EMPirical Distance ASSESSment
EMPD_ASSESS User's Guide
Version 1.1.0

A manual prepared by

Biometrics Northwest LLC
6215 225th Avenue NE
Redmond, WA 98053
Email: info@biometricsnw.com

February 21, 2013

Abstract

A program implementing the nonparametric EMPirical Distance ASSESSment procedures defined in Gehringer (2006) is described. The implementation consists of a stand-alone, Win32 compatible, command line executable program `empd_assess.exe` and this documentation describing the use of the program. A number of examples employing the `empd_assess.exe` program are also provided. The examples demonstrate the correct usage of the program and validate the implementation.

Disclaimer

THIS SOFTWARE IS PROVIDED "AS IS", WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES OR REPRESENTATIONS; INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

IN NO EVENT WILL ANY COPYRIGHT HOLDER BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE SOFTWARE OR PROGRAMS (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Acknowledgements

This work was funded by the Rural Technology Initiative (RTI) in the College of Forest Resources at the University of Washington. Biometrics Northwest LLC would like to thank RTI for providing funding and the opportunity to develop the EMPirical Distance ASSESSment software.

Contents

List of Tables

List of Figures

Chapter 1

Purpose

This User's Guide serves two purposes. First, it describes the use of the program `empd_assess.exe` implementing the nonparametric target definition and assessment procedures described in Gehringer (2006). Second, it provides a demonstration of the correct usage of the `empd_assess.exe` program through its use in several examples, verifying that the program performs as described. The data sets used in the examples are provided with the software and documentation to facilitate an independent verification of the assessment results in the examples.

The target definition and assessment procedures defined in Gehringer (2006) use the empirical distribution of distances from a central value in a reference or target data set, containing quantitative values for one or more forest structure attributes, to perform an EMPirical Distance ASSESSment (hence the name of the `empd_assess.exe` program). The objective of the assessment is to take a set of independent observations of the targeted forest structure attributes and to assess them relative to the reference or target condition. The empirical distance assessment procedures automate the identification of observations that are statistically indistinguishable from a reference or target condition, and those that are statistically different, for some predetermined acceptance level. The usage of the `empd_assess.exe` program, its inputs and outputs, and its command line syntax are described in Chapter 2. Examples demonstrating the use of the `empd_assess.exe` program are provided in Chapter 3.

The `empd_assess.exe` software uses the numerical linear algebra procedures in LAPACK version 3.0 (Anderson et al., 1999) and was implemented using the Fortran 95 programming language.

Chapter 2

Using empd_assess.exe

This chapter describes how to use the empd_assess.exe executable program to perform nonparametric EMPirical Distance ASSESSments. The input and output file formats, the options and arguments to the command line program, error messages and exceptional condition reporting, bug reporting, and known limitations of the software are described. A basic familiarity with the terminology and concepts defined in Gehringer (2006) is assumed. The input and output files are described first in Section 2.1, followed by a description of the options and arguments for the command line program in Section 2.2, a description of the error message formats in Section 2.3, a description of how to report bugs in Section 2.4, and a list of known limitations of the software in Section 2.5.

2.1 Inputs and outputs

The empd_assess.exe program requires two input files: a target data file containing the quantitative data that are used to define the empirical distance distribution for a target, and an observation data file containing the data that are to be assessed relative to the empirical distance distribution of the target data. The empd_assess.exe program also creates or modifies three types of output files: an assessment output file containing the results of performing an assessment on an observation data file, a preprocessed binary target file containing the empirical distance distribution derived from the values in the target data file, and a log file.

File names are specified on the command line using the full or relative path to the individual input or output files, which need not be located in the same directory. File names are assumed to consist of three parts: a path, a base name, and an extension, for example, the file name project1\data\file1.dat has a path equal to project1\data, a base name of file1 and an extension of dat. In this example the path is a relative path, beginning in the current working directory. By convention the trailing path element separator, the backslash (\) preceding the base name, is removed from the path and the file extension does not contain an initial dot (.) character. The forward slash (/) and the backslash (\) are both recognized as path element separators, and they may be used interchangeably, e.g., the path a\b\c\d is equivalent to the path a/b/c/d and the path a/b\c/d.

All of the files except the preprocessed binary target file are ASCII text files. The target data file, the observation data file, and the assessment output file are all comma separated value (CSV) ASCII text files. The log file is a free format ASCII text file. The data contained in the preprocessed binary target file are not directly accessible to a user of the `empd_assess.exe` program, and the file should not be modified.

The target data file and the observation data file are required to have unique column names. Column names are assumed to appear on the first nonblank, noncomment line of the target or observation data files. Column names in the target and observation data files may contain only printable ASCII characters. The column names need not be enclosed between quotes, unless the column name contains a comma or the first character of the first column name is a percent sign. Column names may be enclosed between single quotes, e.g., `'Column 1'` or between double quotes, e.g., `"Column 1"`. Column names are case sensitive, so `'column1'` is not the same as `'Column1'`. Column names in the assessment output file are always quoted using the double quote character (").

The observation data file may contain more columns than the target data file, but it must contain all of the columns that are present in the target data file used in an assessment. In addition, the assessment columns in the target and observation data files must be associated with compatible quantitative values. For example, if the target data file contains a column representing average tree diameter, e.g., `"Avg. Tree Diam."`, then the observation data file must contain the same column name associated with the same kind of data, average tree diameters. The order of the column names within either file is not important, but the target columns are processed from left to right when computing the empirical distance distribution, and the observation data columns are also processed from left to right.

The target and observation data files may also include blank lines and a descriptive header as comments. Comment lines are indicated by a percent sign (%) as the first nonblank character of the line. Comments and blank lines may appear anywhere in the input target and observation data files. Blank lines and comment lines contained within target and observation data files are ignored, and comments and blank lines in an observation data file are, therfore, not transferred to an assessment output file.

In addition to the input and output files, the `empd_assess.exe` program makes use of two other pieces of information to perform an assessment: a center type, used to identify a central value used in the empirical distance target, and an acceptance percent, specifying the percent of the central probability that is used to define the acceptance region of the target and the corresponding accept/reject boundary. The center type for an assessment is specified on the command line using the

```
-c <center type>
```

or

```
-center <center type>
```

option. A valid center type is one of `MEAN`, `MEDIAN`, or `MODE`. The center type is not case sensitive, so `Mode` is equivalent to `MODE`. If the center type is not specified, the default value of `MODE` is used for an assessment.

The acceptance percent is also specified on the command line by using the

```
-a <acceptance percent>
```

or

```
-accept <acceptance percent>
```

option. Values for the acceptance percent are numeric and must be between zero and 100. A percent sign (%) should not be used on the command line when specifying a value for an acceptance percent. If an acceptance percent is not provided on the command line a default value of 95% is used for an assessment.

2.1.1 Target data file

The `empd_assess.exe` program requires a target data file. The target data file contains the data that are used to generate the empirical distance distribution that is used for the assessments, when combined with a center type and an acceptance percent. The target data file is a CSV file whose data columns contain only numeric values. Missing values are not allowed. This is only of concern if the target data file contains more than one data column. Missing values in target data files having only a single column are treated as blank lines, and are transparently ignored. The column names within a target data file must be unique, and they are case sensitive. The target data file name is specified on the command line using the

```
-t <target file name>
```

or

```
-target <target file name>
```

option. If a target data file name is not provided on the command line an error is reported and program execution halts.

Only the path and base name of the target data file are necessary. A file extension of `csv` will be used automatically if a file extension is not provided. If a complete target data file name, including a file extension, is specified on the command line, then that file name, including the extension, will be used.

A sample target data file, `tgt_nqh.csv`, used by one of the examples in Section 3.3 appears in Figure 2.1. This file has three data columns, `TPA_12`, `QMD_12`, and `Avg_HT_12`, representing the stand density in trees per acre (TPA), the quadratic mean diameter (QMD), and the average height, respectively, of trees having a diameter at breast height (DBH) of at least 12 inches for a collection of sample plots. The sample target data file has no comments, indicated by a percent sign (%) as the first nonblank character. The three dots indicate that the middle of the file was removed for display purposes in the figure.

The target data file is used to create a preprocessed binary target file containing the empirical distance distribution for the selected center type, the `MEAN`, `MEDIAN`, or `MODE` of the data contained in the target data file. The preprocessed binary target file name is formed using the path and base name of the target data file by appending a lower case version of the center type name, `_mean`, `_median`, or `_mode`, to the base name, depending upon the center type used, and using a file extension of `tgt`. The path, modified base name, and extension are then concatenated together with the appropriate separators, a backslash (\) between the path and the modified base name and a dot (.) between the modified base name and the extension, to produce the preprocessed binary target file name. For example, the target data file name

```
TPA_12,QMD_12,Avg_HT_12
79.660000,19.503838,108.540673
36.880000,24.498929,106.269523
89.980000,25.345191,134.622583
111.800000,17.013575,89.833810
85.870000,22.697624,132.053336
59.250000,24.207357,134.084726
103.220000,18.011544,86.623716
24.470000,19.561543,82.249694
                .
                .
                .
143.680000,19.424416,106.630707
65.620000,23.466944,112.365133
53.980000,21.246934,97.626528
145.180000,16.380435,92.321532
77.800000,21.406546,95.025707
31.880000,16.510892,83.812422
61.000000,15.848914,93.508525
134.720000,20.204035,103.965261
```

Figure 2.1: A sample target data file used by one of the examples in Section 3.3.

```
project1\data\tgt1.csv
```

would, when used with a center type of `MODE`, produce the file name

```
project1\data\tgt1_mode.tgt
```

for the preprocessed binary target file after the modifications just described.

The preprocessed binary target file is the file that is actually used to perform the empirical distance assessment of the observations in an observation data file. The preprocessed binary target file is created and used for the assessment if it does not exist. The file is updated if the target data file is more recent than an existing preprocessed binary target file. Whether the preprocessed binary target file was created or updated before it was used to perform an assessment is indicated in the log file.

2.1.2 Observation data file

The `empd_assess.exe` program requires an observation data file. The observation data file contains the data that are assessed using the empirical distance distribution generated from the target data file. The observation data file may contain more data columns than those that are to be assessed relative to the target being used in a particular assessment, but all of the columns used to define the target must be present in the observation data file. Further, data columns that are being assessed in the observation data file must contain only numeric values that are consistent with the values used in the target data set. The column names within the observation data file must be unique, and they are case sensitive. The observation data file

name is specified on the command line as the only argument following all of the options. If an observation data file name is not provided on the command line an error is reported and program execution halts.

