Oracle Darwin

Darwin New Features

Release 3.6

February 2000

Part No. A82855-01

ORACLE

Darwin New Features, Release 3.6

Part No. A82855-01

Contents

4 Text Import Wizard

5 Database Import Wizard

6 Database Export Wizard

7 Missing Values Wizard

8 Key Fields Wizard

9 Model Seeker

10 Model Compare Wizard

11 Clustering Wizard

12 Tips for Dealing with Missing Values

13 Files and Datasets for Practice

Index

Send Us Your Comments

Darwin New Features, Release 3.6

Part No. A82855-01

Oracle Corporation welcomes your comments and suggestions on the quality and usefulness of this document. Your input is an important part of the information used for revision.

- Did you find any errors?
- Is the information clearly presented?
- Do you need more information? If so, where?
- Are the examples correct? Do you need more examples?
- What features did you like most?

If you find any errors or have any other suggestions for improvement, please indicate the document title and part number, and the chapter, section, and page number (if available). You can send comments to us in the following ways:

- DARWINDOC@us.oracle.com
- FAX: 781-684-7738. Attn: Oracle Darwin
- Postal service:
  Oracle Corporation
  Oracle Darwin Documentation
  200 Fifth Avenue
  Waltham, Massachusetts 02451
  U.S.A.

If you would like a reply, please give your name, address, telephone number, and (optionally) electronic mail address.

If you have problems with the software, please contact your local Oracle Support Services.

Preface

Darwin is a data mining application designed specifically to handle multiple gigabytes of data, and to provide answers to complex problems of data classification, prediction, and forecasting.

This manual describes the new features added to Darwin at Release 3.5 and Release 3.6.

Intended Audience

This manual is intended for all users of Darwin software.

Structure

This manual contains thirteen chapters:

Chapter 1	Overview
Chapter 2	New Transforms
Chapter 3	Tree Display
Chapter 4	Text Import Wizard
Chapter 5	Database Import Wizard
Chapter 6	Database Export Wizard
Chapter 7	Missing Values Wizard
Chapter 8	Key Fields Wizard
Chapter 9	Model Seeker

Related Documents

The complete Darwin documentation set at Release 3.6 includes the following manuals, available on the documentation CD:

- *Darwin New Features, Release 3.6* (this manual). Describes the features introduced at Release 3.5 and Release 3.6.

  This manual is a revision of *Darwin 3.5 New Features*, which described the functionality introduced at Release 3.5. *Darwin New Features, Release 3.6* contains updated information about the functionality introduced at Release 3.5 plus new material describing functionality introduced at Release 3.6.

  If you are upgrading from 3.5 to 3.6, you can discard the manual *Darwin 3.5 New Features*; the present manual supersedes it.

- *Darwin 3.6 Release Notes for Solaris.* Describes the release, documents any problems or bugs in the software, and describes changes that occurred after the manuals were finished. There are separate release notes for Solaris and for HP-UX.

- For system administrators: *Darwin Installation and Administration, Release 3.6 for Solaris*. There are separate installation / administration guides for Solaris and for HP-UX.

- *Using Darwin, Release 3.0.1*. A how-to manual; describes the user interface and provides detailed instructions for using it. (*Using Darwin* describes all the features available at Release 3.0.1; together with *Darwin New Features*, you have a complete description of the user interface at Release 3.6.)

- *Darwin Reference, Release 3.0.1* (companion volume to *Using Darwin*). Introduces data mining and Darwin; provides background and conceptual material on data mining, Darwin datasets, the Darwin tools and models, and analyses.

Documentation CD

Darwin documentation is distributed on the documentation CD in PDF and HTML formats. You can read or print the documentation directly from the CD.

To view the PDF files, you will need

- Adobe Acrobat Reader 3.0 or later, which you can download from `www.adobe.com`.

To view the HTML files, you will need

- Netscape 2.x or later, or
- Internet Explorer 4.x or later

Darwin Online Help

Darwin includes extensive online help that can be summoned from a list of contents and from Help buttons or the F1 key on dialog windows. For correct display of Darwin's online help, you need Internet Explorer 4.x. If you do not have it, you can download it from `www.microsoft.com`.

Conventions

The following conventions are used in this manual:

Convention	Meaning
boldface	Darwin commands, menu names, menu items, names of dialogs and screens.
Project > New File	Indicates the path for a command. The example shown means on the **Project** menu, click the **New File** command.
`code`	Data fields and values, special characters, etc., examples of files, data, filenames, and pathnames.
`italics`	Argument names and placeholders in command formats.
`% user input` `system output`	In interactive examples, user input is shown in bold typewriter, and system output is shown in regular typewriter.

1

Overview of Release 3.6

Note that Release 3.6 replaces Release 3.5. Release 3.5 and 3.6 replace the *Release 3.0.2 client* and the *Release 3.0.1 server.*

This manual describes the functionality introduced at Release 3.5 and Release 3.6. For each new feature, it provides a general description of the functionality, followed by a detailed description of how to use it. Together with *Using Darwin* and *Darwin Reference*, both published at Release 3.0.1, you have a complete description of the user interface at Release 3.6.

The new functionality at Release 3.6 is as follows:

- Two new dataset transformations, described in Chapter 2:
  - **Rename**
  - **Computed Fields**
- Graphical display of a tree model, described in Chapter 3:
  - **Tree Display**
- Eight new wizards, accessible from the **Options** menu:
  - **Text Import Wizard** (Chapter 4)
  - **Database Import Wizard** (Chapter 5)
  - **Database Export Wizard** (new with Release 3.6) (Chapter 6)
  - **Missing Values Wizard** (Chapter 7)
  - **Key Fields Wizard** (Chapter 8)
  - **Model Seeker** (with linear and logistic regression models, new with Release 3.6) (Chapter 9)

- **Model Compare Wizard** (Chapter 10)
- **Clustering Wizard** (new with Release 3.6) (Chapter 11)

In addition, this manual includes the following supplementary material:

- A discussion of different ways to deal with missing values (Chapter 12).
- A list of datasets for practice (Chapter 13).

2

New Transforms

Darwin Release 3.6 provides two new dataset transformations, accessible by clicking the **Dataset** menu's **Transform** command or by clicking the **Transform Dataset** icon  .

- **Computed Field** (Section 2.1)
- **Rename** (Section 2.2)

> **Note:** As you know from earlier releases of Darwin, each transformation creates a new dataset, which is not, by default, saved. You can save it by clicking the **Save** check box that appears on every transform dialog. The suggested name for the transformed dataset then becomes editable, and you can change it if you like. See Section 3.3.4 of *Using Darwin* and the current release notes to learn which characters are permitted in the names of Darwin objects.

2.1 Computed Field

The **Computed Field** transform provides the data preparation tools that are necessary for most data mining operations. It allows mathematical or string computations between fields in the dataset and creates an additional field to hold the results.

As for all transforms, you first specify the name of the source dataset, click the name of the transform you wish to perform, and then click **Add**. The dialog that appears prompts you for information needed to perform the selected transform.

The **Computed Field** dialog prompts you for the following input:

- **New field name:** Enter a name for the new field (by default, the name **NewField** appears in the text box). The new field is the column that will contain the results of the operation(s) defined by the expression in **New field expression**. This new field may have all entries the same for each record (e.g., standard deviation), or it may have different values for each record (e.g., sine).
- **New field expression:** Input for this text box is created from the text boxes, operator buttons, and functions displayed below the text box. Be sure to specify the elements in the correct order. See **Notes**, below, for other important tips.
  - **Field:** Contains the names of all the dataset fields. Scroll the list to find the name of the field you want to use in the expression. To move the field name to the **New field expression** box, click the up-arrow to the right.
  - **String Constant:** Select or enter a string that you wish to use in the expression. If you enter a string directly, be sure to enclose it in double quotation marks, e.g., "`blue`". To move the string constant to the **New field expression field** box, click the up-arrow to the right.
  - **Numeric Constant:** Enter a numeric value that you wish to use in the expression. To move the numeric constant to the **New field expression** box, click the up-arrow to the right.
  - **Punctuation Buttons:** Click a punctuation button (comma, percent sign, and open and close parentheses) to enter its symbol in the new field expression.
  - **Operations and Functions:** Select an operator or function to perform by clicking the corresponding operator button (+, -, *, /) or by selecting a function from the **Function** list. See Section 2.1.1, below, for a list of supported functions.

  If you make a mistake or change your mind about the expression you are creating, position the cursor in the text box and make the necessary changes.

Notes:

- **Data Types:** Not all functions can be applied to all fields; the field(s) must have an appropriate data type. For example, you can apply mathematical and statistical functions only to fields that are numbers, not strings.
- **Operator Precedence:** The standard rules of operator precedence apply. It is advisable to surround each expression with parentheses; otherwise, some of the operators may not work as expected.

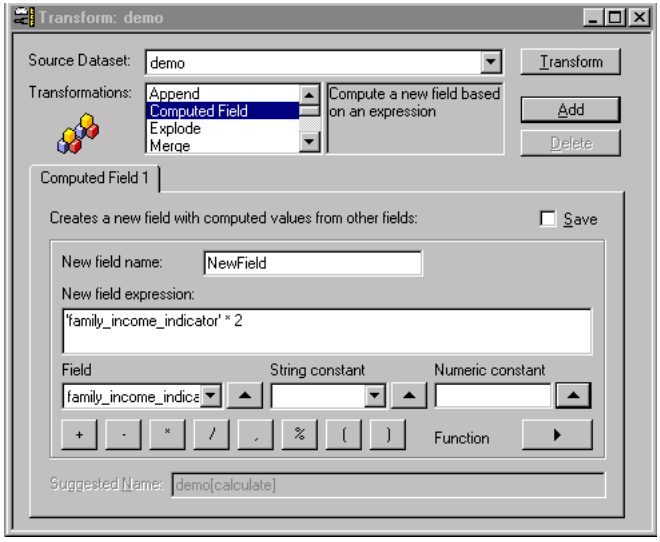
- **Blank Spaces in Data:** Datasets must not contain any blank spaces other than the blank space defined as the delimiter.
- **Multiple Transforms:** Remember that the output dataset of each transform becomes the input dataset for the next transform, so that if you want to perform a second operation on the same dataset that you used for the first operation, you must specify that dataset in the **Source Dataset** box at the top.

If you want to save the dataset, click the **Save** check box. The text box containing the default name for the output dataset then becomes editable, and you can change it if you like.

When you are satisfied that the expression entered in **New field expression** will do what you want, click **Transform** to have Darwin execute the expression and create the new transformation dataset.

The new dataset is then listed in the Darwin Workspace under **Datasets, Transformed**. Its name will be *<fieldname>*`[calculate]` (unless you specified a different name). Double-click the dataset name to display it so you can examine the new field, which will be last column on the right.

The example below shows using **Computed Field** to create a dataset with a new field that holds (for each record) a value that is 2 times the value in `family_income_indicator`.



2.1.1 Functions

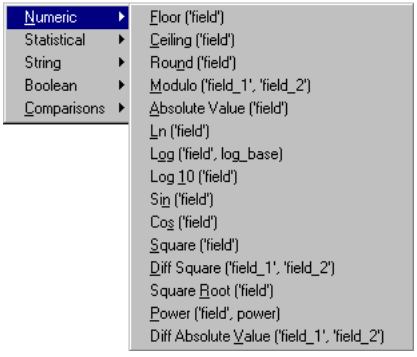
To specify one of the functions displayed on the **Functions** menu, select it from the submenu. For example, click **Function > Numeric > Floor**. Syntax for each function is displayed on the submenu.

For example, to round off the values in a field containing decimal values,

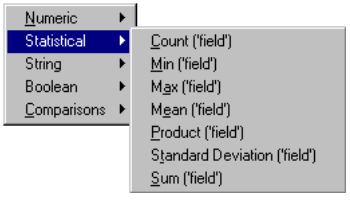
- Specify the function you want to perform (**Function > Numeric > Round**)
- Specify a field that contains decimal values (scroll the drop-down list to find the name of the field, then click the up-arrow to enter that field name in **New field expression**).
- Then click the close parenthesis punctuation symbol to enter it in **New field expression**.
- Click **Transform**.

The functions that you can use with **Computed Field** are as follows:

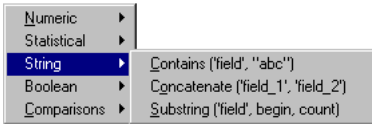
- **Numeric:** Numeric (mathematical) functions for multiple fields (columns): *floor*, *ceiling*, *round*, *modulo*, *absolute value*, *ln*, *log*, *log*$_n$, *sine*, *cosine*, *square*, *difference of squares*, *square root*, *power*, and *difference of absolute values*.



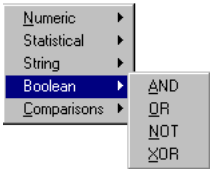
- **Statistical:** Common statistical functions for multiple fields (columns): *count*, *min*, *max*, *mean*, *product*, *standard deviation*, and *sum*. For example, *max* creates a new field that contains, for each record (row), the maximum value of the specified field (column). The new field contains the same value for all records. For example, "sum" results in a new field whose value for each record is the sum of all values in the original field.



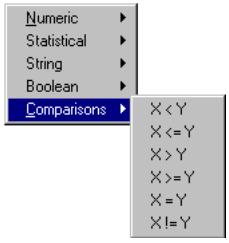
- **String:** String manipulation functions: *contains*, *concatenate*, and *substring*. These functions provide the ability to select records (rows) and compute fields (columns) using string functions and to perform, for example, concatenations of several field string values.



- **Boolean:** Boolean functions: *AND, OR, NOT,* and *XOR.*



- **Comparisons:** *X<Y, X<=Y, X>Y, X>=Y, X=Y,* and *X!=Y*; for example,"age > 45 or blood pressure < 240".



Numeric (Mathematical) Functions

floor
Largest integer smaller than the number; the *floor* of 3.12 is 3.

ceiling
Smallest integer greater than the number; the *ceiling* of 3.12 is 4.

round
round x to the nearest integer; *round(x)* = *floor*(*x*+0.5).

modulo
a modulo b is the remainder that you get when you divide *a* by *b*; for example, 16 *modulo* 3 is 1, because 16/3 is 5, with a remainder of 1.

absolute value
If *x* >= 0, *absolute value* of *x* is *x*; if *x* < 0, *absolute value* of *x* is –*x*.

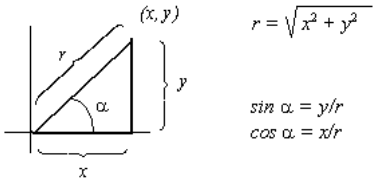
ln
Natural logarithm of a number; *ln a* is the power to which you must raise *e* (a constant) to get *a.*

log
Logarithm base 10 of the number; the power to which you must raise 10 to get the number (log 100 = 2).

$\log_n$
Logarithm to base *n* of a number; the power to which you must raise *n* to get the number (log_2 8 = 3).

sine
The trigonometric function abbreviated *sin* (see diagram below).

cosine
The trigonometric function abbreviated *cos* (see diagram below).



square
square of *a* = *a*a*.

difference of squares
diffsqu is the square of the difference; (a – b)**2..

square root
The number that multiplied by itself gives you *a*; the *square root* of 4 is 2.

power (x, y)
x raised to the *y* power (if *y* is an integer, this means *x* used as a factor *y* times); *power*(2, 3) = 2*2*2 = 8.

difference of absolute values
abs(a) – abs(b); for example,
diff abs val (–10, 5) = *abs*(–10) – *abs*(5) = 10–5=5.

Common Statistical Functions

count
count tabulates the number of nonzero values (in all records) for a given field; the *count* field's value (in every record) is that sum.

min
Smallest number: *min*(–5, 3) = –5.

max
Largest number; *max*(–5, 3) = 3.

product
The number or expression resulting from the multiplication of two or more numbers or expressions.

standard deviation
A measure of variability; specifically, a measure of the dispersion of a frequency distribution that is the square root of the arithmetic mean of the squares of the deviation of each of the class frequencies from the arithmetic mean of the frequency distribution.

sum
The result of adding numbers.

String Manipulation Functions

contains
Returns the boolean value 0 (false) or 1 (true); *contains*('f1', "abc") is 1 if field f1 contains the string "abc", and 0 otherwise.

concatenate
The function *concatenate*('f1', 'f2') creates a new record with the first field concatenated to the first field, second to second, etc. You concatenate two strings by appending the second to the first; for example, "abc" concatenated with "def" is the string "abcdef". "abc" concatenated with itself is "abcabc".

substring
Syntax for *substring* is ('field',0,*n*), which results in a string consisting of the first *n* characters in 'field'. If 'field' contains a string with fewer than n characters, the entire string is selected.

- the *begin* character must be 0, not 1
- the *count* character is the number of characters you want to search on. For example, if the field *marital-status* may have the value *married*, and you enter

  substring ('marital-status',0,4)

  the *substring* will be expressed as *marr*

Boolean Functions

The list below shows what the boolean operators AND, OR, NOT, and XOR do. You can apply these operators only to fields that take on boolean names.

```
0 AND 0 = 0
1 AND 1 = 1
1 AND 0 = 0 AND 1 = 0
0 OR 0 = 0
1 OR 1 = 1
1 OR 0 = 0 OR 1 = 1
0 XOR 0 = 0
1 XOR 1 = 0
1 XOR 0 = 0 XOR 1 = 1
NOT 1 = 0
NOT 0 = 1
```

2.2 Rename

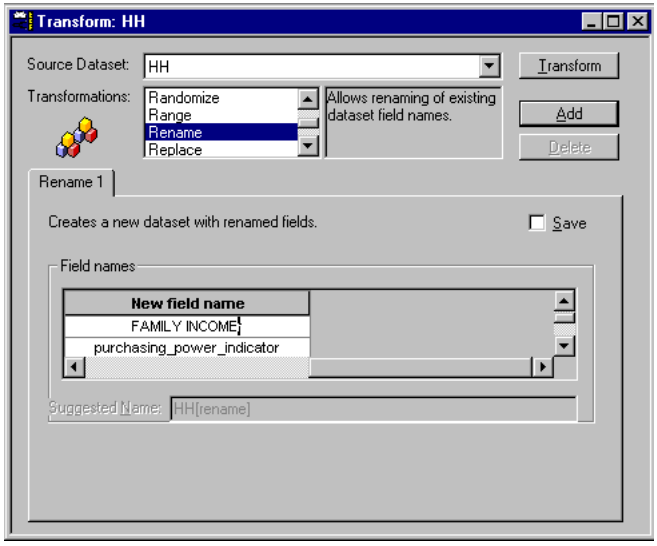
Rename creates a new dataset that contains new names for selected fields. It lets you provide a new name for dataset variables so that on graphs, result tables, and the like, you don't have to use the (usually cryptic) name that is in the database. For example, if the database name for a field is "Cust_Mo_Usage_Param," you might prefer, for display purposes, to rename it something like "CustomerMonths."

Remember that the name must consist of valid characters and must not contain spaces or any of the "special" characters specified in Section 3.3.4 of *Using Darwin*. Also, see the current release notes for any changes in the set of allowed characters.

As for all transforms, you first specify the name of the source dataset, click the name of the transform you wish to perform, and then click **Add**. The dialog prompts you for information specific to the selected transform.

The **Rename** transform dialog displays a two-column window:

- **Current field name:** Displays the current field names.
- **New field name:** Enter the new field name for any field.



Only field names for which a new name is entered on the right will be changed. You can rename several fields at a time.

> **Note:** Remember that different fields in a dataset must have different names. Darwin will not permit you to have two different fields with the same name.

When you have finished providing new names for the fields you want to rename, click **Transform** to have Darwin execute the transformation.

The new dataset is then listed under **Datasets, Transformed**, in the **Workspace**.

3

Tree Display

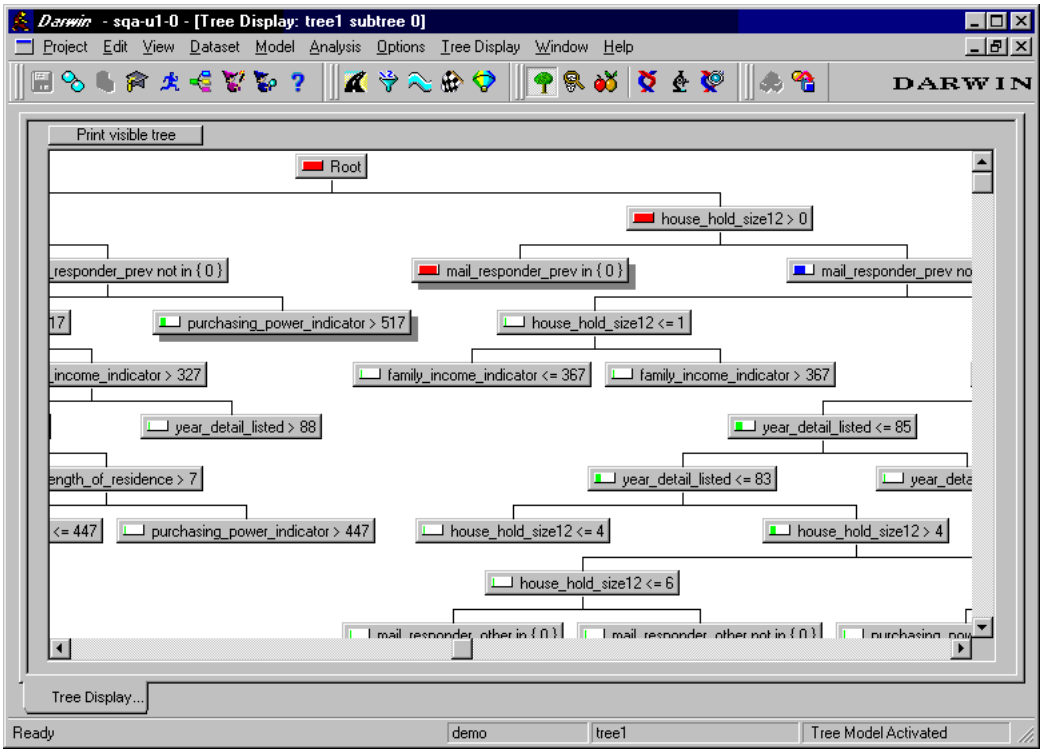
Tree Display, on the **View** menu, provides a graphical representation of a decision tree. It displays the structure of the tree, lets you collapse and expand nodes, and provides basic information about each node, including the rule for the splitting decision at that node. Here's how it works:

- Click **View > Tree Display**.

The dialog that appears prompts you for the following information:

- **Tree Model:** Select the name of the tree model you wish to see displayed; the default is the last tree created.
- You can preselect a tree by clicking its name in the **Workspace** before clicking **View > Tree Display**. (Be sure that you are selecting a tree that is in your current project; if it is not, change your current project to the one that contains the tree you want to display.)
- **Subtree:** Indicate the number of the starting node for the display; the default is node 0.
- Click **OK**. (You can also click **Cancel** to cancel the command, or **Help**, which brings up online help.)

Darwin then displays the tree. A tree display is shown below.



When you invoke **Tree Display**, a new menu item called **Tree Display** appears on the menu bar, between **Options** and **Window**.

The menu provides the following options:

- **Top Down**, which presents the tree branching from top to bottom, i.e., with the root node at the top and splits branching down.
- **Left Right**, which presents the tree branching from left to right, i.e., with the root node at the left and splits branching to the right.
- **Overview**, which produces a second window that displays an overview of the entire tree.

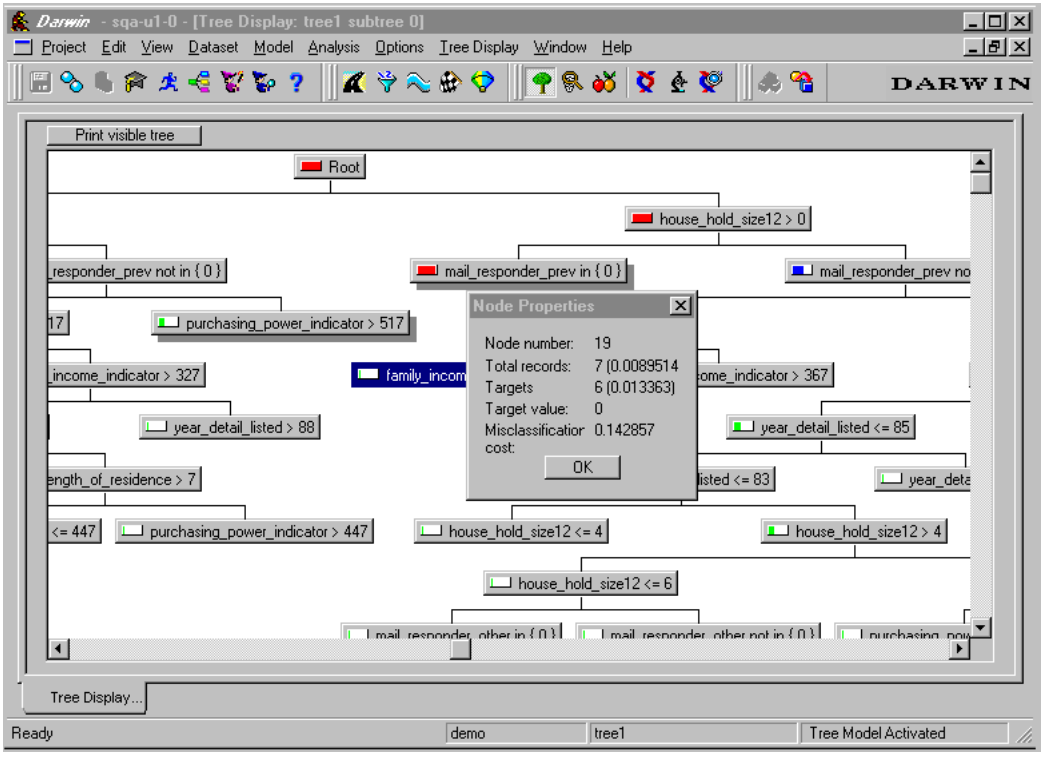
- **Collapse**, which collapses a selected node. See Section 3.1, below, for more information.
- **Expand**, which expands a selected node. See Section 3.1, below, for more information.
- **Full Rule**, which displays the full rule. See Section 3.3, below, for more information.

3.1 Expanding and Collapsing a Node

For categorical trees, like the one shown in the screen captures above and below, each node has a small icon at the left. To expand the node and see the next split, or to collapse a node that is already expanded, click this icon. You can also expand or collapse a node using commands on the context menu (right-click on the node to bring up the context menu).

For regression trees (trees whose target variable contains ordered values), the nodes do not have a clickable icon for expanding and collapsing. To expand a node and see the next split, or to collapse a node that is already expanded, use commands on the context menu.

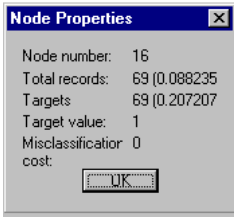
Nodes that have shadows are those that can be expanded.



3.2 Node Properties

Double-clicking a node brings up a property sheet that displays the node's statistics. The statistics displayed depend upon whether the tree is a categorical tree or a regression tree, i.e., whether the target variable is categorical or ordered.

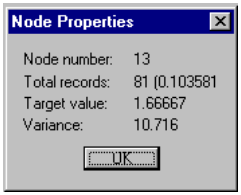
Categorical Trees: The statistics displayed on the node property sheet for a categorical tree are



- node number
- total number of records at this node (expressed both as the actual number and as the fraction of records in the dataset that the number represents)
- target density (the number of records with the target value) and the fraction this represents of records in the dataset with this value

- target value
- misclassification cost

Regression Trees: The statistics displayed on the node property sheet for a regression tree are

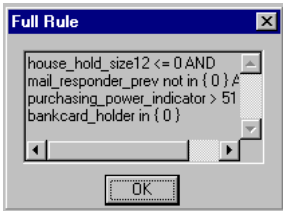


- node number
- total number of records at this node (expressed both as the actual number and as the fraction of records in the dataset that the number represents)
- target value
- variance

3.3 Full Rule

To display the full splitting rule for a node,

- select a node
- right-click to bring up a context menu
- click **Full Rule**



The full rule displayed here is for node 16 (node 16's property sheet is displayed on page 3-16); the full rule is this: household size of 1 or 2 has a probability of less than zero AND previously responded to mail solicitation AND has a purchasing power of more than $51,700 AND does not have a bank card. These rules predict a target value of 1 (see node property sheet above), which is a "yes," i.e., the person is likely to subscribe to the magazine.

See *Darwin Reference*, Chapter 7, for more information on interpreting tree-splitting rules.

4

Text Import Wizard

To invoke the **Text Import** Wizard:

- Click **Options > Text Import**

The **Text Import** Wizard creates a Darwin dataset and a dataset descriptor file from an imported text file.

(Earlier releases of Darwin included `darwinDG`, a UNIX command that created a Darwin descriptor file from a text file. With Darwin 3.5 and 3.6, use of `darwinDG` is deprecated; use the **Text Import** Wizard instead. The **Text Import** Wizard automates the process of creating a Darwin dataset and a descriptor file from an imported text file.)

The Wizard prompts you for the name of the file and provides default names for the dataset and descriptor files. You then provide certain information concerning the text file, such as the delimiter, field names, and the number of unique values a categorical field may have.

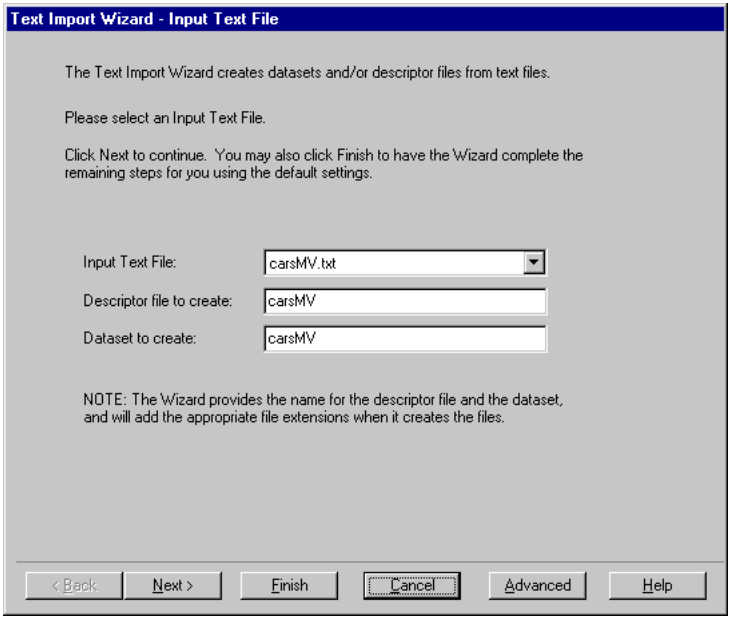
> **Note:** If you are supplying field names from a separate file, you must create that file before you invoke the Wizard. See Section 4.6 for details.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the remaining steps without further input; **Cancel**, which cancels operations; **Help**, which brings up online help; and, on some screens, **Advanced**, which lets you set certain options.

