United States
Environmental Protection
Agency

Customer Complaint Surveillance: Threshold Analysis Tool User Manual

Office of Water (MC-140)
EPA 817- B-13-005
September, 2013

Disclaimer

The Threshold Analysis Tool (TAT) was developed by EPA's Office of Water through its Water Security Division, with support provided under EPA Contract No. EP-C-10-060. The information contained in the TAT and in this Threshold Analysis User Manual is not intended to revise or update EPA policy or guidance on safeguarding the security of drinking water or wastewater systems. Any mention of trade names or commercial products does not constitute endorsement or recommendation for use. For further information about EPA's water security efforts, please see http://water.epa.gov/infrastructure/watersecurity/

The TAT was prepared as an informational tool to assist water utilities in tentatively identifying or developing thresholds for a Customer Complaint Surveillance (CCS) component in a Surveillance and Response System (SRS). It does not purport to provide or produce an accurate, comprehensive, or exhaustive analysis of all potential threshold values from user-provided data, nor does it purport to identify appropriate threshold values for detecting drinking water contamination, to identify instances of actual drinking water contamination or to verify the need for or accuracy of water contamination alarms.

The information contained in the TAT was developed using statistical methodology. It should not be relied upon exclusively in developing SRS alarm thresholds, refining existing thresholds, or verifying actual alarms a utility may have received. This information is also not a substitute for the professional advice of an analytical, environmental or security professional. The information is provided without warranty of any kind, and EPA hereby disclaims any liability for damages arising from the use of the tool, including, without limitation, direct, indirect or consequential damages including personal injury, property loss, loss of revenue, loss of profit, loss of opportunity or other loss. For assistance locating an analytical, environmental, or security professional, please contact an appropriate credentialing authority.

Questions concerning this document should be addressed to:

Nelson Mix, PE, CHMM
U.S. EPA Water Security Division
1200 Pennsylvania Ave, NW
Mail Code 4608T
Washington, DC 20460
(202) 564-7951
Mix.Nelson@epa.gov

Table of Contents

List of Tables

List of Figures

Quick Start Guide

Step 1. Download the Threshold Analysis Tool (TAT) and install it on a computer's hard drive.

Step 2. Go to the Start button on the toolbar; select "All Programs." Click on the TAT program to open the tool.

Step 3. On the main screen of the tool ("Gather Data"), select the "Choose Files" option. This will open the "Please Select a File" Window.

Step 4. From the "Please Select a File" Window, choose a sample data file that has been included with the program in the Sample Data folder, or choose a file that a user has previously created. File formats accepted by the program are .txt, .csv, .xls, .xlsx and .tab. These file formats are automatically detected by the program.

Step 5. After choosing a file, the user will be directed to the Input File Form. While viewing "File Contents," choose the column of data wanted for each data type (date, time, spatial unit and event counts) the tool to use to determine thresholds, and then select the "Process" option located at the bottom of the screen. This will load all of the data from the selected columns into the tool.

Step 6. Once data are loaded, the user will be notified of any errors in data on the main screen of the tool. The user may be able to proceed with errors, but it will affect the accuracy of results. The user will also be able to set both the scan window (1, 2, 3, 4, 5, 6 or 7 days) and algorithm reset type (reset or continuous) by checking "Set Analysis Type." Once the user has made these choices, click the "Set Thresholds" option located on the upper left portion of the main screen.

Step 7. Once the user has selected the "Set Thresholds" option, the user may modify the percentile, standard deviation, and recurrence interval. Alert details and the number of alerts will be displayed on the bottom of the screen under the heading "Alert Details." The resulting threshold will appear in the "Threshold" box immediately above the option modifiers. At this point the user may either return to the main screen to change the analysis type (Step 6), or may create a report (Step 8).

Step 8. Once the user has completed analyzing the data and established a threshold, the user may create a report by exporting the results to a file that can be stored on the user's computer. To do this, check the "Create Report" option in the upper left portion of the main screen. Once the screen has changed, click the "Save File(s) As…" option. This will open a "Browse For Folder" Window that will allow the user to choose the location to have the output file saved. Once the user has selected the location, the file will automatically be saved.*

* All output files are saved in .csv format and are best viewed with Microsoft Excel. In some instances, the user must select this option before opening the file.

Section 1.0: Introduction

The U.S. Environmental Protection Agency (EPA) Water Security Initiative Surveillance and Response System (SRS) detection strategy involves the use of multiple monitoring and surveillance components for timely detection of drinking water contamination incidents. One key component of the SRS is customer complaint surveillance (CCS), which enhances the collection and analysis of calls from customers reporting unusual water quality concerns, and compares trends against an established baseline to detect possible contamination incidents. Automated scan algorithms can provide a simple solution for implementing the CCS component of a SRS, and automated scan statistics allow for an evaluation of a rolling time window, comparing the number of customer complaints (e.g., calls received by the utility) within the window to a preset threshold. When complaints exceed the threshold, an alert is issued.

Identifying appropriate thresholds is a critical step, and several factors can influence threshold selection, including how well customers can detect or sense a particular contaminant and how aware utility personnel are to alert occurrence rates. Since drinking water contamination incidents are rare, formal statistical evaluation of the sensitivity of thresholds is impeded by the lack of actual contamination incident data. Because of this limitation, CCS threshold determination is often based on a utility's analysis of the number of water quality complaints that are indicative of anomalous conditions balanced with the level of effort that may be required to investigate alerts.

When establishing thresholds, utilities may focus on how unusual the volume (number) or customer complaints must be in order for an alert to be generated. Three utilities that participated in EPA Water Security Initiative pilots have established thresholds for their CCS systems, each using a different statistical method to determine thresholds for use in identifying anomalous customer complaint volumes. The three different methods used were based on recurrence intervals, percentile and standard deviation.

The *recurrence interval* is the average time between alerts or how often alerts would occur. This is useful for establishing thresholds based on the level of effort of investigations. Utilities also may apply *percentiles*, which are based on the distribution of the number of alerts generated at each threshold. This approach ensures that the utility does not set thresholds beyond a reasonable expectation of complaints. Utilities may also use *standard deviation from the mean* to establish thresholds. *Standard deviation* is a measure of dispersion, or how much the individual measurements vary across all of the data. While all of these methods are related, none are preferred over the other two for establishing thresholds. The choice of method depends on utility needs and the comfort level of the utility with the statistical approach. Additional details and equations corresponding to these approaches are provided in Appendix C of this User Manual.

To aid historical analysis of customer complaint data, EPA has developed a Threshold Analysis Tool (TAT). The TAT retrospectively applies a configurable scan algorithm to data input by the user, applies one or more of the statistical methods for determining thresholds, and outputs when an alert would have been issued. This tool allows the user to easily:

- Conduct exploratory data analysis of historical customer complaint data in the early development of a CCS system;
- Enter data generated from a water utility's CCS component and assess thresholds based on recurrence interval, percentile or standard deviation;
- Generate a set of statistical analyses that can be manipulated by utilities implementing scan algorithms as part of their CCS system to establish acceptable thresholds; and
- Verify that a CCS component is properly functioning by providing independent identification of alert conditions, which can be cross-referenced with actual alerts received.

Section 2.0: Importing Data

Data may be imported to the tool from the main "Gather Data" page, using the "Choose File" option (see **Figure 2.1**).

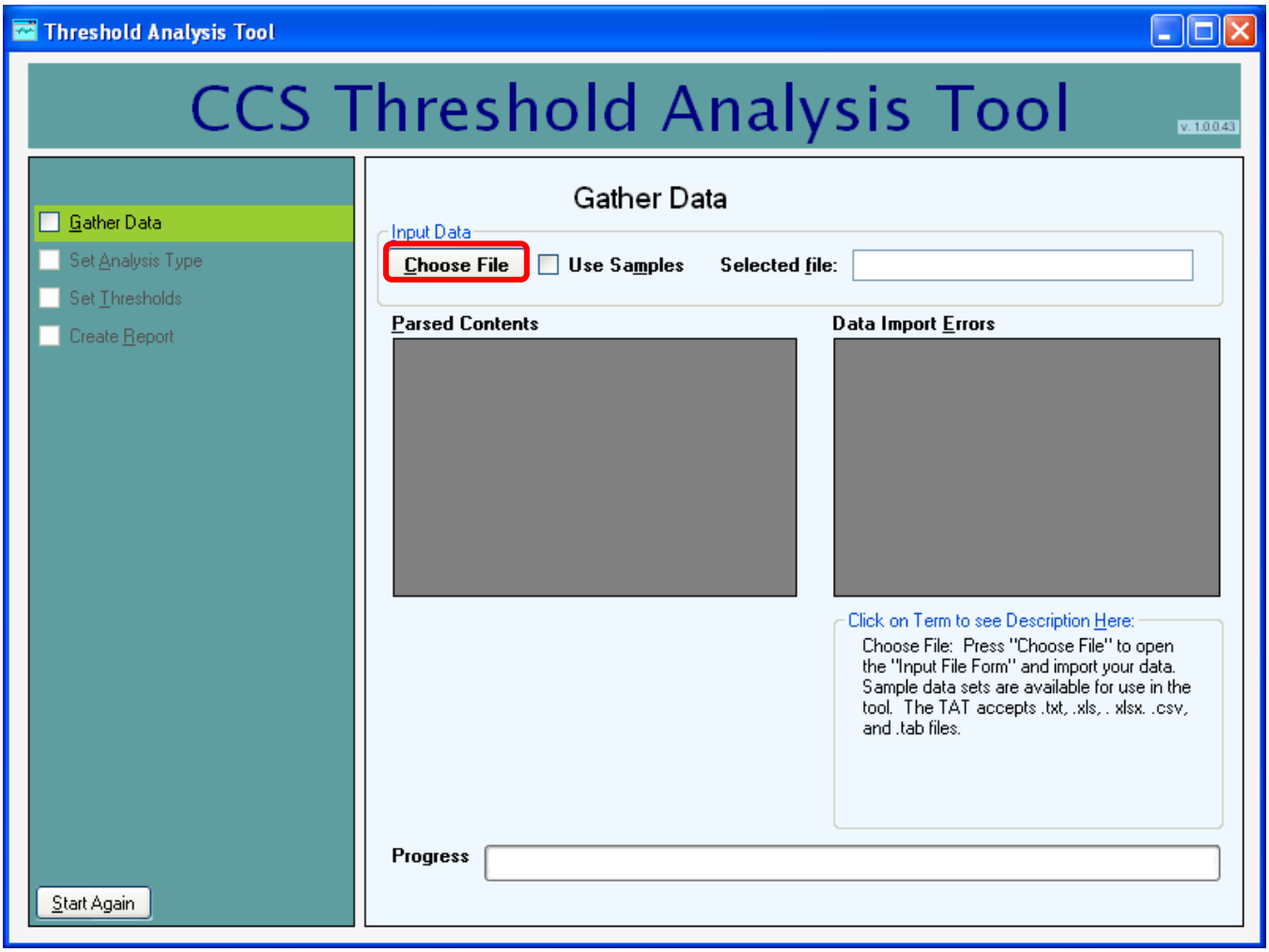

Figure 2.1. Main Screen

The user may either select the "Choose File," button, or check the "Use Samples" box and then select the "Choose File" button. Checking the "Use Samples" option will direct the user to sample files that have been included for convenience. Once the user has selected the "Choose File" button, a "Please Select a File" window will be opened (see **Figure 2.2**). This will allow the user to browse through folders to select data.