Missing values are allowed in the assessment data columns of an observation data file. If a missing value is encountered when processing a row of the observation data file, then a message indicating the data row number and the data column number of the first missing value on the row is displayed or logged. The row number reported in the log message may not be the same as the line number in the observation data file, since blank and comment lines in the file are skipped. The row number reported refers only to the data lines in an observation data file.

Only the path and base name are necessary when specifying an observation data file name, a file extension of `csv` will be used automatically if the file extension is not provided. If a complete observation data file name, including the file extension, is specified on the command line, then that file name, including the extension, will be used.

A sample observation data file, `obs_nqh.csv`, used by one of the examples in Section 3.3 appears in Figure 2.2. This file has three data columns, `TPA_12`, `QMD_12`, and `Avg_HT_12`, representing the stand density in trees per acre (TPA), the quadratic mean diameter (QMD), and the average height, respectively, of trees having a diameter at breast height (DBH) of at least 12 inches for a collection of sample plots. These are the same columns that were used in the target file for this example. This sample observation data file was minimal, it has no comments, indicated by a percent sign (%) as the first nonblank character of a line, and no additional data columns. The three dots indicate that the middle of the file was removed for display purposes in the figure.

```
TPA_12,QMD_12,Avg_HT_12
8.280000,26.775835,125.292271
44.530000,23.744906,123.978442
61.930000,28.300255,149.539803
103.310000,20.726745,88.559288
62.700000,26.340217,153.722169
47.740000,15.299457,62.865731
33.710000,25.727205,101.054287
36.390000,23.792192,104.889530
                .
                .
                .
16.960000,18.796160,75.687500
82.960000,26.456808,120.398023
58.060000,19.851819,110.760248
46.640000,23.544585,123.772727
48.820000,19.251695,91.902089
35.800000,18.689109,83.637430
102.340000,17.201916,97.968927
85.820000,13.834240,85.264973
```

Figure 2.2: A sample observation data file used by one of the examples in Section 3.3.

Additional, nonassessment data columns may also appear in an observation data file, as in Figure 2.3, where the data columns `Stand`, `Loc`, `Age_Cls`, `Elev`, and `Resid?`, have been added to the assessment data

columns from the observation data file in Figure 2.2. Three of the additional columns, `Stand`, `Age_Cls`, and `Elev`, contain integer or whole number values, the `Loc` data column contains character data, and the `Resid?` data column contains logical data values, i.e., a `True` or `False` value. All of the logical values for this sample are `False` since only sample plots having no residual stand trees were selected for this example. For more information on these data see the description in Section 3.3.

Character data values in an observation data file are limited a length of 64 characters. Character data values that are longer than 64 characters are truncated to a length of 64 characters. Leading and trailing blanks in a character data column are ignored, unless the value is enclosed between matching quote characters. When a character data value is enclosed between matching quote characters, leading blanks are preserved, but training blanks are not. In addition, if a character data value contains a comma (`,`) or if it is in the first data column and it contains a percent sign (%) as its first nonblank character, then it must be enclosed between matching quote characters, either a single quote (`'`) or a double quote character (`"`). A single quote character or a double quote character may appear in a quoted string enclosed with the other quote character, e.g., `"User's Guide"` or `'double quotes "work here"'`. To get a quote character of the same kind as the enclosing quotes within a quoted character data value, the internal quote character is repeated twice, e.g.,

```
"double a double quote "" to get it to appear"
```

represents the character data value

```
double a double quote " to get it to appear
```

in a character data column. An individual quote character appearing in a data value by itself is an error; quote characters must always appear in pairs. Therefore a character data value may only contain a quote character if the data value itself is enclosed between matching quote characters.

2.1.3 Assessment output file

The `empd_assess.exe` program creates an assessment output file containing the results of the assessment if the program execution was successful. The assessment output file contains all of the data columns that were in the observation data file, including any additional or nonassessment data columns, with two additional data columns added on the right, a logical valued data column `Accept` and a numeric valued data column `p-EMPD`.

The `Accept` data column has a value of `TRUE` if the assessment data columns for the observation on that row were acceptable relative to the empirical distance target for the acceptance percent used, and a value of `FALSE` if the observation was not acceptable. Any rows that have missing values for an assessment data column are not assessed. If a missing value is encountered in an assessment data column, then the value of the `Accept` assessment data column will be blank. In addition, the presence of the first missing value on a data row is logged, identifying the data row number and data column number where the missing value was located.

The `p-EMPD` data column contains approximate p-values derived using the empirical distribution of distances from the chosen central value. The `p-EMPD` values are all between zero and one, `p-EMPD` $\in [0, 1]$, and

```
Stand,Loc,Age_Cls,Elev,Resid?,TPA_12,QMD_12,Avg_HT_12
95,western Oregon,9,932,False,8.280000,26.775835,125.292271
365,western Oregon,10,1401,False,44.530000,23.744906,123.978442
373,western Oregon,13,1099,False,61.930000,28.300255,149.539803
555,western Oregon,10,1299,False,103.310000,20.726745,88.559288
719,western Oregon,10,1001,False,62.700000,26.340217,153.722169
842,western Oregon,11,1001,False,47.740000,15.299457,62.865731
915,western Oregon,22,1499,False,33.710000,25.727205,101.054287
1154,western Oregon,21,2201,False,36.390000,23.792192,104.889530
                                .
                                .
                                .
28492,western Washington,21,141,False,16.960000,18.796160,75.687500
28496,western Washington,15,259,False,82.960000,26.456808,120.398023
28658,western Washington,10,180,False,58.060000,19.851819,110.760248
28695,western Washington,9,180,False,46.640000,23.544585,123.772727
28724,western Washington,16,299,False,48.820000,19.251695,91.902089
28726,western Washington,9,440,False,35.800000,18.689109,83.637430
28753,western Washington,9,171,False,102.340000,17.201916,97.968927
28757,western Washington,11,249,False,85.820000,13.834240,85.264973
```

Figure 2.3: A sample observation data file with additional, nontarget data columns. This file may be used with one of the examples in Section 3.3.

they indicate the degree of similarity with the central value for each observation. A `p-EMPD` value that is near one indicates a high degree of similarity to the central value, while a `p-EMPD` value near zero indicates a low degree if similarity. The `p-EMPD` values may be interpreted as if they were p-values in a statistical hypothesis test. As for the `Accept` data column, if missing values are encountered in an assessment data column, the `p-EMPD` value for that row will be blank, indicating that the row was not assessed.

An assessment output file name may be specified on the command line using the

```
-o <output file name>
```

or

```
-outfile <output file name>
```

option. If an assessment output file name is not provided on the command line, a default output file name that is derived from the name of the observation data file is used.

The default assessment output file name is formed using the path and base name of the observation data file by appending `_assessed` to the base name and using the file extension of the specified observation data file or a default of `csv`. The path, modified base name, and extension are then concatenated together with the appropriate separators, a backslash (\) between the path and the modified base name and a dot (.) between the modified base name and the extension, to produce the default assessment output file name. For example, the observation file name

```
project1\data\obs1.csv
```

would produce a default output file name of

```
project1\data\obs1_assessed.csv
```

after the modifications just described.

Only the path and base name are necessary when specifying an assessment output file name, a file extension of `csv` will be used automatically if a file extension is not provided. If a complete assessment output file name, including the file extension, is specified on the command line, then that file name, including the extension, will be used.

A sample assessment output file, `obs_nqh_assessed.csv`, appears in Figure 2.4. This file was created by performing an assessment of the observation data file `obs_nqh.csv`, see Figure 2.2. The observation data file had three data columns, `TPA_12`, `QMD_12`, and `Avg_HT_12`, and the assessment output file added the columns `Accept` and `p-EMPD`. These are the same columns that were used in the target file for this example. The sample observation data file was minimal, having no comments and no additional data columns. The three dots indicate that the middle of the file was removed for display purposes in the figure.

```
"TPA_12","QMD_12","Avg_HT_12","Accept","p-EMPD"
8.280000,26.775835,125.292271,TRUE,0.2645
44.530000,23.744906,123.978442,TRUE,0.4764
61.930000,28.300255,149.539803,TRUE,0.1255
103.310000,20.726745,88.559288,TRUE,0.2497
62.700000,26.340217,153.722169,FALSE,0.0378
47.740000,15.299457,62.865731,FALSE,0.0769
33.710000,25.727205,101.054287,TRUE,0.5128
36.390000,23.792192,104.889530,TRUE,0.8259
                    .
                    .
                    .
16.960000,18.796160,75.687500,TRUE,0.1296
82.960000,26.456808,120.398023,TRUE,0.7733
58.060000,19.851819,110.760248,TRUE,0.5169
46.640000,23.544585,123.772727,TRUE,0.4845
48.820000,19.251695,91.902089,TRUE,0.6532
35.800000,18.689109,83.637430,TRUE,0.3576
102.340000,17.201916,97.968927,TRUE,0.4548
85.820000,13.834240,85.264973,TRUE,0.2537
```

Figure 2.4: A sample assessment output file created by one of the examples in Section 3.3.