4.1 Text Import Wizard – Input Text File

The first dialog introduces the Wizard, describes what it does, and requests the following information:

- **Input Text File:** Name of the input file. This is the name of a text file in the current project.
- **Descriptor file to create:** Based on the name of the input text file, the Wizard offers a name for the descriptor file that will be created. The name appears without an extension; the extension is added when the Wizard creates the file.
- **Dataset to create:** Based on the name of the input text file, the Wizard offers a name for the new dataset that will be created. The name appears without an extension; the extension is added when the Wizard creates the file.
- Click **Next** to proceed. You can also click **Finish** to have the Wizard perform all the remaining steps and return you to the Darwin client.



4.2 Text Import Wizard – Input Parameters

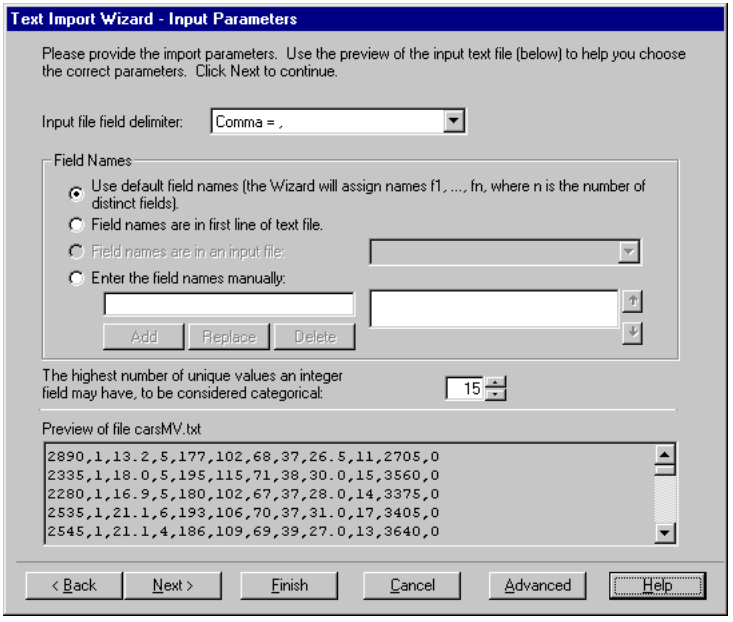
In this step, you specify the parameters of the input file and examine a preview of the text file. If it looks okay to import it as is, click **Next** to proceed. You can also click **Finish** to have the Wizard complete the remaining steps.

If the input parameters require changes before importing the file, make the changes here. You can change the delimiter, specify field names, and you can set a limit on the number of values a categorical field may have.

- **Input file field delimiter:** If the delimiter (field separator) displayed is incorrect, select the correct delimiter from the list.
- **Field names:** Indicate how field names are to be provided; there are four options:
  - Use default field names. The Wizard assigns default names `f1`, `f2`, . . . , `f`*n*, where *n* is the number of distinct fields.
  - Field names are in the first line of the text file.
  - Specify an input file that contains the field names.

    Note: This file must exist before you invoke the wizard; see Section 4.6 for details.
  - Enter the field names by hand.
- Specify the upper limit on the number of unique values a categorical field may have. If an integer field has fewer values than this number, it is considered categorical; otherwise it is considered ordered.

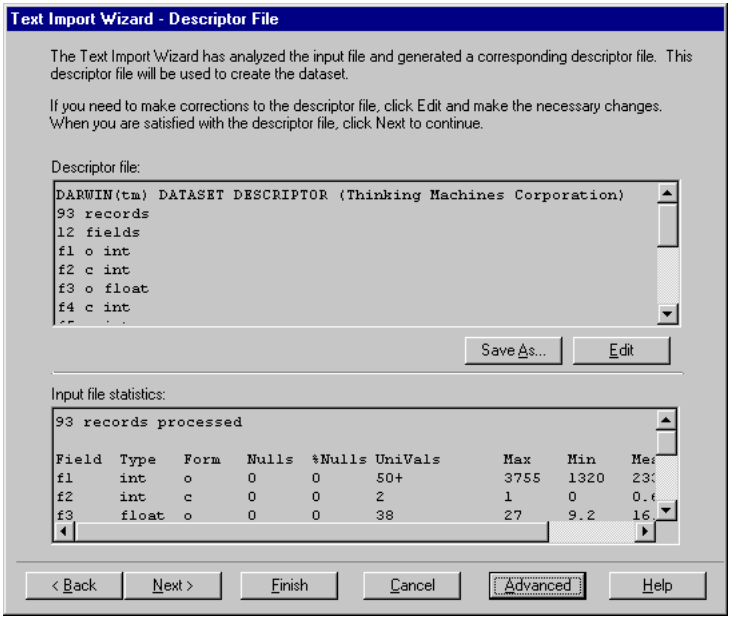
Click **Next** to continue. You can also click **Finish** to have the Wizard complete the remaining steps and return you to the Darwin client.



4.3 Text Import Wizard – Descriptor File

This dialog displays the descriptor file the Wizard has created for the input text file. If you need to make changes to the descriptor file, click **Edit** and make the corrections. When you are satisfied with the descriptor file, save it by clicking **Save As** and giving it a name; then click **Next** to continue.

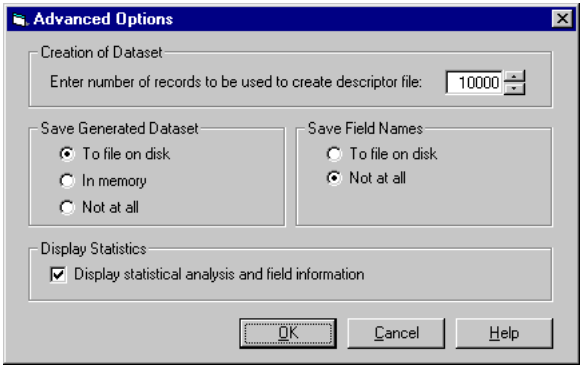
You can also click **Finish** to have the Wizard complete the remaining steps and return you to the Darwin client.



4.4 Advanced Options

Each screen has an **Advanced** button, which takes you to a dialog on which you can specify the following:

- The number of records to be used to create the descriptor file
- How you want the generated dataset saved:
  - to file
  - in memory
  - don't save it
- How you want the field names saved:
  - to file
  - in memory (not an option for field name files)
  - don't save them
- Whether you want statistical analysis and field information displayed.



4.5 Text Import Wizard – Finished

The last screen announces that the Wizard has completed its task, says that "the following descriptor file and dataset were created," and displays the names of the newly created files.

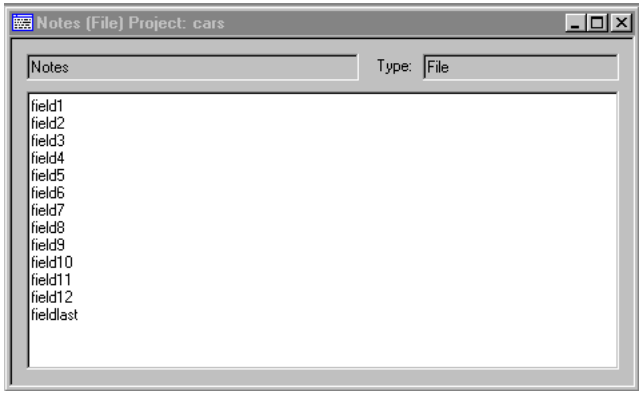
Click **Finish** to update the Darwin **Workspace** and return to the Darwin client. The names of the newly created files appear with the other datasets and descriptor files for this project.

4.6 Supplying Field Names from a File

On the **Input Parameters** dialog, one of the ways to supply field names is to provide the name of an input file that contains the field names. The input file is a Notes file, with a `.nts` extension.

Here's how to create a Notes file containing the field names:

- On the Darwin main window, click **Project > New File**
- The **File Handler** dialog appears. Click **Darwin File**, and the list of Darwin files becomes available.
- From the list of Darwin files, click **Notes**, and then click **OK**. An empty Notes file appears.
- Enter the field names, one to a line. The finished file should contain nothing but a list of the field names, one to a line, as in the example below.



- Click **Project > Save**. A **Save As** dialog appears and offers a default name for the file, which you can accept or change. The default name includes a `.nts` extension. (The file will work properly even without the `.nts` extension.)
- Click **Yes** to save the file.

5

Database Import Wizard

To invoke the **Database Import** Wizard:

- Click **Options > Database Import**

The **Database Import** Wizard creates a Darwin dataset from an Oracle database table. The Wizard takes care of connecting to and disconnecting from the database.

> **Note:** For the **Database Import** Wizard to work, Darwin must be configured for database connectivity. See your system administrator if you have any questions.

The Wizard performs the following functions:

- connects you to the database.
- permits you to log in to the database
- displays available tables in the database
- converts the selected table to a Darwin dataset
- closes the connection to the database

> **Note:** You can only import tables that exist in the database. You cannot create or modify a database table with the Wizard.

You can cancel the import operation at any time.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the

remaining steps without further input; **Cancel**, which cancels operations; and **Help**, which brings up online help.

5.1 Converting Oracle Tables to Darwin Datasets

Converting an Oracle table to a Darwin dataset requires converting Oracle data types to Darwin data types and setting the form of each field in the dataset.

5.1.1 Converting Oracle Data Types

Oracle and Darwin do not support the same data types; for example, Darwin does not support date fields or fields containing national characters. Even when Darwin and Oracle support the same data type, they may not both support the same range of values. You cannot import fields into a Darwin dataset that have a data type that is not fully supported by Darwin.

Table 5–1 summarizes how the Database Import Wizard imports Oracle 8 data types into Darwin.

Table 5–1 Darwin Support for Oracle 8 Data Types

Oracle 8 Data Type	Darwin Support
VARCHAR2	Full range supported; becomes a Darwin string.
NVARCHAR2	Fields containing ASCII characters supported; becomes a Darwin string.
NUMBER	Full range supported; becomes a Darwin double.
LONG	Not supported; import fails with "You have selected a database field of a data type that is not supported in Darwin."
DATE	Imports only the date portion of the field; becomes a Darwin string.
RAW	Darwin does not support a binary data type. The field becomes a Darwin string. You can compare two such imported fields to see if they match exactly; otherwise, it is not clear how to use them in Darwin.
LONG RAW	Not supported; import fails with "You have selected a database field of a data type that is not supported in Darwin."
ROW ID	Supported; becomes Darwin string.
CHAR	Supported; becomes Darwin string.
NCHAR	Fields containing ASCII characters supported; becomes a Darwin string.
MLSLABEL	Not supported.

Table 5–1 Darwin Support for Oracle 8 Data Types

Oracle 8 Data Type	Darwin Support
CLOB	Not supported; import fails with "You have selected a database field of a data type that is not supported in Darwin."
NCLOB	Not supported; import fails with "You have selected a database field of a data type that is not supported in Darwin."
BLOB	Not supported; import fails with "You have selected a database field of a data type that is not supported in Darwin."
BFILE	Not supported.

5.1.2 Setting the Form of Imported Fields

The Database Import Wizard sets the form of fields as follows:

- For numeric items (such as those of type NUMBER in the Oracle table), the form is set to ordered.
- For character items (such as those of type VCHAR2 or CHAR in the Oracle table), the form is set to categorical.

Before you use the dataset in Darwin, check that the assignment of form is correct. It is important that the target variable have the correct form; it may not be necessary for other fields to have the correct form.

If you need to change the form of a field, use the **Set Form** dataset transformation.

5.2 Using the Database Import Wizard

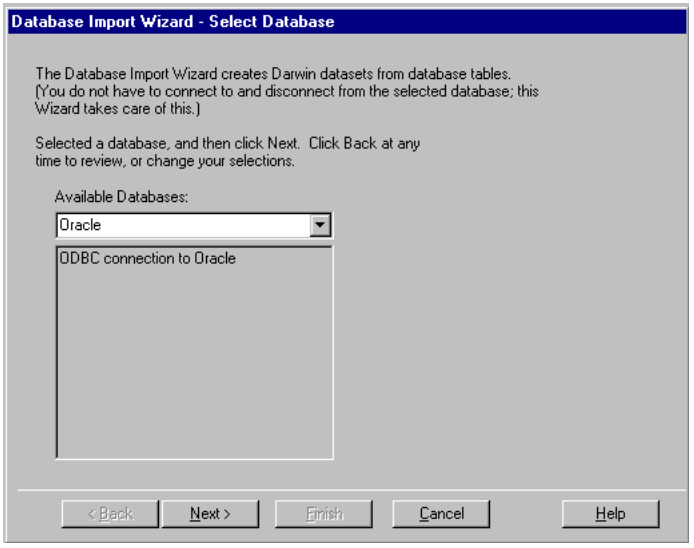
Follow these steps to import an Oracle table to a Darwin dataset:

1. Select the database that contains the table.
2. Select the table to import.
3. Specify a name for the dataset; select the fields to include in the dataset; perform the import.
4. Finish the operation.

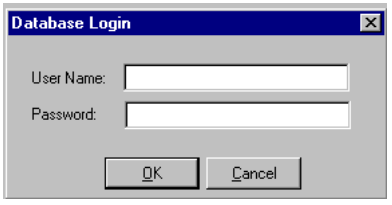
You can cancel the import operation at any step.

5.2.1 Database Import Wizard — Select Database

The first dialog describes what the **Database Import** Wizard does, and displays a list of available Oracle databases. Select a database by clicking its name, and then click **Next**.



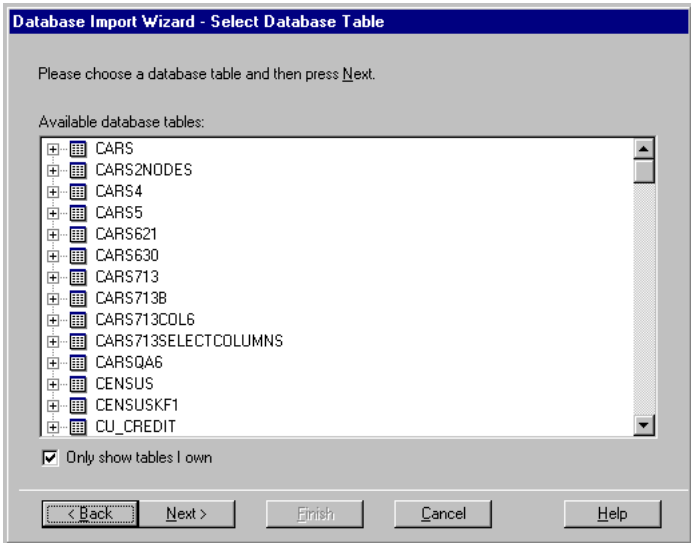
A **Database Login** dialog appears, and prompts you for your login name and password. Enter your login and password, and then click **OK** or press RETURN (if you're not sure what login and password to use, see your database administrator):



5.2.2 Database Import Wizard — Select Database Table

The **Select Database Table** dialog displays the a list of the tables available in the database you selected. The **Only show tables I own** checkbox at the bottom of the

displays controls which tables are displayed; the box is checked by default. Select a database table by clicking its name, and then click **Next**.



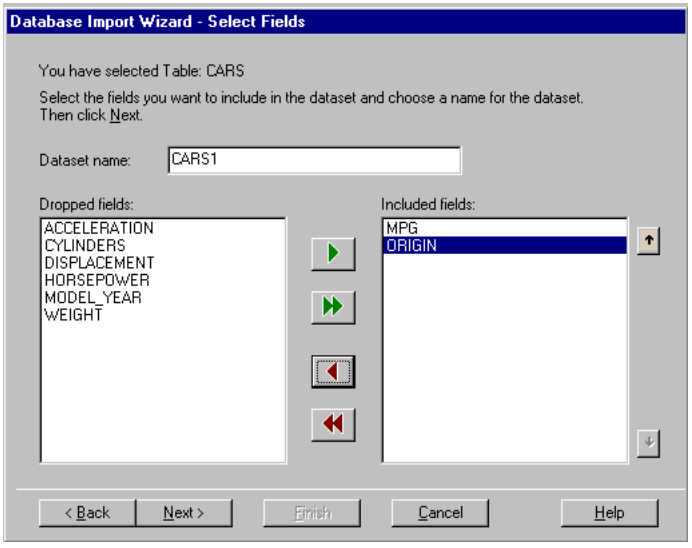
5.2.3 Database Import Wizard — Select Fields

The **Select Fields** dialog confirms your database table selection, offers a name for the dataset, and lets you select fields to be included in the dataset:

- **Dataset Name:** The Wizard offers a default name for the new dataset; you can accept or change the name, as you like.
- **Dropped Fields:** This box will contain fields you want to exclude from the new dataset.
- **Included Fields:** This list initially includes all the fields in the database table. To move a field from the list of included fields to the list of dropped fields, click its name and then click the left arrow. To move all fields, click the double arrow.

  You can move a field up or down in the list by using the up and down arrows at the top and bottom of the right side of the list.

After selecting fields from the CARS database table and dropping all fields except MPG and ORIGIN, the dialog looks like this:



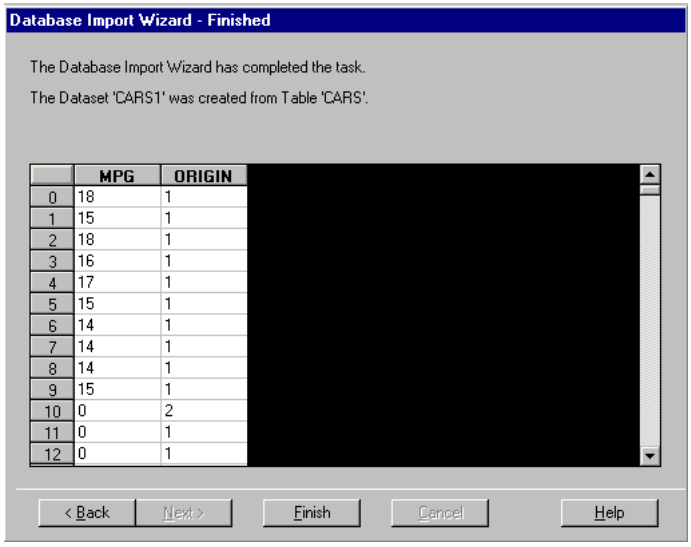
Click **Next** to proceed; the Wizard then creates the new dataset.

5.2.4 Database Import Wizard — Finished

The Wizard reports on the progress of the command while the import takes place.

The last dialog displays the new dataset that the Wizard has created.

Click **Finish** to save the dataset, update the Darwin **Workspace**, and return to the Darwin client.


Database Import Wizard - Finished
The Database Import Wizard has completed the task.
The Dataset 'CARS1' was created from Table 'CARS'.
MPG
ORIGIN
0 18 1
1 15 1
2 18 1
3 16 1
4 17 1
5 15 1
6 14 1
7 14 1
8 14 1
9 15 1
10 0 2
11 0 1
12 0 1
< Back
Next >
Finish
Cancel
Help

6

Database Export Wizard

To invoke the **Database Export** Wizard:

- Click **Options > Database Export**

The **Database Export** Wizard lets you export a Darwin dataset or result table to a table in an Oracle database.

The Wizard performs the following functions:

- displays the list of datasets or result tables in the current active project
- permits you to select particular fields to export
- connects you to the database that you select
- permits you to specify how to create the target table
- writes the selected dataset or result table to a table in the database
- closes the connection to the database and returns you to the Darwin client

You can cancel the export operation at any time. If the operation is cancelled after it is underway, the cancel may take a long time to complete because Oracle has to restore the database to a consistent state.

> **Note:** For the **Database Export** Wizard to work, Darwin must be configured for database connectivity. See your system administrator if you have any questions.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the

remaining steps without further input; **Cancel**, which cancels operations; and **Help**, which brings up online help.

6.1 Database Export Wizard Limitations

Darwin datasets and Oracle tables have different size limitations and data types. Because of these differences, a Darwin dataset and the corresponding exported table are not identical.

6.1.1 Data Type Differences

Darwin and Oracle do not support the same data types. For example, Darwin supports an integer data type which Oracle does not support. For a Darwin dataset field that is of type integer, the corresponding field in the exported Oracle table will be of type NUMBER, a floating-point number.

Each field in a Darwin dataset has a form (ordered or categorical); there is no corresponding field property in an Oracle table.

6.1.2 Result Table Limitation

The **Database Export** Wizard exports all fields in a result table as character strings, even if they appear to be numeric. This limitation does not apply to datasets.

6.1.3 Limit on Number of Fields

Oracle 7.x supports a maximum of 256 fields. Darwin datasets can contain more than 256 fields. You cannot export a Darwin dataset with more than 256 fields to an Oracle table. You can work around this limitation by dividing the Darwin dataset into subsets of 256 fields and exporting each subset to a separate table; the tables will be most useful if each table includes all key fields.

Similarly, Oracle 8.x supports a maximum of 1000 fields. You cannot export a Darwin dataset with more than 1000 fields to an Oracle 8 database. You can work around this limitation by dividing the Darwin dataset into subsets of 1000 fields and exporting each subset to a separate table; the tables will be most useful if each table includes all key fields.

6.2 Database Export Wizard – Select Data

This first dialog tells what the Wizard does and displays a list of datasets in the current active project (the default) or a list of result tables in the current active project. To see a list of result tables, click the option button next to **Result table**.

You can export only one dataset or result table at a time.

To select the dataset or result table to export, click its name and then click **Next**.

6.3 Database Export Wizard – Select Database

The **Select Database** dialog displays a list of **Available Databases** to which you can export the selected dataset or result table. Only Oracle databases referenced in the `.odbc.ini` file in your UNIX home directory are displayed.

Select the database by clicking its name; then click **Next**.

The **Database Login** dialog appears. Enter your login name and password for the database, and then click **OK**.

6.4 Database Export Wizard – Select Database Table

You can write the selected dataset or result table to a new table (the default), replace an existing table, append to an existing table, or update an existing table.

6.4.1 Select Tables to Display

The checkbox below the list of tables lets you view either

- all tables owned by the login name that you used when you connected to the database (the default)
- all tables in the database that you have permission to view. (Tables that you don't have permission to view will not appear in the list.)

6.4.2 Write to a New Table

To write to a new table, specify the name of the new table.

Click **Next**. The Wizard displays the **Select Fields** dialog described in Section 6.5.

6.4.3 Replace or Append to an Existing Table

To replace an existing table or to append to an existing table, click the appropriate option button, and select the table to overwrite or append to.

Click **Next**. The Wizard displays the **Select Fields** dialog described in Section 6.5.

Darwin does no processing to the dataset or result table other than exporting it. An append operation will succeed only when the data and the target table are compatible; the data being exported and the existing table must consist of identically named fields with equivalent data types.

6.4.4 Update an Existing Table

To update an existing table, click the appropriate option button, and then click the name of the table in the list.

Click **Next**. The Wizard displays the **Select Update Fields** dialog.

6.4.4.1 Select Update Fields

When you are updating an existing table, you must specify one or more keys that identify the records to update; you must also specify the field(s) to update.

Notes: You must specify key fields that *uniquely* identify the records to update. If the records are not uniquely specified, you may corrupt the table that you are updating.

We *strongly* recommend that users create an index on the key columns. (Consult your database administrator.)

Specify at least one key field and at least one updated field by clicking the field name(s) and the appropriate arrow.

Click **Next**.

If you are exporting at least one field that contains numeric values, the Wizard displays the **Set Numeric Handling** dialog, described in Section 6.6. Note that you do not set numeric handling when you export a Darwin result table, because all fields in result tables are exported as strings.

6.4.4.2 Troubleshooting Updates

If the update seems to take a disproportionately long time, you should confirm that there is an index on the key fields that are specified and that the key fields appear in

the selected list in the same order as they appear in the index. You should then modify the `.odbc.ini` file on UNIX to enable ODBC tracing. Then run the command again. Open the ODBC trace file and find one of the UPDATE lines that write to the table you selected. Then issue an ANALYZE PLAN from `sqlplus`, and see if the index is actually being used. You may need the assistance of your database administrator.

6.5 Database Export Wizard – Select Fields

You can export all fields or just specific fields of the dataset or result table. The default is to export all fields. Move the fields that you do not wish to export from the **Selected fields** list to the **Available fields** list using the arrow buttons between the fields. To move all fields, use the appropriate double arrow; to move one field, click it and use the appropriate single arrow. When you are finished selecting the fields to export, click **Next**.

If you are exporting at least one field that contains numeric values, the Wizard displays the **Set Numeric Handling** dialog, described in Section 6.6. Note that you do not set numeric handling when you export a Darwin result table, because all fields in result tables are exported as strings.

6.6 Set Numeric Handling

Darwin datasets support a larger range of floating-point numbers than Oracle databases do. The Set Numeric Handling dialog permits you to specify how to handle floating-point values that Oracle databases do not support.

Note that you do not set numeric handling when you export a Darwin result table, because all fields in result tables are exported as strings.

You can select one of the following ways to handle out-of range values:

- Constrain the value to the database limit (the default)
- Skip (do not export) any row containing an out-of-range value
- Cancel the export

After you make a selection, click **Finish** to perform the export.

6.7 Finish the Export

The Wizard performs the export and displays progress during the export operation.

If there are problems with the export, the Wizard displays a message describing the problem.

If the export succeeds, the message “Data export completed successfully” is displayed. Click **Close**. The Wizard disconnects you from the database and returns you to Darwin.

7

Missing Values Wizard

To invoke the **Missing Values** Wizard:

- **Click Options > Missing Values**

The **Missing Values** Wizard finds all instances of missing values and gives you several options for treating them. Treatment options are categorized as record-wise or field-wise, depending on whether the focus of the treatment is a record or a field. The outcome of the process is a new dataset; the original dataset is not changed.

Note that a field value is considered missing if the field value is <null>, that is, if there is no value in the field at all. Darwin does not permit you specify that some value other than <null> indicates a missing value. For general information about missing values and some tips on different ways to deal with them, see Chapter 12.

Record-wise treatment has to do with *dropping* a record. Record-wise treatment locates all instances of missing values and lets you specify the rules that determine whether to drop a particular record that has missing values. For example, you can specify that a record is to be dropped if it is blank in every field, or if it is blank in specified fields, or if it is blank in either of certain specified fields. The record-wise treatment produces a new dataset plus a report showing the number of input records and the number of output records.

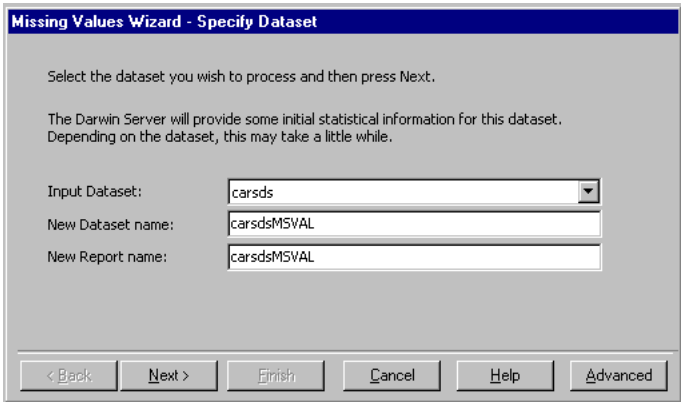
Field-wise treatment has to do with *replacing* missing values. Field-wise treatment locates all instances of missing values and lets you substitute various values (maximum, minimum, mean, or a user-specified value) for the missing value. You can also specify a default treatment for all fields of a particular type, e.g., all categorical numeric fields, all categorical string fields, and all ordered fields. The default treatment is *no* treatment, i.e., leave the field value <null>. The field-wise treatment produces a new dataset plus a report showing which fields were treated and by which option.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the remaining steps without further input; **Cancel**, which cancels operations; **Help**, which brings up online help; and, on some screens, **Advanced**, which lets you set certain options.

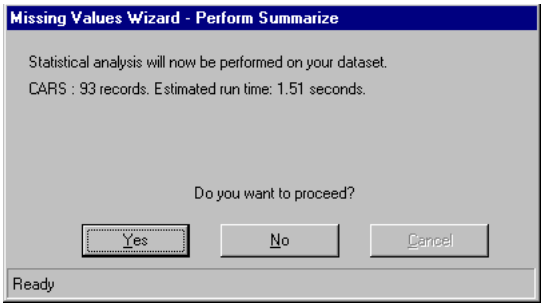
7.1 Missing Values Wizard – Specify Dataset

The first step is to specify the dataset whose missing values you want to treat. By default, the dataset specified in the **Input Dataset** box on the dialog is taken from your current project. You can preselect a specific dataset by clicking its name in the Darwin **Workspace** before starting the wizard.

The **New Dataset name** and **New Report name** boxes contain suggested names for the output. You can accept the suggested names or change them, as you wish.



Click **Next** to proceed. A dialog announces that statistical analysis will be performed on your dataset, and gives you an estimate of how long this will take. Click **Yes** to proceed (or **No** to return to the first dialog).



If you click **Yes**, the Wizard performs the analysis (Summarize). When the analysis is complete, the **Select Treatment** dialog appears.

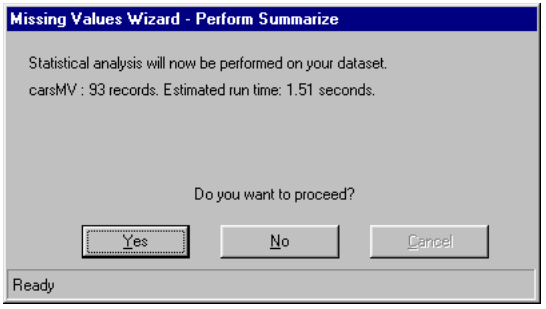
7.1.1 Missing Values Wizard — Advanced Options

An **Advanced Options** command button appears on several of the dialogs. Clicking this button brings up a dialog that lets you specify whether and how you want results saved. Options are as follows:

- Save Generated Dataset
  - To file on disk
  - In memory
  - Not at all
- Save Report File
  - To file on disk
  - Not at all

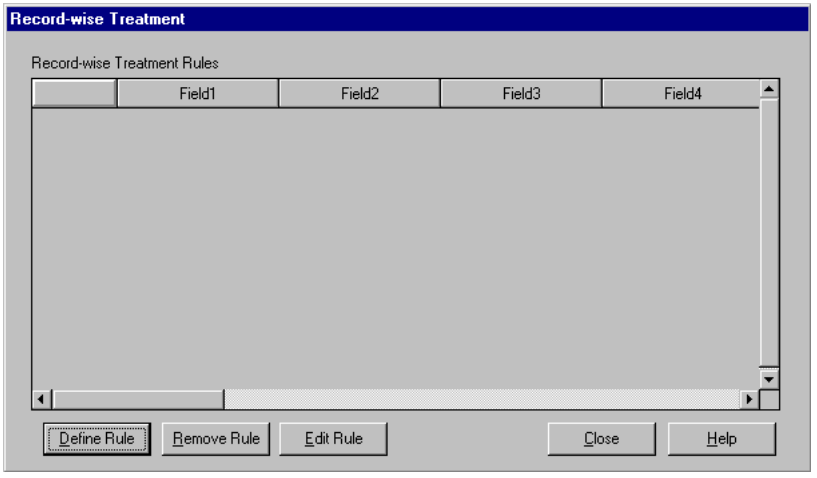
7.2 Missing Values Wizard – Select Treatment

The **Select Treatment** dialog lets you select record-wise or field-wise treatment. Click **Record-wise** or **Field-wise** to bring up a dialog on which you specify the details of the treatment. You'll then return to this dialog to either select another treatment or to click **Next** to have the Wizard perform the treatment(s) you have specified.