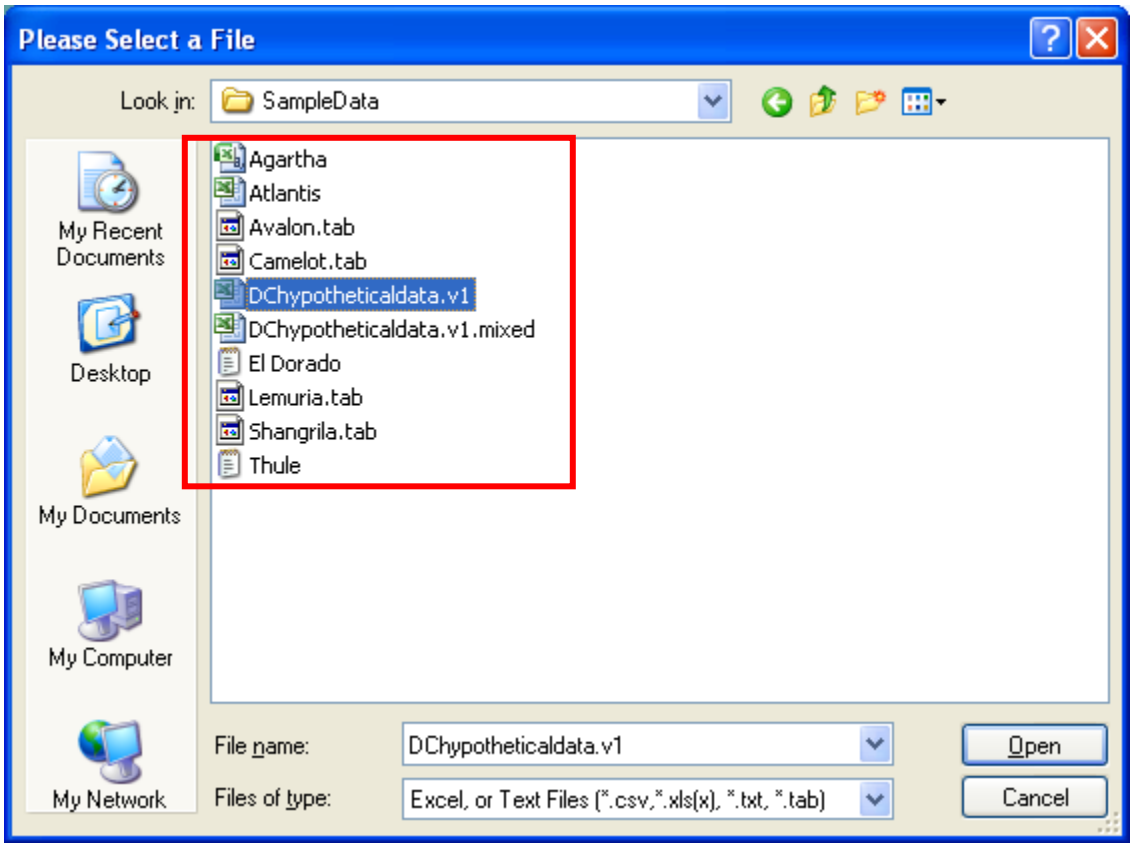

Figure 2.2. "Please Select a File" Window

The file selection window shown in **Figure 2.2** allows the user to browse the folders on a computer or network and select a file to open. File types accepted by the system are described in **Table 2-1**, and include .txt, .csv, .xls, .xlsx and .tab formats. The dialog box will display only these types of files and will not read other file types. These file formats and related terminology are also described in the Glossary, located in Appendix A of this User Manual.

Table 2-1. File Formats Accepted by the Threshold Analysis Tool

File Extension	File Delimiter	Description
.txt	Comma	A .txt, or text, file is structured as a sequence of lines. The end of a text file is often denoted by one or more special characters, known as an end-of-file marker, after the last line.
.xls, .xlsx	Excel	Excel files are produced by Microsoft Excel, a commercial spreadsheet application written and distributed by Microsoft for Microsoft Windows and Mac OS X. .xls files are produced by older versions of Excel, whereas .xlsx files are produced by Microsoft Excel 2007 and newer versions. **Note: Excel spreadsheets that contain macros will not work with this program.**
.csv	Comma	A .csv, or comma-separated values file format is a set of file formats used to store tabular numbers and text data in plain-text that can be easily written and read in a text editor. Traditionally, lines in the text file represent rows in a table, and commas separate the columns.

File Extension	File Delimiter	Description
.tab	Tab	A .tab, or tab-separated values file is a simple text format for a database table. Each record in the table is one line of the text file and each field value of a record is separated from the next by a tab stop character.

After navigating to the desired data file, select the file and click “Open”. This will open the “Input File Form,” which is described in the next section.

Section 3.0: Using the Input File Form

Once the user has chosen a file and clicked "Open", the tool will automatically direct the user to the "Input File Form," as shown in **Figure 3.1**, for an Excel file. The "Input File Form" shows a preview of the file selected in the "File Contents" box to aid in selection of the correct data for analysis. **Note: In an Excel file that contains multiple spreadsheets, the tool defaults to the first spreadsheet.**

The file chosen is read and parsed, or analyzed, based on selection criteria provided by the user. The tool will attempt to automatically select the correct file delimiter or format. If the tool is unable to determine the file format, the user will be allowed to choose a format from the "File delimiter" drop-down menu (as shown in **Figure 3.2**).

3.1. User Selections

The "Input File Form" initially imports the data and identifies the columns based on the file type chosen. The default settings for the example Excel file, "DCHypotheticaldata.v1," are shown in **Figure 3.1**.

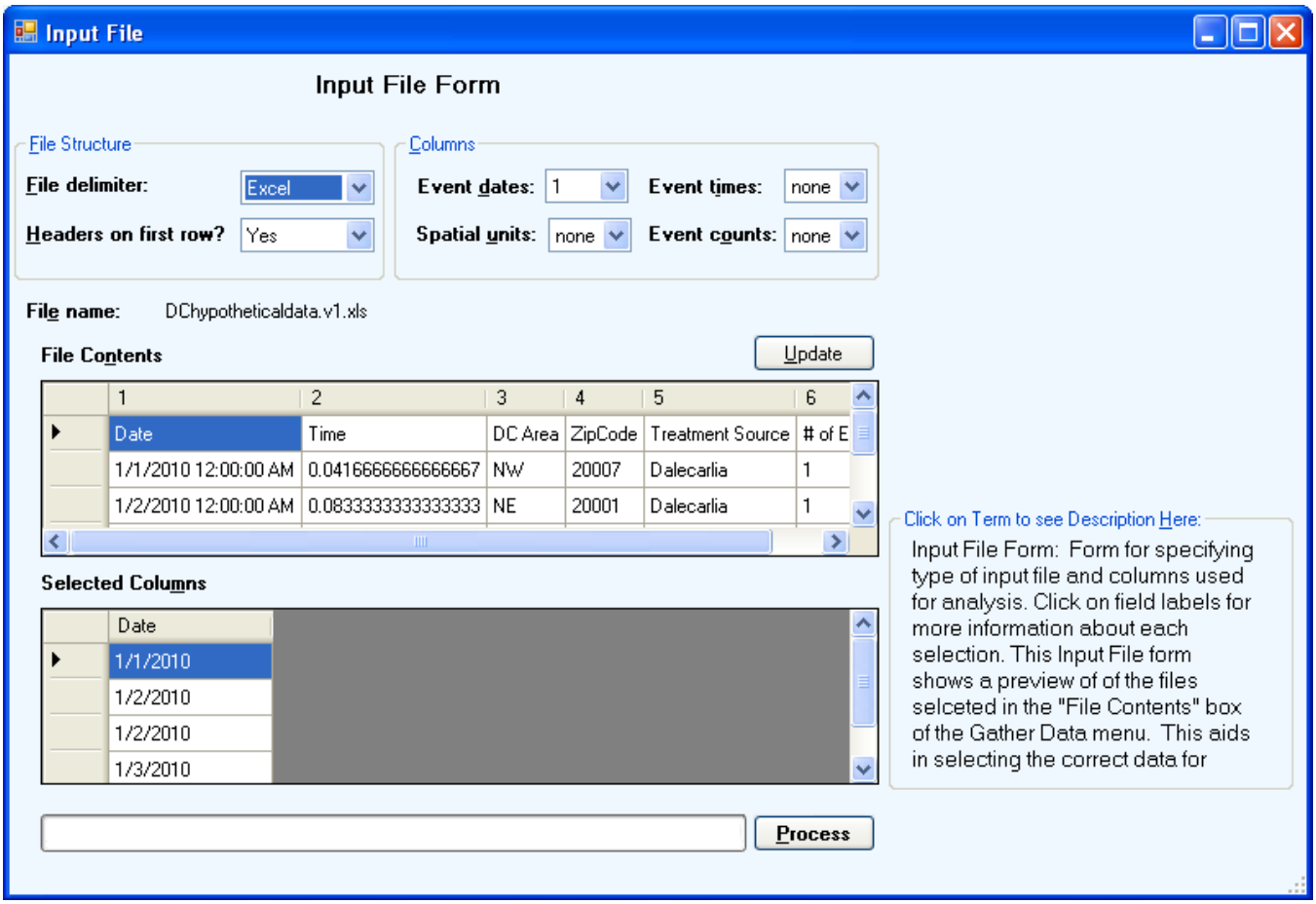

Figure 3.1. Default Input File Form Settings for an Example Excel File

The user has the ability to modify the file delimiter and column choices to better fit the data in the file chosen to import. These options will analyze and sort the user's data into a final format to be entered into the system.