Additional, nonassessment data columns may also appear in an observation data file, as in Figure 2.3, where the data columns `Stand`, `Loc`, `Age_Cls`, `Elev`, and `Resid?`, were added to the observation data file in Figure 2.2 to produce the file `obs_nqh_extra_cols.csv`. The assessment output file created by an assessment of the file `obs_nqh_extra_cols.csv` containing the additional data columns is shown in Figure 2.5.

```
"Stand","Loc","Age_Cls","Elev","Resid?","TPA_12","QMD_12","Avg_HT_12","Accept","p-EMPD"
95,"western Oregon",9,932,FALSE,8.280000,26.775835,125.292271,TRUE,0.2645
365,"western Oregon",10,1401,FALSE,44.530000,23.744906,123.978442,TRUE,0.4764
373,"western Oregon",13,1099,FALSE,61.930000,28.300255,149.539803,TRUE,0.1255
555,"western Oregon",10,1299,FALSE,103.310000,20.726745,88.559288,TRUE,0.2497
719,"western Oregon",10,1001,FALSE,62.700000,26.340217,153.722169,FALSE,0.0378
842,"western Oregon",11,1001,FALSE,47.740000,15.299457,62.865731,FALSE,0.0769
915,"western Oregon",22,1499,FALSE,33.710000,25.727205,101.054287,TRUE,0.5128
1154,"western Oregon",21,2201,FALSE,36.390000,23.792192,104.889530,TRUE,0.8259
                                        .
                                        .
                                        .
28492,"western Washington",21,141,FALSE,16.960000,18.796160,75.687500,TRUE,0.1296
28496,"western Washington",15,259,FALSE,82.960000,26.456808,120.398023,TRUE,0.7733
28658,"western Washington",10,180,FALSE,58.060000,19.851819,110.760248,TRUE,0.5169
28695,"western Washington",9,180,FALSE,46.640000,23.544585,123.772727,TRUE,0.4845
28724,"western Washington",16,299,FALSE,48.820000,19.251695,91.902089,TRUE,0.6532
28726,"western Washington",9,440,FALSE,35.800000,18.689109,83.637430,TRUE,0.3576
28753,"western Washington",9,171,FALSE,102.340000,17.201916,97.968927,TRUE,0.4548
28757,"western Washington",11,249,FALSE,85.820000,13.834240,85.264973,TRUE,0.2537
```

Figure 2.5: A sample assessment output file with additional, nontarget data columns. This file was created by one of the examples in Section 3.3.

If an assessment output file already exists, an error will be reported and program execution will halt. The `empd_assess.exe` program will not overwrite existing assessment output files. If an existing assessment output file is to be replaced by performing a new assessment the `-r` or `-replace` command line option must be specified. This option instructs the `empd_assess.exe` program to replace an existing assessment output file. If the assessment output file does not exist, this option has no effect.

2.1.4 Log file

The `empd_assess.exe` program maintains a log file containing information, e.g., the current date and time, the target data file name, the observation data file name, the central value used in the target etc., for each run of the program. A log file name may be specified using the

```
-l <log file name>
```

or

```
-logfile <log file name>
```

option. If a log file name is not provided on the command line, a default log file name that is derived from the name of the observation data file is used.

The default log file name is formed using the path and base name of the observation data file by appending _assessment to the base name and using a file extension of log. The path, modified base name, and extension are then concatenated together with the appropriate separators, a backslash (\) between the path and the base name and a dot (.) between the modified base name and the extension, to produce the default log file name. For example, the observation file name

```
project1\data\obs1.csv
```

would produce a default log file name of

```
project1\data\obs1_assessment.log
```

after the modifications just described.

Only the path and base name are necessary when specifying a log file name, a file extension of log will be used automatically if a file extension is not provided. If a complete log file name, including the file extension, is specified on the command line, then that file name, including the extension, will be used.

A sample log file created by one of the examples in Section 3.2 appears in Figure 2.6. The program name, the software copyright, the current date and time, and information about the target, observation, and assessment output files, as well as the center name and the acceptance percent for the assessment being performed are written to the log file. Whether a new preprocessed binary target file, indicated by Binary target file in the log file, is created and used or an existing preprocessed binary target file is used for an assessment is also reported, followed by an indication that the assessment is being performed. If the same log file name, whether the default or specified on the command line, is used for multiple assessments, the log information for each assessment is appended to the end of the file in the sequence that the assessments were performed. The empd_assess.exe program does not overwrite the log file information.

If a missing value is encountered in one of the assessment data columns during an assessment of an observation data file, a message indicating the data row number and the data column number containing the missing value will be written to the terminal screen or console window and the log file. If the program is being run silently, minimizing output to the terminal screen or console window, by using the -s or -silent option, then the missing value message is only written to the log file. The missing value message is

```
A missing assessment value was found in column M of row N. This row was not assessed.
```

where N is replaced by the number of the data row and M is replaced by the number of the data column containing the missing value. If there are multiple missing values, a message will be logged for each data row containing a missing value.

2.2 Command line options and arguments

The command line options and arguments for the empd_assess.exe program are described in this section. The option and argument descriptions are identical to those from the program help obtained by using the

```
*****************************

EMPD_ASSESS 1.0.0
Copyright 2005-2006 Biometrics Northwest LLC

No part of this software may be copied, distributed, or used without
the written consent of Biometrics Northwest LLC, Redmond, Washington.

Assessment date and time   : 2006-04-20 10:51:04.296

Assessment target file     : tgt_2.csv
    Binary target file     : tgt_2_mean.tgt
Assessment observation file: obs_2.csv
Assessment output file     : assess_2_67.csv
   Replace output file     : FALSE
Assessment log file        : assess_2_67.log
Assessment center name     : MEAN
Assessment acceptance %    : 67.00

The observation file was created or modified on  : 2006-04-18 11:38:01.000
The target file was created or modified on       : 2006-04-18 11:38:01.000
The binary target file was created or modified on: 2006-04-18 11:38:43.000

The target file is older than the binary target file
   The existing binary target file will be used for the assessment

Performing the assessment, creating the output file
```

Figure 2.6: A log file created by one of the examples in Section 3.2.

-h or -help option. The complete command line syntax for the empd_assess.exe program is given in Figure 2.7.

```
EMPD_ASSESS [-h | -help]                                              ...
      [-ver | -version]                                               ...
      [-s | -silent]                                                  ...
      [-v | -verbose]                                                 ...
      [-r | -replace]                                                 ...
      [-c <center type> | -center <center type>]                      ...
      [-a <acceptance percent> | -accept <acceptance percent>]        ...
      [-l <log file name > | -logfile <log file name >]               ...
      [-o <output file name > | -outfile <output file name >]         ...
      -t <target file name > | -target <target file name >            ...
      <observation file name>
```

Figure 2.7: Command line syntax for empd_assess.exe.

The square brackets, [and], indicate optional command line options and any associated values. Angle brackets, < and > indicate required values for the options when specified and required command line arguments. Ellipses, ..., indicate that the line break is to be ignored, and the text on the second and subsequent lines should be typed on the same line as the rest of the empd_assess.exe command. The vertical bar, |, indicates that there are synonyms for an option, either of which is valid. Before running the empd_assess.exe program be sure to make a backup of any files that are to be modified. After they are modified the data in the original files will be lost.

Options for the empd_assess.exe program are indicated by a dash or minus sign (-) immediately preceding the option name with no intervening spaces. Options may appear in any order but all options must appear before the command line argument. Some options have associated values which must immediately follow the option, separated from the option name by one or more blanks or spaces. The case of the command line option names is not important. The case of filenames, paths, etc., that are specified on the command line may be important, depending on the operating system in use.

2.2.1 Command line options

The empd_assess.exe program has a number of command line options. All of the options may be used at the same time to create a valid command line, except the help options, -h and -help, and the version options, -ver and -version, which must appear as the only option on the command line.

-a <acceptance percent> or -accept <acceptance percent> Specify the acceptance percent used in an assessment. The acceptance percent must be between 0 and 100. An acceptance percent of zero accepts only observations whose distance from the chosen central value derived from the target data set is zero, i.e., only observations that equal the central value will be accepted. An acceptance percent of 100 accepts all observations having distances from the central value that are less than the maximum distance in the target data set. If this option is not present, the default value of 95% will be used as the acceptance percent.

`-c <center type>` or `-center <center type>` Specify the center type to be used for an assessment. Valid center types are `MEAN`, `MEDIAN`, and `MODE`. The center type is not case sensitive, so a center type of `Mode` is equivalent to `MODE`. The default center type is `MODE`.

`-h` or `-help` Display the help message on the standard output, typically a terminal screen or a console window, and exit. This option must appear by itself on the command line.

`-l <log file name>` or `-logfile <log file name>` Specify the name, including the full or relative path, to the log file. Only the path and base name of the log file are required; a file extension of `log` is assumed. If the file extension for the log file is something other than `log` then it must be included as a part of the log file name.

If a log file name is not specified using the `-l` or `-logfile` options, then the default log file name will be used. The default log file name is given by the path and base name of the observation data file with `_assessment` appended to the base name and an extension of `log`. The default log file, then, appears in the same location as the observation data file that is being assessed.

If the log file already exists, then new information is appended to the existing file, otherwise the log file is created.

`-o <output file name>` or `-outfile <output file name>` Specify the name, including the full or relative path, to the assessment output file. Only the path and base name of the assessment output file are required; a file extension of `csv` is assumed. If the file extension for the assessment output file is something other than `csv` then it must be included as a part of the assessment output file name.

If an assessment output file name is not specified using the `-o` or `-outfile` option, then the default assessment output file name will be used. The default assessment output file name is given by the path and base name of the observation data file with the string `_assessed` appended to the end of the base name and an extension of `csv`. The default assessment output file, then, appears in the same location as the observation data file.

If the assessment output file already exists, it is an error. An existing assessment output file may not be overwritten unless the `-r` or `-replace` option is used, indicating that the assessment output file is to be overwritten.

`-r` or `-replace` This option instructs `empd_assess.exe` to replace the assessment output file if it already exists, rather than signaling an error. If the assessment output file does not exist, the option has no effect. The `empd_assess.exe` program will not overwrite an assessment output file unless it has been instructed to do so.

`-s` or `-silent` This option suppresses normal and verbose output to the standard output device, typically the terminal screen or a console window. Error messages, if an error occurs while the program is running, are however, reported. This option does not affect the information written to the log file.

`-t <target file name>` or `-target <target file name>` Specify the name, including the full or relative path, to the target data file. Only the path and base name of the target data file are required; a file extension of `csv` is assumed. If the file extension for the target data file is something other than `csv` then it must be included as a part of the target data file name. This option is required if an assessment is to be performed.

`-v` or `-verbose` This option instructs `empd_assess.exe` to display additional information as well as the normal output to the standard output device, typically the terminal screen or a console window, and write the information to the log file.

`-ver` or `-version` Display the program version on the standard output, typically the terminal screen or a console window, and exit. This option must appear by itself on the command line.

2.2.2 Command line arguments

The `empd_assess.exe` program has only one command line argument, the name of the observation data file. The name of the observation data file must be the last item that appears on the command line.

`<observation file name>` The name, including the full or relative path, to the observation data file that is to be assessed relative to the specified target data file. Only the path and base name of the observation data file are required; a file extension of `csv` is assumed. If the file extension for the observation data file is something other than `csv` then it must be included as a part of the observation data file name. This argument is required if an assessment is to be performed.