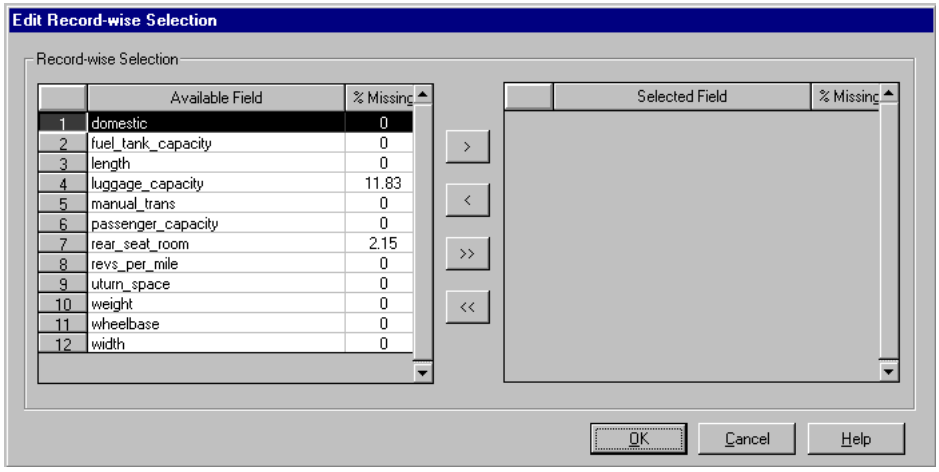
7.2.1 Record-wise Treatment

Clicking **Record-wise** brings up the **Record-wise Treatment** dialog, where you specify the rules for dropping records with missing values.



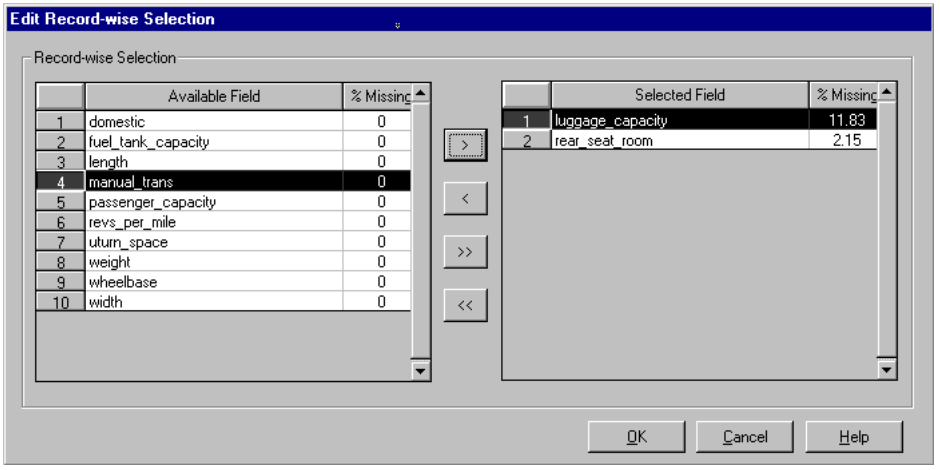
To specify a rule, click **Define Rule**. The **Edit Record-wise Selection** dialog appears, displaying the fields (in alphabetical order) and, for each field, the

percentage of missing values. For example, in the screen capture below, two fields have missing values.

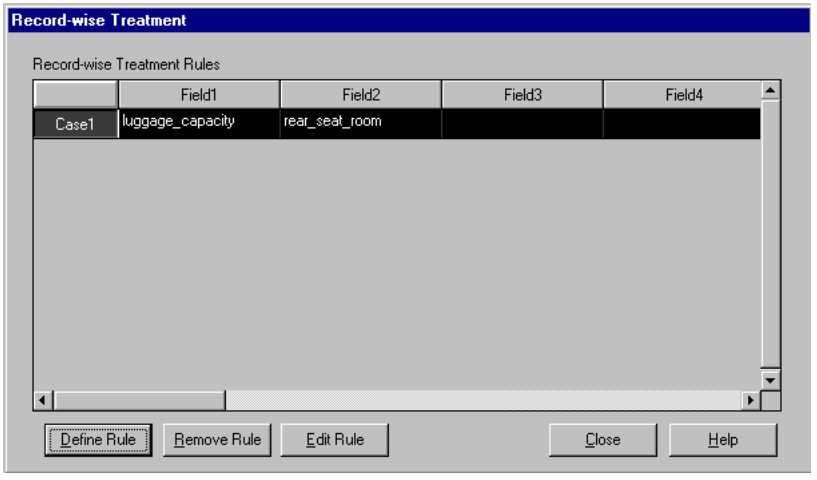


To specify the rule that a record with missing values in a particular field will be dropped, click the field's name and then click the arrow to move it to the **Selected Field** list.

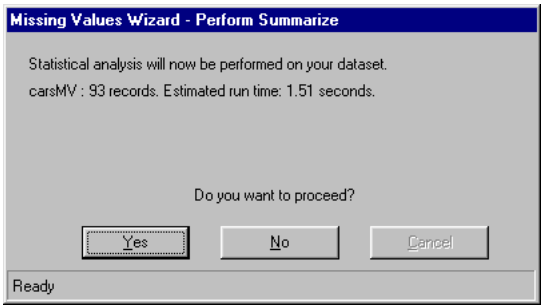
Specifying two or more fields in a rule means that the Wizard is to drop records that have missing values in *both* selected fields. To specify that the Wizard drop a record if it has missing values in one field *or* another field, define two rules, one for each field.



When you have finished specifying the rules for record-wise treatment, click **OK**. The R**ecord-wise Treatment** dialog appears, displaying the rule(s) you have specified. Here we have specified one rule, which the Wizard calls **Case 1**.



Click **Close** to return to the **Select Treatment** dialog.



On the **Select Treatment** dialog, you can either click **Field-wise** to specify field-wise missing values treatment(s) (see Section 7.2.2) or, if there are no field-wise treatments you want to specify, and you have finished specifying treatments, click **Next** to have the Wizard perform the treatment(s) you have specified (see Section 7.3).

7.2.2 Field-wise Treatment

Field-wise treatment lets you treat missing values by field. Treatment consists of replacing a missing value with another value, a value that you specify. Clicking **Field-wise** on the **Select Treatment** dialog brings up the **Field-wise Treatment** dialog.

You can specify treatments by *field type* and by *field*, i.e., you can establish default treatments for field values of a particular type, and you can specify individual treatments field by field.

At the top of the dialog, in the **Field-wise Treatment Defaults** box, are three edit boxes in which you specify the default treatment for three field types:

- For *categorical numeric* fields, options for the default treatment are **Average**, **Minimum**, **Maximum**, **Plurality**, and **None** (default).
- For *categorical string* fields, options for the default treatment are **Plurality** and **None** (default).
- For *ordered* fields, options for the default treatment are **Average**, **Minimum**, **Maximum**, and **None** (default).

Below this is a table that displays all the fields in the input dataset, along with some information about each field. Field(s) with missing values are at the top of the list; the percentage of missing values is shown in the third column.

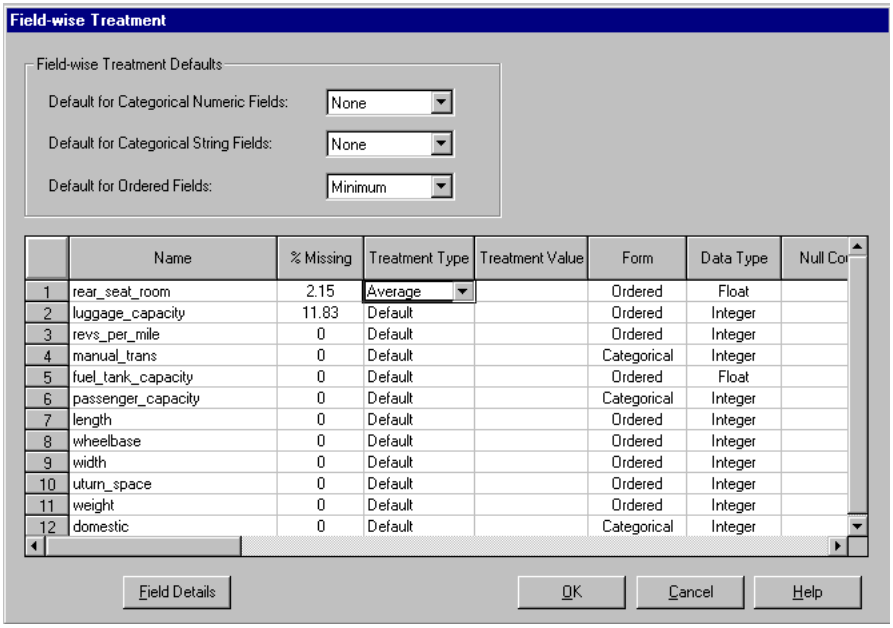
The fourth column, **Treatment Type**, displays the treatment for each field. Note that the treatment entered for each is **Default**, which refers to the default treatment settings specified for the three field types in the **Field-wise Treatment Defaults** boxes at the top.

The default treatment for all three types is, by default, **None**, which means no treatment. Thus, when the **Treatment Type** for a *particular* field is **Default** and the **Default Treatment** for the field's *type* is **None**, nothing is done to the field, i.e., its contents remain the same (`<null>`) as in the original dataset.

You can change the entries in the **Treatment Type** column so that they do not refer to the default settings for field type. Double-clicking an entry produces a list of options from which you can choose the replacement value for that field (the options available depend on the field type):

- **Average:** A missing value will be replaced by the average value for the field.
- **Minimum:** A missing value will be replaced by the minimum value for the field.
- **Maximum:** A missing value will be replaced by the maximum value for the field.
- **User Value:** A missing value will be replaced by a value you enter in the next column, **Treatment Value**. (If you forget to enter a value in the **Treatment Value** column, the Wizard will remind you.)
- **None:** A missing value will not be replaced (it will remain `<null>`).

The screen capture below shows **Average** as the treatment type selected for one of the fields, `rear_seat_room`. For the second field, `luggage_capacity`, the entry remains **Default**; however, the default for this field type (ordered) has been changed to **Minimum**. When the treatment is performed, the average value for the field will replace a missing value in the first field, and the minimum value for the field will replace a missing value in the second field.



Field Details

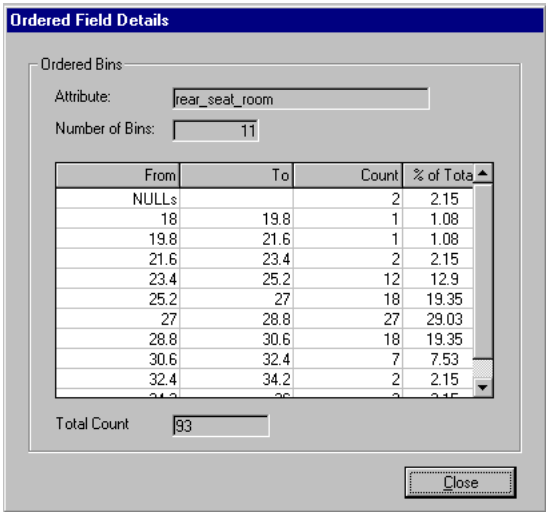
To see details of a selected field, click **Field Details**. The content of the next dialog depends on whether the field contains ordered or categorical values:

- If the selected field contains ordered values, the dialog announces that binning will now be performed for the selected field, and gives you an estimate of the run time.
- If the selected field contains categorical values, the dialog announces that a histogram will be produced for the selected field, and gives you an estimate of the run time. (A maximum of 200 bins are created for a categorical value. This prevents server error if there are thousands of values.)

To proceed, click **Yes**.

The Wizard then displays details about the field. For categorical fields, the information includes the frequency and percentage of each value. For ordered fields, the information includes the number of bins, the starting and ending value for each bin, the frequency in each bin, and the percentage of the total for each bin.

The screen capture below shows field details for an ordered field, `rear_seat_room`.



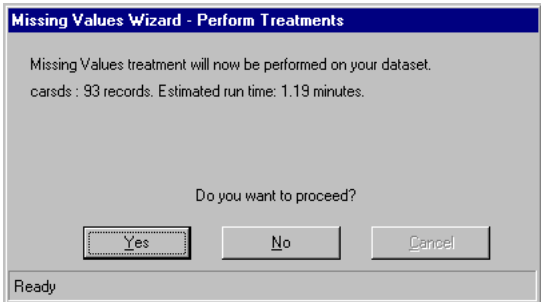
Click **Close** to return to the **Field-wise Treatment** dialog. There you can select another field and click **Field Details** to view details of that field.

When you are finished examining the details of individual fields, you can (on the **Field-wise Treatment** dialog) click **OK** to return to the **Select Treatment** dialog.

On the **Select Treatment** dialog, you can select **Field-wise** or **Record-wise** to specify additional missing values treatments, or, if you have finished specifying treatments, click **Next** to proceed and have the Wizard perform the missing values treatments you have specified.

7.3 Missing Values Wizard – Perform Treatments

The **Perform Treatments** dialog announces that the selected missing values treatments will now be performed, and gives you an estimate of the run time. Click **Yes** to proceed.

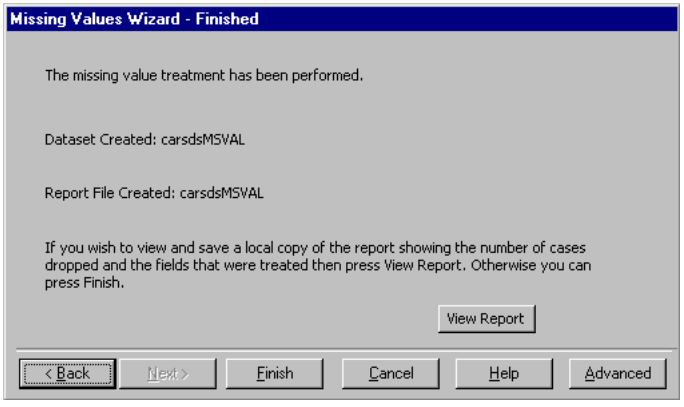


The dialog displays a progress bar; with details displayed in the status bar at the bottom.

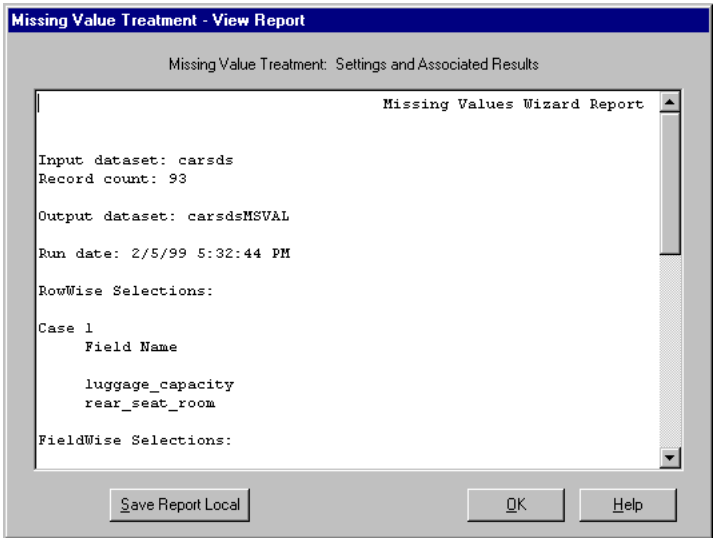
Note that if you have specified both record- and field-wise treatments, record-wise treatments are performed first.

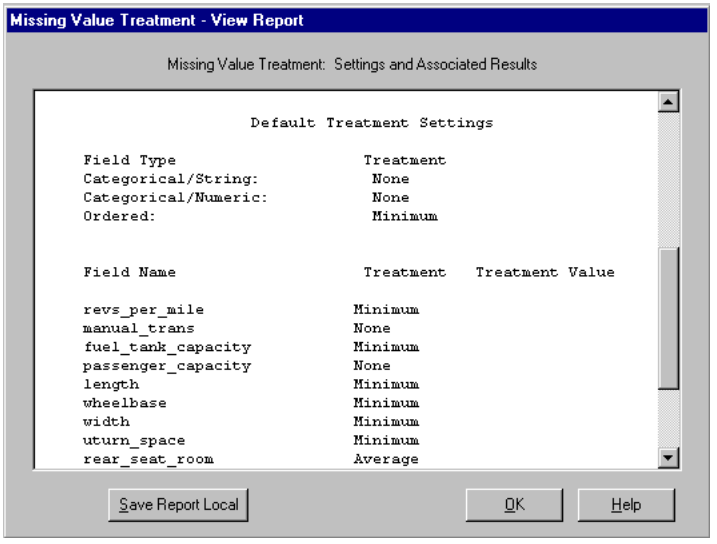
7.4 Missing Values Wizard – Finished

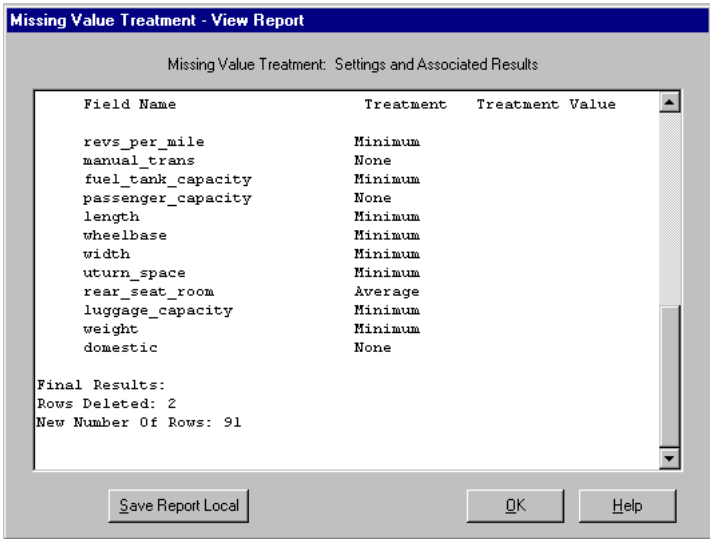
When the Wizard has finished performing the specified missing values treatments, the **Finished** dialog announces this, and displays the name of the output dataset and report file. The Wizard also gives you the option of viewing and saving a local copy of the report that summarizes the results of the treatment.



To view this report, click **View Report**.







Click **OK** to return to the **Finished** dialog.

On the **Finished** dialog, click **Finish** to update the Darwin **Workspace** and return to the Darwin client. When you click **Finish**, a dialog announces that the treated dataset will now be saved to disk, and gives an estimate of how long this will take. Click **Yes** to proceed.

The new dataset is listed in the **Workspace** under **Datasets, Transformed**, with MSVAL appended to its name. If you have run the **Missing Values** Wizard on this dataset more than once, there will be a number following MSVAL

8

Key Fields Wizard

To invoke the **Key Fields** Wizard:

- Click **Options > Key Fields**

The **Key Fields** Wizard identifies the dataset fields that are likely to be most important in a model's performance, then creates a new dataset that contains only those fields.

The Wizard accomplishes this by using a user-specified dataset to build C&RT models iteratively, identifying the important fields. At the next iteration, it removes those fields, and builds a new model with the remaining fields. The Wizard repeats this process until the remaining fields are equally important or unimportant. This process uncovers fields whose importance is masked by fields with greater importance. For example, *age* may be important, but if *income* is more important, *age*'s importance may be hidden until *income* is removed.

When this process is complete, the Wizard displays a list of all fields in the dataset, showing for each a value representing its relative importance and a recommendation to keep or drop the field when creating the new dataset. You can override a recommendation to drop or keep any field.

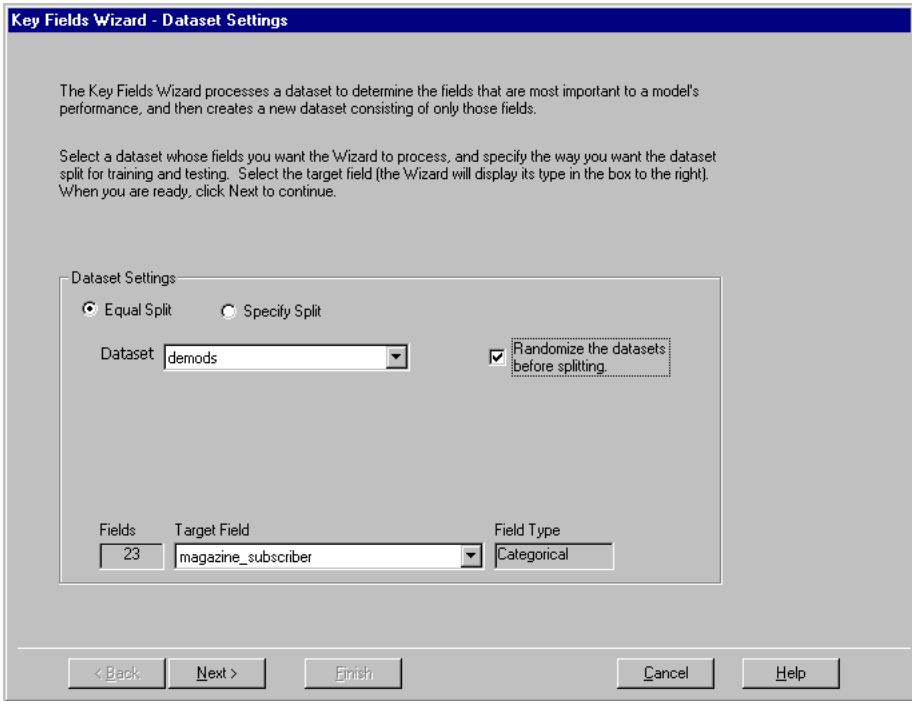
You can expect that a model built with the new dataset will give you faster and more accurate predictions.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the remaining steps without further input; **Cancel**, which cancels operations; **Help**, which brings up online help; and, on some screens, **Advanced**, which lets you set certain options.

8.1 Key Fields Wizard – Dataset Settings

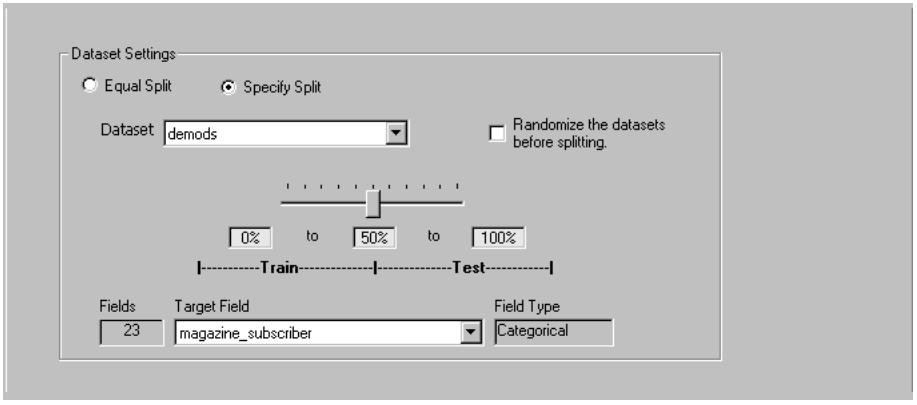
The first screen introduces the Wizard, tells you what it does, and prompts you for the following information:

- **Dataset:** Specify the dataset whose fields you want the Wizard to process.
- **Randomizing:** Click the check box to have the Wizard randomize the dataset before splitting.
- **Equal Split:** Clicking this option means the input dataset will be divided into two equal parts, to be used as the Train and Test datasets. The screen capture below shows this option selected.
- **Specify Split:** Clicking this option lets you indicate the proportions to be used as the Train and Test datasets. The next screen capture shows this option selected.



- **Number of Fields:** This is a read-only box that displays the number of fields in the dataset.
- **Target Field:** Specify the target field.
- **Field type:** This is a read-only box that displays the target field's type (categorical or ordered), once you specify the target field.

The screen capture below shows the dialog with the **Specify Split** option selected. Use the slider bar to indicate the proportions to be used for the Train and Test datasets.



When you have supplied the required information, click **Next** to proceed.

8.2 Key Fields Wizard – Determine Importance

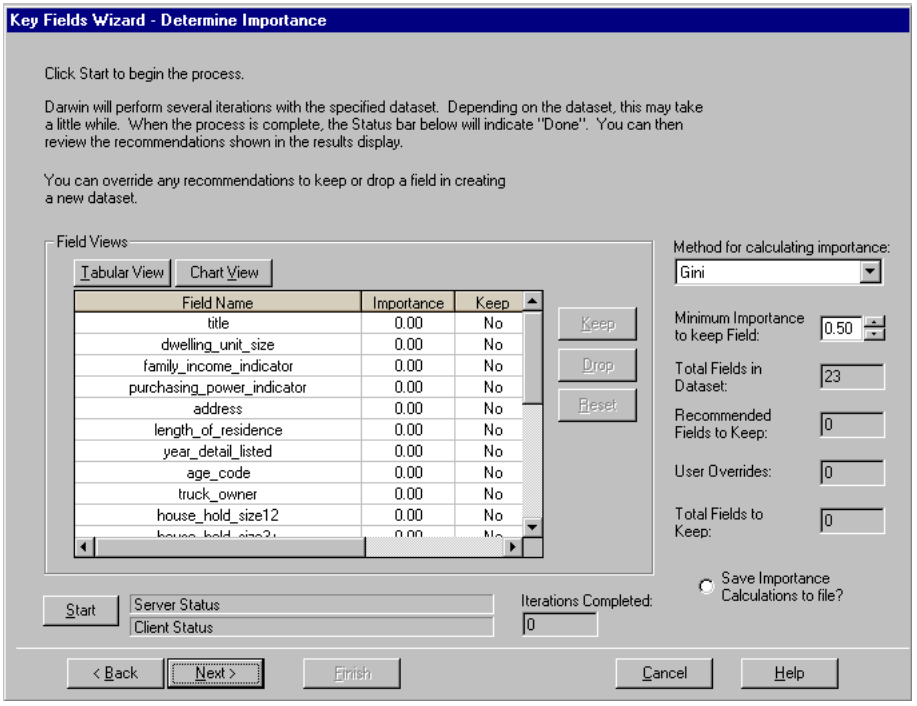
In this step, the Wizard performs several iterations with the specified dataset — as many iterations as are necessary to determine which fields are important, how important they are, and to get to the point where the remaining fields are equally important or unimportant. Depending on the dataset, this process can take a little while.

The screen capture below shows what this dialog looks like before the process has started. The table in the middle lists all fields in the dataset; each has a starting importance value of 0.00, and no field is yet recommended to keep.

To the right are several bits of useful information:

- The method used for calculating importance; the default is Gini. The alternative is Cost.
- The minimum importance value for keeping a field. The default value is 0.50; you can adjust this value downward or upward, to keep or drop more fields.
- The total number of fields in the dataset.
- The number of fields the Wizard recommends to keep.
- The number of Wizard recommendations you have elected to override.
- The total number of fields to keep in creating the new dataset.

If you want to save the importance calculations to a file, click the option button.



8.2.1 Start

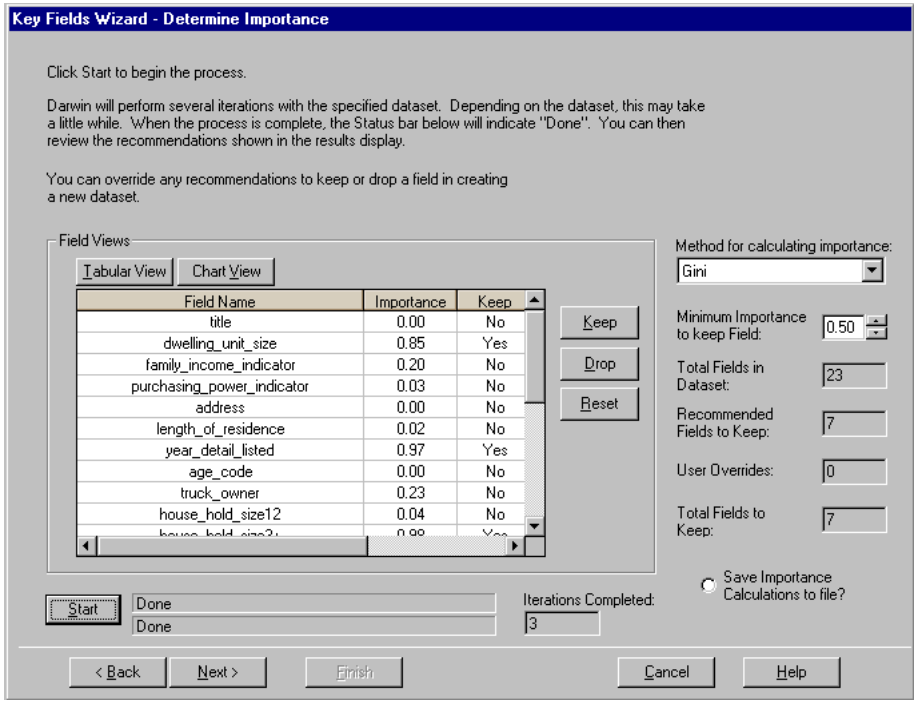
Click **Next** or **Start** to have the Wizard begin the process. The label on the command button changes to **Stop**; when the process is complete, the label changes back to **Start**.

Note the status reports in the status bars and the number of iterations completed. When the Wizard has completed all the iterations necessary, the status bars read "Done."

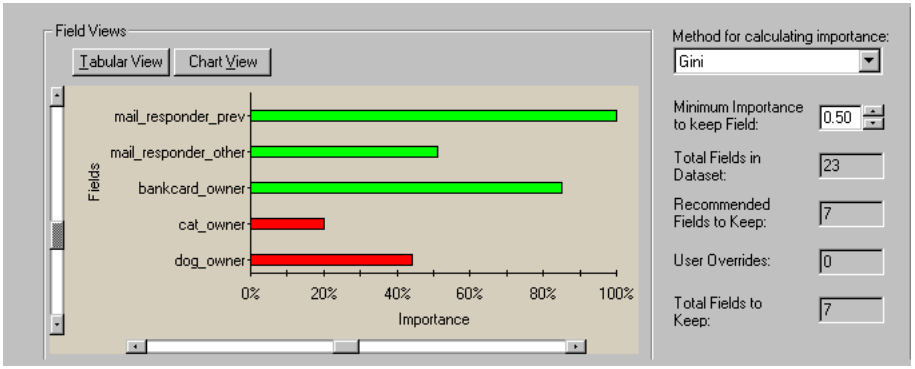
The screen capture below shows the results.