All options available for modification on the "Input File Form" are listed in the sections that follow.

3.1.1 File Delimiter

A file delimiter is a character that marks a boundary between data items. For example, a comma-delimited file places commas between each complaint date and corresponding complaint time. This is sometimes synonymous with the file type (see **Table 2-1**). If the system does not automatically select the correct delimiter, the user can change the delimiter with this option. File delimiter choices are displayed on a drop-down "File delimiter" menu and are limited to comma, tab, single line and Excel (see **Figure 3.2**).

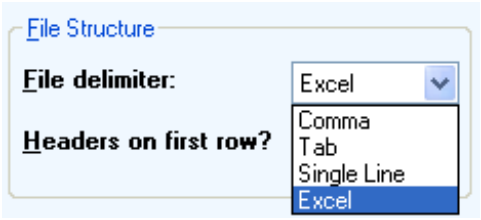

Figure 3.2. File Delimiter Drop-down Menu

Any changes in the file delimiter selection will be reflected immediately in the "File Contents" box (see **Figure 3.1**). Intentionally selecting an incorrect file form for a data file may cause an "Unhandled Exception" error and the program may shut down.

3.1.2 Headers on First Row?

If the first row of the data file contains column header labels, selecting the "Yes" option in the "Headers on first row?" drop-down menu, causes the program to ignore the first row during data processing (see **Figure 3.3**). If the data file has no column header, ensure that the "No" option is selected. If the user does not select "Yes" and the header row is included in the data, an error will be displayed when the data are processed.

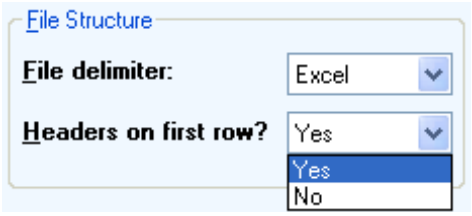

Figure 3.3. Headers On First Row? Drop-down Menu

Any changes in this option will be reflected immediately in the "Selected Columns" box on the "Input File Form" (**Figure 3.1**). Intentionally processing data that has headers using the selection "Headers on first row?" then "No," will cause a data read error and may also yield an application error.

3.1.3 Event Dates

The "Event dates" option located under "Columns" allows the user to choose the column containing event dates in the data file (see **Figure 3.4**). Event dates are the dates on which customer complaints occurred. (Note that in this section, and all subsequent sections, events refer to customer complaints received and documented by the drinking water utility.) An event date may be a record of a water quality complaint call date or the date a work order related to water quality was generated. If a column is not selected by the user, the tool will by default choose the first column in the data file for the date. It is recommended that the user preview "File Contents" and choose the correct column containing these data. The user will be able to choose from any column. Note that

while Excel uses letters to denote data columns, the tool recognizes multiple file types and uses drop-down menus with numbers to denote data columns.

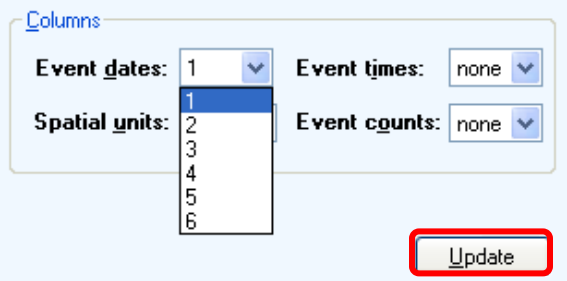

Figure 3.4. Event Dates Drop-down Menu

The tool uses Windows Routines to parse dates for standard date formats, in addition to recognizing YYYYMMDD and other Excel formats. Any changes in this option will be reflected in the "Selected Columns" box once the "Update" button has been clicked.

NOTE: "Event dates" is the only user option on this form required for the tool to operate.

Program Notes:

It is not necessary to have dates and times located in separate columns. If data are in MMDDYY HH:MM:SS (AM/PM) format, nothing will be affected.

If the original file contains a date and time in separate columns, the two fields will be combined into one in the format of MM/DD/YY HH:MM:SS (AM/PM).

If data do not contain a column for time, the time portion of the date will be set to 12:00:00 AM.

3.1.4 Event Times

Choice of event times (see **Figure 3.4**) allows the user to designate the column containing event times, if part of the data set. While the tool algorithms need both date and time data to operate, inclusion of event times is not required for analysis because the TAT will automatically use a default time of 12:00 AM if none is identified.

Any changes to the event times selection will be reflected in the "Selected Columns" box once the "Update" button has been clicked. Note that the tool may read time as a fraction of a day, similar to how Microsoft views date and time (e.g. 8:00 AM is displayed as 0.33333). The correct time format [HH:MM:SS (AM/PM)] will be displayed in the "Selected Columns" window when updated.

3.1.5 Spatial Units

Customer calls may be assigned to specific spatial units designated by the utility, such as a zip code or pressure zone (see "Spatial units" drop-down box in **Figure 3.4**). This choice allows the user to specify the column that contains data for the spatial unit in which an event occurred. Selecting a spatial unit column is necessary only if the user wants to set separate thresholds for each spatial unit. Any changes to this option will be reflected in the "File Contents" box once the "Update" button has been selected.

Additionally, this column can also be used to describe temporal units to parse the data by another descriptor, such as weekday or weekend. This will separate the data by described units of time, allowing the user to set thresholds for each unit. Note that the algorithm still analyzes the data as if there is no discontinuity in the timeline (e.g., the tool will still run the analysis on weekdays when a weekend temporal unit is selected, even if there are no data for weekdays). The tool does not distinguish between zip codes, days of the week, or any other descriptive text; it simply sorts the spatial units data, and then analyzes all data based on the units identified in the spatial units column. Users should ensure that the data modeled by the tool accurately depict any existing or desired CCS event detection algorithms when using the "Spatial unit" selection to identify temporal units.

3.1.6 Event Counts

Event counts may be used when a data set has multiple events (complaints) for a specific date (e.g., 01/02/2011, 4), as opposed to multiple date listings, with each representing a single event (see "Event counts" drop-down box in **Figure 3.4**). The event counts option allows the user to designate the column that contains the event counts in the data file. A utility may not have a specific time for each complaint, but may have the total number of complaints received in an hour, day or week/weekend. Event counts should be used for this situation. The user will be able to choose from any column. Any changes to this option will be reflected in the "Selected Columns" box once the "Update" button has been selected. A user dialog pop up box will appear when selecting an event count, as shown in **Figure 3.5**.

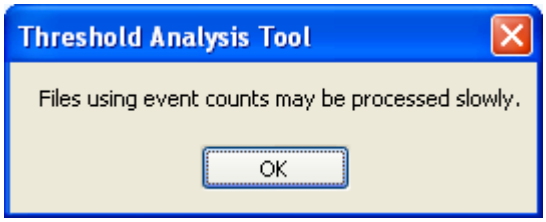

Figure 3.5. Event Counts Pop-up

3.1.7 Update Button

Each time the user selects a column choice, the user may click the "Update" button to view selections and changes in the "Selected Columns" box (see **Figure 3.6**). Failure to select "Update" will not affect data processing.