2.3 Errors and error messages

The `empd_assess.exe` program detects a variety of potential problems, reporting any errors that it detects. All errors that are detected by the EMPD_ASSESS software are considered to be fatal errors, that is, the execution of the program is halted when an error occurs. If an error occurs, an error message is generated and displayed on the standard error device, typically the screen or a console window. Error messages are also written to the log file, prior to halting the program execution. To indicate the state of the program when execution halts, if an error occurs, the program returns a nonzero exit code to the operating system, otherwise a zero value is returned to the operating system, indicating a successful program run.

Error messages contain information intended to be helpful in resolving the problem which caused the error to occur. An error message consists of two parts, an error type and an error description. The error type provides an indication of the general class of error that occurred, e.g., `InvalidInput` or `FileDoesNotExist`. The error description provides the relevant context, e.g., missing values, invalid values, file presence or absence, etc., and other pertinent information. The error description is intended to help determine exactly what caused the error to occur. The template used to format error messages is given in Figure 2.8, and an actual error message is given in Figure 2.9.

2.4 Bug reporting

In the event that a bug is discovered when using the `empd_assess.exe` software, a problem or bug report may be sent to `bugs@biometricsnw.com` with a complete description of the problem. Before sending the bug or problem report, please use the following short check-list to help prepare the bug or problem report. This will enable a more rapid resolution of the problem.

```
------------------------------------------------------------------------

   Error(<error type>) --

   <error description>

------------------------------------------------------------------------
```

Figure 2.8: Error message format.

```
------------------------------------------------------------------------

   Error(FileAlreadytExists) --

   The output file 'observations_assessed.csv' already exists. To
   replace the file use the '-r' or '-replace' option.

------------------------------------------------------------------------
```

Figure 2.9: Example error message.

1. Be sure that the problem to be reported is real and not simply an erroneous input value or some other simple problem that is not actually a bug or defect in the software.
2. Attempt to fully and concisely document the circumstances under which the problem occurred. Be sure to include the following items in any message reporting a possible bug.
   - Your name and contact information, email address and phone number in particular.
   - The name and version of the program that was used.
   - The specific command line that was used.
   - All input or output files that were used when the problem occurred.
   - Error messages, if any, reported by the software or the computer when the problem occurred.
3. The subject line of the email should clearly indicate that the message is a bug report or problem report and give the name of the program involved.

To facilitate the reporting of possible bugs or problems encountered when using the `empd_assess.exe` program, an execution traceback is provided when a run-time error occurs. Run-time errors are errors that are not detected and signaled by the `empd_assess.exe` program. This traceback should be provided in any bug or problem report.

In addition, a debug version of the program is also provided. The debug version of the executable has `_d` appended to the base name of the executable, i.e., the debug version of the program has the file name `empd_assess_d.exe`. The debug version of the program performs a variety of internal integrity checks as well as providing an execution traceback for run-time errors detected and signaled by the `empd_assess.exe` program. If there is a problem with the detection and reporting of errors by the `empd_assess.exe` program

the debug version may be used to generate additional information which should be provided in any bug or problem report.

Contact Biometrics Northwest LLC at the address below to request additional information or to report problems and possible bugs.

Biometrics Northwest LLC
6215 225th Avenue NE
Redmond, WA 98053

email:
Bug reports: bugs@biometricsnw.com
Information requests: info@biometricsnw.com

Home Page: http://www.biometricsnw.com

2.5 Limitations

The assessment methods used in the `empd_assess.exe` program assume that the distribution of the data in the reference or target data set is unimodal, continuous, and symmetric. The following limitations exist when performing an assessment using this program.

- No checks are performed at this time to determine whether the target data distribution is consistent with the assumption of unimodality.
- Use of this program with a reference or target data set that is not unimodal will produce misleading or incorrect results. This use is, therefore, not recommended. Any data set that is to be used as a reference or target data set must be screened for unimodality prior to its use with `empd_assess.exe`.
- The data columns used to define the reference or target data set and to perform an assessment of an observation data set must be quantitative values representing data that were collected as a sample from a continuous distribution. Numerical values that represent categories are inappropriate for use with this program. As a rule of thumb, if it makes sense to use the mean value of a variable, then it may be reasonable to use that variable in an assessment using `empd_assess.exe`.
- The use of this program with a reference or target data set that is not symmetric should be done with care. The assessment methods, as implemented, are robust to small deviations from the symmetry assumption, but this should be confirmed for each reference or target data set prior to performing an assessment.

In addition to the data distribution limitations just mentioned, there are also a number of operational limitations. The operational limitations specify maximum lengths for lines in the target and observation data files, maximum lengths for file names, constraints on the kinds of data that may appear in the input files, and constraints on the sizes of files that may be processed.

- File names may contain only printable ASCII characters, and are subject to the file naming conventions and restrictions imposed by the operating environment.

- Column names in the target data file and observation data file may contain only printable ASCII characters.
- When performing an assessment, all of the data columns from the target data file must appear in the observation data file and have the same interpretation. For example, if the column `Mean tree diameter` is in the target data file, then it must appear in the observation data file, and the meaning of the data in each file should be the same, that is, each data column should represent mean tree diameter values.
- File names, including the path and extension, may be up to 2048 characters in length.
- The command line has a maximum length of 16384 characters.
- The maximum length of a data column name is 64 characters. Data column names may contain only printable ASCII characters.
- The maximum length of a character data value within a character data column is 64 characters. Character data values may contain only printable ASCII characters.
- There is a maximum line length of 8192 characters for the target and observation data files and the assessment output file.
- The number of additional data columns, data columns that are not being assessed relative to a target, that may appear in an observation data file depends on the size of the observation data file and the maximum line length of 8192 characters. If the observation file is too large to be assessed, it may be divided into smaller files that can be assessed. If the length of lines in an observation file exceeds the maximum allowed, then a smaller observation data file may be obtained by extracting only the data columns being assessed.
- The entire target and observation data files are read into memory prior to processing. A maximum of two GB of data, preprocessed target data, observation data that are to be assessed, and temporary storage may be held in memory. If the size of an observation data file combined with the preprocessed target data exceeds the 2 GB limit a run time error will occur. To perform the assessment, divide the observation data file into smaller files that will fit into memory.
- The minimum detectable difference in file creation times is one second.

Chapter 3

Examples

This chapter presents three examples of using the `empd_assess.exe` program to perform empirical distance based assessments. The examples use the `empd_assess.exe` program to assess an observation data set relative to the empirical distance distribution obtained from a target data set for: (1) target and observation data sets derived from a standard normal distribution $N(0, 1)$ in Section 3.1, (2) target and observation data sets derived from a two dimensional standard normal distribution $N(0, I_2)$ in Section 3.2, and (3) target and observation data sets derived from a forestry data set in Section 3.3.

The first two examples demonstrate the use of the `empd_assess.exe` program in the case where its assumptions are all valid by using simulated data sets generated from standard normal and standard multivariate normal distributions. The normal and multivariate normal distributions are the quintessential continuous, unimodal, symmetric distributions. The third example demonstrates the use of the `empd_assess.exe` program on a real data set to define a target using the forest structure attributes stand density (TPA) and quadratic mean diameter (QMD) computed for trees having diameter at breast height (DBH) values of at least 12 inches. A target developed in a similar manner could, for example, be used to directly specify the range and variability of forest structures that are associated with the habitat for one or more species.

All of the examples were run using the command lines as they appear. The execution of the command lines assumes that the current directory is the `examples` directory, and that it is in its default location within the software distribution package, `empd_assess\examples`. Data summaries and figures were produced using Matlab (MathWorks, 2006) and the requisite target data files, observation data files, and assessment output files.

3.1 Example 1: Standard normal distribution

This example uses simulated target and observation data from the one-dimensional standard normal distribution $N(0, 1)$. For this example, two independent random samples of size $N = 100$ were generated from the $N(0, 1)$ distribution to obtain target and observation data files, `tgt_1.csv` and `obs_1.csv`, respectively. Summary statistics for the target and observation data sets are presented in Table 3.1. The variation row

Table 3.1: Summary statistics for the 1-D target and observation data sets. The 1-D example data consisted of independently generated target and observation data sets containing 100 points from a standard normal distribution $N(0, 1)$.

Statistic	Target data	Observation data
Mean	-0.0079	0.0498
Std. Dev.	1.0268	1.1131
Minimum	-2.5930	-2.3438
Median	-0.1208	-0.0450
Maximum	2.4000	2.4120
Mode	0.0162	-0.0363
Variation	1.0550	1.2464

in the table measures the variability about the estimate of the mode, and is similar to the variance. For a symmetric distribution, like the normal distribution, the true mean is equal to the mode, so the mode estimate and the variation should be nearly equal to the mean and variance for each data set. This will not be the case for nonsymmetric data distributions.

The `empd_assess.exe` program was used to assess the individual points in the observation data set relative to the empirical distance distribution obtained from the target data set using the mean as the central value. Acceptance percentages of 67% and 95% were used to control the extent of the target. These two targets, then, represent, approximately, the mean value plus or minus one and two standard deviations within the target data set, $\bar{x}_t \pm s_t$ and $\bar{x}_t \pm 2s_t$, respectively. The assessment at 67% was performed using the command line

```
..\empd_assess -c mean -a 67 -t tgt_1.csv -l assess_1_67.log -o assess_1_67.csv obs_1.csv
```

where `-c mean` indicates that the mean value of the target data set was used as the center, `-a 67` indicates the desired acceptance percent of 67%, `-t tgt_1.csv` identifies the target data file, `-l assess_1_67.log` identifies the log file, `-o assess_1_67.csv` identifies the assessment output file, and `obs_1.csv` is the observation data file. The results of the 67% assessment appear in Figure 3.1.

The 95% assessment was performed using the command line

```
..\empd_assess -c mean -t tgt_1.csv -l assess_1_default.log -o assess_1_default.csv obs_1.csv
```

where all of the options have the same meanings as before, but the log and assessment output file names have been changed. The `-a` option to define the acceptance percent was not used in this example since an acceptance percent of 95% is the default if no acceptance percent is specified. The results from the 95% assessment appear in Figure 3.2.

At 67% acceptance for this target, 59% of the observations were accepted and 41% of the observations were rejected. The difference of -8% between the target acceptance level of 67% and the acceptance of 59% of the observations may, at first glance, seem somewhat large, given that the two samples were drawn from the same distribution, but this is not the case. The difference of -8% is readily explained by considering the characteristics of the target and observation samples, the assessment procedure, and the sample size.