- **Fields, Importance, and Keep:** Each field now has a value between 0 and 1 representing its relative importance, and a recommendation to keep or drop the field when creating the new dataset. You can override any recommendation by selecting the entry in the **Keep** column and clicking **Keep** or **Drop**. Click **Reset** to undo your changes and restore the original entries.

  Note that the importance calculations show you not only which fields are important, but also show you their *relative* importance.



- **Chart View:** You can see the results displayed in graph form by clicking **Chart View**. The Wizard displays a bar chart showing the importance for each field.



  Scroll down the *y* axis to bring all fields into view. Fields that the Wizard recommends keeping are shown in green; fields the Wizard recommends dropping are shown in red.

  You can change a field's keep / drop status by clicking its bar. The corresponding change is made in the **Keep** column of the tabular view.

- Click **Tabular View** to return to the results table.

When you are ready, click **Next** to have the Wizard create the new dataset that contains only the fields with **Yes** in the **Keep** column.

8.3 Key Fields Wizard – Finished

This dialog tells you that the **Key Fields** Wizard has completed its task, has created and saved the new dataset, and displays the name under which the dataset is saved.

If the Wizard did not find any difference in the importance of the dataset's fields, it reports this to you and does not create a new dataset.

Click **Finish** to update the Darwin **Workspace** and return to the Darwin client.

The new dataset is listed in the **Workspace**, under **Datasets, Transformed**, with `KeyFields` appended to its name. If you have run the Key Fields Wizard on this dataset more than once, there will be a number following `KeyFields`.

9

Model Seeker

To invoke **Model Seeker**:

- Click **Options > Model Seeker**

Note that Linear and Logistic Regression models (Section 9.1.2) are new with Release 3.6 (they were not part of Release 3.5).

Model Seeker allows you to set combinations of specifications for different models, build the models, and compare them. You can then save the model(s) that perform the best.

If you have used Darwin before, you are probably aware that Darwin uses the training and testing datasets differently in building the different types of model. With **Model Seeker**, these differences are specifically as follows:

- **Tree:** For tree models, **Model Seeker** uses the training dataset to create the tree and the testing dataset to test or evaluate the tree to determine the best subtree.
- **Net:** For net models, **Model Seeker** uses the training and testing datasets in the Train and Test option to build and test the net in one operation.
- **Match:** With match models, the training dataset becomes part of the model, and the testing dataset is used to optimize the weights.

From there, the process is the same for all models: The evaluation dataset (sometimes called the prediction dataset) is used for a practice prediction. The input and output datasets of the practice prediction are merged, and the resulting merged dataset is used to analyze the model's performance.

If the target field is categorical, the merged dataset is also used to calculate a Lift results table. The data in this table is used to populate the last column of the grid and to plot lines on the chart.

Note: The evaluation dataset must contain at least 100 records for each node of the server; for example, if the server has four nodes, the dataset should contain 400 records.

Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the remaining steps without further input; **Cancel**, which cancels operations; **Help**, which brings up online help; and, on some screens, **Advanced**, which lets you set certain options.

9.1 Darwin Model Seeker – Settings

Use the **Settings** dialog to specify dataset settings and to specify the variations on the models you want **Model Seeker** to build.

Model Seeker offers a default name prefix, displayed in the **Common Model Name Prefix** box, which is used as a basis for the names of the models and their test results tables. The default name prefix is a 12-character string beginning with `MS`; the next 6 characters are the date in `YYMMDD` format; the last 4 characters are the time of day in military (24-hour clock) `HHMM` format.

9.1.1 Dataset Settings

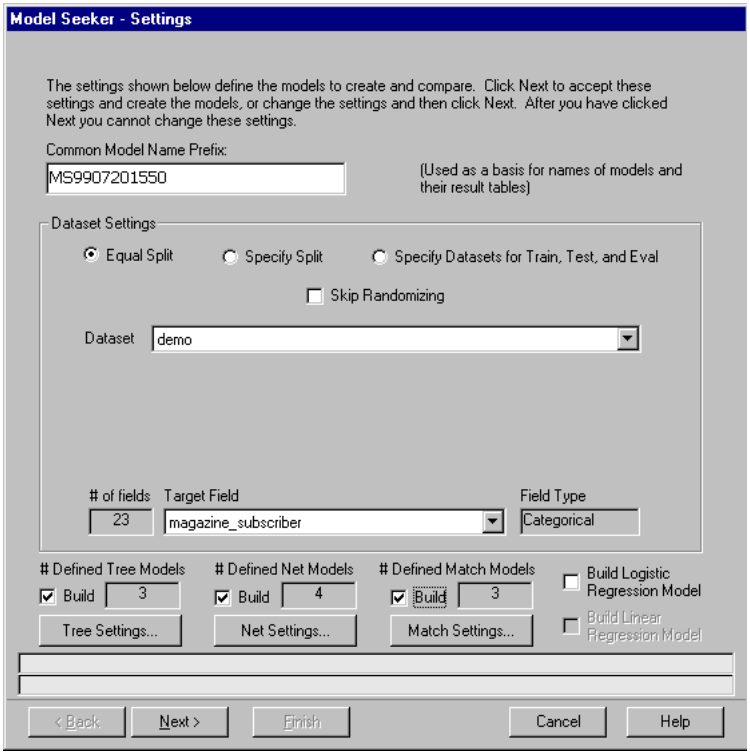
For dataset settings, specify the dataset to be used for building the models, how it is to be split into subsets, and the target field.

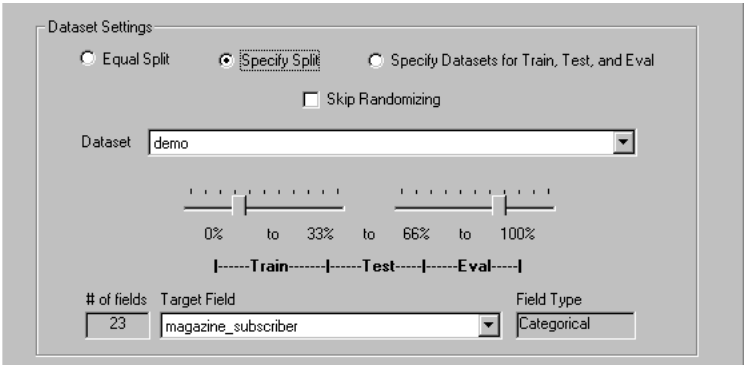
- **Dataset:** Specify the input dataset. The dataset name that appears by default is taken from your current project.
- **Skip Randomizing:** Randomizing is by default performed before splitting the dataset. If you want to skip randomizing, click this check box.
- **Equal Split:** Divides the input dataset into three roughly equal parts (33%, 33%, and 34%).
- **Specify Split:** Use the two sliders to specify the proportions for the three parts.
- **Specify Datasets for Train, Test, and Eval:** Specify the names of three datasets to be used for the three phases. This option is convenient if you have already split the input dataset into three parts for training, testing, and evaluation.

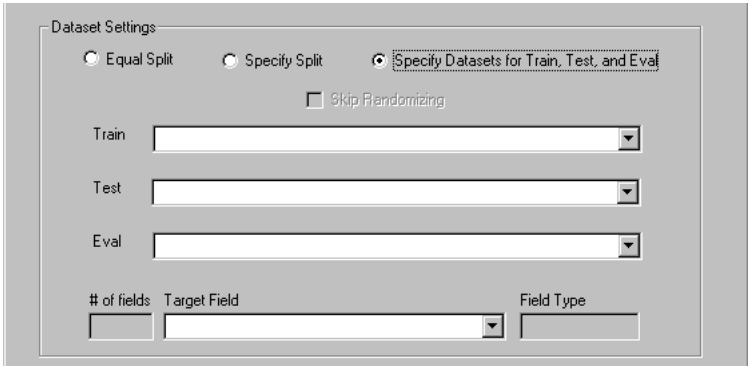
Below these settings are three boxes for field information:

- **Number of Fields:** A read-only box that displays the number of fields in the dataset.
- **Target Field:** Specify the target field. The text box lists all fields in the dataset.
- **Field Type:** Automatically displays the field type (categorical or ordered) of the target field specified.

The screen capture below shows the **Settings** dialog with **Equal Split** selected. The two screen captures that follow show the other splitting options.



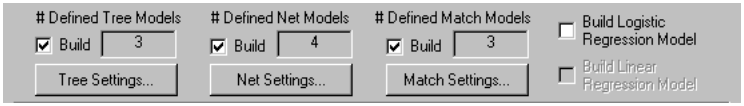




9.1.2 Model Settings

Below the dataset settings are command buttons for the model settings. These command buttons are activated when you click the **Build** check box for a model type. Each command button takes you to a dialog that displays the settings that define the default variations for that type of model. You can change the settings to provide for different variations and for additional variations.

The read-only box to the right of the **Build** check box displays the number of models defined by the settings.

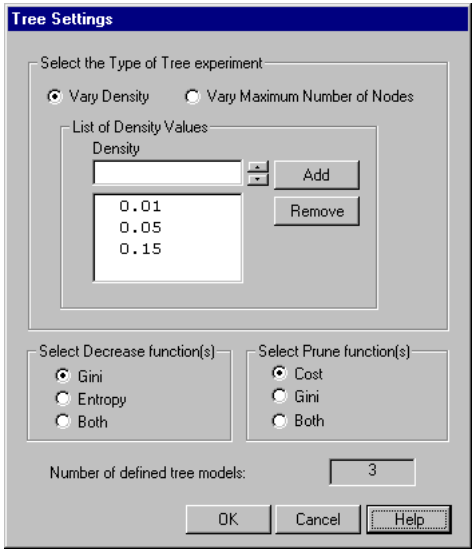


Tree Settings

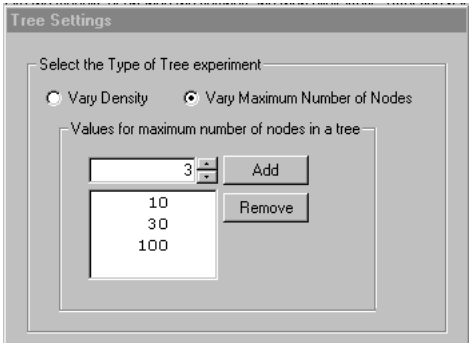
To see the default settings for tree models, click **Tree Settings**. By default, **Model Seeker** defines three tree models, by specifying three variations on the parameters relevant to tree models.

At the top of the **Tree Settings** dialog, you can choose to vary either density or the maximum number of nodes. Varying one means leaving the other at a default setting. The default choice is to vary density values.

- **Varying Density Values:** Density values reflect the minimum fraction of records for a node; lower values allow splitting to continue longer, and therefore produce larger trees. You can specify values between 0% and 100% for density. The default values are 1%, 5%, and 15%. If you vary density values, the maximum number of nodes is unbounded.

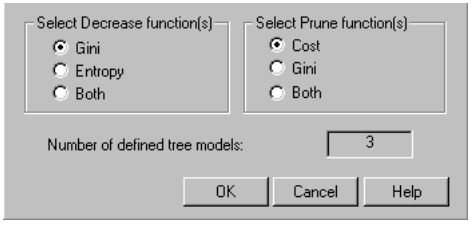


- **Varying Maximum Number of Nodes:** A value here sets a limit on the maximum number of nodes the tree can have. Trees whose growth is halted by this method are biased; however, this option can be useful in the earliest stages of experimental model building. The default values are 10, 30, and 100; the allowable range is 3 to 30,000. If you choose to vary the maximum number of nodes, density defaults to 0.001.



At the bottom of this dialog are settings for two additional parameters, decrease and prune functions, which apply only to categorical target fields:

- **Decrease function(s):** Decrease functions are internal functions for measuring the degree of impurity in a split. Darwin has two decrease functions: gini (the default) and entropy. You can select either or both.
- **Prune function(s):** Prune functions dictate how a tree is pruned (split) into subtrees. Darwin has two pruning functions: cost (the default) and gini. You can select either or both.



See *Darwin Reference*, Chapter 7, for more information about these parameters.

Number of defined tree models: At the bottom of the dialog is a read-only box that displays the number of tree models defined by the default specifications or by your selections. This is the number of tree models that the Wizard will build.

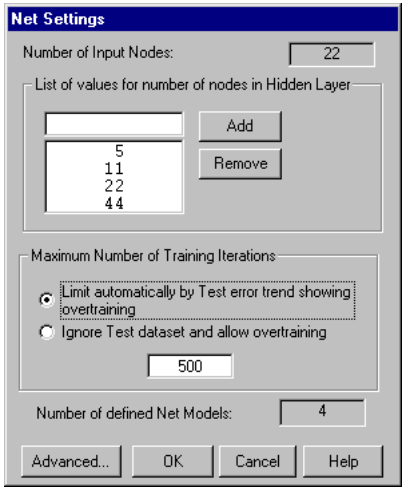
When you are finished with the **Tree Settings** dialog, click **OK** to return to the **Settings** dialog. You can then view or change settings for net or match models, or you can click **Next** to have the Wizard create the tree models defined.

Net Settings

To build net models, click the **Build** check box for net models (at the bottom of the **Settings** dialog). The number 4 appears in the read-only box. By default, the Wizard defines four net models, based on four values for the number of nodes in the hidden layer. To see the default settings for net models, click **Net Settings**.

- **Number of Input Nodes:** A read-only box that displays the number of input nodes. This is the number of non-target fields (*N*), or one less than the total number of fields. The four default values displayed in the list of values are derived from *N*.
- **List of values for number of nodes in Hidden Layer:** This list box contains four default values, which are derived from the number of input nodes:
  - square root of *N* (rounded up)
  - *N* / 2 (rounded up)
  - *N*
  - 2**N*

Note that if you are building net models with 6 or fewer input nodes, you will have duplicate models. Duplication occurs because of the way the four default values are derived (*N*, *N* / 2 (rounded up), 2**N*, and square root of *N* (rounded up)). For example, if *N*=6, *N* / 2 and square root of *N* (rounded up) are the same.



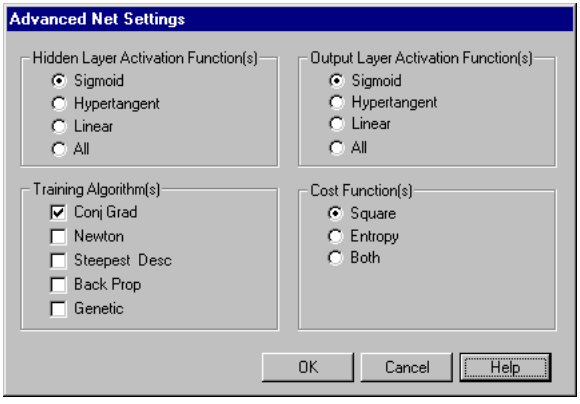
You can also add or remove values from the list of value for number of nodes in the hidden layer:

- To add a value, enter it in the edit box and click **Add**.
- To remove a value, select it in the list and click **Remove**.

- **Maximum Number of Training Iterations:** This is the number of times the dataset is passed through the net, with weights being adjusted at each iteration. There are two options:
  - **Limit automatically by Test error trend showing overtraining:** This option, the default, means that the net model process limits itself. That is, it trains and tests (using the training and testing datasets) until it detects the beginning of overtraining, or until it reaches the specified number of iterations, whichever happens first, and then iterations stop. The default number of iterations is 500; you can set a different limit if you wish.
  - **Ignore Test dataset and allow overtraining:** For this option, the net model process continues to train until it converges or until it reaches the maximum number of iterations specified, whichever happens first, and then iterations stop. The default is 100; you can set a different limit if you wish. With this option, the net model uses only the training dataset; it ignores the testing dataset.This option specifically allows the net to overtrain.
- **Number of defined net models:** At the bottom of the dialog is a read-only box that displays the number of net models defined by the default selections or by

your selections. This is the number of net models that the Wizard will build.

- **Advanced:** Click the **Advanced** button to view or change settings for the following parameters:
  - **Hidden Layer Activation Function(s):** Select sigmoid (default), hypertangent, linear, or all.
  - **Output Layer Activation Function(s):** Select sigmoid (the default for categorical fields), hypertangent, linear (the default for ordered fields), or all.
  - **Training Algorithm(s):** Select any combination of conjugant gradient (default), modified Newton, steepest descent, back-propagation, and genetic.
  - **Cost Function(s):** Select square (default), entropy, or both.



See *Darwin Reference*, Chapter 8, for more information about these parameters.

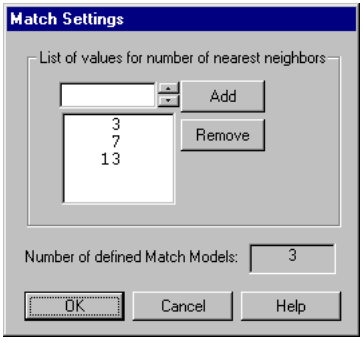
When you are finished with the **Net Settings** dialog, click **OK** to return to the **Settings** dialog.

Match Settings

To build match models, click the **Build** check box for match models (on the **Settings** dialog). The number 3 appears in the read-only box. By default, the Wizard defines three match models, based on three values for the

number of nearest neighbors. To see the default settings for match models, click **Match Settings**.

- **List of values for number of nearest neighbors:** The default values are 3, 7, and 13. To add a value, enter it in the top box and click Add. To remove a value, click it and then click **Remove**.
- **Number of defined match models:** At the bottom of the dialog is a read-only box that displays the number of match models defined by the default selections or by your selections. This is the number of match models that the Wizard will build.



When you are finished with the **Match Settings** dialog, click **OK** to return to the **Settings** dialog.

Linear/Logistic Regression

To the right of the settings buttons for the tree, net, and match models are two check boxes for regression models, only one of which is enabled at any time:

- **Build Logistic Regression Model** (enabled for categorical target variables)
- **Build Linear Regression Model** (enabled for ordered target variables)

The following parameters for a **Logistic Regression** model are set automatically:

- **Number of Hidden Layers:** 0
- **Iterations:** 10000 (the intent is to iterate to convergence).
- **Training Mode:** Simple training.
- **Training Algorithm:** Modified Newton.
- **Cost Function:** Square.

- **Dataset:** Same as the training dataset used for all other models.
- **Output Activation Function:** Sigmoid.
- **Weight:** 1

Parameters for a **Linear Regression** model are the same as those for a **Logistic Regression** model except that the output activation function is **Linear** (instead of **Sigmoid**).

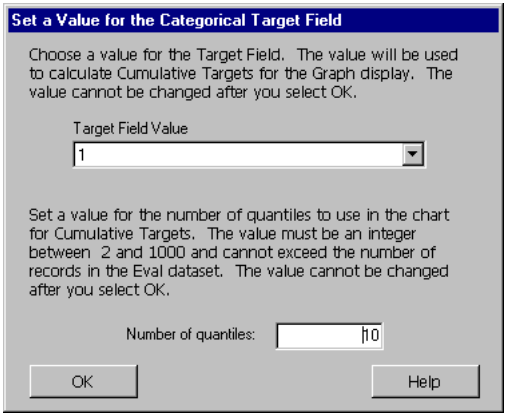
Note that regression models are listed in the **Workspace** as net models.

Next: Build the Models

After you have specified all the values for the models you want to build, click **Next** to have **Model Seeker** build the models.

Note that after you have clicked **Next**, you cannot change the model settings, although you can click **Back** to review the settings while **Model Seeker** is building the models. You cannot change any settings while **Model Seeker** is building the models

After you click **Next**, **Model Seeker** will, if the target field is categorical, prompt you for the target value, which is used for calculating cumulative targets in the grid and graph displays.



You also have the option of specifying the number of quantiles to use in the chart for cumulative targets. The default is 10; the permissible range is 2–1000.

9.2 Darwin Model Seeker – Monitor Progress

This dialog shows you what is happening as it happens. As each model is created, information about it is displayed. The dialog displayed when working with categorical target fields is different from the one displayed when working with ordered target fields.

9.2.1 Categorical Target Values

The **Monitor Progress** dialog displays the following information when the target variable contains categorical values:

- **Number (#):** Number of model. Models are listed in reverse order.
- **Model Name Suffix:** All models built with **Model Seeker** have the same name prefix, which was shown in the first text box on the **Settings** dialog. The suffix shown in this column indicates the type of model and its position in the series.
- **Show:** A check mark in this column means to show this model on the graph.
- **Save:** A check mark in this column means keep the model and its result tables when you finish the Wizard and return to Darwin. The model with the check mark here is the one with the best value in the **Accuracy** column.
- **Build Time (Hours):** How long, in hours, it took to build and analyze the model.
- **Accuracy:** Accuracy is measured by percentage of correct predictions. The best model (highest accuracy) is automatically checked in the **Save** column when **Model Seeker** completes all its work.
- **Total Cum Targets:** This is the average value for cumulative targets, averaged over all quantiles; it is the area under the curve plotted on the graph.
- **Cum Targets at 20% of Population:** This column displays for each model one data point that is plotted on the graph, where the x axis is the percentage of the population and the *y* axis is cumulated targets. The values in this column are the values of *y* for each model at a specified value of *x*, the default being 20%.

  You can change the default percentage value. At the top of the dialog, in the **Population %** box, select the value and click **Apply**. The column heading and the values displayed change accordingly.

You can sort the information in each column by clicking the column head. Clicking again reverses the order.

Note the two status bars at the bottom of the dialog: The first shows you what **Model Seeker** is doing; the second shows you what the Darwin server is doing.

Stop and **Continue**: These command buttons let you interrupt the server and then continue. If you want to stop the server, click **Stop**. A dialog appears, tells you that you have clicked **Stop**, and asks if you want to interrupt the server. You have three options:

- Click **Yes** if you want to interrupt the server and have it stop as soon as possible. If you do this, the current model will be skipped. (Any skipped models are grayed out in the table on the **Monitor Progress** dialog; Section 9.2.2 contains a screen capture that shows what this looks like.)

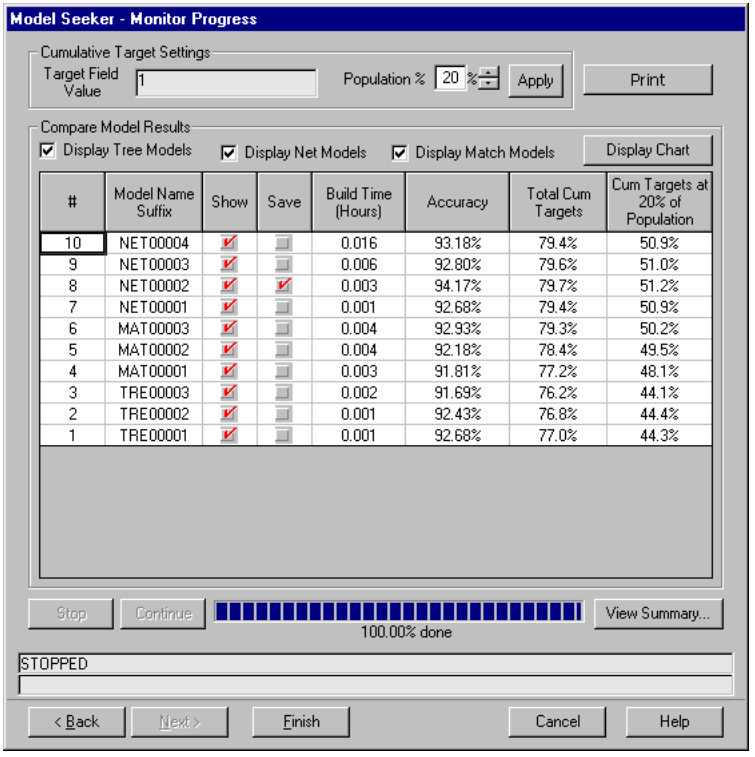
  After the server stops, the **Continue** button will be enabled. Click **Continue** if you want to continue building any remaining models.

- Click **No** if you want the server to complete all work for the current model before stopping. After the server stops, the **Continue** button will be enabled. Click **Continue** if you want to continue building any remaining models.

- Click **Cancel** to let the server continue without stopping.

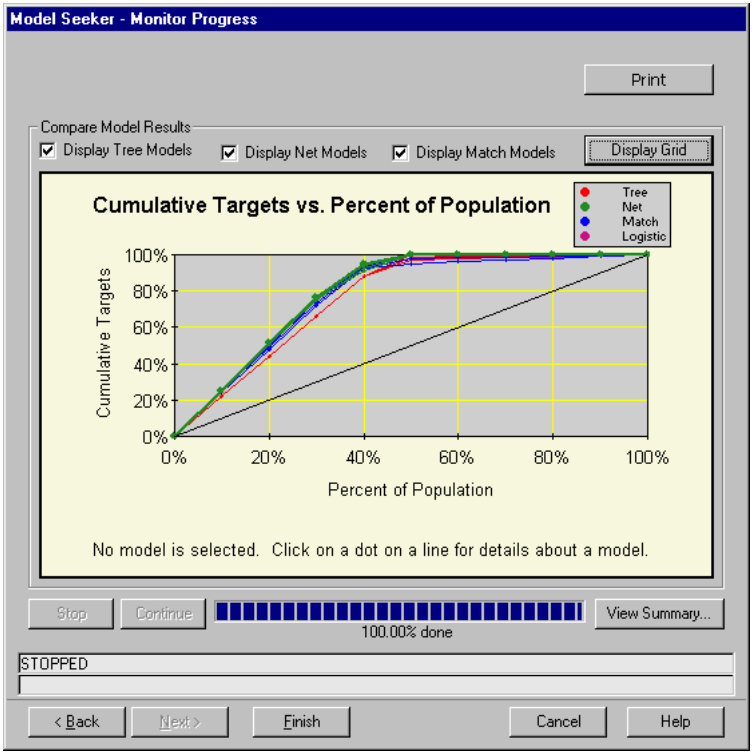
If Darwin encounters a problem during processing, you receive an error message and you may be presented with these same three options.

The screen capture below shows the **Monitor Progress** dialog after the wizard has finished building all the models. This is what it looks like when the target field contains categorical values. It looks a little different if the target field contains ordered values (see Section 9.2.2).



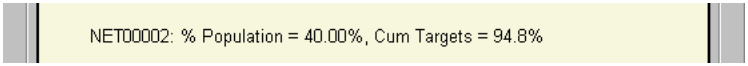
Print: Click **Print** to send a results summary of the Wizard's processing to your printer. You can print the summary at any point during processing.

Display Chart: For categorical target fields, results are plotted on a graph. To display the graph, click **Display Chart**.



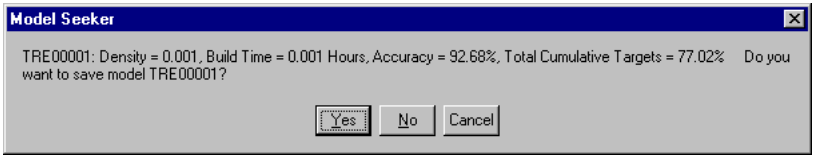
The diagonal line on the graph represents the results you could expect from a completely random prediction. As you can see, all the models in this example did significantly better than that.

For details about a particular model, position the cursor on one of the dots for that model. The information appears at the bottom of the graph, where it says "No model is selected. Click on . . . ." The example below shows what is displayed with the cursor on one of the dots for one of the net models. This information changes as you move the cursor along the plotted line.



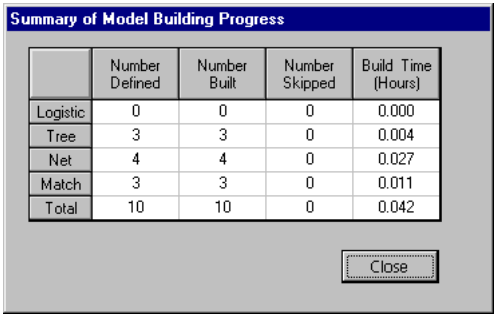
With the cursor still on the line, click the left mouse button for additional information about the model. A message box displays the information and also

gives you the opportunity to save the model. If you click **Yes**, the corresponding check box on the grid is automatically checked.



To return to the tabular display, click **Display Grid**.

View Summary: To see a summary of the experiment's results, click **View Summary**. The summary table shows how many models of each type were defined, how many were built, how many were skipped, and how long it took to build them. If you click **View Summary** while **Model Seeker** is still working, the grid displays a summary of work completed thus far.



Click **Close** to dismiss the **Summary** dialog.

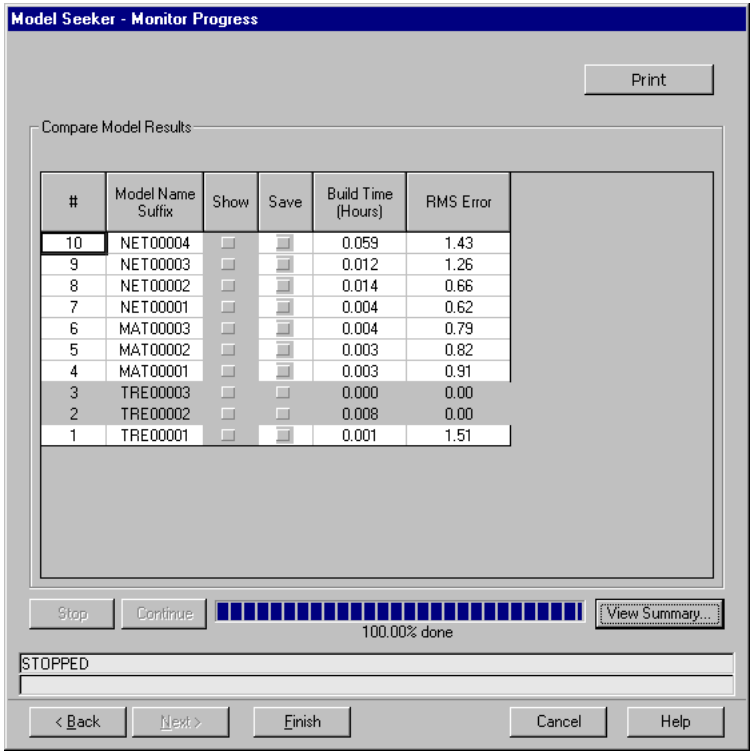
Click **Finish** to update the information in the **Workspace** and return to the Darwin client.

9.2.2 Ordered Target Values

When the target field contains ordered values, the **Monitor Progress** dialog looks a bit different because the columns and data having to do with Lift are absent. The information displayed is as follows:

- **Number (#):** Number of model. Note that models are listed in reverse order.

