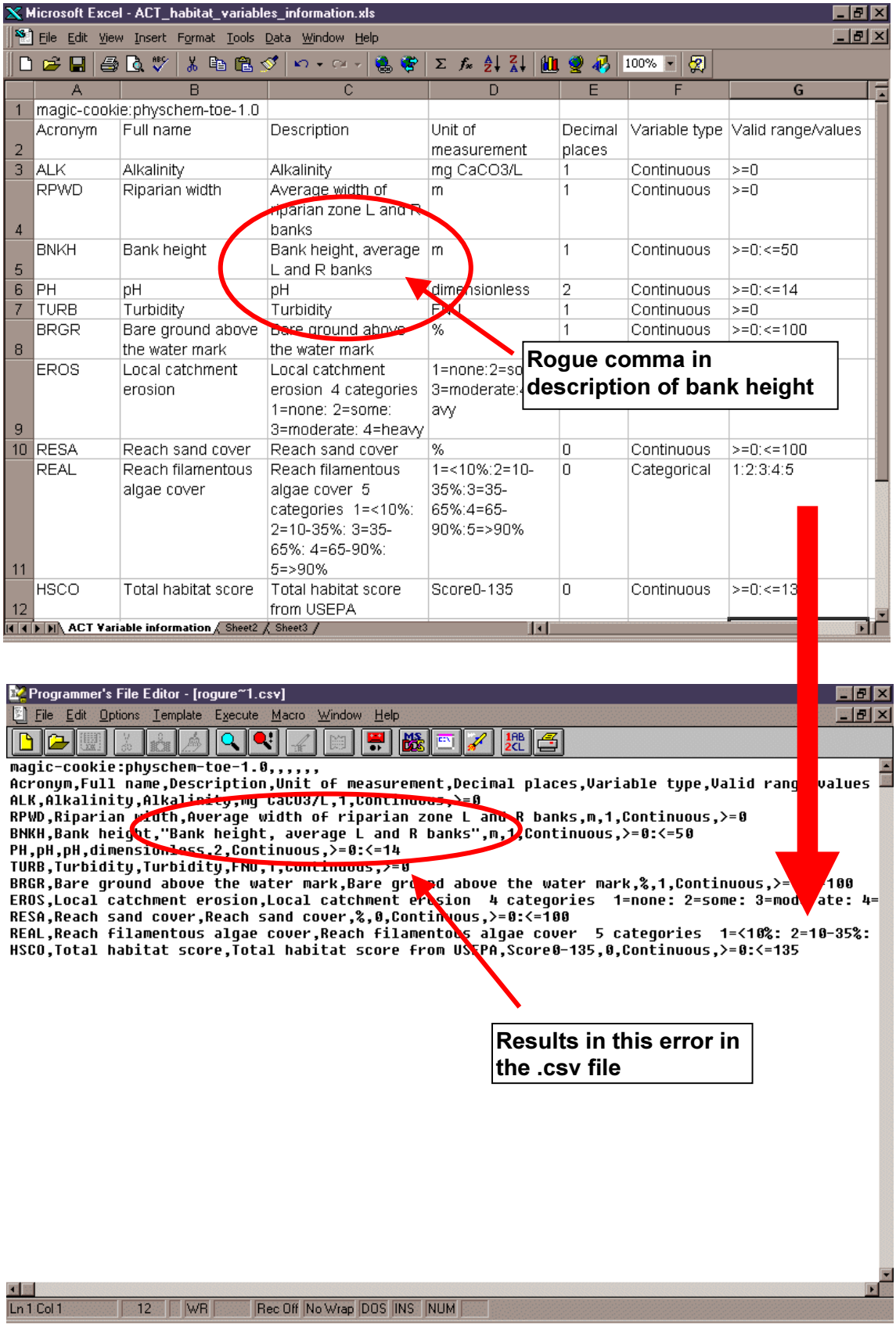
7.3.2.2 Problems encountered when converting spreadsheets to .csv files

Extra commas in .the csv file

Saving from Excel format into .csv format may result in placement of extra commas at the ends of rows or columns. **Aside from the first row, which is a special case outlined above**, these commas should be removed from the reference variables information sheet .csv file. The procedures for doing this are outlined in Section 7.2.2.2.

Rogue commas in the Excel spreadsheet

Insertion of a rogue comma in the Excel spreadsheet will result in a major program error because the program relies on reading comma delimited files, in which each data item is separated by a comma. **It is critical that there are no rogue commas within the entries on the reference variables information sheet.** For example:



The solution to this problem is to search the Excel spreadsheet and remove rogue commas before saving to .csv format. To do this, choose the "Find" function from the

Edit menu. Enter a comma as the item to be found but change the "Look in" box to "Values" rather than "Formulas". Do not click the "Find entire cells only" box.

7.4 Checklist

At the end of setting up a reference site data set and reference variables information sheet you should have the following items:

1. a reference data set for each model, saved in .csv format with NO ERRORS! For example, a State with six available models will have a set of six corresponding .csv reference data files.
2. one reference variables information sheet containing entries on all the variables included in all the reference data files. This should be saved in .csv format with NO ERRORS!

These .csv files are then forwarded to the system administrator and slotted into the physical and chemical assessment software.