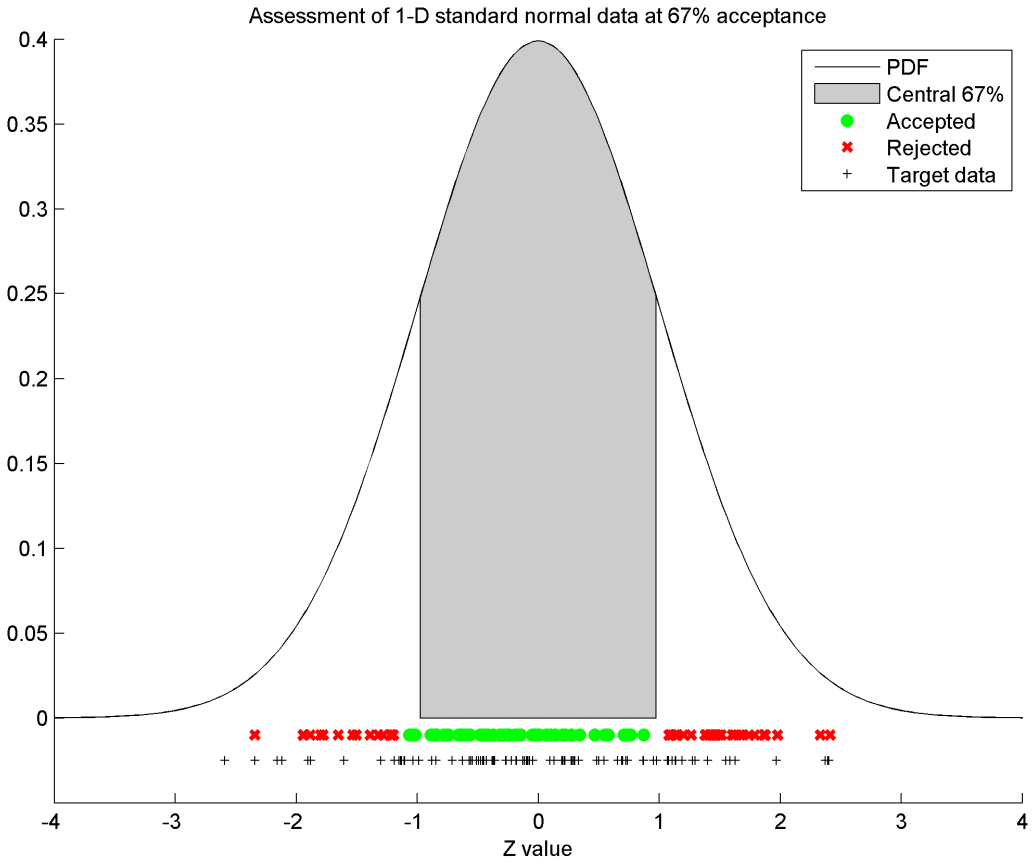

Figure 3.1: Assessment results at 67% acceptance for the 1-D example data. Included in the figure are the PDF of the standard normal distribution, the theoretical 67% target region, the target data, and the results of performing an assessment of the observations using the mean as the central value. The shaded area represents the central 67% of the probability, and the vertical boundary lines define the theoretical target region.

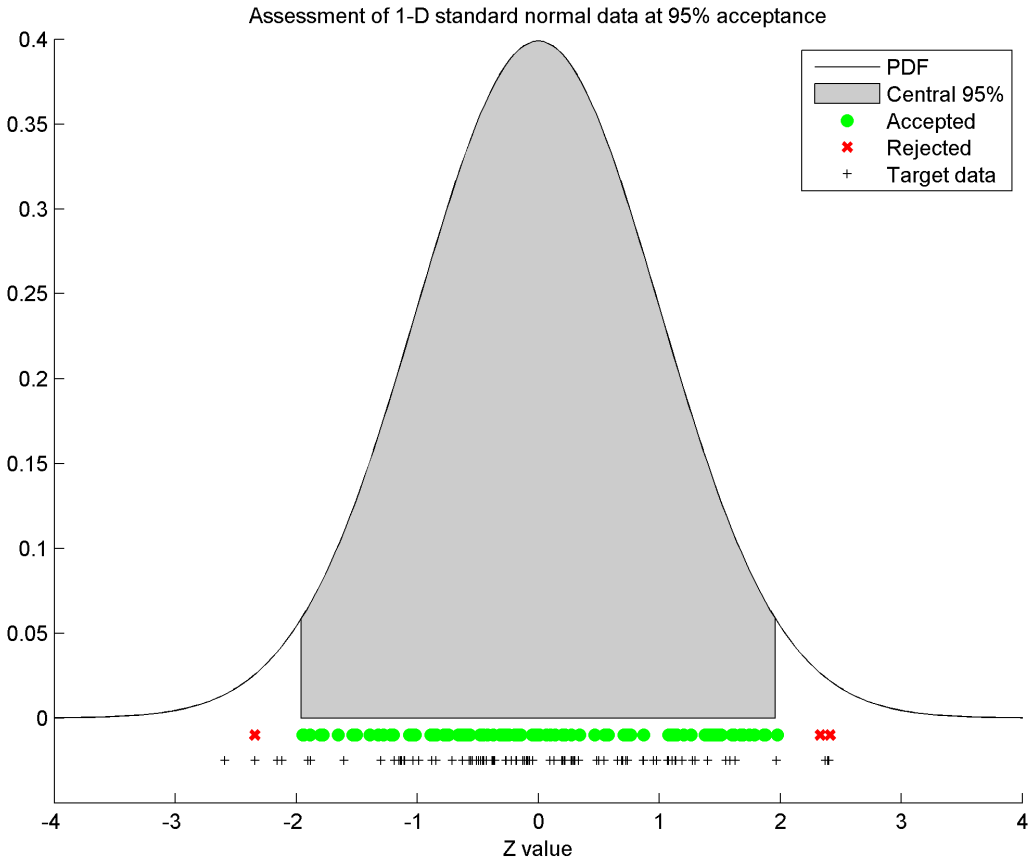

Figure 3.2: Assessment results at 95% acceptance for the 1-D example data. Included in the figure are the PDF of the standard normal distribution, the theoretical 95% target region, the target data, and the results of performing an assessment of the observations using the mean as the central value. The shaded area represents the central 95% of the probability, and the vertical boundary lines define the theoretical target region.

First, the target and observation data sets were generated independently and they, therefore, have different statistical properties, as indicated by the summary statistics in Table 3.1. Second, the observation data set has a mean value that is approximately 0.06 units greater than the mean of the target data set, as well as having a higher degree of variability, having a standard deviation of 1.1131 *vs.* 1.0268 for the target data set. Third, the assessment procedure essentially identifies outliers relative to the target data set for a specified acceptance percent, and the greater variability and bias of the observation data, relative to the target data, would, therefore, tend to produce a greater number of outliers than would be expected on average. This behavior is particularly noticeable for smaller acceptance percentages, where small differences in location and spread can have a significant impact on the degree of overlap between the target and observation data.

At 95% acceptance for this target, 97% of the observations were accepted and 3% of the observations were rejected. The difference of 2% between the target acceptance level of 95% and the acceptance of 97% is quite reasonable, even given the differences in the statistical properties of the target and observation data sets already described. The explanation for the better agreement in this case is that as the desired acceptance percent increases, more of the range of possible values is spanned by the acceptance region of the target. There is, therefore, a greater chance of overlap between the target and observation data, when they are from the same distribution, as in this example.

3.2 Example 2: Two-dimensional standard normal distribution

This example uses simulated target and observation data from the two-dimensional standard normal distribution $N(0, I_2)$, where I_2 is the 2×2 identity matrix and 0 is the two-dimensional zero vector. For this example, two independent random samples of size $N = 1000$ were generated from the $N(0, I_2)$ distribution to obtain target and observation data files, `tgt_2.csv` and `obs_2.csv`, respectively. A larger sample size was used for this example to demonstrate the improvement that a larger sample size can have when using smaller acceptance percentages, as well as to provide an adequate sample size for the two-dimensional example. The *curse of dimensionality* requires more data in higher dimensions to obtain a similar degree of approximation to the data distribution (Silverman, 1986). Summary statistics for the target and observation data sets are presented in Table 3.2. The variation measures the variability about an estimate of the mode, and is similar to the covariance. As for the one-dimensional case, a symmetric distribution, like the normal distribution, has a mean value that is equal to its mode, so the mode estimate and the variation matrix should be nearly equal to the mean and covariance matrix for each data set. This will not be the case for nonsymmetric data distributions.

The two-dimensional target data are displayed in Figure 3.3. Also shown in the figure are the true or theoretical 67% and 95% acceptance region boundaries for the two-dimensional standard normal distribution, and the respective empirical distance distribution approximations. The acceptance region boundaries obtained from the empirical distance distribution are nearly coincident with the theoretical acceptance region boundaries for both acceptance percentages, demonstrating the consistency of the procedures used to compute the empirical distance distribution and the acceptance regions in multiple dimensions.