- **Model Name Suffix:** All models built with **Model Seeker** have the same name prefix, which was shown in the first text box on the Settings dialog. The suffix shown in this column indicates the type of model and its position in the series.
- **Show:** This column is dimmed because it refers to showing the model on the chart, which applies only to categorical target fields.
- **Save:** A check mark in this column means keep the model and its result tables when you finish the Wizard and return to Darwin. The model with the check mark here is the one with the lowest score in the **RMS Error** column.
- **Build Time:** How long, in hours, it took to build and analyze the model.
- **RMS Error:** The best model (lowest Root Mean Square error) is automatically checked in the **Save** column when **Model Seeker** completes its work.



The grayed-out models in the display above are models that were skipped during processing because of problems encountered with them during the build.

10

Model Compare Wizard

To invoke the **Model Compare** Wizard:

- Click **Options > Model Compare**

The **Model Compare** Wizard lets you compare different models built by any method — using the **Modeling Wizard**, **Model Seeker**, or models you have built directly in Darwin. You can compare models that are "compatible" i.e., that have the same evaluation dataset and the same target field.

Your current project determines which models are available for comparing. Any model in the current project is available unless it is

- a tree model whose best subtree is the entire tree
- a model whose training dataset has been deleted

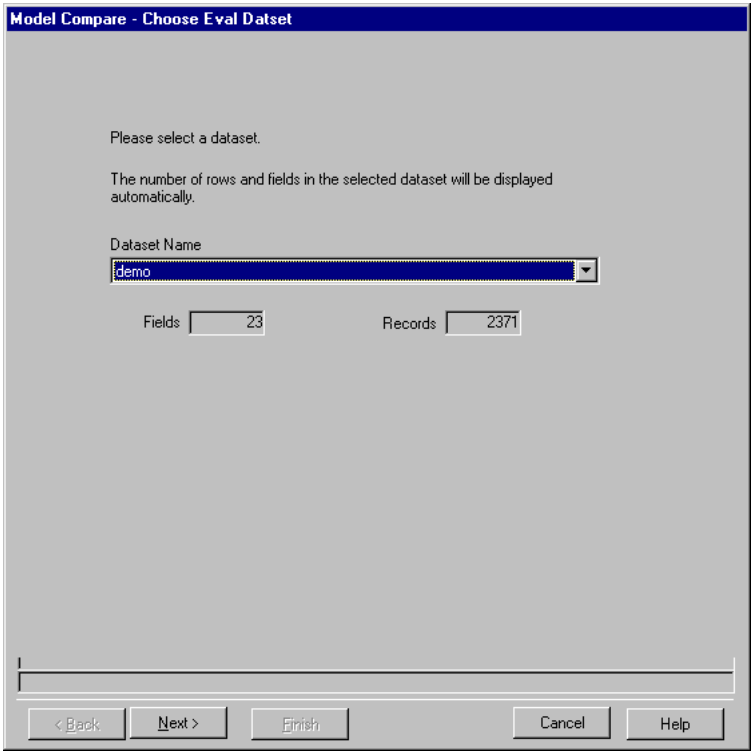
Navigation: Each wizard dialog contains the standard navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Finish**, which is unavailable until you are at a point where Darwin can perform the remaining steps without further input; **Cancel**, which cancels operations; **Help**, which brings up online help; and, on some screens, **Advanced**, which lets you set certain options.

10.1 Darwin Model Compare – Choose Eval Dataset

Specify the dataset to be used in the comparisons. Keep in mind the following requirements:

- The dataset should contain at least 100 records for each server node, e.g., if the server has four nodes, the dataset should contain 400 records.
- The dataset must be compatible with the dataset used to build the models to be compared, i.e., they must have the same fields (field names), in the same order; they must have the same number of fields; the data types must be the same; and the form (categorical or ordered) must be the same.

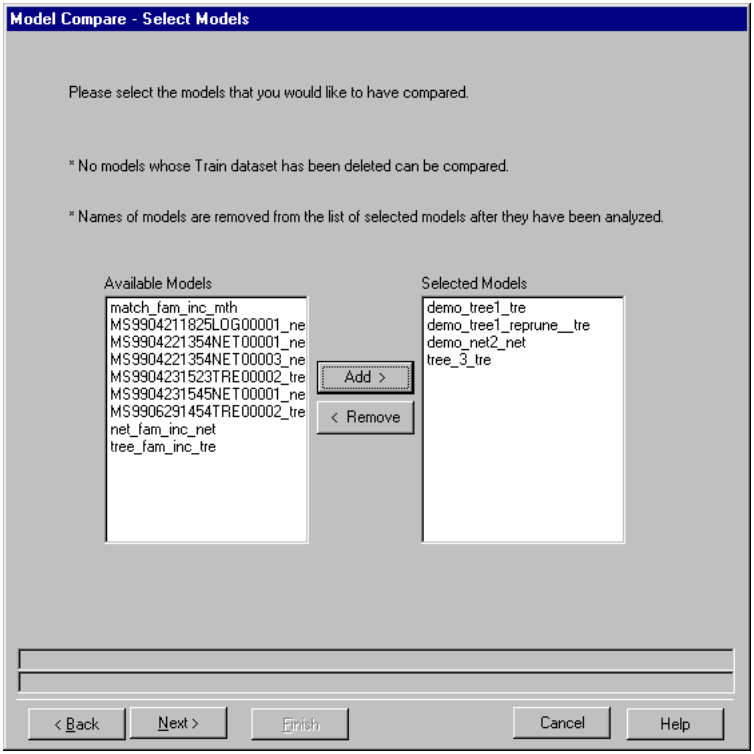
When you specify the dataset, the number of records (rows) and fields (columns) are automatically filled in.



10.2 Darwin Model Compare – Select Models

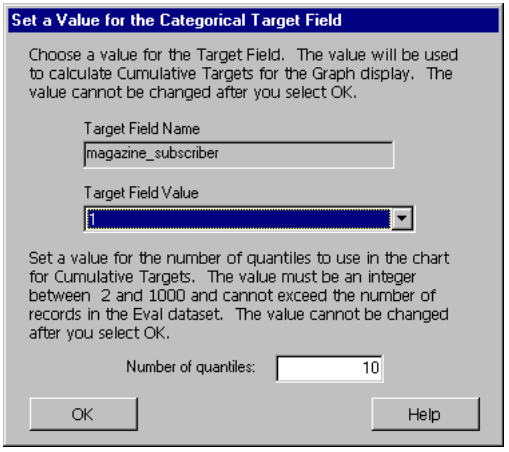
The **Select Models** dialog displays a column listing models available for comparison and a column for selected models. To move a model from the list of available models to the list of selected models, click its name and then click **Add**. To deselect a model, click its name and then click **Remove**.

The screen capture below shows the dialog with all three model names moved from the list of available models to the column for selected models.



When the list of selected models contains the names of all the models you wish to compare, click **Next**.

After you click **Next**, the Wizard will, if the target field is categorical, prompt you for the target value, which is used for calculating cumulative targets in the grid and graph displays.



You also have the option of specifying the number of quantiles to use in the chart for cumulative targets. The default is 10; the permissible range is 2–1000.

10.3 Darwin Model Compare – Monitor Progress

This dialog shows you what is happening as it happens. As each model's performance is evaluated, information about it is displayed. The dialog displayed when working with categorical target fields is different from the one displayed when working with ordered target fields.

10.3.1 Categorical Target Values

The **Monitor Progress** dialog displays the following information when the target variable contains categorical values:

- **Number (#):** Number of model. Models are listed in reverse order.
- **Model Name:** This column displays the names of the models.
- **Show:** A check mark in this column means to show the model on the graph.
- **Save:** A check mark in this column means save the result tables generated for this model when you finish the Wizard and return to Darwin.
- **Accuracy:** Accuracy is measured by percentage of correct predictions. The best model is the one with the highest accuracy.

- **Total Cum Targets:** This is the average value for cumulative targets, averaged over all quantiles; it is the area under the curve plotted on the graph.
- **Cum Targets at 20% of Population:** This column displays for each model one data point that is plotted on the graph, where the *x* axis is the percentage of the population and the *y* axis is cumulated targets. The values in this column are the values of *y* for each model at a specified value of *x*, the default being 20%.

  You can change the default percentage value. At the top of the dialog, in the **Population %** box, select the value and click **Apply**. The column heading and the values displayed change accordingly.

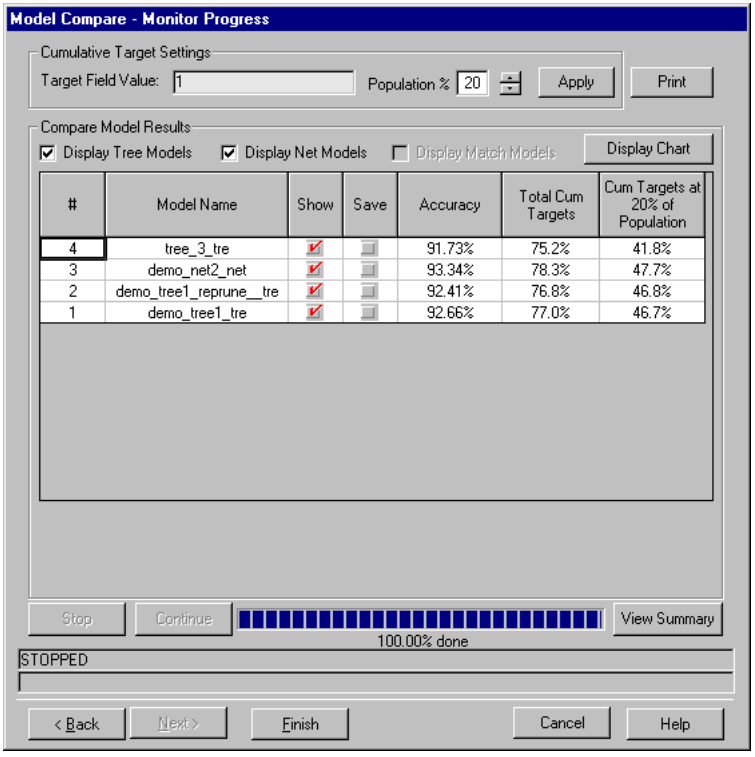
You can sort the information in each column by clicking the column head. Clicking again reverses the order.

Note the two status bars at the bottom of the dialog: The first shows you what the Wizard is doing; the second shows you what the Darwin server is doing.

Stop and **Continue**: These command buttons let you interrupt the server and then continue. If you want to stop the server, click **Stop**. A dialog appears, tells you that you have clicked **Stop**, and asks if you want to interrupt the server. You have three options:

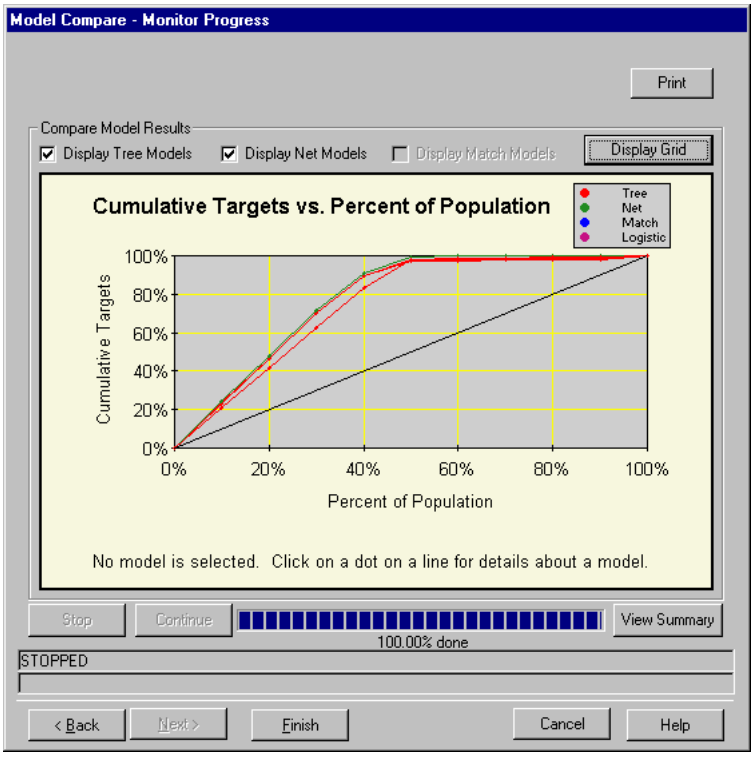
- Click **Yes** if you want to interrupt the server and have it stop as soon as possible. If you do this, the current model will be skipped. (Any skipped models are grayed out in the table on the **Monitor Progress** dialog.) After the server stops, the **Continue** button will be enabled. Click **Continue** if you want to continue building any remaining models.
- Click **No** if you want the server to complelte all work for the current model before stopping. After the server stops, the **Continue** button will be enabled. Click **Continue** if you want to continue building any remaining models.
- Click **Cancel** to let the server continue without stopping.

The screen capture below shows the **Monitor Progress** dialog after the wizard has finished comparing all the models. This is what it looks like when the target field contains categorical values. For target fields containing ordered values, see Section 10.3.2.



Print: Click **Print** to send a results summary of the Wizard's processing to your printer. You can print the summary at any point during processing.

Display Chart: For categorical target fields, results are plotted on a graph. To display the graph, click **Display Chart**.

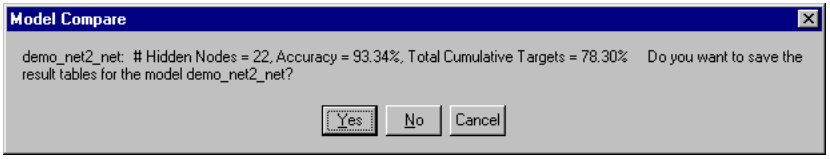


The diagonal line on the graph represents the results you could expect from a completely random prediction. As you can see, all the models in this example did significantly better than that.

For details about a particular model, position the cursor on one of the dots for that model. The information appears at the bottom of the graph, where it says "No model is selected. Click on . . . ." The example below shows what is displayed with the cursor on one of the dots for one of the net models. This information changes as you move the cursor along the plotted line.

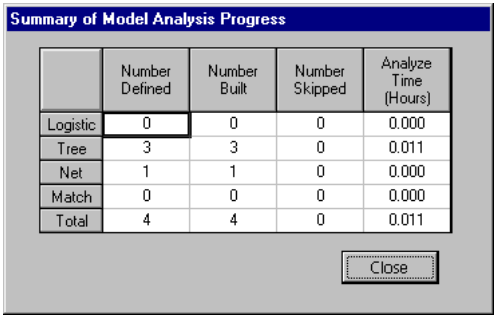
With the cursor still on the line, click the left mouse button for additional information about the model. A message box displays the information and also

gives you the opportunity to save the result tables for the model. If you click **Yes**, the corresponding check box in the **Save** column on the grid is automatically checked.



To return to the tabular display, click **Display Grid**.

View Summary: To see a summary of the experiment's results, click **View Summary**. The summary table shows how many models of each type were defined, how many were built, how many were skipped, and how long it took to build them. If you click **View Summary** while **Model Compare** is still working, the grid displays a summary of work completed thus far.



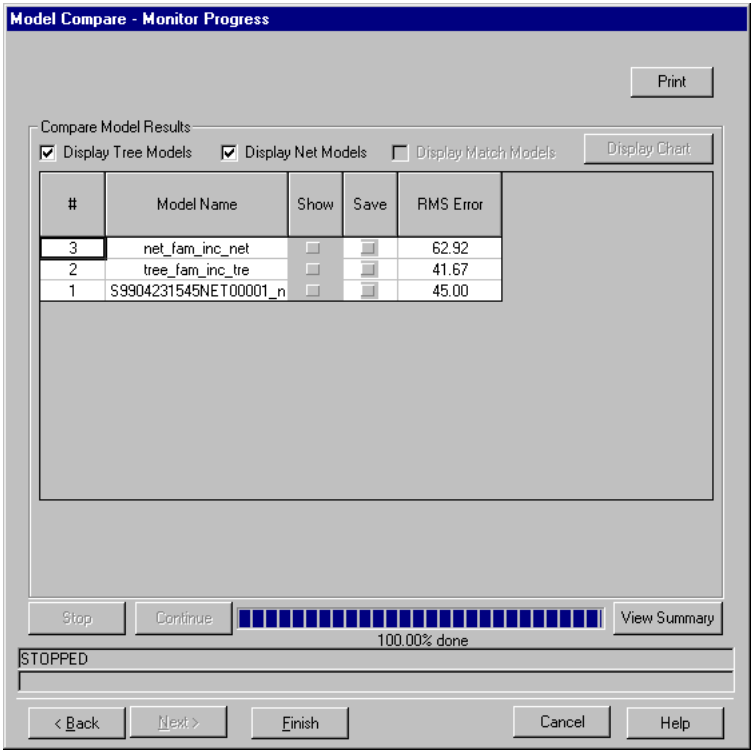
Click **Close** to dismiss the **Summary** dialog.

Click **Finish** to update information in the **Workspace** and return to the Darwin client.

10.3.2 Ordered Target Values

When the target field contains ordered values, the **Monitor Progress** dialog looks different because the columns and data having to do with Lift are absent. The information displayed is as follows:

- **Number (#):** Number of model. Note that models are listed in reverse order.
- **Model Name:** This column displays the name of the models.
- **Show:** This colum is dimmed because it refers to showing the model on the chart, which applies only to categorical target fields.
- **Save:** A check mark in this column means keep the model's result tables when you finish the Wizard and return to Darwin. The model with the check mark here is the one with the lowest score in the **RMS Error** column.
- **RMS Error:** The best model is the one with the lowest value in this column (lowest Root Mean Square error).



11

Clustering Wizard

This chapter provides a brief introduction to clustering, describes how clustering models are different from predictive models, discusses the appropriate data for a clustering model, and introduces the Darwin Clustering wizard.

11.1 What Is Clustering?

Clustering is a useful technique for initially exploring and visualizing data. It is particularly helpful in situations where you have lots of records, no idea what natural groupings might be there, and you would like the data mining software to find whatever natural groupings may exist.

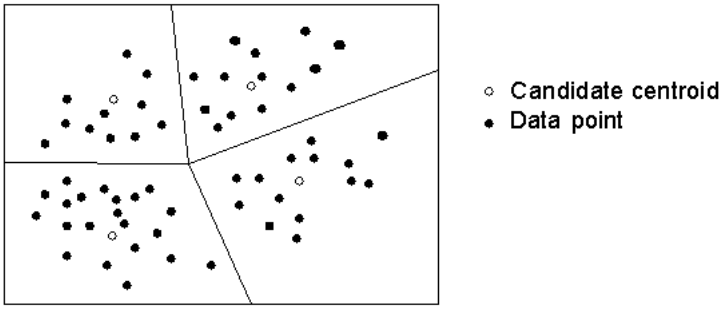
Clustering also serves as a useful data preprocessing step to identify homogeneous groups on which to build predictive models such as trees or neural networks.

A clustering model is different from predictive models in that the outcome of the process is not guided by a known result, that is, there is no target variable. Predictive models predict values for a target variable, and an error rate between the target and predicted values can be calculated to guide model building. With clustering models, the density itself drives the process to a final (clustered) state.

This difference is the distinction between *supervised* and *unsupervised* learning algorithms (sometimes referred to as *directed* and *undirected* learning). Unsupervised learning methods are typically used when you are interested in finding the intrinsic structure, correlations, or affinities in a body of data but no classes or labels are assigned a priori. Examples of unsupervised learning algorithms include self-organizing maps (SOM), *k*-means, competitive learning, and association rules.

Darwin 3.6 implements a *k*-means algorithm. Clustering begins by assigning, at random, an initial candidate *centroid* (a "most central" node) in each of *k* areas of the data space. The user chooses *k*, the number of clusters. Clustering proceeds by assigning each record to the cluster with the centroid closest to it.

The diagram below shows a scattergram of data points partitioned into four clusters. In practice, things are more complicated than this diagram can indicate -- you are typically dealing with many more dimensions than two.



When all records are assigned, the algorithm calculates the average distance between each record and its centroid. After each round of calculations, the algorithm compares the distances between each record and its centroid, adjusts the position of the centroid to reduce average distances, and recalculates.

This process repeats for each record until the adjustments to be made become so small as to be insignificant or until the maximum number of iterations specified is reached. The values in these calculations are normalized at the beginning, so that all comparisons are based on a common scale.

To gain some understanding of the clusters themselves, clustering is typically followed by applying classification and regression tree (C&RT) rules. After your clustering model is built, the Wizard lets you backfit a C&RT algorithm to extract a set of rules to help interpret the clusters.

Clustering works best with ordered or interval data. If your dataset contains categorical fields, see Section 11.1.2 for suggestions.

11.1.1 Clustering in Data Mining

The outcome of clustering is a set of clusters, the members of which are more like each other than they are like the members of other clusters. Now that you have all these clusters, what do you do with them? Nobody is much interested in clusters as ends in themselves. Clustering is useful because we assume that records that are similar reflect people who are similar, and that people who are similar are likely to behave in similar ways.

Clustering is typically followed by applying other data mining tools:

- Trees, to spell out the rules by which clustering decisions were made.
- Trees, to predict membership in clusters.
- Trees or nets, to predict responses of certain clusters to the variable(s) of your choice. For example, you might use certain clusters to identify the groups of people likely to buy a certain product, or file fraudulent insurance claims, or respond to an unsolicited mailing.
- More clustering, with different valus of *k*. Ideally, you want disjoint clusters — aim for a set of clusters that are distinct, that are least confusable with one another.

You can think of clustering is a *data reduction* or sampling technique that can help you select representative records or subsets. For example, you might build 100 clusters, select one record from each cluster, and then use this 100-record dataset to create other kinds of models. Alternatively, you might create a set of clusters, and then build one model per cluster and apply the model to the records in that cluster.

Before you create clusters, consider the following:

- How many clusters do you want to find? A very large or very small number of clusters may not give useful information for your problem.
- Do you want to cluster according to all fields or just certain fields? You should restrict the set of features so that the geometry of the clusters makes sense.

11.1.2 Clustering with Categorical Fields

Clustering works best with ordered or interval data. Interpreting results of clustering with categorical fields is inherently problematic.

The notion of "centroid" makes geometric sense when you are dealing with ordered or interval values, and measuring distances between records and their centroids makes sense. However, it is not at all clear what "most central" means when you are dealing with categorical values, especially multiclass categorical values. For example, if one of your variables is color, and red turns out to be "closer" to the centroid than blue -- what does this mean?

There are three approaches with multiclass categorical values:

- Do not cluster on multiclass categorical fields. Simply exclude them from your clustering model.

- Cluster on multiclass categorical fields, leaving them as is. The challenge is in their interpretation.
- Before clustering, explode multiclass categoricals into binary categorical fields. (See **Dataset > Transform >Explode**. The outcome will be a set of binary fields, one for each of the values in the original multivalued field. Then use the **Project** transformation to remove the original multiclass categorical field.)

  Note that the total number of fields in your dataset is increased by the number of new binary fields minus the original multiclass field.

If you choose to include multiclass categorical fields in your clustering model, just remember that divining their meaning is difficult if not impossible, and you may be better off ignoring them.

11.2 Darwin Clustering Wizard

To invoke the **Clustering** Wizard:

- Click **Options > Clustering**

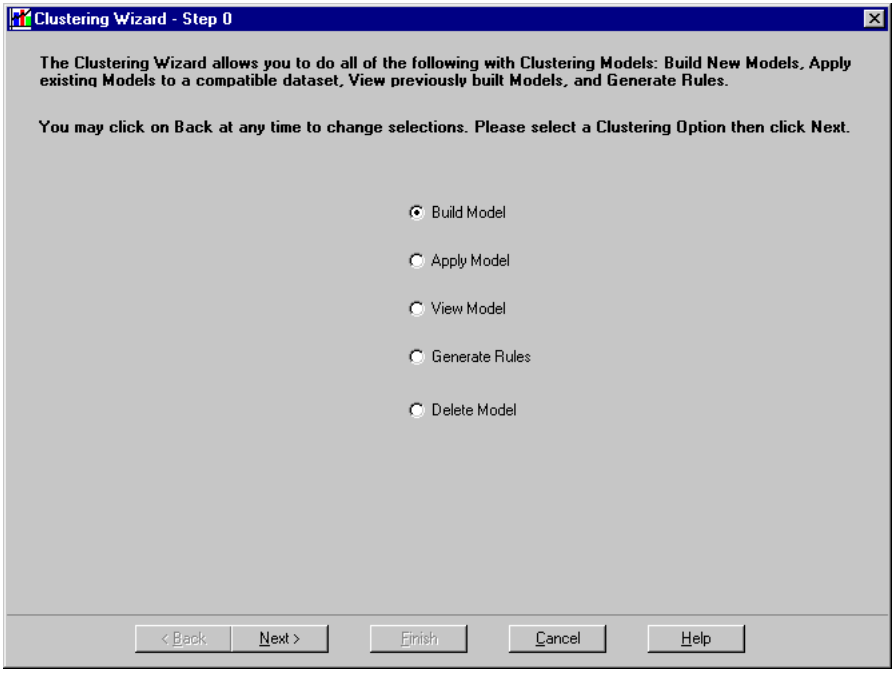
Important: Before you invoke the Clustering Wizard, set your active Darwin project to the project that contains the dataset you want to cluster. You cannot change project from within the Clustering Wizard.

Navigation: Each wizard dialog contains the following navigational controls: **Back**, which takes you to the previous screen; **Next**, which takes you to the next screen; **Step 0**, which returns you to the **Step 0** dialog, where you can select one of the five functional paths (the **Step 0** button is disabled until the last dialog of each path); **Cancel**, which cancels operations and exits the Wizard (if the Wizard is in the middle of processing, you will get a message telling you so, and asking whether you really want to exit); **Finish**, which exits the wizard after everything is saved, and returns you to the Darwin client (**Finish** appears on the last dialog of each path, in place of **Cancel**); and **Help**, which brings up online help.

11.3 Darwin Clustering — Step 0: Select Function

The **Clustering Wizard — Step 0** dialog presents you with five options, representing five functional paths:

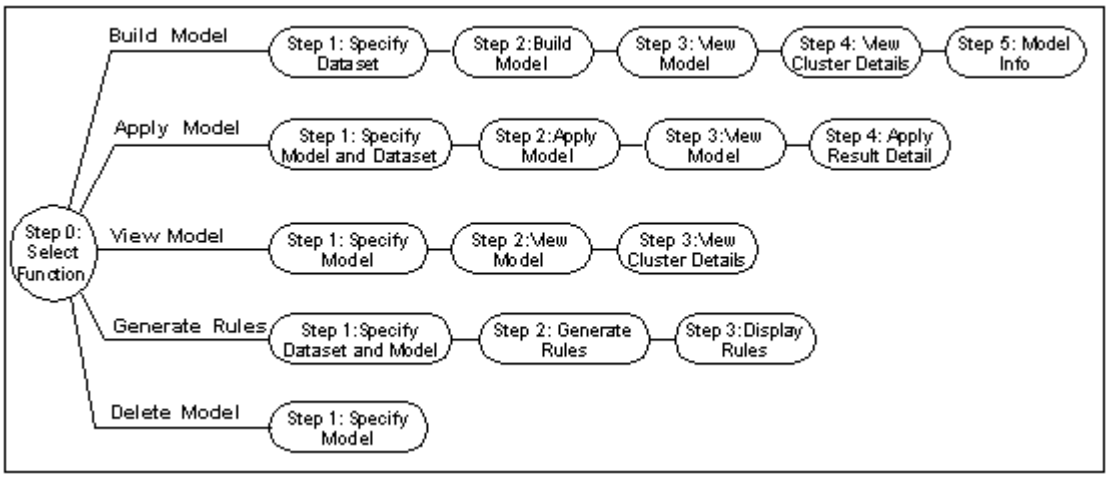
- **Build Model** — allows you to build a new model and customize it
- **Apply Model** — allows you to apply an existing model to a different dataset
- **View Model** — displays an existing model and allows you to customize it
- **Generate Rules** — generates and displays C&RT rules that help interpret the individual clusters
- **Delete Model** — allows you to delete a clustering model



After you have selected a path, click **Next** to go to the next step in that path. If you want to change to a different path, click **Back** until you get to **Step 0**, where you can select a different path.

The steps in each functional path are diagrammed below. Side trips, such as, for example, editing a weights file in **Step 1** of the **Build Model** path, are not included in the diagram.

At the end of each functional path, you can click **Step 0** to return to the **Step 0** dialog, where you can select a different path, or you can click **Finish** to exit the wizard and return to the Darwin client.



Clustering models are not listed with other models in the **Workspace** listing, because clustering models are not managed by the Darwin client **Workspace**. To see a list of all saved clustering models in the current project, select **Apply Model** or **View Model** on the **Step 0** dialog. The **Step 1** dialog of the **Apply** or **View** paths displays (in the **Model Name** list box) the names of all available clustering models.

11.4 Darwin Clustering — Build Model

Choose the **Build Model** option to build and customize a new clustering model.

The output of the **Build** path is a new 2-field dataset with as many records as the input dataset:

- The first field holds the cluster identification integer that of the cluster to which each record belongs.
- The second field holds a value representing the distance between each record and the cluster's centroid.

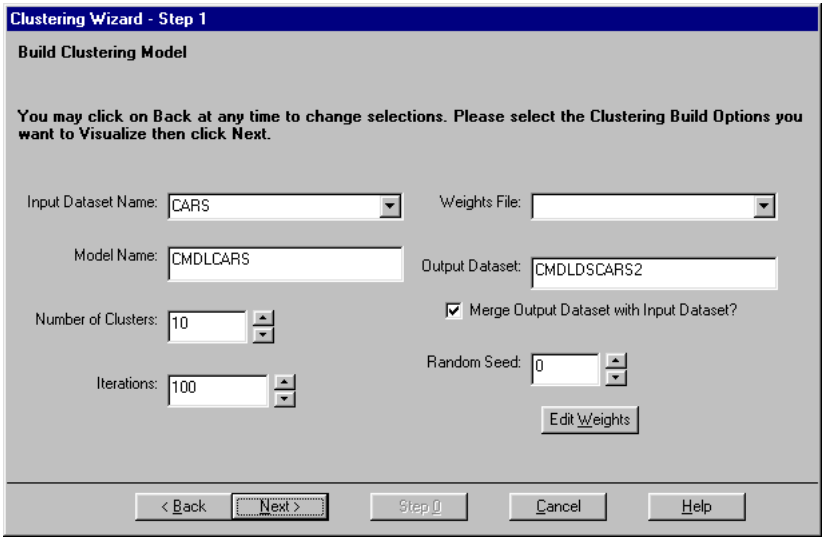
By default, the output dataset is merged with the input dataset (see below).

11.4.1 Clustering — Build Model: Step 1

Build Model | Step 1: Specify Dataset | Step 2:Build Model | Step 3: View Model | Step 4: View Cluster Details | Step 5: Model Info

The **Step 1** dialog prompts you for the following information:

- **Input Dataset Name:** Specify the name of the dataset that contains the records to be clustered.
- **Model Name:** Specify the name of the new model. A default name is offered, based on the name of the dataset specified in the **Input Dataset Name** box. You can change the model name if you wish.