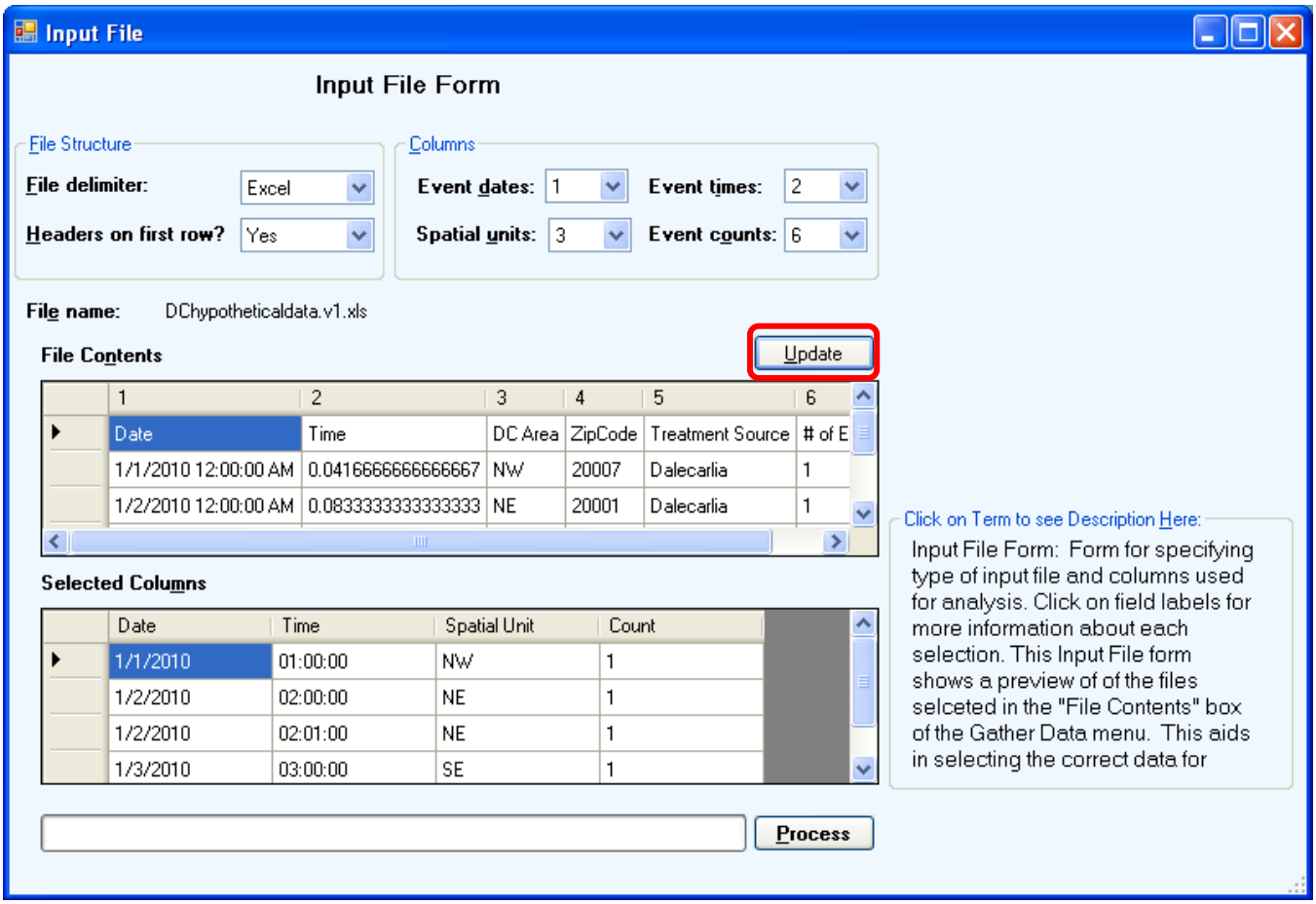

Figure 3.6. Input File Form with User Selections

3.2 Process Data

Once the user has chosen columns of data to analyze, the next step is to process the data. By selecting the "Process" button shown in **Figure 3.7**, the "Input File Form" will close and the main screen will be displayed with the processed columns in the parsed "File Contents" table.

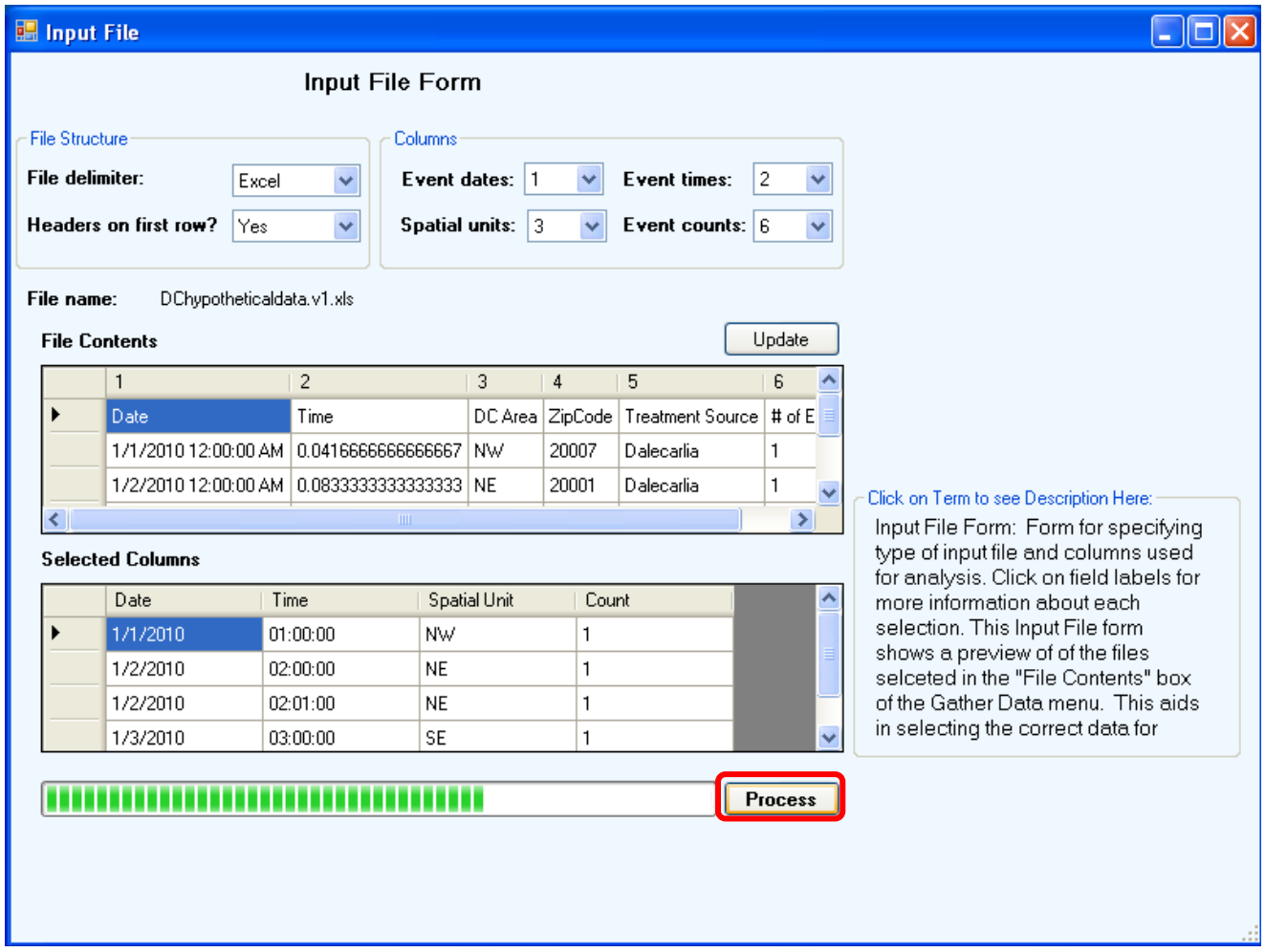

Figure 3.7. File Import Form Process Data Option

Once the user has processed the data and it has been loaded to the tool, the user will receive the message shown in **Figure 3.8.** Select "OK" and proceed to the main screen.

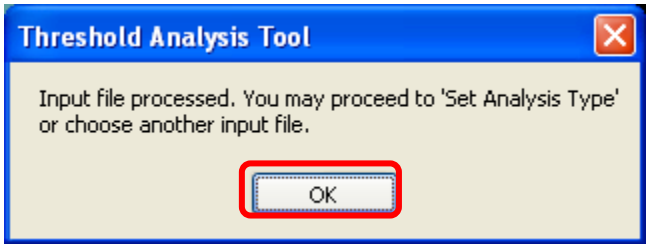

Figure 3.8. Threshold Analysis Tool Analysis Pop-up

3.2.1 Process Data Errors

The TAT contains over 100 custom user dialog boxes that have been strategically coded to assist with effectively using the tool. Issues may be identified in pop-up boxes, or in the tables in the tool itself. If any rows cannot be

processed, error details will be displayed in the "Data Import Errors" table (see **Figure 3.9**). To resolve data import errors, the user must go back through the "Please Select a File" window and enter the "Input File Form" to make changes to the "File Structure" or column selections, or alter the data outside of the tool and import the revised data. The tool is designed to accept multiple data types. However, the user may have to do some data preparation if the tool does not accept the original data configuration. If there are import errors, the data may be incomplete, leading to unreliable results. The tool can analyze a maximum of 40,000 data points. If there are more than 200 errors, the tool will return an error message and will not analyze the data. Once the user has made appropriate corrections to the data and/or selections, the user will be able to move forward with data analysis.