The `empd_assess.exe` program was used to assess the individual points in the observation data set relative to the empirical distance distribution obtained from the target data set using the mean as the central value. Acceptance percentages of 67% and 95% were used to control the extent of the target. These two targets, then, represent, concentric circles centered at the mean of the target data set, with the 67%

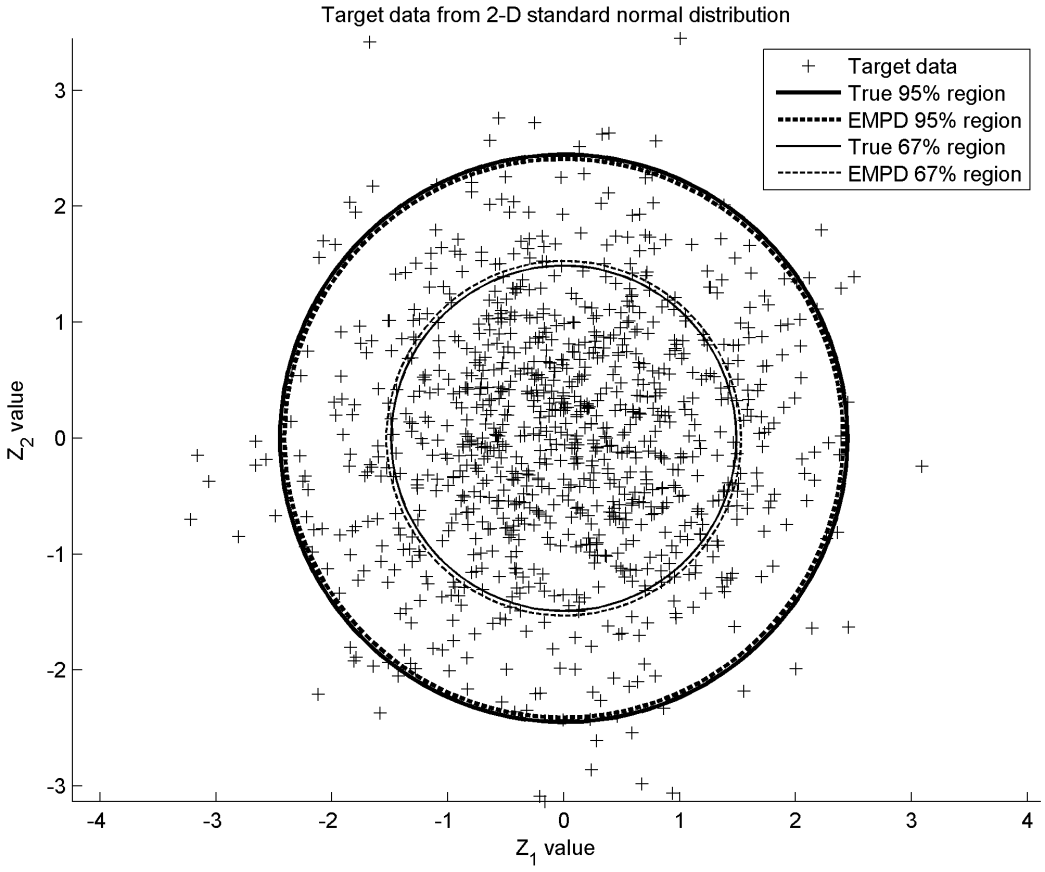

Figure 3.3: Target data set for the 2-D example data. The data were obtained from a 2-D standard normal distribution $N(0, I_2)$. Included in the figure are the theoretical 67% and 95% target acceptance region boundaries and the approximate empirical distance target acceptance region boundaries based on using the mean as the central value.

Table 3.2: Summary statistics for the 2-D example target and observation data sets. The 2-D example data consisted of independently generated target and observation data sets containing 1000 points from a standard 2-D normal distribution $N(0, I_2)$.

	Target data		Observation data	
Statistic	Z_1	Z_2	Z_1	Z_2
Mean	-0.0122	0.0127	0.0400	0.0230
Std. Dev.	1.0117	1.0164	0.9944	1.0047
Minimum	-3.2191	-3.1363	-3.1977	-3.7162
Median	-0.0060	0.0181	0.0606	0.0517
Maximum	3.0903	3.4495	3.2592	2.9239
Covariance	1.0236	0.0431	0.9888	-0.0473
	0.0431	1.0331	-0.0473	1.0094
Mode	-0.2109	0.2120	0.2060	-0.0276
Variation	1.0631	0.0034	1.0164	-0.0557
	0.0034	1.0729	-0.0557	1.0119

acceptance region boundary located at approximately one standardized distance unit radially from the mean and the 95% acceptance region boundary located approximately two standardized distance units radially from the mean. The assessment at 67% was performed using the command line

```
..\empd_assess -c mean -a 67 -t tgt_2.csv -l assess_2_67.log -o assess_2_67.csv obs_2.csv
```

where `-c mean` indicates that the mean value of the target data set was used as the center, `-a 67` indicates the desired acceptance percent, `-t tgt_2.csv` identifies the target data file, `-l assess_2_67.log` identifies the log file, `-o assess_2_67.csv` identifies the assessment output file, and `obs_2.csv` is the observation data file. The results of the 67% assessment appear in Figure 3.4.

The 95% assessment for the two-dimensional data was performed using the command line

```
..\empd_assess -c mean -t tgt_2.csv -l assess_2_default.log -o assess_2_default.csv obs_2.csv
```

where all of the options have the same meanings as before, but the log and assessment output file names have been changed. The `-a` option to define the acceptance percent was not used in this example since an acceptance percent of 95% is the default if no acceptance percent is specified. The results from the 95% assessment appear in Figure 3.5.

At 67% acceptance for the two-dimensional target, 69.6% of the observations were accepted and 30.4% of the observations were rejected. The difference of 2.6% between the target acceptance level of 67% and the acceptance of 69.6% of the observations is quite reasonable, given that the independence of the target and observation data sets implies that there should be some differences between the properties of the two data sets. Similar to the one-dimensional assessments, a greater degree of variability in the overlap between the target and observation data is expected for smaller acceptance percentages, which can, in turn, affect the observed acceptance percentage for any particular target and observation data sets.

At 95% acceptance for the two-dimensional target, 95.3% of the observations were accepted and 4.7%

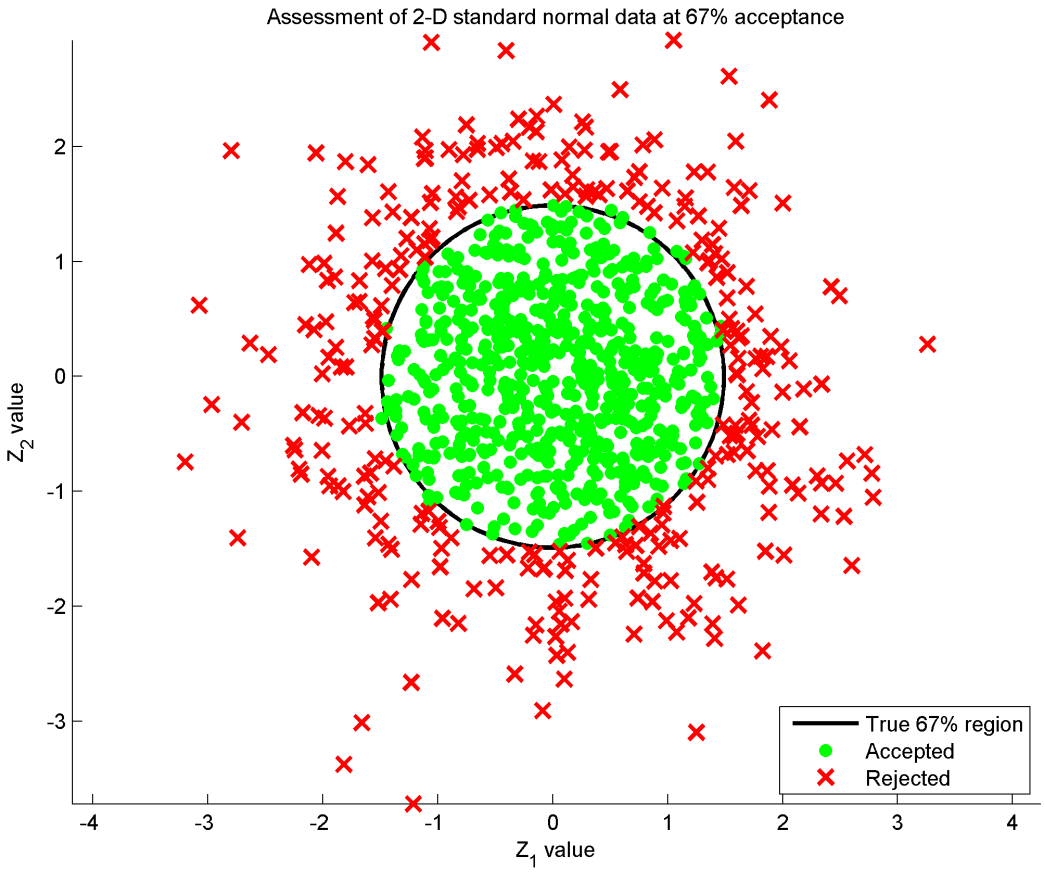

Figure 3.4: Assessment results at a 67% acceptance for the 2-D example data. Included in the figure are the theoretical 67% target region boundary and the results of performing an assessment of the observations using the mean as the central value.

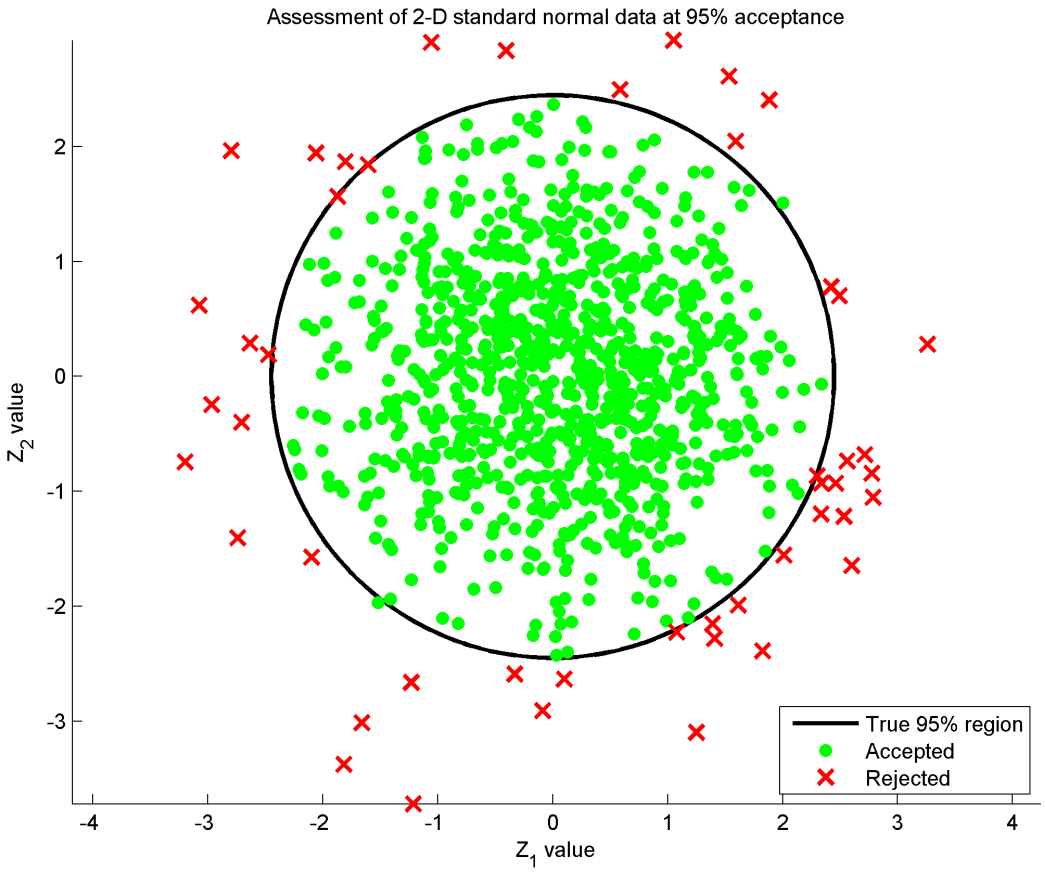

Figure 3.5: Assessment results at a 95% acceptance for the 2-D example data. Included in the figure are the theoretical 95% target region boundary and the results of performing an assessment of the observations using the mean as the central value.

of the observations were rejected. The difference of 0.3% between the target acceptance level of 95% and the acceptance rate of 95.3% for the assessed observations is quite reasonable, given the similarity in the statistical properties of the target and observation data sets given in Table 3.2 and the sample size. Again, the stronger agreement between the observed acceptance rate and the desired acceptance percent at 95% acceptance is caused by the fact that more of the range of possible values is spanned by the acceptance region of the target, giving a greater chance of overlap between the target observation data, when they are from the same distribution, as in this example.