- **Number of Clusters:** Enter the number of clusters desired. The default is 10; the permissible range is 2–100. The greater the number of clusters, the finer the level of granularity.
- **Iterations:** Enter the number of iterations desired. The default is 100; you can set a different limit if you wish.
- **Weights File:** If there is a weights file that you want to use, click the drop-down arrow to find its name here. If there is no weights file that you want to use, leave this box blank. If you are creating a weights file, leave this box blank also; you will be prompted for a name on the **Weights File Contents** dialog (displayed when you click **Edit Weights**). See **Edit Weights**, below.

- **Output Dataset:** The name of the output dataset, which is a new 2-field dataset containing the following information:
  - Cluster assignments, that is, for each record, the identifying integer of the cluster to which the record is assigned.
  - Distance from centroid, that is, for each record, the distance between the cluster's centroid and the record.

  A default name is offered for the output dataset, based on the name of the model. You can change the name if you wish.
- **Merge Output Dataset with Input Dataset?** The default is to merge, as indicated by the checkmark in the box. If you do not want them merged, click the checkbox to uncheck it. The output dataset will then be the 2-field output dataset (cluster assignment and distance from centroid).

  A 2-field output dataset is convenient if, for example, you want to merge it with a different dataset that has only the record identifiers. A merged output dataset is convenient if, for example, you want to pull out the records belonging to a given cluster to use with one of the predictive models.
- **Random Seed:** The random seed determines the positions of the initial candidate centroids. A default value of 0 is offered; you can change it if you wish, to any positive integer.

  Some usage tips:
  - If you want to do another build using the same dataset but with a different randomization, use different random seeds on the two builds, or use a random seed of zero on both builds.
  - If you want to get the same output using the same dataset on another build, use the same *non-zero* random seed value on both builds.
  - If you use zero or different random seed values, you are likely to get different results.

In addition to the navigation buttons, the **Step 1** dialog has an **Edit Weights** command button. See **Edit Weights**, below.

When you are ready to proceed, click **Next**.

Edit Weights

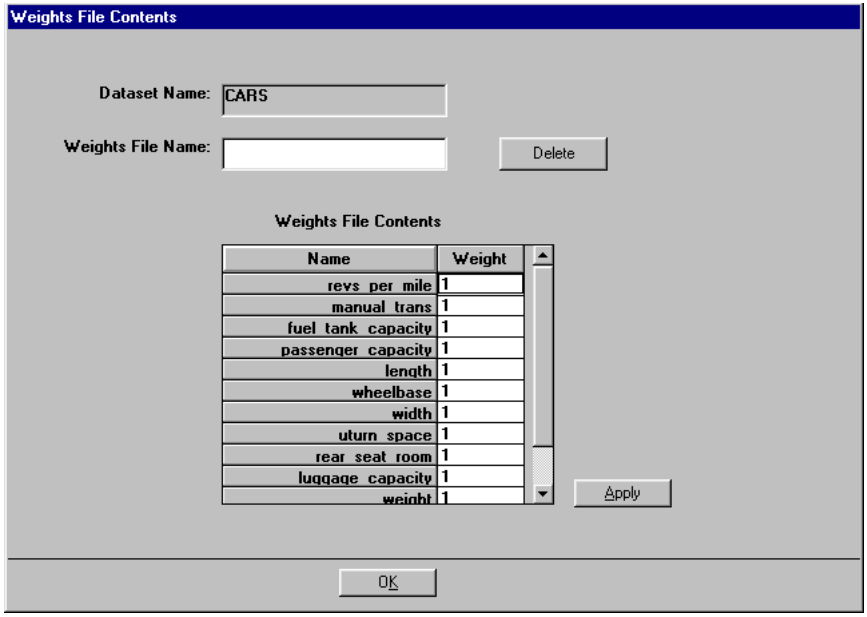
Click **Edit Weights** to bring up the **Weights File Contents** dialog, which you can use to edit an existing weights file or create a new one.

You can use a weights file to emphasize the importance of some fields over others in the clustering build process by giving them higher weight values.

A weight can be any value, including zero. Zero means ignore the field; values other than zero reflect the relative importance you wish to assign to the fields.

The **Weights File Contents** dialog displays the following:

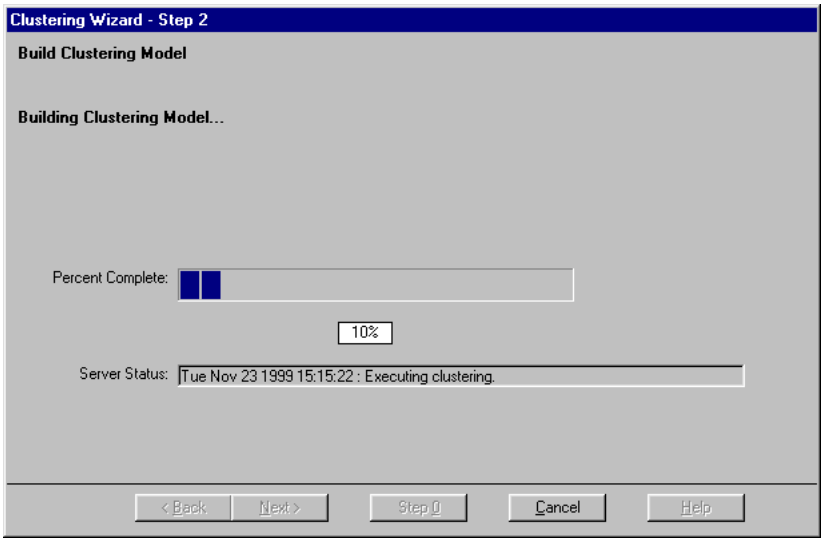
- **Dataset Name:** Name of the dataset.
- **Weights File Name:** If you specified a weights file on the **Step 1** dialog, its name appears here. If you are creating a new weights file, enter a name for it here. The file is automatically saved; it is listed in the Darwin **Workspace**, in the current project, under Other (if you don't see it right away, click V**iew > Refresh** to update the display).
- **Weights File Contents:** Displays the contents of the weights file, if one is specified. Otherwise, displays a list of fields in the dataset, with their default weight values (1), which you can change.



When you have made the desired changes to the weights, click **Apply** to apply the weights file and return to the **Step 1** dialog. If you decide not to change any weights, click **OK** to return to the **Step 1** dialog. To delete the weights file, click **Delete**.

11.4.2 Clustering — Build Model: Step 2

The **Step 2** dialog displays a progress bar that indicates the status of processing, that is, building the model and scoring the input dataset.



When processing is complete, the Wizard proceeds to the next step.

Note: Processing proceeds by functional groups, which means that if you decide to cancel operations (by clicking **Cancel**), there could be a significant delay before processing actually stops.

11.4.3 Clustering — Build Model: Step 3

Build Model | Step 1: Specify Dataset | Step 2:Build Model | Step 3: View Model | Step 4: View Cluster Details | Step 5: Model Info

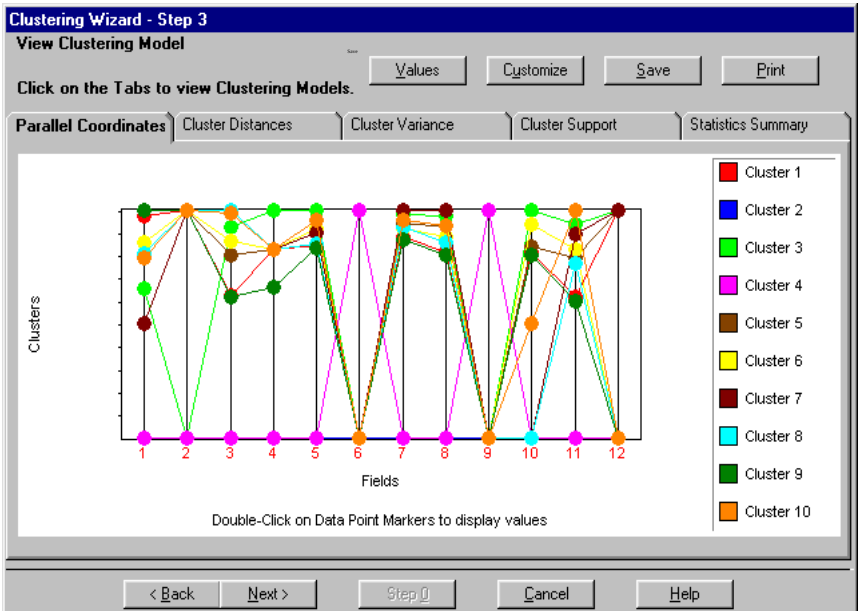
- The **Step 3** dialog, **View Clustering Model**, presents five visualization panels: Parallel Coordinates
- Cluster Distances
- Cluster Variance
- Cluster Support
- Statistics Summary

Click the tabs to go from one panel to another.

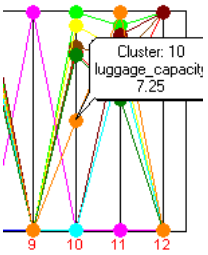
Parallel Coordinates

The first visualization panel is the **Parallel Coordinates** graph, which compares the clusters field by field and displays the values for the fields that constitute each cluster. This graph lets you see at a glance differences between clusters on any given attribute, and gives an overall picture of how the clusters compare.

For example, most of the clusters in the screen capture below follow a similar pattern except for one or two. You might want to take a closer look at some of the clusters to see whether their patterns appear to be related to anything interesting.



Each cluster is represented by a specific color. Color assignments are arbitrary (but consistent), and you can modify them if you wish (see **Customize**, below).



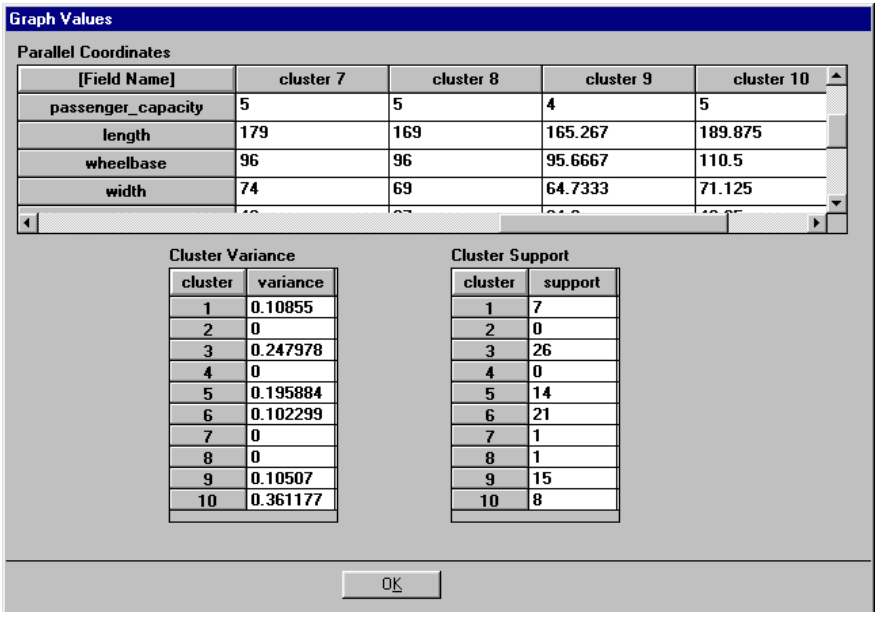
Each cluster line intersects a field line at the relative position for that cluster's centroid at that field. You can display the value at any data point by double-clicking the data point marker. Ordered field values are normalized for purposes of display; categorical field values are evenly distributed from top to bottom.

If the dataset on which the model is based has more fields than can be displayed at one time, a scroll bar appears at the bottom of the chart to let you bring the remaining fields into view.

Values, Customize, Save, and Print

Note the four command buttons at the top right of the parallel coordinates panel; they appear on all five visualization panels:

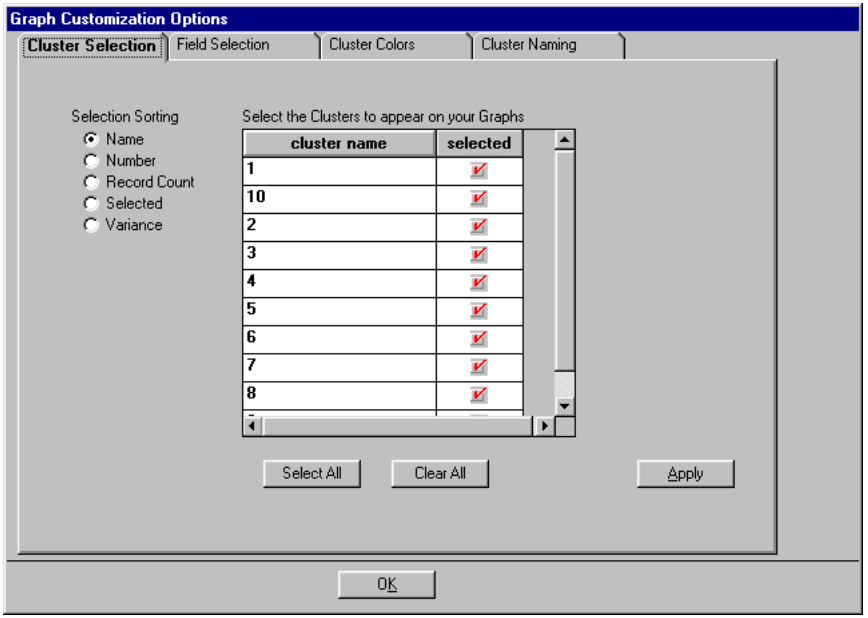
- **Values:** Displays in tabular form the values underlying the visualization panel graphs (parallel coordinates, cluster variance, and cluster support).



- **Customize:** Allows you to customize aspects of clusters and fields, including letting you focus on specific clusters or fields. The customization selections affect all graphs and tables that display cluster information. Clicking **Customize** brings up a dialog with four tabs. Click the tabs to go from one to another:

  Note: For Release 3.6, customization settings are not saved with the model. This means that if you customize a model while you are building it and later view it using the **View Model** path, you will not see it with the customization settings you applied during the build. If you think you might want to recreate the customizations, make a note of them so that you can apply them again during the **View Model** path.

  – **Cluster Selection:** The first tab shows the default cluster selection, that is, all clusters selected, as indicated by a check mark beside each cluster listed. To be selected means to be included in all graphs and tables.



  To focus on a subset of clusters, leave those clusters selected and deselect the others. To deselect a cluster, click its check box to remove the check

mark; that cluster is then excluded from the analysis and from the graphs and tables.

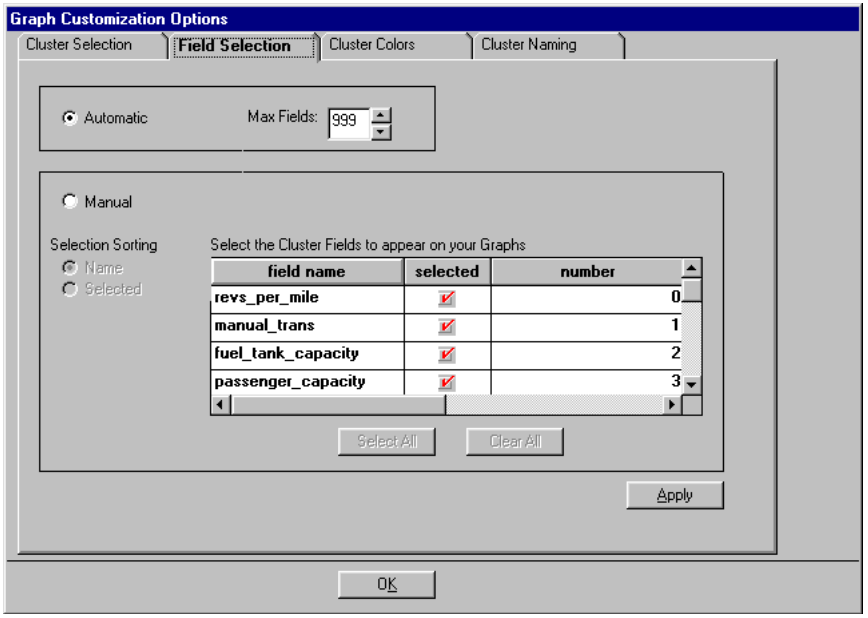
Click **Select All** or **Clear All** to select or deselect all clusters.

In addition, you can sort the clusters by any of the following: Name, Number (cluster identifier), Record Count, Selected (since all are by default selected, you would use this option to deselect the ones you are not interested in, leaving selected only those of interest), and Variance, by clicking an option button under **Selection Sorting**. Clicking a sorting option does the following: (1) adds a third column to the display (except for sorting by Name and Selected), (2) labels the column for the chosen sorting option, (3) lists the sorting option value for each selected field, and (4) sorts selected fields according to their values for the sorting option.

Click **Apply** to apply your settings. Click the next tab (**Field Selection**) to proceed with customizing, or click **OK** to return to the **Step 3** dialog (visualization panels).

– **Field Selection:** On the **Fields Selection** tab, you have two options, **Automatic** and **Manual**:

  * **Automatic** uses all the fields and displays up to the number of fields specified in **Max Fields** (default is 999, which is also the maximum).

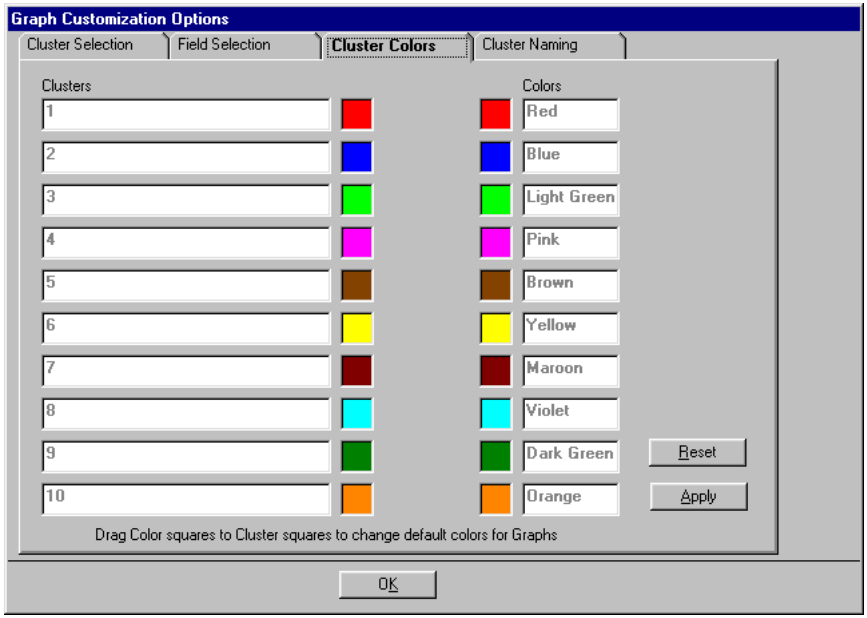


* **Manual** lets you specify the fields you want to include or exclude from the graphs and tables and lets you sort the fields by Name or by Selected.

  Select All and **Clear All** are enabled in the **Manual** option. Use them to select or deselect all fields. By default, all fields are selected, as indicated by a check mark beside each field name in the list. To exclude a field from the graphs and tables, click its check box to deselect it.

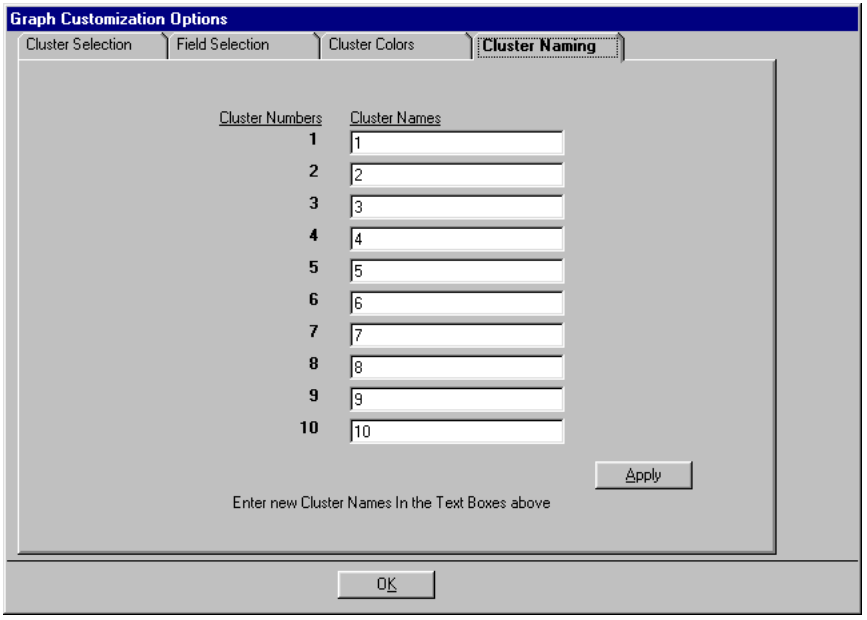
Click **Apply** to apply your settings. Click the next tab (**Cluster Colors**) to continue with customizing, or click **OK** to return to **Step 3** dialog (visualization panels).

– **Cluster Colors:** To change a color assignment, drag a color from the right-hand column to the left-hand column. Click **Reset** to restore the default assignments.



Click **Apply** to apply your settings. Click the next tab (**Cluster Naming**) to continue with customizing, or click **OK** to return to **Step 3** dialog (visualization panels).

– **Cluster Naming:** The **Cluster Naming** tab lets you assign names to clusters. Enter the names in the text boxes.



Click **Apply** to apply your settings. This is the fourth and last customization tab. Click **OK** to return to the **Step 3** dialog (visualization panels).

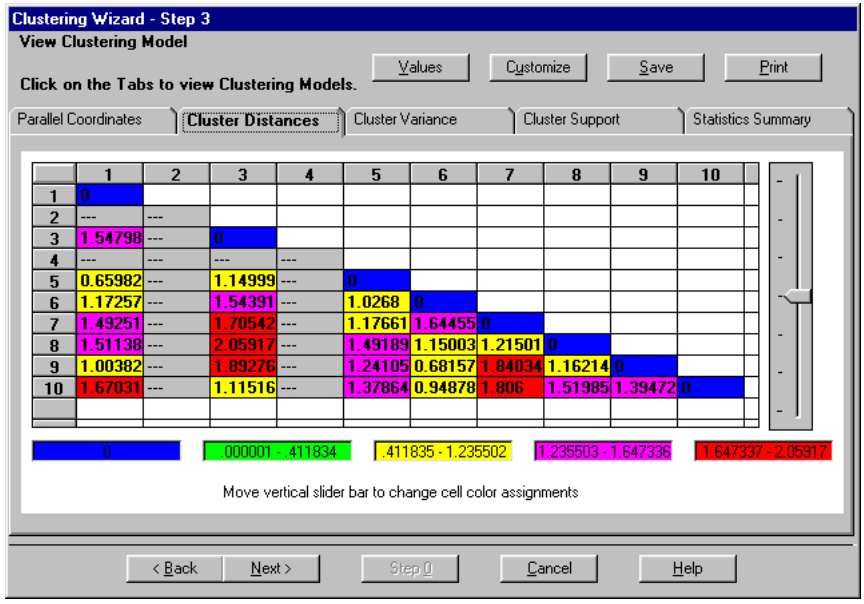
The last two command buttons at the top right of the visualization panels are **Save** and **Print:**

- **Save:** Brings up a **Save As** dialog, on which you can enter a name for the file and specify where it is to be saved. Charts and graphs are saved to bitmap files; tables are saved to text files (with a `.txt` extension).
- **Print:** Sends the displayed graph or table to your printer. There are some differences between the online display of the graphs and tables and their hardcopy printout.

Cluster Distances

Cluster Distances is the second visualization panel in the **Step 3** dialog.

This table displays distances between all pairs of cluster centroids. For example, the value 1.14999 at the intersection of cluster 3 and 5 represents the distance between the centroids for clusters 3 and 5. The greater the distance between clusters, the more different the clusters are.



Dashes in a table cell mean that the cluster is empty, i.e., has no members. See **Statistics Summary** for an explanation of empty clusters.

Below the matrix is a row showing the color assigned to each range of cluster distances. Blue is assigned to least distant; red is assigned to most distant.

The slider at the right lets you change color groupings. Moving the slider up lowers the threshold and includes more clusters in the blue space (low values, less distance). Moving the slider down raises the threshold and includes more clusters in the red space (high values, greater distance).

You can save or print any of the visualization panels (see page 11-16). The printout of **Cluster Distances** does not include the slider bar at the right or the range values below the grid.

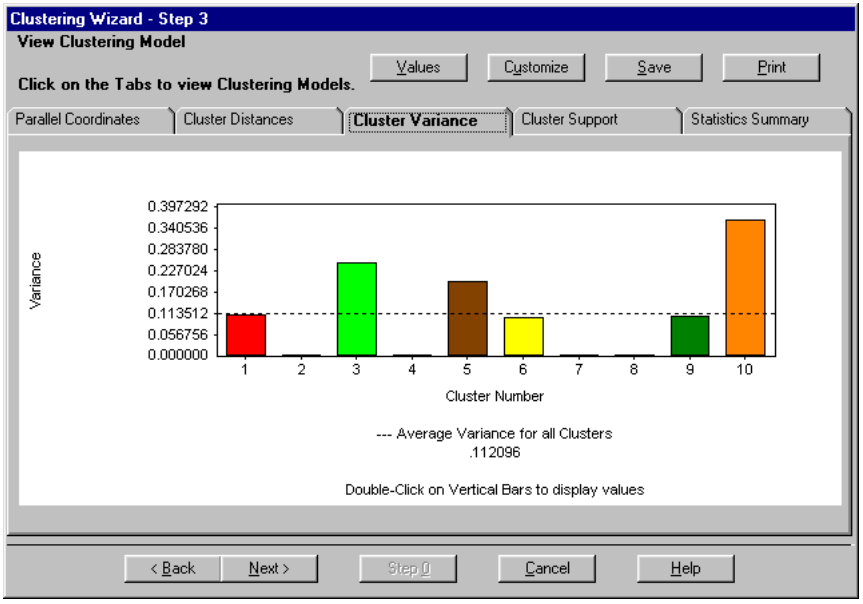
Cluster Variance

Cluster Variance is the third visualization panel in the **Step 3** dialog.

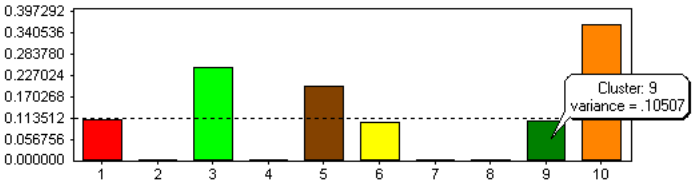
Cluster variance is shown in a bar chart, with the bars representing the variance for each cluster, considering each record in the cluster with respect to the centroid. The dotted line represents the average variance of all clusters displayed.

Clusters with low variability are more homogeneous than clusters with more variability:

- low variance: tighter, more homogeneous cluster, better defined
- high variance: looser, less homogeneous cluster, less well defined



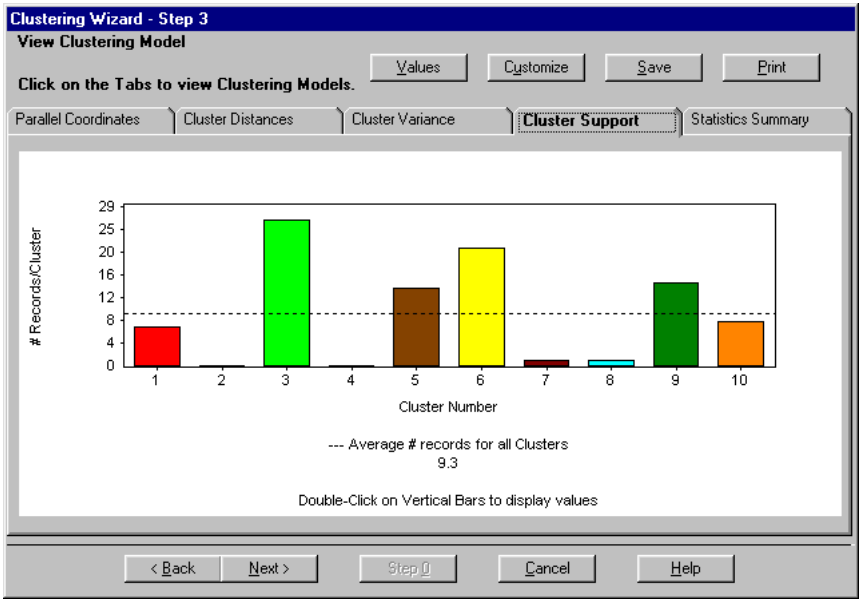
Double-click a bar to display the variance value for that cluster (or click **Values** to display the values underlying all the graphs). When you double-click a bar whose value is zero, nothing happens (clusters 2, 4, 7, and 8 here have zero variance).



Cluster Support

Cluster Support is the fourth visualization panel in the **Step 3** dialog.

Cluster support is shown in a bar chart, with the bars representing the number of records from the training dataset that are assigned to each cluster. The dotted line represents the average number of records for all clusters displayed.



Double-click a bar to display the number of records in that cluster.

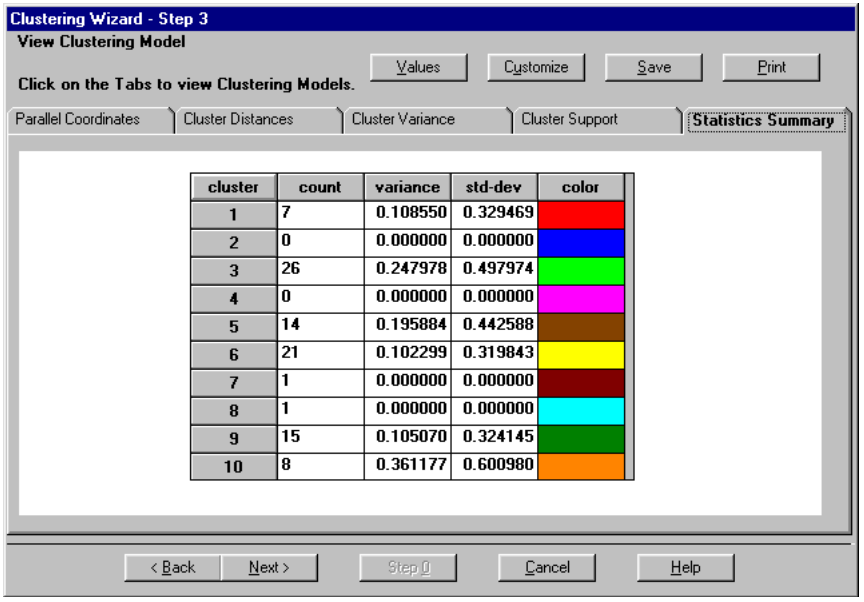
Clusters 2 and 4 have no support (value = 0), and clusters 7 and 8 have one record each. This explains why these clusters have zero variance (see **Cluster Variance**).