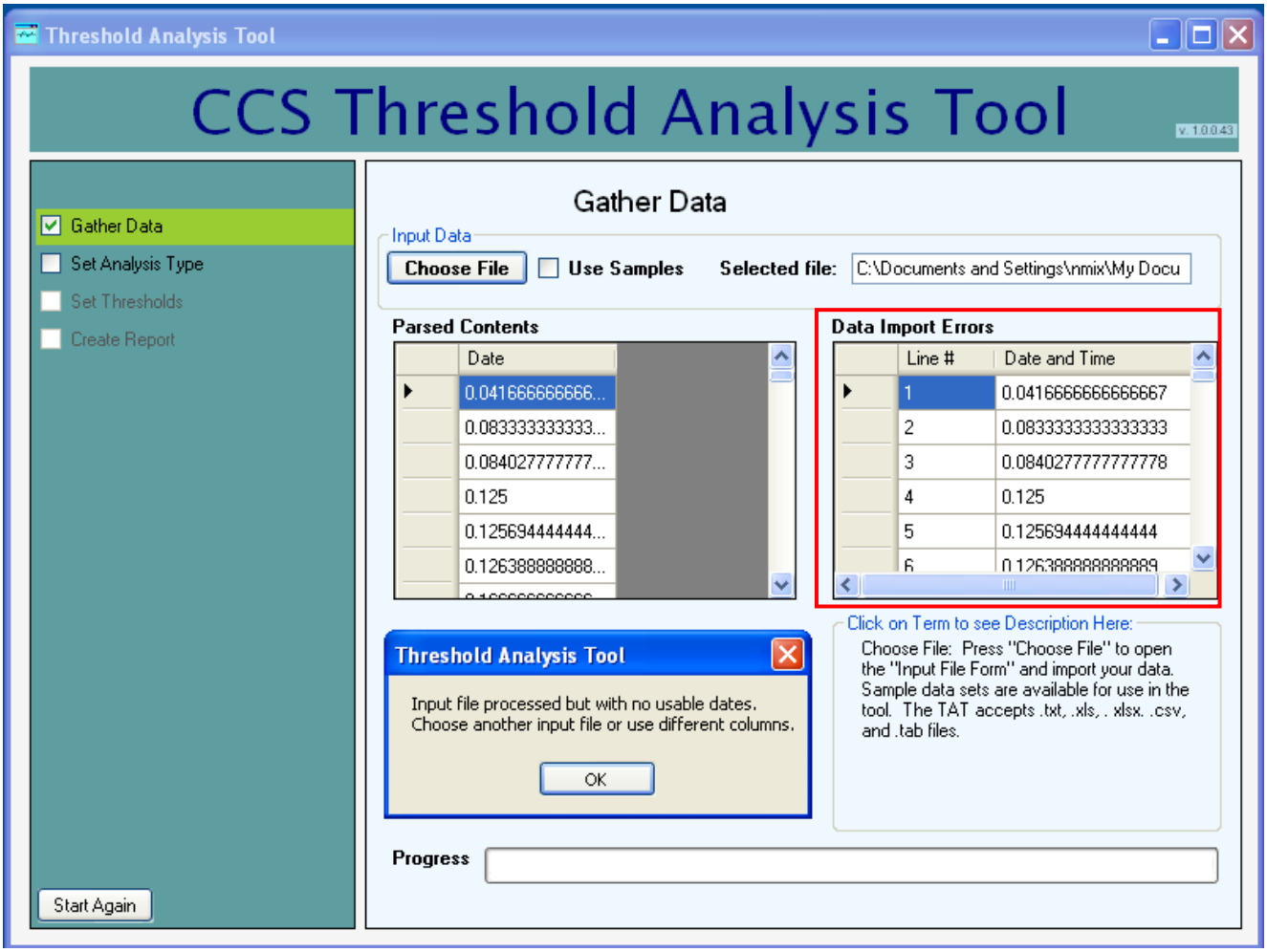

Figure 3.9. Data Import Errors Display

It is also important to note that the "Selected file" box on the main screen will display the specific file path for the data file that the user is analyzing (see **Figure 3.10**), which may be helpful for future reference when creating a report (see **Section 6**).

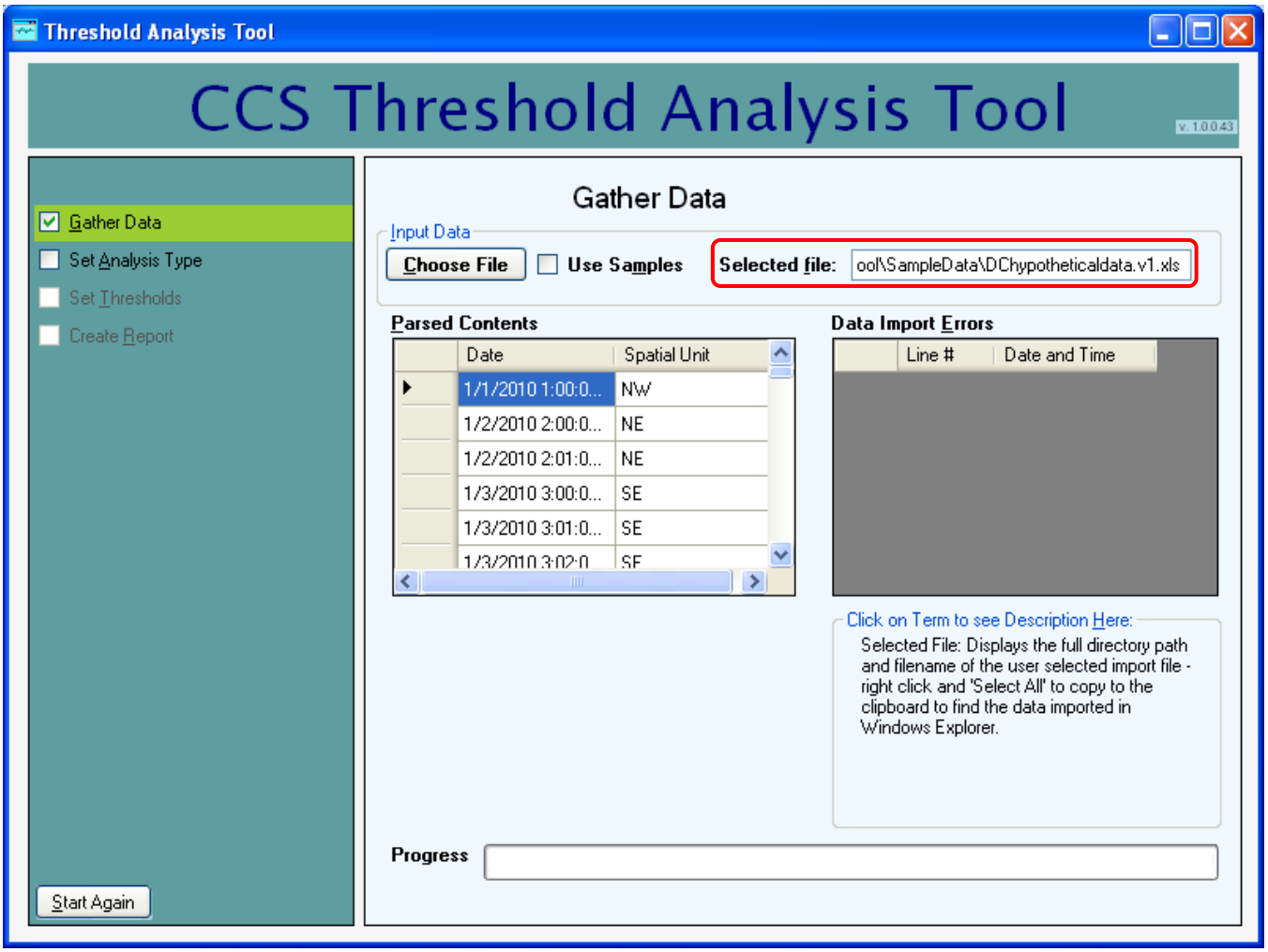

Figure 3.10. "Selected File" File Path

The data that will be analyzed by the tool can be viewed in the "Parsed Contents" window. In the example shown in **Figure 3.10**, the "Event Counts" column has been selected, and the tool has created individual records for each event for the day. The original "DChypotheticaldta.v1.xls" Excel file artificially had all events counts equal to one, so the parsed data equal the original data. However, if the original data only had one row of data with an event count equal to two on January 2nd (representing two customer complaints) the parsed data appears as 2 separate entries on January 2nd in the TAT. Once the user has ensured that all errors have been corrected, the user can proceed to "Set Analysis Type."

Section 4.0: Set Analysis Type

The choice of analysis type must be completed before proceeding to "Set Thresholds" and is discussed in the following sections. **Figure 4.1** shows the "Set Analysis Type" window.

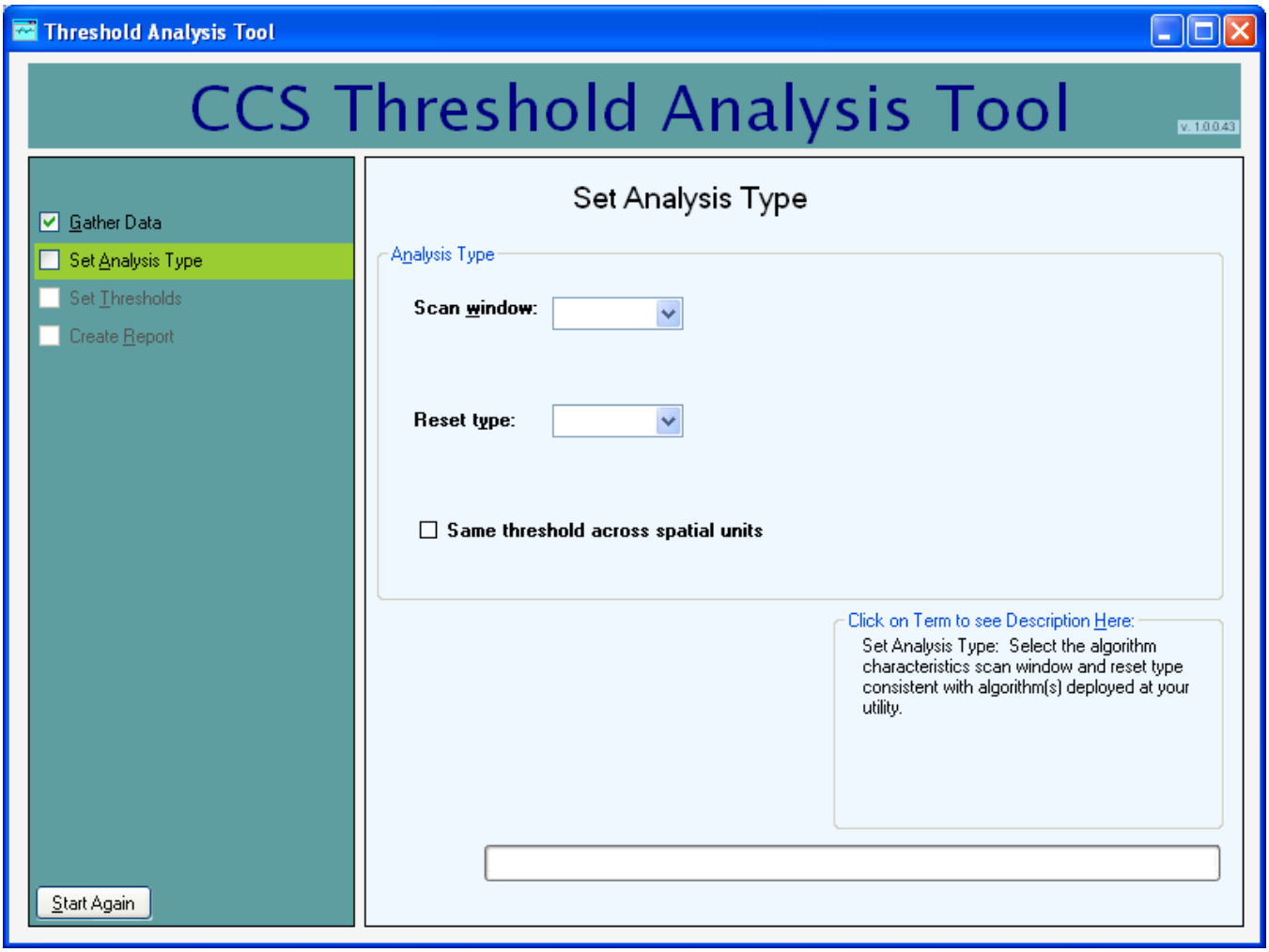

Figure 4.1. Set Analysis Type Window

4.1 Scan Window (Days)

The "Scan window" is the interval (number of days) in which all data are analyzed. The window may be set at 1, 2, 3, 4, 5, 6 or 7 days (see **Figure 4.2**), and should correspond with the algorithm duration (in days) that the utility uses or will use for the event detection system. This selection is required.