3.3 Example 3: Assessment of forest structures

For the third example, the target and observation data sets were derived from a forestry data set comprised of stand summaries obtained using data from sample plots contained in the PNW FIA integrated database (IDB) version 1.4 (Hiserote and Waddell, 2004). The IDB, produced by the Pacific Northwest Forest Research Station Forest Inventory and Analysis program of the U.S. Forest Service (Hiserote and Waddell, 2004), contains inventory data for California, Oregon, and Washington collected by the Forest Service and the Bureau of Land Management, including data from the Forest Inventory and Analysis program of the Pacific Northwest Research Station (PNWFIA), the Continuous Vegetation Survey program of the Pacific Northwest Region (R6, Region 6), the Forest Inventory program of the Pacific Southwest Region (R5, Region 5), and the Natural Resource Inventory program of the Bureau of Land Management (BLMWO, western Oregon districts only) (Hiserote and Waddell, 2004). The inventory data from these sources were standardized by the FIA to include a uniform set of attributes and were combined within the IDB to provide a high quality comprehensive database of forest inventory information for these states.

The target and observation data sets created for this example were intended to provide a real world application of the empirical distance assessment procedures. The objective of this example is to identify a reference condition of forest structure attributes, the target, and to assess a different set of observed values for those attributes relative to the target. The forest structure attributes used here, as well as the target and observation data files, were solely for demonstration purposes and they should not be used in other contexts where their use has not been validated. The following criteria were used to select stands from the IDB.

- Sample plots were located in western Washington and western Oregon.
- Sample plots were classified as forest land, using the `GLC` code 20.
- Elevations of the sample plots were less than 2500 feet.
- Ages of the sample plots were at least 80 years. Ages were obtained using the FIA age class codes associated with the data column `STAND_AGE`.
- Sample plots were from Douglas-fir dominated stands, FIA species code 202. The stand type was determined using the codes in the `FOREST_TYPE` column.
- The sample plots did not contain any residual overstory trees, as indicated by the `STAND_POS` column. Residual overstory trees were defined by the FIA as

Live trees growing within the residual overstory component of a stand. Residual overstory trees are generally older trees that remain after a stand altering disturbance has occurred (such as harvest, fire, or windstorm).

- Sample plots contained trees having DBH values of at least 12 inches.

These data selection criteria yielded a total of 1482 sample plots distributed throughout western Washington and Oregon. The sample plots obtained were assumed to represent independent Douglas-fir stands. Stand summaries consisting of stand density, measured in trees per acre (TPA), and average tree size, measured as quadratic mean diameter (QMD) in inches, and average tree height in feet, were computed for each stand. Only trees having diameter at breast height (DBH) measurements of at least 12 inches were used to compute the stand summaries. This lower DBH limit was used to reduce the variability in stand density values, as well as to emphasize the moderate and larger sized trees in the targeted reference condition. Forests containing moderate to larger sized trees are generally considered to be more desirable, for aesthetic and economic reasons, and potentially more difficult to produce than forests having very large numbers of small trees.

Target and observation data sets for this example were created using the TPA and QMD data from the sample stands by assigning the odd numbered stand summaries to the target data set and the even numbered stand summaries to the observation data set. Summary statistics for target and observation data sets were computed, and appear in Table 3.3. A visual inspection of the target and observation data sets was also conducted, see Figure 3.6, to ensure that there was sufficient overlap between the two data sets to make the assessment worthwhile.

The summary statistics and the plot of the TPA and QMD data indicated that the partitioning of the larger data set was reasonable. The variation, again, measures the variability about an estimate of the mode, and is similar to the covariance. As for the normal distribution examples, the target and observation data sets are somewhat symmetric, so their respective mean and mode values are very similar, as are the variation matrix and the covariance matrix for each data set.

The `empd_assess.exe` program was used to assess the TPA-QMD points in the observation data set relative to the empirical distance distribution computed using the target data set and an estimate of the mode as the central value. A 90% acceptance percent was used for the assessment. An estimate of the mode was used rather then the mean since the TPA-QMD data are not radially symmetric, like the normal distributions, and the assessment procedures are based on using the most likely point, the mode, as the central value when computing the empirical distance distribution. The assessment was performed using the command line

```
..\empd_assess -a 90 -c mode -t tgt_nq.csv obs_nq.csv
```

where `-c mode` indicates that an estimate of the mode of the target data set was used as the center, `-a 90` indicates the desired acceptance percent, `-t tgt_nq.csv` identifies the target data file, and `obs_nq.csv` is the observation data file. The use of the `-c mode` option was not required, as the mode is the default, but was included for clarity. Also, in this example the default log and assessment output file names, `obs_nq_assessment.log` and `obs_nq_assessed.log`, respectively, were used. The default log and assessment output file names were derived from the observation file name. The results of the 90% assessment appear in Figure 3.7.

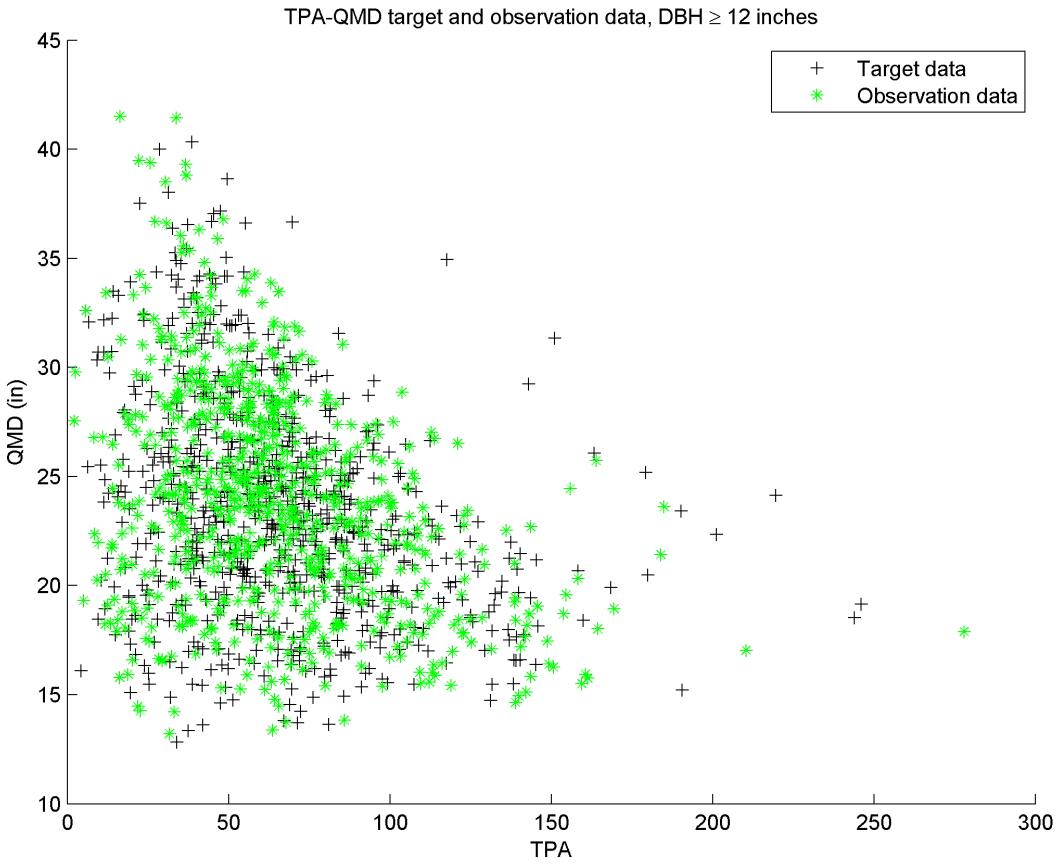

Figure 3.6: TPA-QMD target data set.