Cluster 3 has good support (26) but somewhat high variance (.248). Cluster 6 has relatively good support (21) and lower variance (.1023); it is tighter cluster. These two clusters have the most support, and might be worth analyzing further. For example, it would be interesting to know the rules that determined them.

Statistics Summary

Statistics Summary is the fifth visualization panel in the **Step 3** dialog. It displays the following statistics for each cluster:

- count
- variance
- standard deviation



cluster	count	variance	std-dev	color
1	7	0.108550	0.329469	
2	0	0.000000	0.000000	
3	26	0.247978	0.497974	
4	0	0.000000	0.000000	
5	14	0.195884	0.442588	
6	21	0.102299	0.319843	
7	1	0.000000	0.000000	
8	1	0.000000	0.000000	
9	15	0.105070	0.324145	
10	8	0.361177	0.600980	

As we noted before, clusters 2 and 4 are empty. Empty clusters can happen because the algorithm (*k*-means) used to create these clustering models is a form of competitive learning, in which it is possible to have a centroid that is never determined to be the "winner" for record assignment. The random seed that initialized the cluster centroids can affect how records initially get assigned to centroids. Hence, some clusters can have zero records assigned.

You can save or print any of the visualization panels (see page 11-16). The printout of **Statistics Summary** does not include the **Color** column.

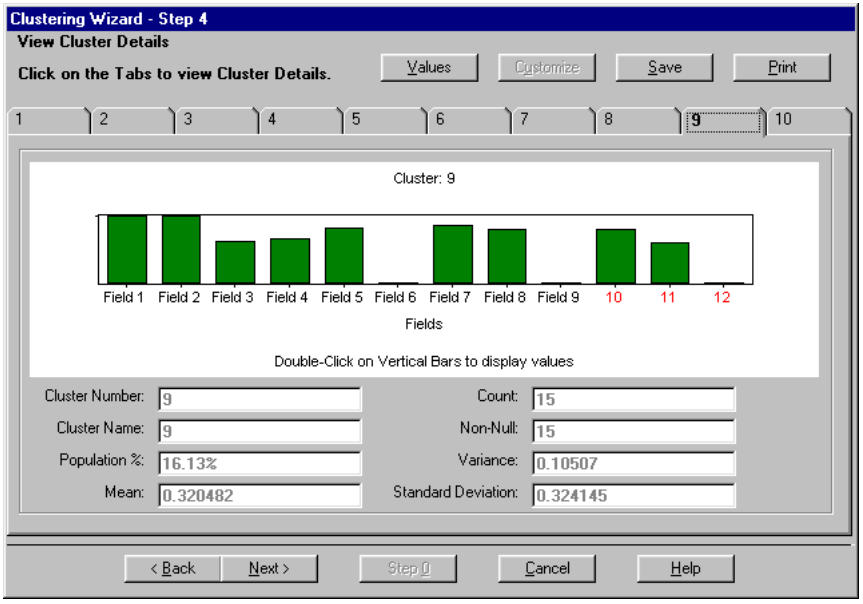
When you have finished examining the graphs and tables displayed on the **Step 3** dialog, **View Clustering Model**, click **Next** to proceed.

Next takes you to graphs and tables that present the details of each cluster.

11.4.4 Clustering — Build Model: Step 4

Build Model | Step 1: Specify Dataset | Step 2:Build Model | Step 3: View Model | Step 4: View Cluster Details | Step 5: Model Info

The **Step 4** dialog, **View Cluster Details**, presents the details for each cluster. This dialog has a tab for each cluster; it can display ten clusters at a time. For each cluster, the Wizard displays a bar chart, with the bars representing fields, and the statistics for that cluster. The bar graph below displays field details for cluster 9.



Bar heights are relative, based on normalized values. In interpreting this graph, be aware that a very tall bar may have a low value, and a short bar can have a high value — the bar heights are relative for the *field*, that is, the tallest bar in field 1 for cluster 1 means that this is the cluster in which this field has the highest value.

Color assignments are arbitrary (but consistent), and you can change them. (If you want to change them, click **Back** to return to the previous step, where **Customize** is enabled.)

The table below the bar chart displays the cluster's defining values: cluster number, cluster name, population (percentage of records from the dataset population that belongs to this cluster), mean, count, non-null count, variance, and standard deviation.

Values, Save, and Print

Note the command buttons at the top right, **Values**, **Save**, and **Print**. (**Customize** is not available at this step; any customizing must be performed in the previous step.)

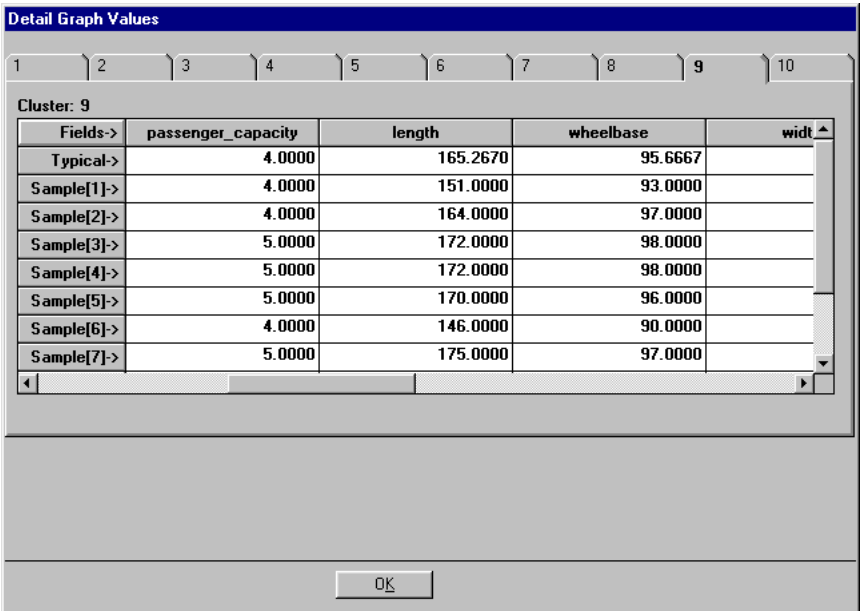
- **Values:** Clicking **Values** brings up a table that shows individual record values for all fields for each cluster. Click the tabs to move from one cluster to another.

  The tables list 11 records:

  – The first record is a "typical" record, that is, the centroid.

  – Records 2 through 11 are sample records from the dataset.

  Use the scroll bars to scroll through all the sample records and all the fields.



  These details can be useful to examine if a field's values appear to be too far from the typical, which may indicate that this particular field is not valuable to this cluster's definition.

- **Save:** Brings up a **Save As** dialog, on which you can enter a name for the file and specify where it is to be saved. Charts and graphs are saved to bitmap files; tables are saved to text files (with a `.txt` extension).

- **Print:** Sends the displayed graph or table to your printer.

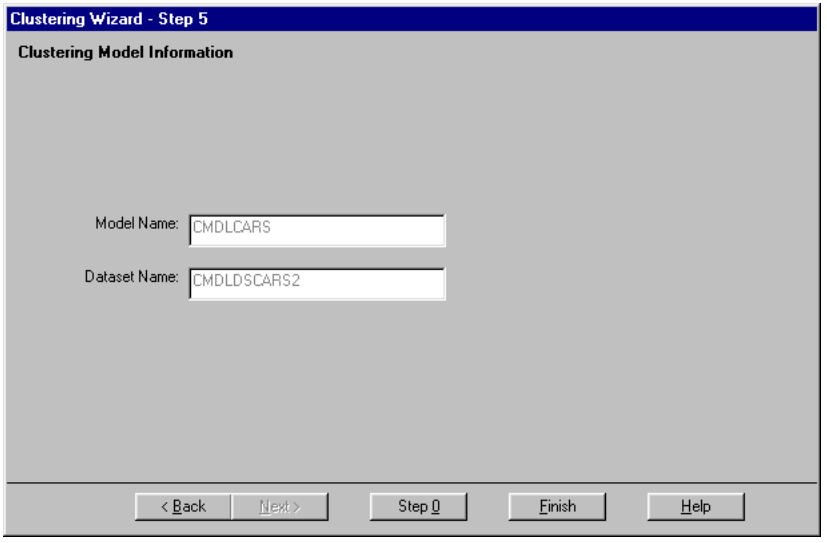
When you have finished examining the details of each cluster and are ready to proceed, click **Next**. **Next** takes you to **Step 5**, the **Clustering Model Information** dialog.

11.4.5 Clustering — Build Model: Step 5

Step 5, **Clustering Model Information**, is the last step in the **Build Model** path. This dialog displays the model's name and the name of the output dataset. The output dataset is either the new 2-field dataset (the fields are cluster assignment and distance of record from centroid) or the merged dataset (merged output and input datasets), depending on the choice you made on the **Step 1** dialog.

Clustering models and the output datasets are saved automatically (and separately):

- Clustering models are saved to your UNIX `darwin` directory, in the subdirectory for the particular project. They are not listed with other Darwin models in the Darwin client **Workspace** listing.
- Clustering output datasets are listed in the Darwin **Workspace** under Datasets, Created, if it is the 2-field output dataset. If it is the *merged* dataset, it is listed under Datasets, Transformed.



To see -- from within Darwin -- a list of all saved clustering models, start the Clustering Wizard and, on the **Step 0** dialog, select **Apply Model** or **View Model**.

The drop-down list for **Select Model Name** lists all saved Darwin clustering models in the current Darwin project.

(Remember that you must be in the right Darwin project before you invoke the Clustering Wizard, i.e., once you have started the Wizard, you have access only to datasets and models that are in your current Darwin project.)

Click **Step 0** to return to the **Step 0** dialog, where you can select a different path, or click **Finish** to exit the Wizard and return to the Darwin client.

11.5 Darwin Clustering — Apply Model

Apply Model is the second functional path on the **Step 0** dialog.

Choose this option if you want to apply an existing clustering model to a different (but compatible) dataset. (See Section 11.5.1, below, 2d bullet, "Important").

The output of the **Apply Model** path is a new 2-field dataset with as many records as the input dataset:

- The first field holds the cluster identification integer that defines the cluster to which each record belongs.
- The second field holds a value representing the distance between each record and the cluster's centroid.

By default, the output dataset is merged with the input dataset (see below).

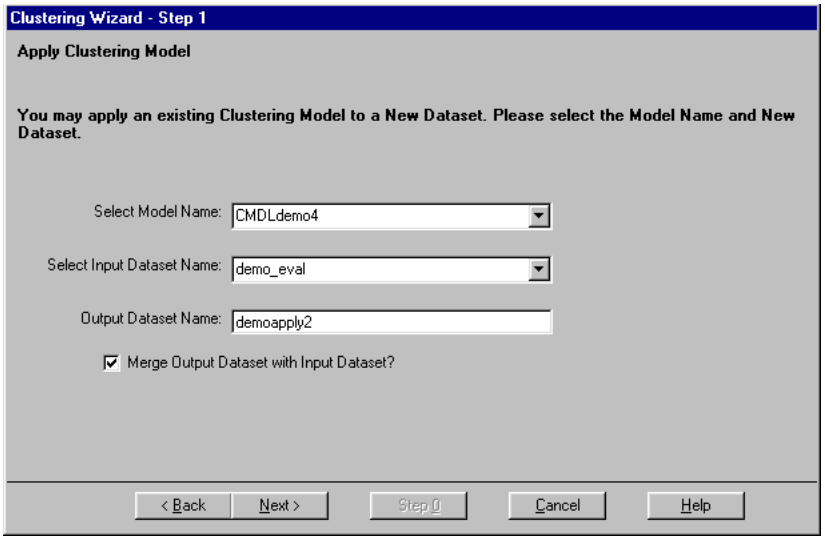
11.5.1 Clustering — Apply Model: Step 1

The **Step 1** dialog prompts you for the following information:

- **Select Model Name:** Scroll the drop-down list to find the name of the clustering model you want to use to score the new dataset.
- **Select Dataset Name:** Scroll the drop-down list to find the dataset that you wish to apply the model to. The dataset name that appears by default is not necessarily a dataset that is associated with the model.

  Important: The original and new datasets must be compatible, i.e., they must have the same dataset descriptor and must have undergone the same transformations in the same order.
- **Output Dataset Name:** Enter a name for the output dataset, which is a new 2-field dataset containing the following information:
  - Cluster assignments, that is, for each record, the identifying integer of the cluster to which the record is assigned.
  - Distance from centroid, that is, for each record, the distance between the cluster's centroid and the record.
- **Merge Output Dataset with Input Dataset?** The default is to merge, as indicated by the checkmark in the box. If you do not want them merged, click the checkbox to uncheck it. The output dataset will then be the 2-field output dataset (cluster assignment and distance from centroid).

A 2-field output dataset might be convenient if, for example, you want to merge it with a different dataset that has only the record identifiers. A merged output dataset is convenient if, for example, you want to separate the records belonging to a given cluster to use with one of the predictive models.

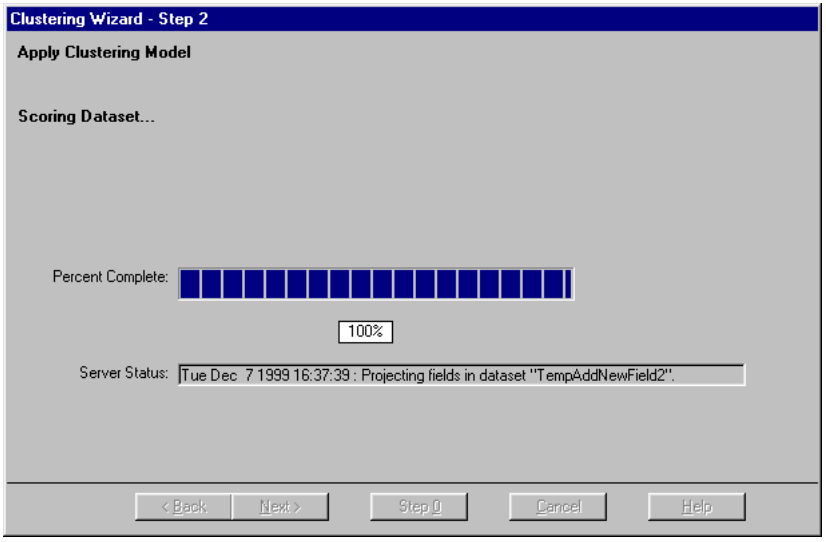


When all the necessary information is entered, click **Next** to proceed.

11.5.2 Clustering — Apply Model: Step 2

The **Step 2** dialog displays a progress bar that indicates the status of the processing, i.e., scoring the new dataset.

Processing occurs by functional groups, which means that if you decide to cancel operations (by clicking **Cancel**), there could be a significant delay before processing actually stops.



When processing is complete, the Wizard proceeds to the next step.

11.5.3 Clustering — Apply Model: Step 3

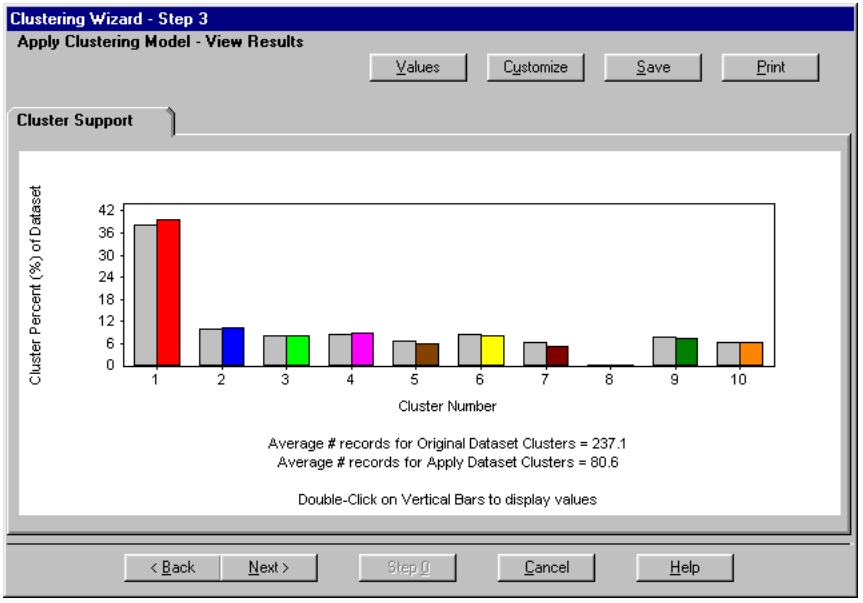
The **Step 3** dialog, **View Results**, presents one visualization panel:

- Cluster Support

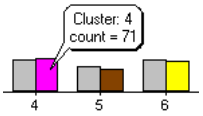
Cluster Support

The cluster support graph compares the support of the model's original clusters with the support of the clusters that result from applying the model to the new dataset.

In each pair of bars, the grey bar on the left represents the original dataset; the colored bar on the right represents the new dataset. The values plotted on this bar are the *percentages* of records that belong to the particular cluster in the original and new dataset.



Double-click a bar to display its underlying value:

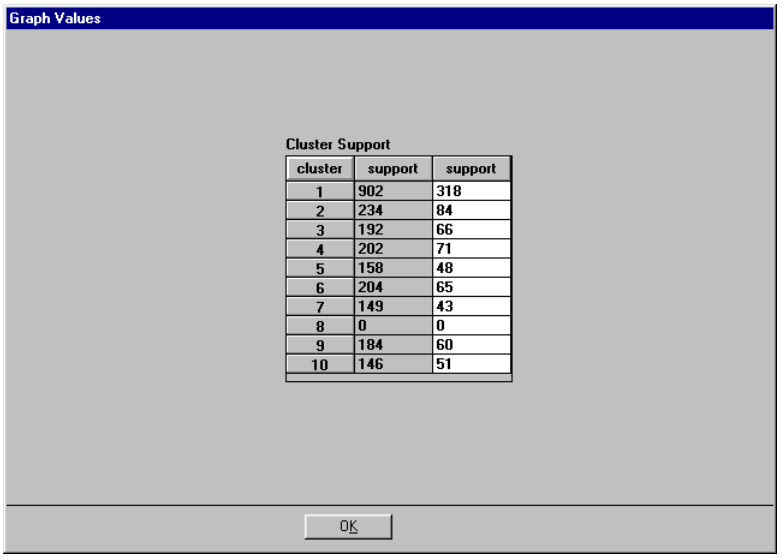


Values, Save, and Print

Note the command buttons at the top right of the visualization panel:

- **Values:** Displays in tabular form the values underlying the cluster support graph. These values are converted to percentages for plotting on the bar graph.

  The grey column on the left holds the values for the original dataset; the white column on the right holds the values for the new dataset.



cluster	support	support
1	902	318
2	234	84
3	192	66
4	202	71
5	158	48
6	204	65
7	149	43
8	0	0
9	184	60
10	146	51

- **Customize:** The only customization option available in the **Apply** path is Cluster Selection; the other customizing options are not available at Release 3.6

  By default, all clusters are selected. To focus on a subset of clusters, leave those clusters selected and deselect the others by clicking their checkboxes. Click **Apply** to apply your settings; then click **OK** to return to the graph, where you will see the results of your choices.

- **Save:** Brings up a **Save As** dialog, on which you can enter a name for the file and specify where the file is to be saved. Charts are saved to bitmap files; tables are saved to text files (with a `.txt` extension).

- **Print:** Sends the displayed graph or table to your printer.

Click **OK** to return to the **Step 3** dialog (visualization panel). Then click **Next** to proceed.

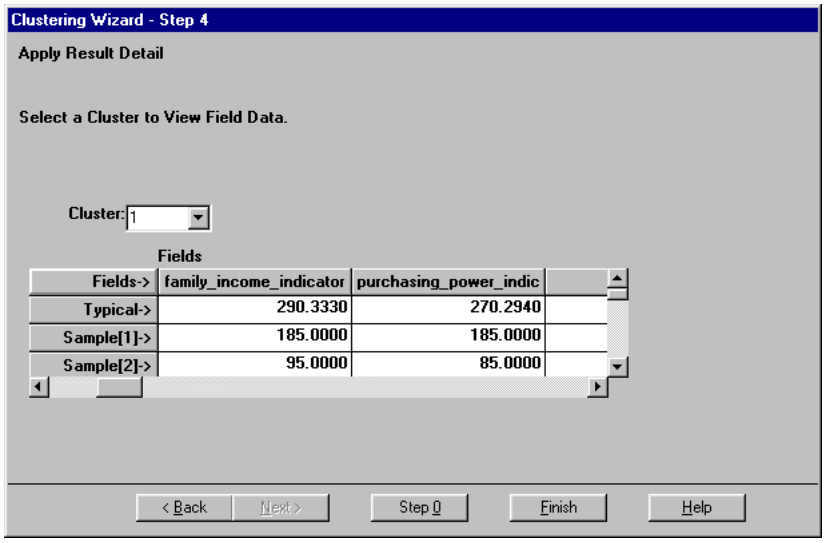
11.5.4 Clustering — Apply Model: Step 4

The **Step 4** dialog, **Apply Result Detail**, is the last step in the **Apply Model** path. It displays a grid containing individual record values for all fields for each cluster you specify. Use the **Cluster** list box to specify a cluster.

The grid lists 11 records:

- The first record is a “typical” record, that is, the centroid.
- Records 2 through 11 are sample records from the dataset.

Use the scroll bars to scroll through all the sample records and all the fields.



The result details can be useful to examine manually if a field’s values appear to be too far from the typical. This may indicate that this particular field is not valuable to this cluster’s definition.

When you have finished examining the result details, click **Step 0** to return to the **Step 0** dialog, where you can specify another path, or click **Finish** to exit the Wizard and return to the Darwin client.

The output dataset is listed in the Darwin **Workspace** under Datasets, Created, if it is the 2-field output dataset. If it is the *merged* dataset, it is listed under Datasets, Transformed.

11.6 Darwin Clustering — View Model

View Model is the third functional path on the **Step 0** dialog.

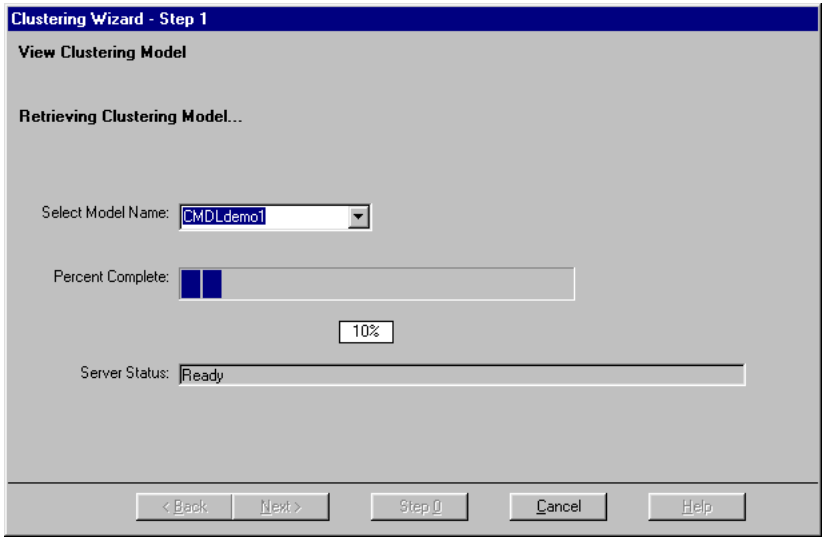
Choose this option to retrieve and view a previously built clustering model and to customize it.

11.6.1 Clustering — View Model: Step 1

This dialog contains the following:

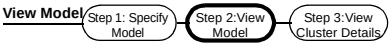
- **Select Model Name:** Scroll the list to find the name of the model you want to view.
- **Percent Complete:** A progress bar indicating status of the retrieval.

Click **Next** to retrieve the model. The progress bar indicates the status of the retrieval. When processing is complete, the Wizard proceeds to the next step.



If you change your mind at any point and click **Cancel**, be aware that processing occurs by functional groups, and that therefore when you click **Cancel**, there could be a significant delay before processing actually stops.

11.6.2 Clustering — View Model: Step 2



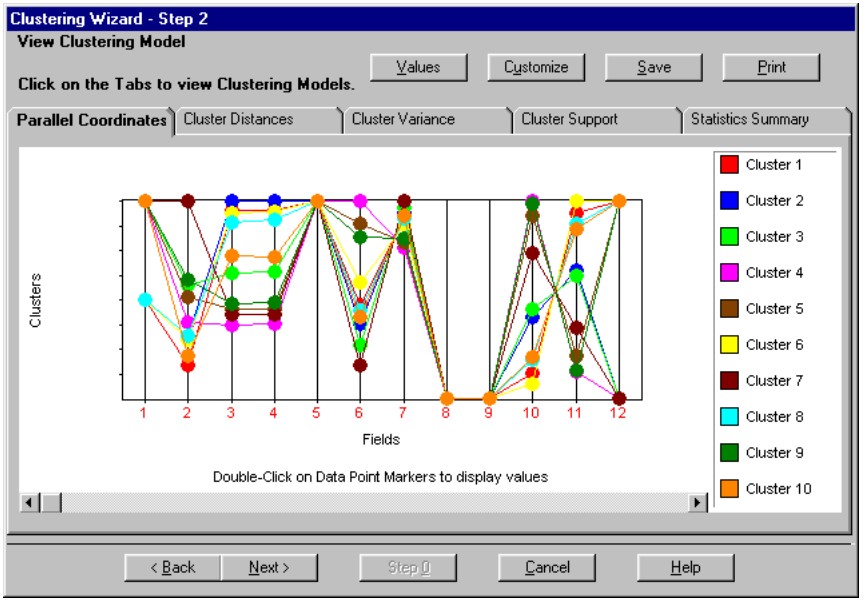
The **Step 2** dialog presents five visualization panels:

- Parallel Coordinates
- Cluster Distance
- Cluster Variance
- Cluster Support
- Statistics Summary

Click the tabs to go from one visualization panel to another. (These are the same visualization panels that are displayed in the **Build Model** functional path.)

Parallel Coordinates

The first panel is the **Parallel Coordinates** graph, which presents a field-by-field comparison of all the clusters, and displays the values for the fields that constitute each cluster. This graph lets you see at a glance how the clusters compare — how similar their patterns are, how different they are on any given attribute, whether there are any that are markedly different from all the others, etc.



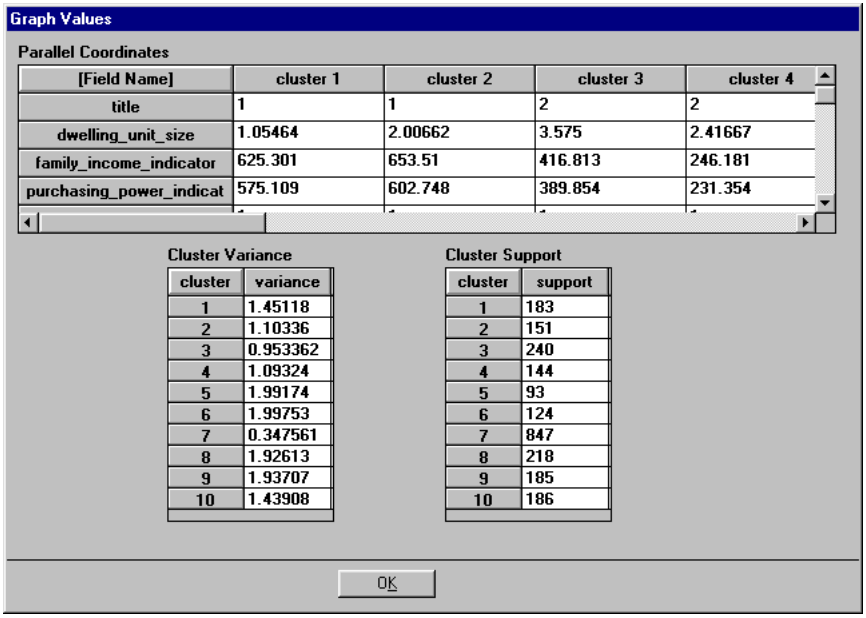
Double-click a data point to display the value for that point.

Note the scroll bar at the bottom of the chart. This scroll bar appears whenever you are viewing a model that is based on a dataset with more than 12 fields. In this case, the dataset has 23 fields. Use the scroll bar to display the rest of the parallel coordinates chart.

Values, Customize, Save, and Print

Note the command buttons at the top right of the parallel coordinates panel; they apply to all five visualization panels:

- **Values:** Displays in tabular form the values underlying the graphs (parallel coordinates, cluster variance, and cluster support).



- **Customize**: Allows you to customize aspects of clusters and fields, including letting you focus on specific clusters or fields. The customization selections affect all graphs and tables that display cluster information. **Customize** brings up a dialog with four tabs. Click the tabs to go from one to another. See Section 11.4.3 in the **Build Model** path for a description of the four customization options. For Release 3.6, customization settings are not saved with the model.

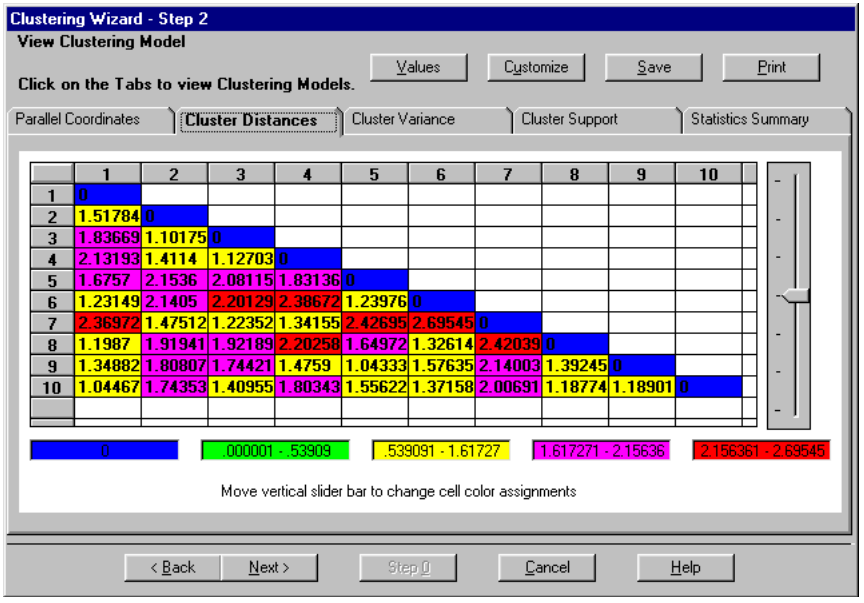
- **Save:** Brings up a **Save As** dialog, on which you can enter a name for the file and specify where the file is to be saved. Charts and graphs are saved to bitmap files; tables are saved to text files (with a `.txt` extension).
- **Print:** Sends the displayed graph or table to your printer.