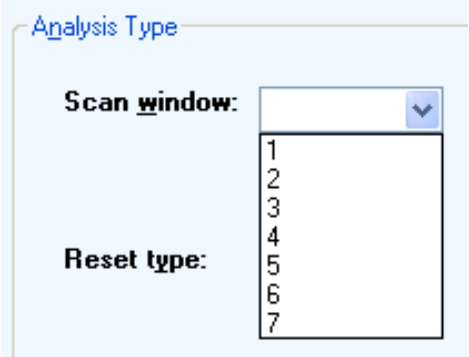

Figure 4.2. Scan Window Drop-down Menu

4.1.1 Reset Type

The "Reset type" selection determines how the data are analyzed once a threshold has been reached. There are two algorithm choices: "Reset" and "Continuous" (see **Figure 4.3**). The "Reset" algorithm option sets the total number of alerts back to zero once a threshold is reached. The "Continuous" algorithm option stops adding alerts until the number of alerts needed to exceed the threshold dips below the threshold and then surpasses the threshold again. The choice should correspond with the type of algorithm the utility uses or will use for tits event detection system.

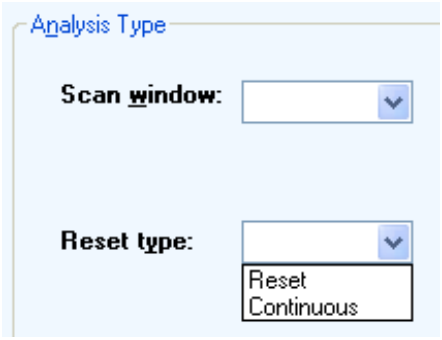

Figure 4.3. Reset Type Drop-down Menu

4.1.2 Same Threshold Across Spatial Units

The "Same threshold across spatial units" option (see **Figure 4.4**) is available only for data sets that contain multiple spatial units. Selection of this option identifies thresholds across a common percentile, standard deviation, or recurrence interval for all spatial units. If not selected, the user will analyze each spatial unit separately. Thresholds are discussed in detail in Section 5.

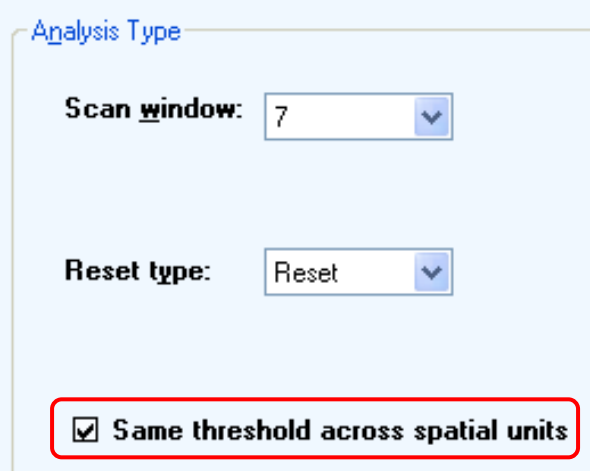

Figure 4.4. Same Threshold Across Spatial Units Option

Section 5.0: Setting Thresholds

There are three possible scenarios for use in setting thresholds: (1) no spatial units; (2) spatial units with different thresholds and (3) spatial units with the same thresholds.

These options are selected from the "Set Analysis Type" initial screen and are discussed in Sections 5.1, 5.2 and 5.3, respectively.

5.1 Setting Thresholds: No Spatial Units

This scenario means that the user did not specify a column to distinguish data by spatial unit. The "Same threshold across spatial units" selection is disabled, and the analysis is based on the interrelations of various threshold possibilities across the entire data set for the given "Scan window" and "Reset type" (see **Figure 5.1**).

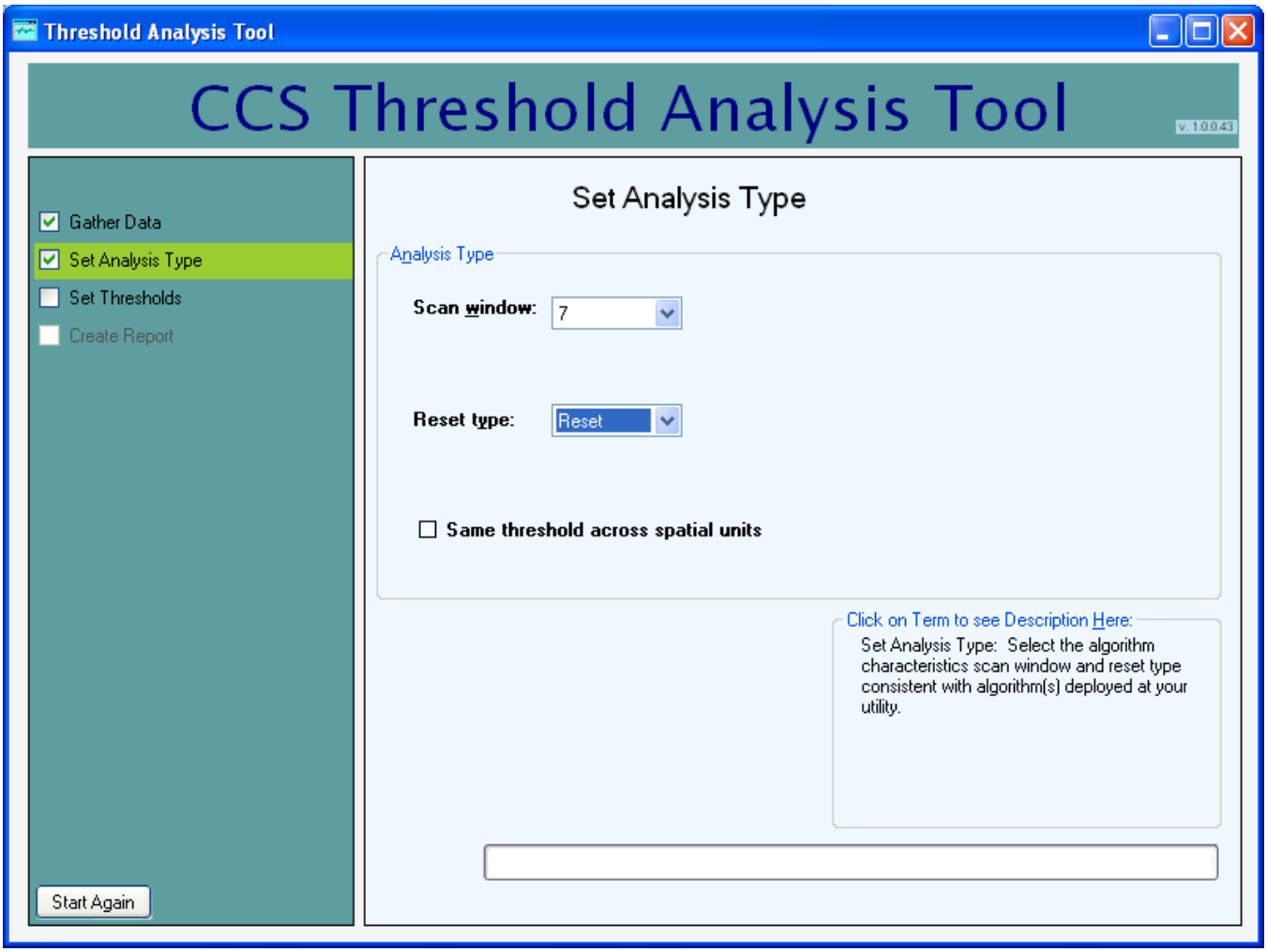

Figure 5.1. No Spatial Units, Main Screen

From the "Set Analysis Type" Screen, the user will then select "Set Thresholds" on the navigational panel to the left. Once this is selected, the user will see a screen which is devoted to providing consistent thresholds according to the following criteria (see **Figure 5.2**):

- Percentile,
- Standard deviation, and
- Recurrence interval.

The formulas underlying these three statistical criteria are presented and discussed in Appendix C.