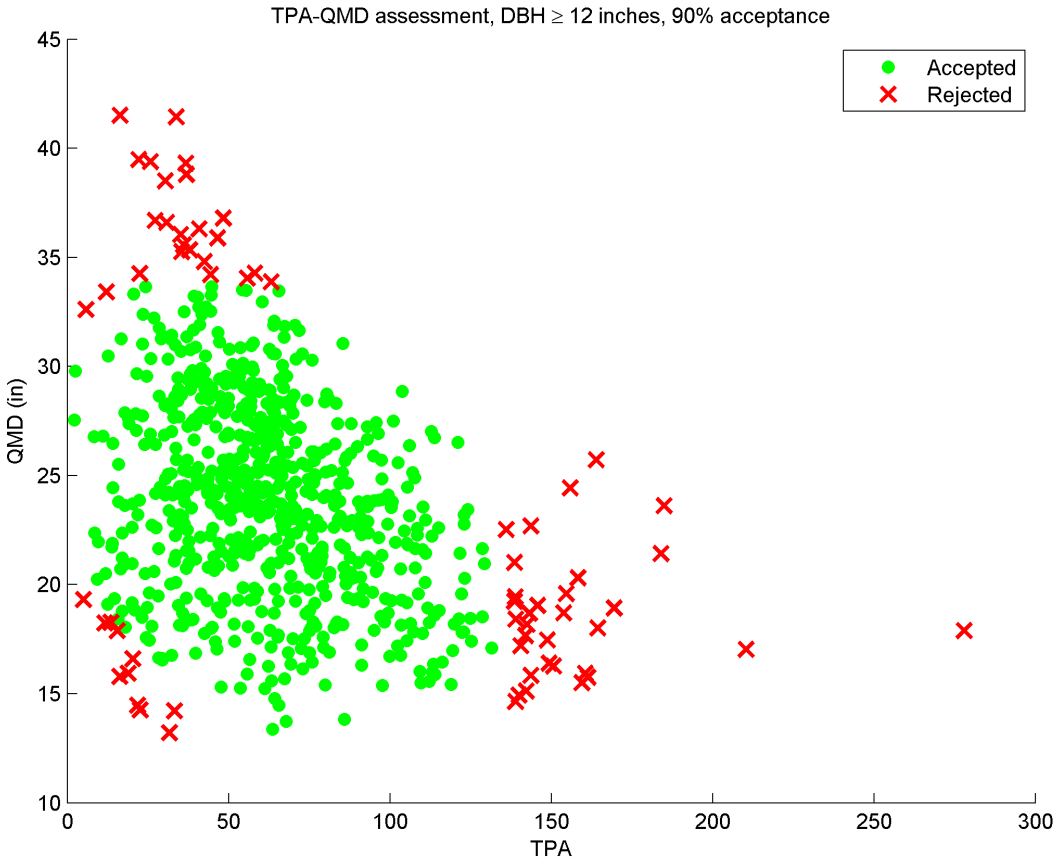

Figure 3.7: TPA-QMD assessment results for 90% acceptance using an estimate of the mode as the central value.

Table 3.3: Summary statistics for the TPA-QMD target and observation data sets. The TPA-QMD data consisted of 1482 points, with the odd numbered points assigned to the target data set and the even numbered points assigned to the observation data set.

	Target data		Observation data	
Statistic	TPA	QMD	TPA	QMD
Mean	65.8	23.6	64.7	23.9
Std. Dev.	33.8	5.0	33.4	5.0
Minimum	4.3	12.8	2.1	13.2
Median	61.6	23.0	59.9	23.6
Maximum	246.1	40.3	277.9	41.5
Covariance	1143.8	-52.5	1117.0	-59.9
	-52.5	25.3	-59.9	25.2
Mode	65.0	23.4	61.4	25.4
Variation	1144.5	-52.4	1127.9	-64.8
	-52.4	25.3	-64.8	27.4

The assessment results presented in Figure 3.7 clearly indicate that the central portion of the observation data set was accepted, and that the tails extending along the QMD and TPA axes were rejected, as well as some points nearer to the origin. This is exactly the behavior expected of the assessment procedures, as the TPA-QMD points in the tails and those nearer to the origin are furthest from the mode estimate of 65.0 TPA with a QMD of 23.4 inches. The assessment accepted 91.0% of the observations, rejecting 9.0% of the observations. If this target were to be used in practice, an 80% acceptance percent, producing a more restrictive target, may be more appropriate, as it would reject some of the very low density stands along the left edge of the data, as well as rejecting some of the stands having higher density and smaller trees along the lower edge of the data.

As a final demonstration of the assessment procedures a three dimensional assessment is performed, using the forestry data set previously described. The three-dimensional target and observation data sets consisted of TPA, QMD, and average height values for the selected forest stands. The stands were assigned as before. Summary statistics for the three-dimensional forest structure target and observation data sets appear in Table 3.4.

The `empd_assess.exe` program was used to assess the three dimensional TPA-QMD-average height points in the observation data set relative to the empirical distance distribution computed using the target data set using an estimate of the mode as the central value. A 90% acceptance percent was used here as well. An estimate of the mode was used rather then the mean for the same reasons just given for the two-dimensional TPA-QMD example. The assessment was performed using the command line

```
..\empd_assess -a 90 -c mode -t tgt_nqh.csv obs_nqh.csv
```

where `-c mode` indicates that an estimate of the mode of the target data set was used as the center, `-a 90` indicates the desired acceptance percent, `-t tgt_nqh.csv` identifies the target data file, and `obs_nqh.csv` is the observation data file. The use of the `-c mode` option was not required, as the mode is the default, but was included for clarity. Also, in this example the default log and assessment output file names, `obs_nqh_assessment.log` and `obs_nqh_assessed.log`, respectively, were used. The default log and assess-

Table 3.4: Summary statistics for the TPA-QMD-average height target and observation data sets. The TPA-QMD-average height data consisted of 1482 points, with the odd numbered points assigned to the target data set and the even numbered points assigned to the observation data set.

	Target data			Observation data		
Statistic	TPA	QMD	Avg. Height	TPA	QMD	Avg. Height
Mean	65.8	23.6	110.8	64.7	23.9	111.4
Std. Dev.	33.8	5.0	18.9	33.4	5.0	18.6
Minimum	4.3	12.8	49.3	2.1	13.2	52.4
Median	61.6	23.0	108.6	59.9	23.6	110.8
Maximum	246.1	40.3	174.5	277.9	41.5	199.2
Covariance	1143.8	-52.5	-23.5	1117.0	-59.9	-2.8
	-52.5	25.3	73.6	-59.9	25.2	69.6
	-23.5	73.6	357.8	-2.8	69.6	346.1
Mode	64.0	23.6	110.7	56.9	25.3	113.3
Variation	1147.3	-52.6	-23.2	1177.9	-70.8	-18.1
	-52.6	25.3	73.6	-70.8	27.1	72.3
	-23.2	73.6	357.8	-70.8	27.1	72.3

ment output file names were derived from the observation file name. The results of the 90% assessment appear in Figure 3.8.

The assessment results presented in Figure 3.8, again, clearly indicate that the central portion of the observation data set was accepted, and that the tails extending along the QMD and TPA axes were rejected, as well as some points near the origin. this is consistent with the nature of the distributions of the three structure attributes: average height is more symmetrically distributed than were TPA and QMD and the influence of the TPA and QMD tails dominated the assessment, in terms of distance from the mode, causing the tail points to be rejected. The assessment accepted 89.9% of the observations, rejecting 10.1% of the observations. As before, if this target were to be used in practice, an 80% acceptance percent, producing a more restrictive target, may be more appropriate.

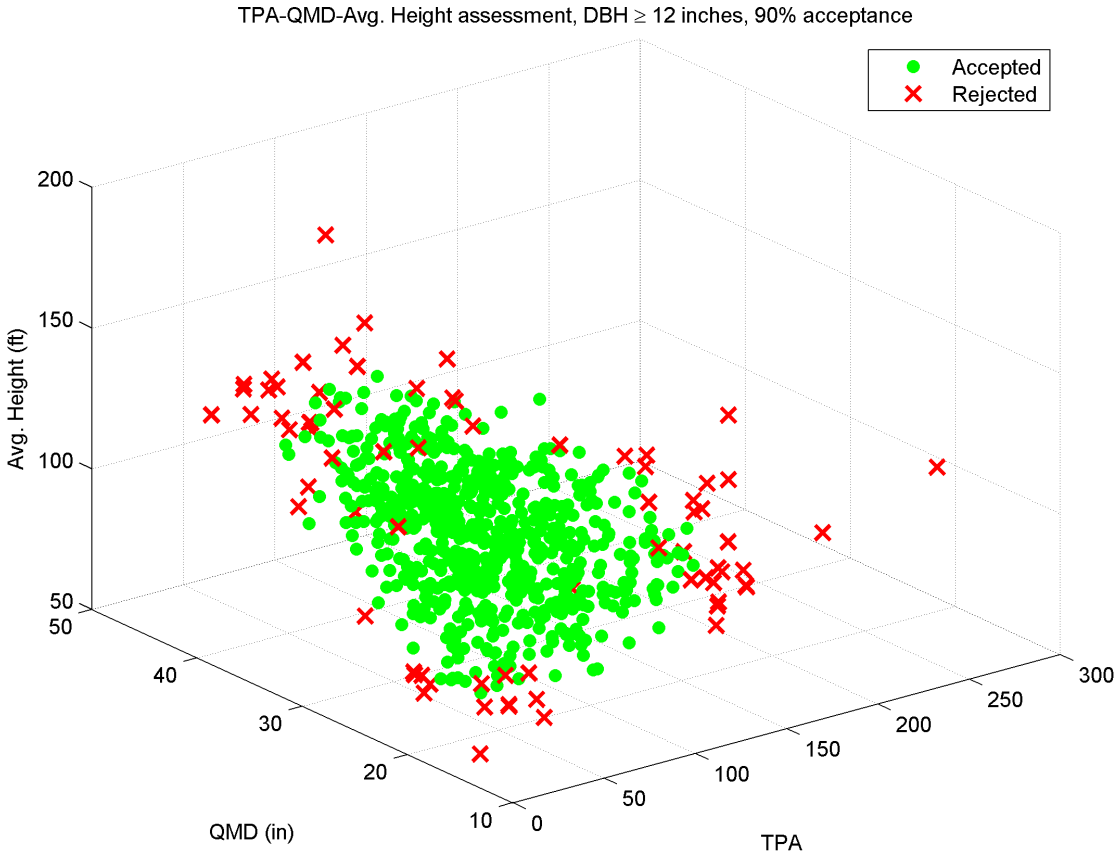

Figure 3.8: TPA-QMD-Avg. height assessment results for 90% acceptance using an estimate of the mode as the central value.

Bibliography

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK User's Guide*. Softare, Environments, Tools. SIAM, Philadelphia, Third edition, 1999.

Kevin R. Gehringer. Structure-based nonparametric target definition and assessment procedures with an application to riparian forest management. *Forest Ecology and Management*, 223:125–138, 2006.

Bruce Hiserote and Karen Waddell. *The PNW-FIA Integrated Database User Guide: A Database of Forest Inventory Information for California, Oregon, and Washington*. Forest Inventory and Analysis Program, Pacific Northwest Research Station, Portland, Oregon, v 1.4 edition, April 2004.

The MathWorks. *MATLAB: The language of technical computing*. The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA, 01760-2098, version 7 edition, 2006.

B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability 26. Chapman & Hall/CRC, 1986.