Click **OK** to return to the **Step 2** dialog (visualization panels).

Cluster Distances

Cluster Distances is the second panel in the **Step 2** dialog.

Cluster Distances is a table that shows the distances between all pairs of cluster centroids. For example, the value 1.12703 at the intersection of cluster 3 and 4 represents the distance between the centroids for clusters 3 and 4. The greater the distance between clusters, the more different the clusters are.



Dashes in a table cell mean the cluster is empty, i.e., has no members (see **Statistics Summary** for an explanation of empty clusters).

Below the matrix is a row showing the color assigned to each range of cluster distances. Blue is assigned to least distant; red is assigned to most distant.

The slider at the right lets you change color groupings. Moving the slider up lowers the threshold and includes more clusters in the blue space (low values, less

distance). Moving the slider down raises the threshold and includes more clusters in the red space (high values, greater distance).

Cluster Variance

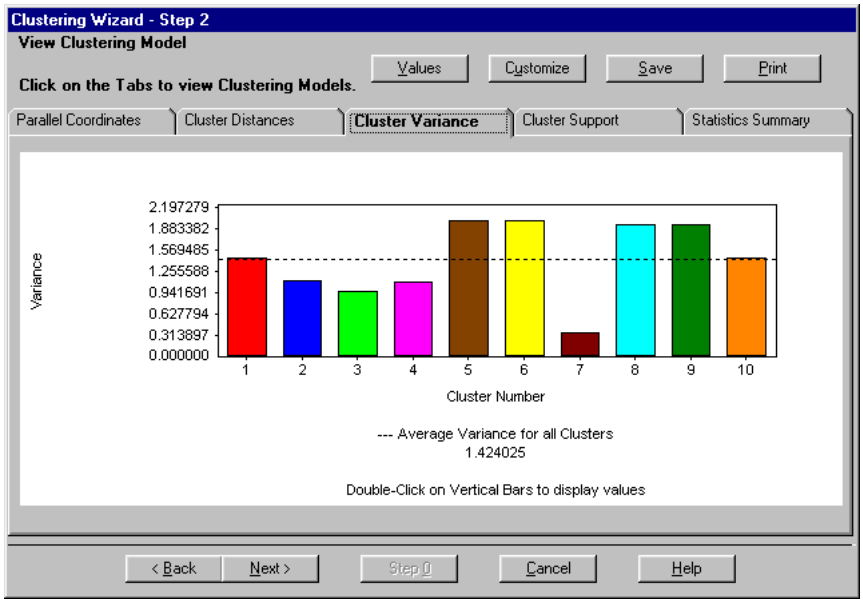
Cluster Variance is the third panel in the **Step 2** dialog.

Cluster variance is shown in a bar chart, with the bars representing the variance for each cluster, considering each record in the cluster with respect to the centroid. The dotted line represents the average variance of all clusters displayed.

Clusters with low variability are more homogeneous than clusters with more variability:

- low variance: tighter, more homogeneous cluster
- high variance: looser, less homogeneous cluster

Double-click a bar to display the specific variance for that cluster. If the value is 0, nothing happens when you double-click the bar.

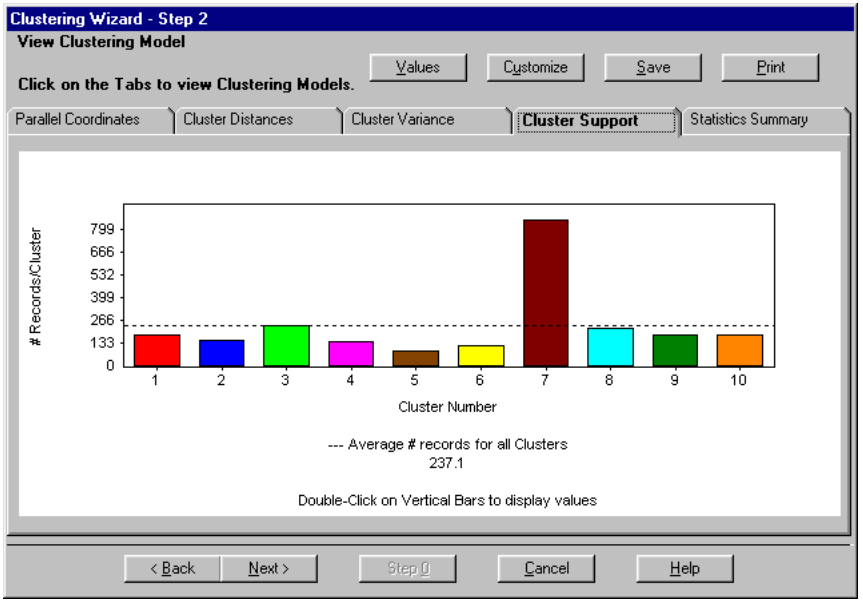


Cluster Support

Cluster Support is the fourth panel in the **Step 2** dialog.

Cluster Support is shown in a bar chart, with the bars representing the number of records in each cluster. The dotted line represents the average number of records for all clusters displayed.

Double-click a bar to display the number of records in that cluster. If there are no records in a particular cluster, nothing happens when you double-click the bar.

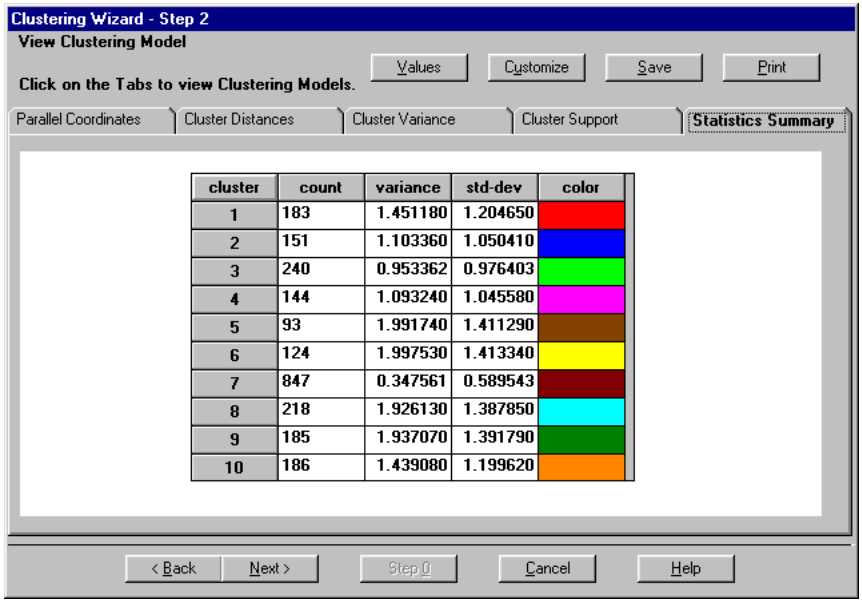


Statistics Summary

Statistics summary is the fifth and last panel in the **Step 2** dialog.

The statistics summary table displays the following statistics for each cluster:

- count
- variance
- standard deviation



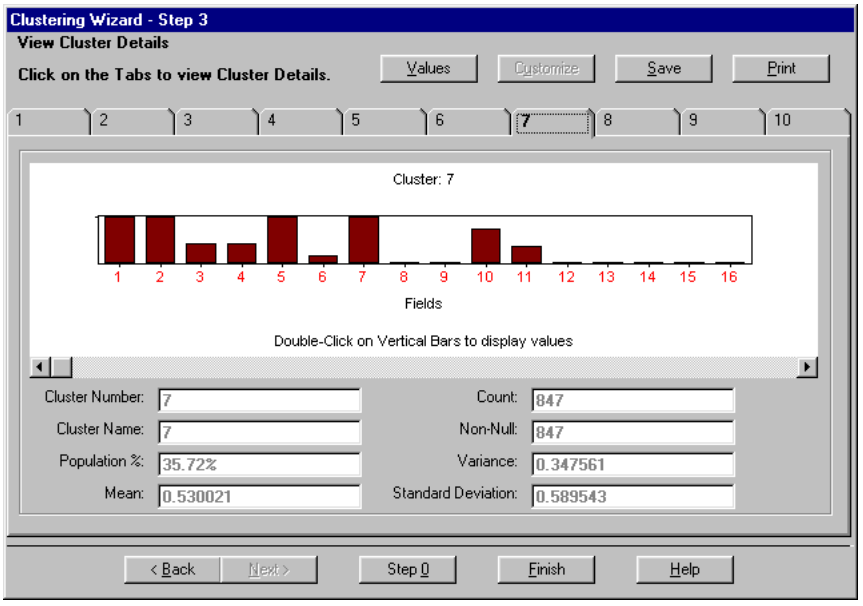
cluster	count	variance	std-dev	color
1	183	1.451180	1.204650	
2	151	1.103360	1.050410	
3	240	0.953362	0.976403	
4	144	1.093240	1.045580	
5	93	1.991740	1.411290	
6	124	1.997530	1.413340	
7	847	0.347561	0.589543	
8	218	1.926130	1.387850	
9	185	1.937070	1.391790	
10	186	1.439080	1.199620	

If you find empty clusters, you may wonder how they can happen. Empty clusters can happen because the algorithm used to create these clustering models (the *k*-means algorithm) is a form of competitive learning, in which it is possible to have a centroid that is never determined to be the "winner" for record assignment. The "seed" chosen to initialize the cluster centroid can affect how records initially get assigned to centroids. Hence, some clusters can have zero records assigned.

When you have finished examining the graphs and table displayed on the **View Clustering Model** dialog, click **Next** to proceed. **Next** takes you to graphs and tables that present the details of each cluster.

11.6.3 Clustering — View Model: Step 3

The **Step 3** dialog, **View Cluster Details**, presents the details for each cluster. This dialog has a tab for each cluster; it can display ten clusters at a time. For each cluster, the Wizard displays a bar chart, with the bars representing fields, and the statistics for that cluster. The screen capture below displays field details for cluster 7.



Bar heights are relative, based on normalized values. In interpreting this graph, be aware that a tall bar may have a low value, and a short bar can have a high value — the bar heights are relative for the *field*, that is, the tallest bar in field 1 for cluster 1 means that this is the cluster in which this field has the highest value.

The table below the bar chart displays the cluster's defining values: cluster number, cluster name, population (percentage of records from the dataset population that belongs to this cluster), mean, count, non-null count, variance, and standard deviation.

Values, Save, and Print

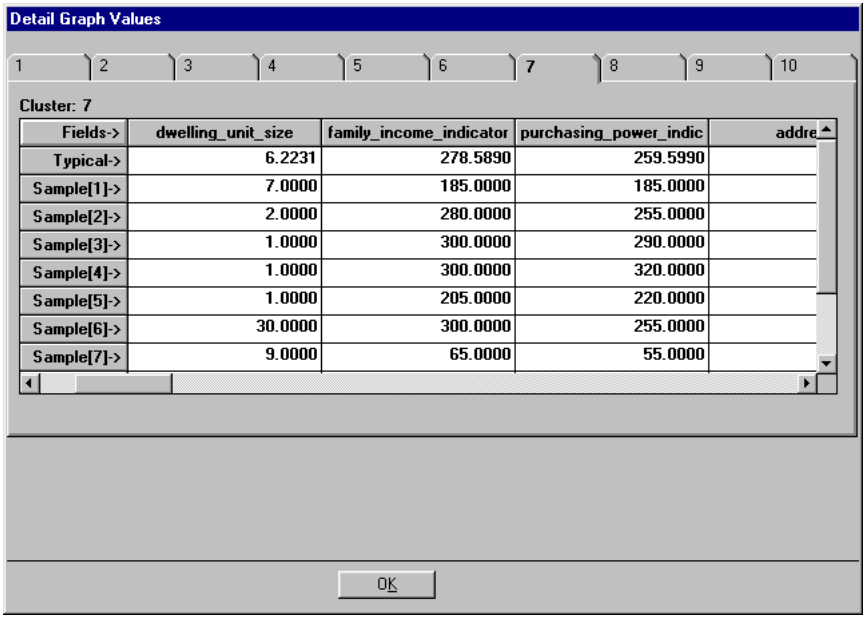
Note the command buttons at the top right, **Values**, **Save**, and **Print**. (**Customize** is not available in **Step 3** of the **View Model** path.)

- **Values:** Clicking **Values** brings up a table for each cluster that contains individual record values for all fields for that cluster. Click the tabs to move from one cluster to another. The tables list 11 records: the first record is a "typical" record, that is, the centroid; records 2 through 11 are sample records from the dataset.

  Use the scroll bars to scroll through all the sample records and all the fields.

- **Save:** Brings up a **Save As** dialog, on which you can enter a name for the file and specify where it is to be saved. Charts and graphs are saved to bitmap files; tables are saved to text files (with a `.txt` extension).

- **Print:** Sends the displayed graph or table to your printer.



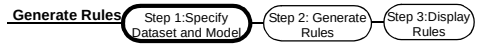
Step 3 is the last step in the **View Model** path. Click **Step 0** to return to the **Step 0** dialog, where you can select another path, or you can click **Finish** to exit the Wizard and return to the Darwin client.

11.7 Darwin Clustering — Generate Rules

Generate Rules is the fourth path on the **Step 0** dialog.

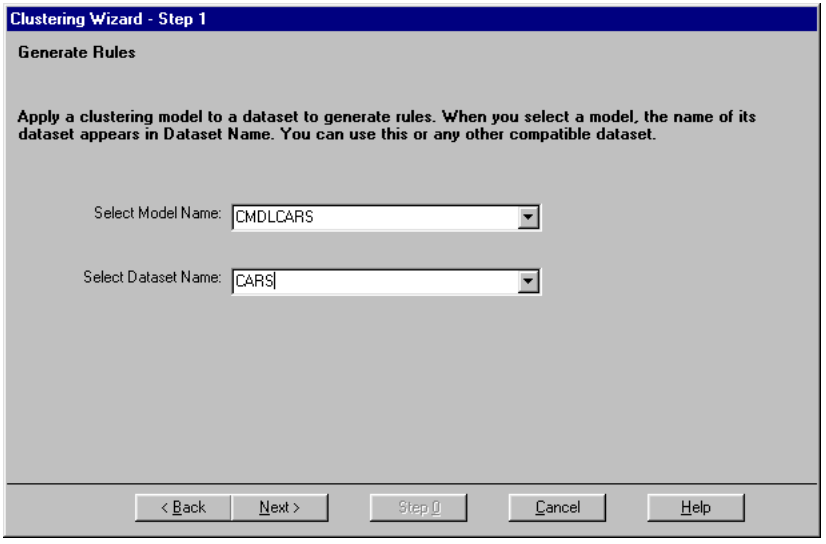
This option lets you generate the rules that determine the clusters in a specified model. The rules can sometimes provide help in understanding the meaning of the clusters.

11.7.1 Clustering — Generate Rules: Step 1



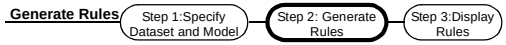
The **Step 1** dialog prompts you for the name of a model and dataset.

- **Select Model Name:** Scroll the drop-down list to find the name of the model whose rules you wish to generate.
- **Select Dataset Name:** When you select a model name in **Select Model Name**, the name of the dataset on which it was built appears in **Select Dataset Name**. You can generate rules for this dataset or for any other compatible dataset. (A compatible dataset is one that has the same dataset descriptor and has undergone the same transformations in the same order.)



Click **Next** to proceed.

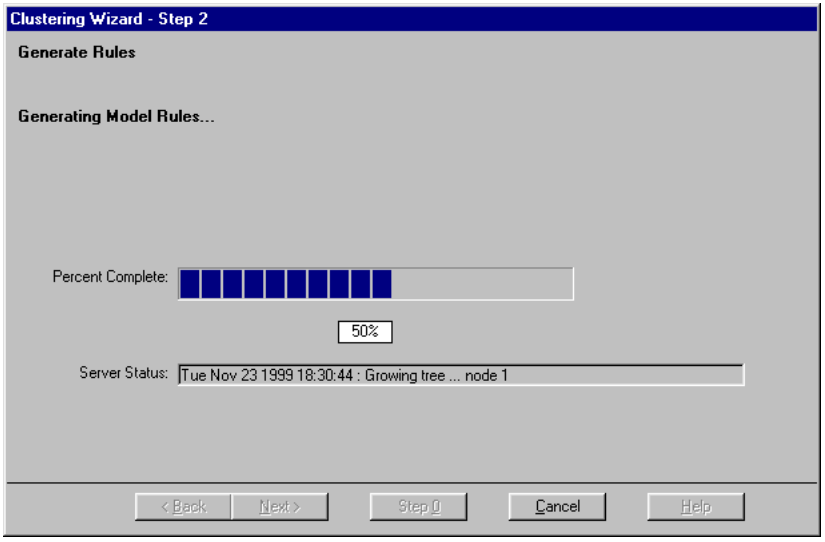
11.7.2 Clustering — Generate Rules: Step 2



The **Step 2** dialog displays a progress bar that indicates the status of the operation.

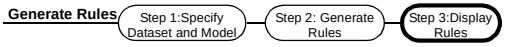
Click **Next** to begin processing.

When processing is completed, the Wizard takes you to the next step.



If you change your mind and click **Cancel**, be aware that there could be a significant delay before processing actually stops, because the operation proceeds by functional groups.

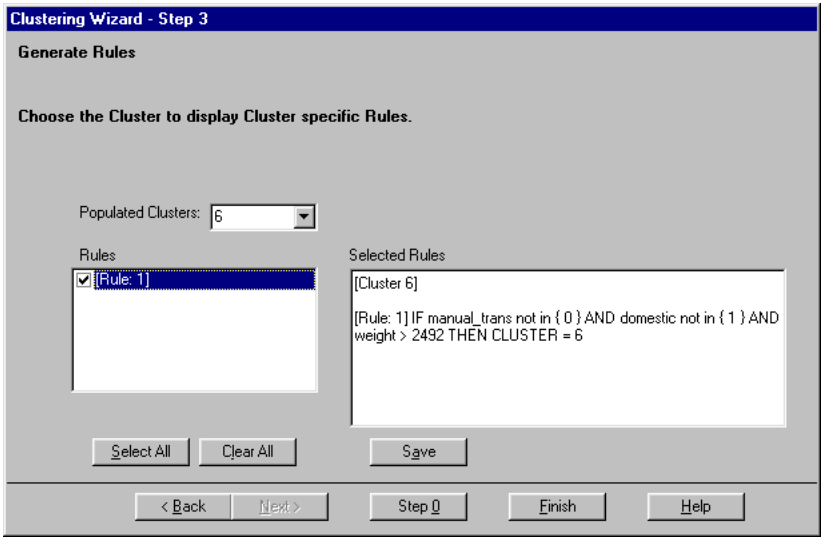
11.7.3 Clustering — Generate Rules: Step 3



The **Step 3** dialog displays the rules that define a specified cluster.

You can specify any of the available clusters that were derived from the model and dataset pairing that was used to generate the set of rules.

When you specify a cluster, the **Rules** box displays the list of rules for that cluster, and the **Selected Rules** box displays the ASCII text for the rule(s).



Reading the rules correctly in most cases requires that you know how the values for that variable were coded. For example, a tree rule that states "IF origin = 1" is meaningless unless you know that this is a multivalued categorical variable for "country of origin," and that it uses the following values: 0, 1, 2, and 9, with 0 = U.S., 1 = Japan, 2 = Germany, and all others = 9. Similarly, the tree rule shown in the screen capture above reads "IF manual_trans not in {0}, which is pretty mysterious until you learn that this field is binary, to be interpreted as true / false, with 0 = false and 1 = true, so that "manual_trans not in {0}" means "manual transmission not false"; in other words, this vehicle has a manual transmission.

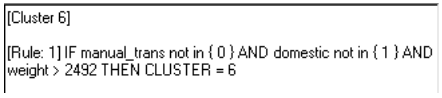
Some tree rules are more straightforward. For example, an ordered variable for wheelbase might contain number of inches, and if a tree rule states "IF wheelbase <= 99" you know it is saying "if this vehicle's wheelbase is less than 99 inches (then what? then this vehicle belongs in cluster x)." Although you can read this rule more easily, it may have nothing to do with anything of interest to you. Ordered variables

can of course also be coded; you might see a tree rule that reads "IF OQ14z => 4"; you'll need to know how responses were coded.

In general, the rules offer you your best chance of understanding the clusters that the wizard found. The rules might make some sense to you, or they may make no sense whatsoever. Their advantage is that they are, empirically, the characteristics that define the clusters. But remember, knowledge of the data is important.

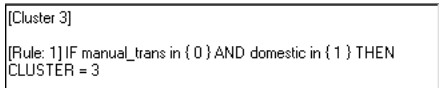
In the example shown here, we've generated the rules for clusters 6 and 3 (clusters with the most support). Each cluster is defined by only one rule (this is not typical, and may be an artifact of the dataset's small size (93 records, 12 fields) or some other characteristic of the dataset).

Rule 1 for cluster 6:



If the car has a manual transmission, is foreign, and weighs over 2492 pounds, it belongs in cluster 6.

Rule 1 for cluster 3:



If the car has an automatic transmission and is domestic, it belongs in cluster 3.

A next step might be to create separate datasets for the two clusters and use them with a predictive model. The people whose records are in cluster 6 -- manual transmission, foreign, weighs over 2492 pounds -- may be people who enjoy the feel of a large, performance car of European or Japanese origin. The people in cluster 3 have rather different preferences -- domestic cars with automatic transmissions. You might want to tailor a marketing approach to the preferences of each group, and test the approaches using a predictive model with the datasets built on each cluster.

You can save the rules for a specified cluster to a text file by clicking **Save**; you'll be prompted for a name and location.

The **Select All** button selects all the rules and displays their ASCII text. The **Clear All** button clears both the **Rules** box and the **Selected Rules** box.

Click **Step 0** to return to **Step 0**, where you can select a different functional path, or click **Finish** to exit the wizard and return to the Darwin client.

11.8 Darwin Clustering — Delete Model

Choose this path to delete a clustering model.

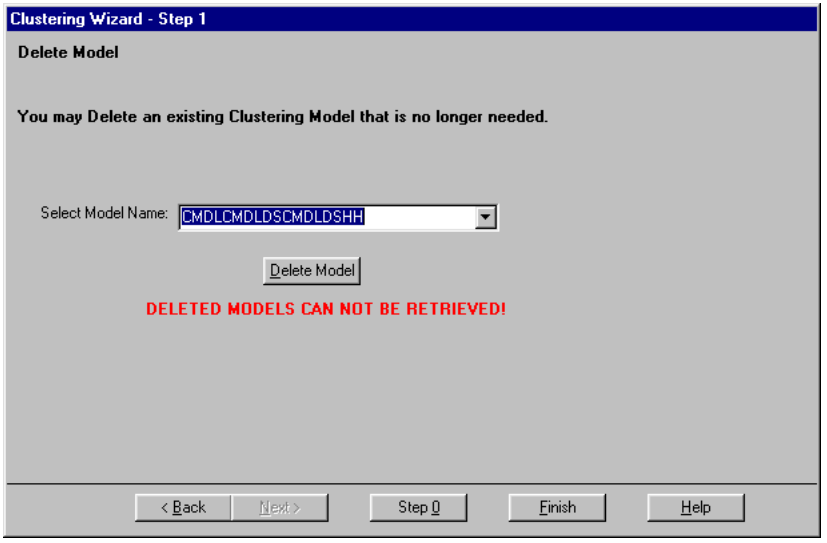
11.8.1 Clustering — Delete Model: Step 1

Delete Model → Step 1: Specify Model

There is only one dialog in the **Delete Model** path.

The **Step 1** dialog prompts you to specify the name of the clustering model you want to delete. You can delete one model at a time.

To delete the model, click **Delete**. Once a model is deleted, it is gone forever; you cannot retrieve it.



When you have finished deleting any models you don't want to keep, click **Step 0** to return to the **Step 0** dialog, where you can select another path, or click **Finish** to exit the Wizard and return to the Darwin client.

12

Tips for Dealing with Missing Values

This chapter describes several ways to deal with missing values in Darwin datasets.

If your dataset contains fields that do not have values and you want to use Darwin to deal with the missing values, you have several options:

- Use the **Missing** and **Select** transforms to remove all records containing the missing value.
- Use the **Replace** transform to replace the missing value with a typical value.
- Use the **Missing Values** Wizard to detect and treat the missing values.
- Use a Darwin model to predict missing values.

12.1 Detecting and Replacing Missing Values

12.1.1 How Darwin Detects Missing Values

The only missing value that Darwin detects is a null value, that is, no value in the field at all. You cannot specify that some non-null value (for example, "*") indicates a missing value. Also, if the delimiter (separator) in a dataset is " " (space), Darwin interprets " " (two spaces) as one space, not as two separators with no value between them.

In summary, to use any of Darwin's missing value facilities:

- the dataset must have a delimiter that is not " " (space)
- missing values must be indicated as null values (for example, if the delimiter is "," then ",," indicates a missing value)

12.1.2 Replacing Values

If you want to substitute another value for a missing value, you must decide what value to substitute. You may have information about the dataset that permits you to pick some value to substitute for missing ones, such as the maximum, minimum, or average of the existing values. You can also create a Darwin model that predicts the missing values.

12.1.3 Replacing the Missing Value With a Typical Value

The **Replace** transform lets you replace each null value with a typical value (the average of the values in the field.) To replace null values in field **f1** with the average value of **f1**, select the **Replace** transform and specify the following items in the **Field Information** box:

- **Replaced: f1**
- **Compared: f1**
- **Comparator: Equal** (the default)

Leave the **Values** both blank. Click **Transform**. The new dataset will have the average value of **f1** substituted for any occurrence of null.

12.2 Removing Records Containing the Missing Value

The **Missing** transform identifies missing values in a field **f1**. The transform creates a new dataset containing the field **f1_m**, which has two values, 0 and 1; 1 indicates that the field **f1** in the same record has an null value. Once you've located all missing values, you can use the **Select** transform to remove all records from the transformed dataset that have a missing value in field **f1**; do this by selecting all records where **f1_m** is equal to 0. As a last step, you may want to use the **Project** transformation to remove the **f1_m** values.

12.3 Using the Missing Values Wizard

The **Missing Values** Wizard permits you to identify and treat record-wise and field-wise missing data; that is, you can either remove records for which data is missing in specified fields, or you can assign a value to fields that have missing values. When you assign values to fields, you substitute the same value in all null fields; for example, you might substitute the maximum value. For more information, see Chapter 7.

12.4 Using a Model to Predict Missing Values

Suppose that the field **f1** in the dataset **Original.ds** has some missing values. You can use a Darwin model to predict missing values as follows:

5. From **Original.ds**, create two datasets, **NotMissing.ds**, consisting of those records in which **f1** has a value; and **Missing.ds**, consisting of those records in which the value of **f1** is missing. You can create these datasets in one of the following ways:
   - edit the text file on UNIX and use the **Text Import** Wizard to create Darwin datasets
   - apply the **Missing** and **Select** transforms to **Original.ds**, as described in Section 12.2
6. Use **NotMissing.ds** to create a Match model with the target field **f1**. You may want to use the **Modeling Wizard** to create the model.
7. When you have a model that predicts values for **f1** with an appropriate level of confidence, use that model to predict the values of **f1** in **Missing.ds**. The prediction results in a dataset that has has a prediction **f1_p** for each values of **f1** in **Missing.ds** and a confidence for each prediciton. Use these values to replace the missing values in **Original.ds**.

13

Files and Datasets for Practice

We provide several files and datasets for you to use in experimenting and practicing with Darwin. All are in the directory

`/opt/TMCdarwin/demo/darwin/demo.proj`

Before proceeding, you should study the online help and the documentation (*Using Darwin* and *Darwin Reference*).

13.1 Setting Up the Practice Files and Datasets

Follow the steps outlined below.

- After the Darwin server software has been installed, someone logged in as superuser (root) can enable the Darwin demo by entering the following command:

  `# sh /opt/TMCdarwin/demo/demo-setup`

  (If you reboot your system, the system administrator will have to re-execute this command.)

  When you see the message "Server started," you are ready to begin. If the message is not displayed, enter the command

  `# /opt/TMCdarwin/etc/darwinconfig start demo-config`

- Now look at the configuration file:

  `# cat /opt/TMCdarwin/etc/demo-config`

  and note these entries:

  ```
  name
  description
  server
  port
  ```

You will need this information when you install the Darwin client software (see *Darwin Installation and Administration, Release 3.6 for Solaris*).

Now anyone logged in as an ordinary user can start Darwin, create a project, and begin experimenting with these practice files and datasets.

13.2 Using the Practice Files and Datasets

Note: Do not set `demo.proj` as the active project. It is created for the sole purpose of storing files to be shared by all users. You will drag files and datasets from it and drop them in your newly created practice project.

The first seven entries in Table 1 come in pairs — for each, there is a text file (with a `.txt` or `.ascii` extension) and a descriptor file (with a `.des` extension). For example, for the first item listed, the two files are `cars2.txt` and `cars2.des`. You can use these files to create a Darwin dataset. There are two methods:

- **Use the Text Import Wizard:** Drag and drop only the text data file to your new project, and use the **Text Import** Wizard to create the Darwin dataset. (You don't need the descriptor file; the **Text Import** Wizard creates one from the information in the text file; this is one of the advantages of using the Wizard.)
- **Use Create Dataset:** Drag and drop the text file and descriptor file to your new project, and use the **Create Dataset** command.

  Note: If you create the dataset by hand, you may have to specify the delimiter used in the text file to separate fields. The default delimiter is a comma. If the text file you are working with uses a delimiter other than a comma, you'll need to specify that delimiter before proceeding. Go to **Options -> Advanced**, and click the **Datasets** tab, where you can specify the delimiter used in the file.

Entries 8 and 9 in Table 1, **CD_buyers** and **Churners**, are already Darwin datasets, and have the extension `.ds` . You can use them as is for dataset transformations and for building models.

Table 13–1 Files and Datasets for practice

	Name	Number of Records	Number of Fields	Delimiter	Target Field	Comment
1	car2	93	12	space	domestic	binary target
2	cars	406	8	space	origin	multiclass target, string
3	cars1	406	8	space	origin	multiclass target, integer
4	german	1000	21	space	a21	credit risk: 1 = good

Table 13–1 Files and Datasets for practice

	Name	Number of Records	Number of Fields	Delimiter	Target Field	Comment
5	house	406	14	space	MEDV	ordered target
6	mush	8124	23	,	edib-pois	string target
7	salary	3000	15	,	class	1 = high salary
8	CD_buyers	3000	15	,	CD_ buyer	This file is a Darwin dataset
9	Churners	2070	26	,	Churner	This file is a Darwin dataset

Index

A

C

D

R

S

T

U

V

W