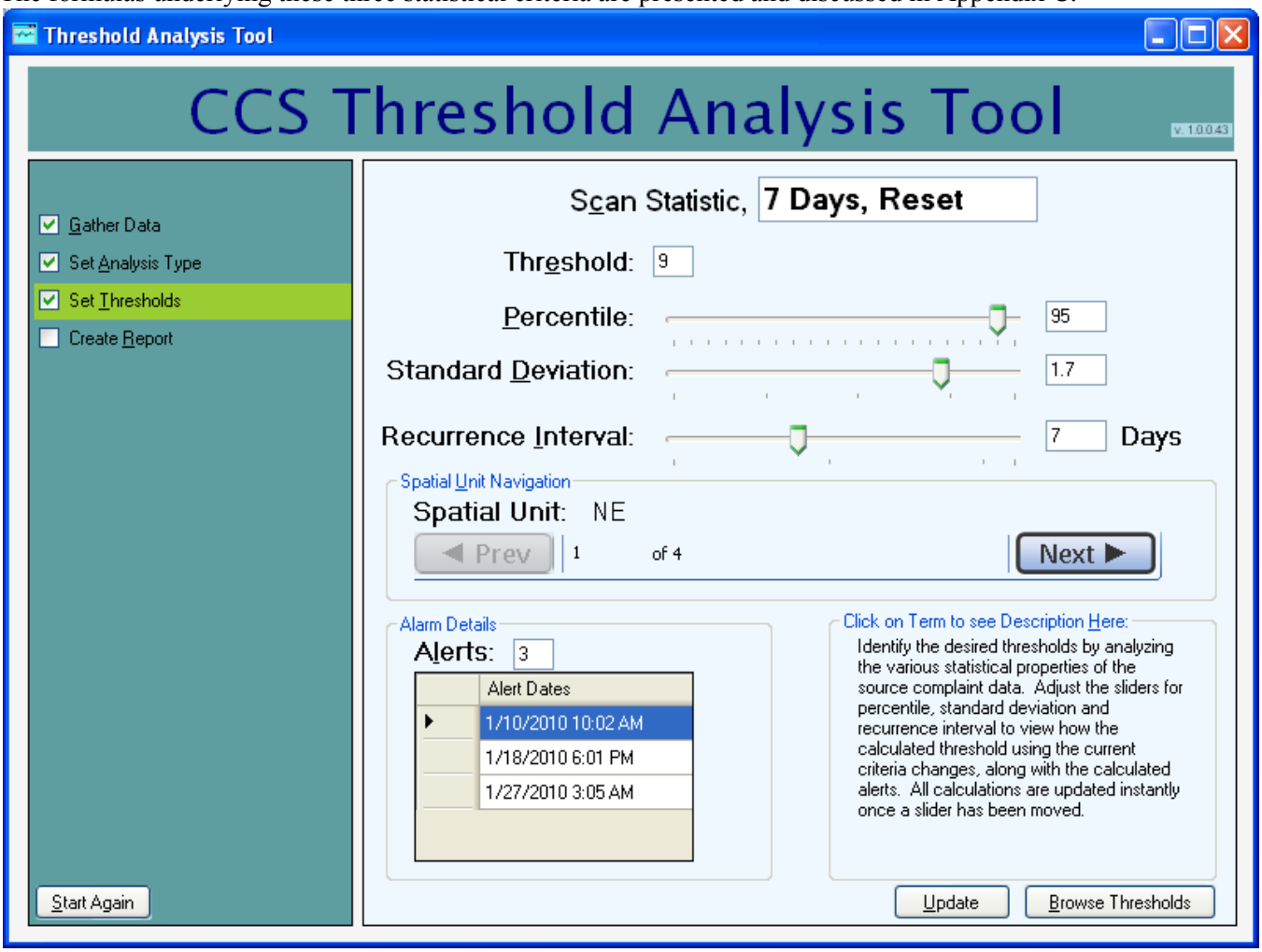

Figure 5.2. No Spatial Units, Threshold Screen

When the threshold screen first loads, the tool defaults to an initial threshold that is consistent with the 95th percentile. All values (for percentile, standard deviation and recurrence interval) can be specified for each threshold.

When the user slides the "Percentile" selector (or edit it using the edit box), the system finds the smallest threshold consistent with the slider and sets the values of the standard deviation and minimum recurrence interval based on that threshold. Similar calculations occur when the user selects "Standard Deviation" or "Recurrence Interval." As the slider is moved, the alerts generated are listed in the alarm details "Alerts" grid.

Once the user has specified the preferred percentile, standard deviation or minimum recurrence interval, the user may select "Create Report" on the navigational bar to export results.

5.2 Setting Thresholds: Spatial Units with Different Thresholds

In this scenario, the user has specified a column to distinguish spatial units, and the user did not check the "Same thresholds across spatial units" box in the "Set Analysis Type" window, as shown in **Figure 5.3**. If the user chooses not to set thresholds across various units, the user will have the opportunity to set the threshold for each individual spatial unit.

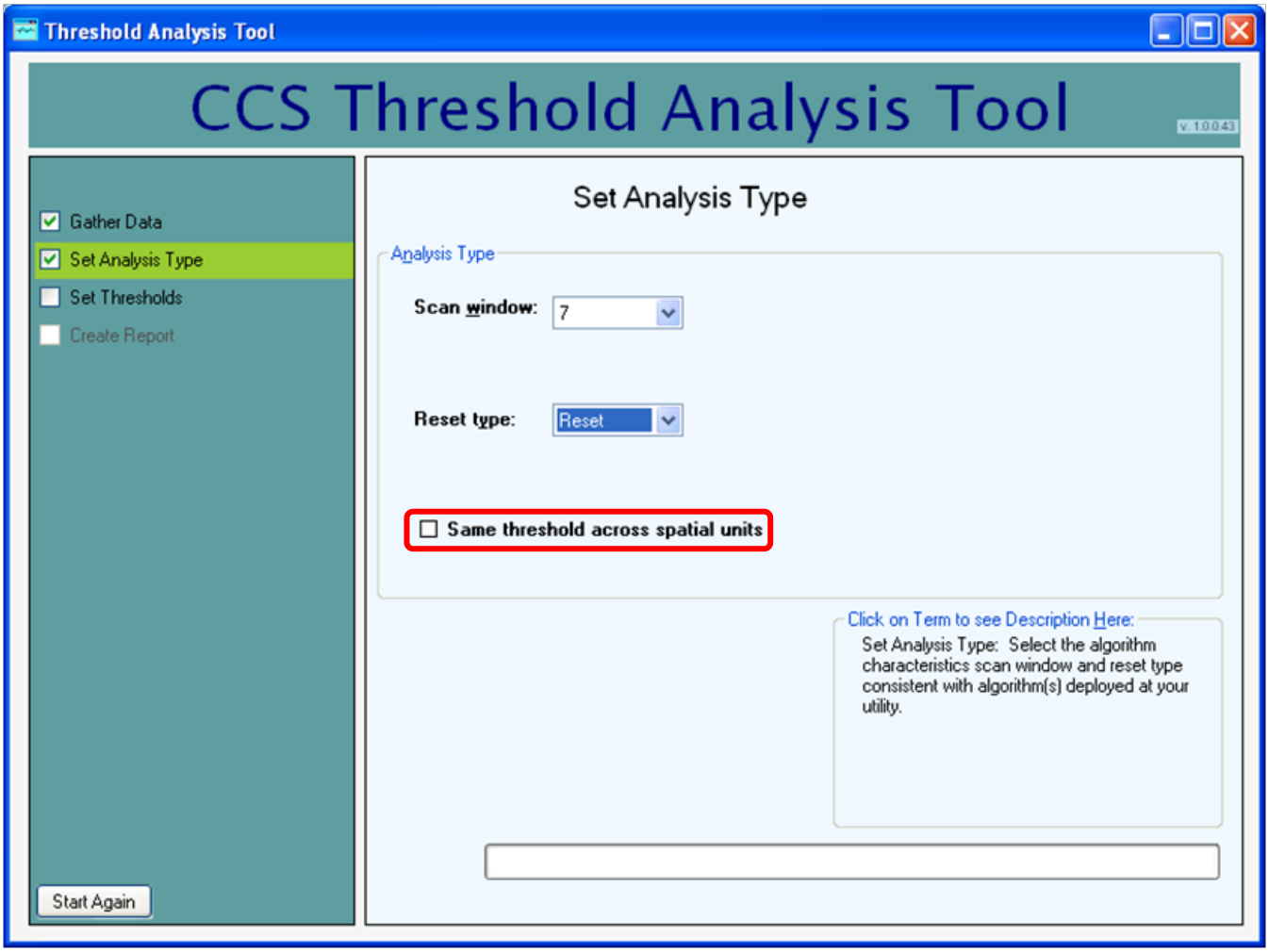

Figure 5.3. Spatial Units with Different Thresholds

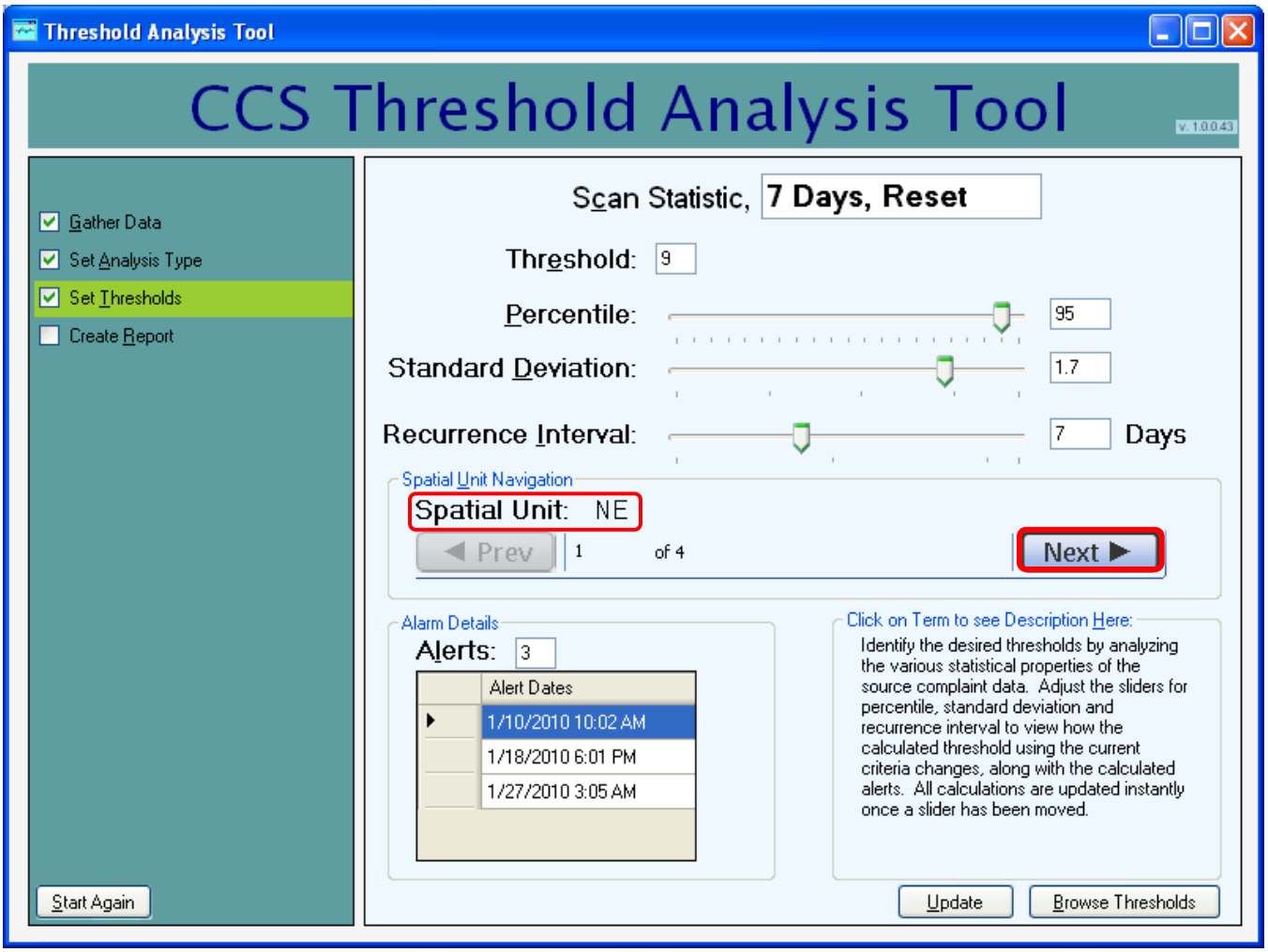

Figure 5.4. Spatial Units with Different Thresholds, Threshold Screen (Spatial Unit NE)

The user can set thresholds, as in the no-spatial-unit case, but these thresholds only apply to the events in the given spatial unit. The screenshot in **Figure 5.4** displays setting thresholds for the Northeast (NE) spatial unit.

The "Next" button allows the user to navigate to the next "Spatial Unit" (e.g., NW) which is depicted in **Figure 5.5**.

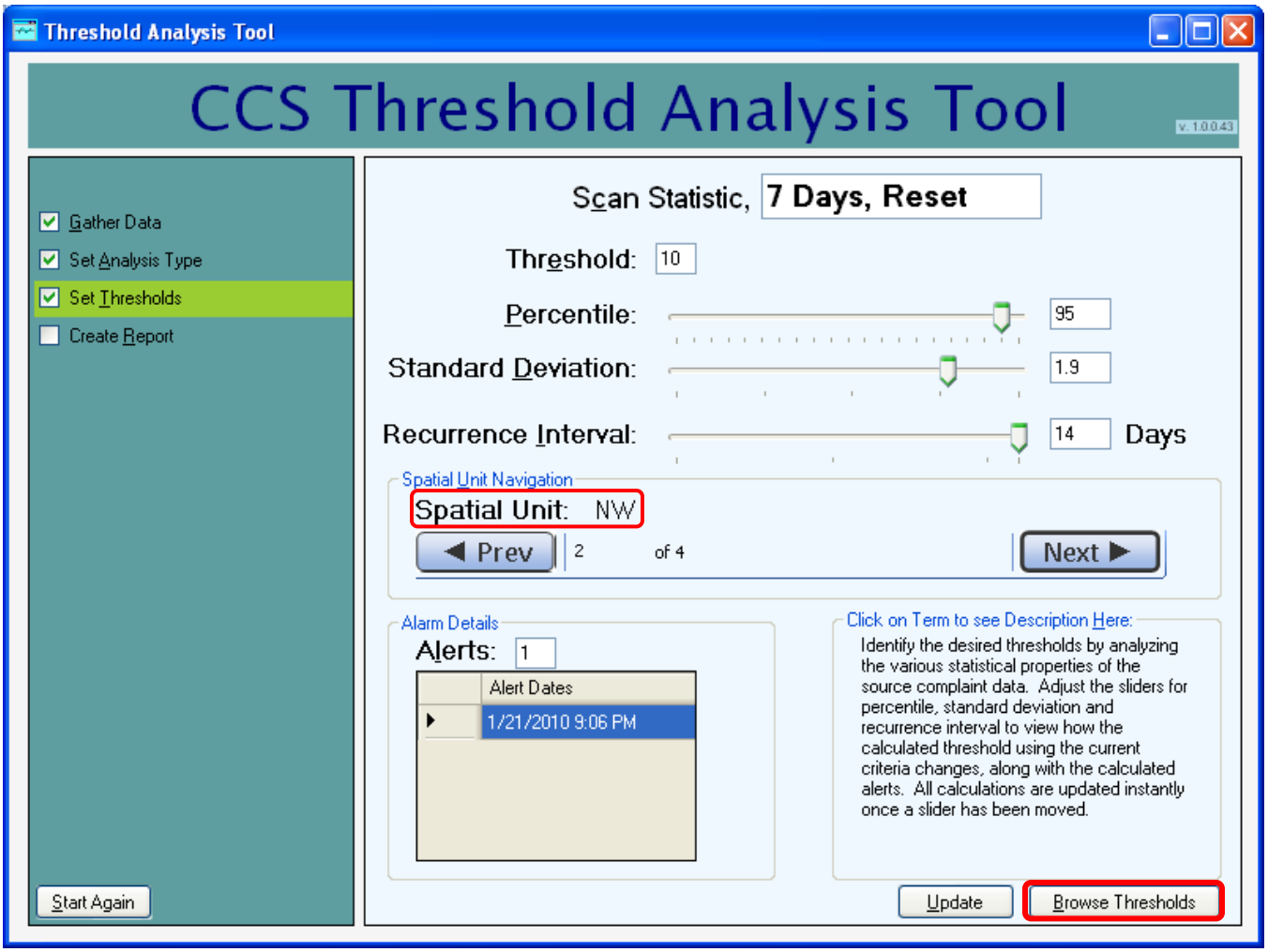

Figure 5.5. Spatial Units with Different Thresholds, Threshold Screen (Spatial Unit NW)

The user can set thresholds and continue to navigate using the "Next" and "Previous" buttons until the user has addressed all spatial units or decided that the 95% percentile default is adequate.

Click "Create Report" and specify a folder to export results. There will be one file generated for each spatial unit (in this example, one each for NE, NW, SE and SW).

The "Browse Thresholds" button allows the user to review all selections as displayed in **Figure 5.6**:

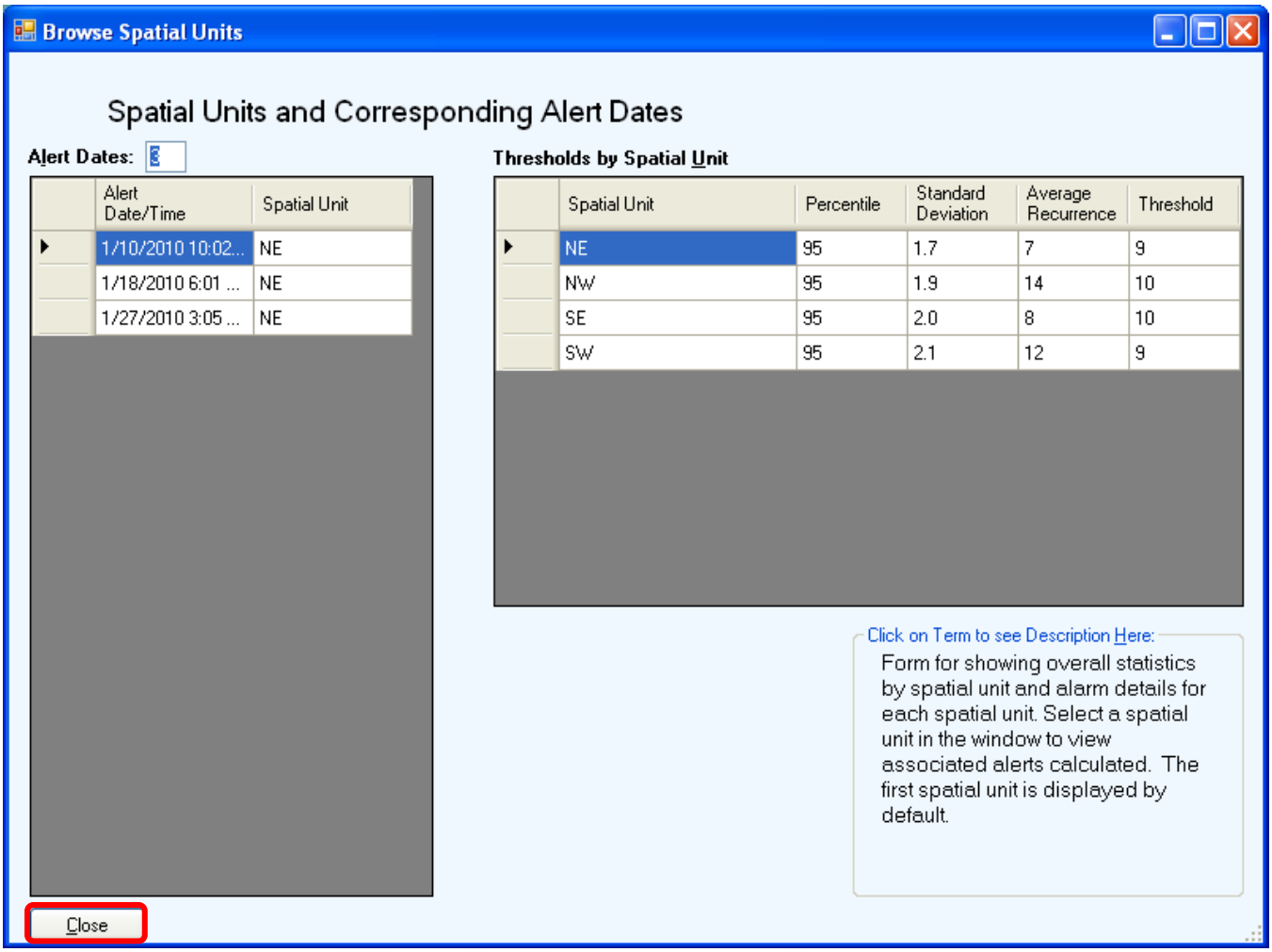

Figure 5.6. Browse Thresholds

The user can click in the "Thresholds by Spatial Unit" grid and navigate using the arrow keys or by clicking a spatial unit of interest. The "Alert Date/Time" column on the left will adjust to match that spatial unit. This is a dialog box, and the user must use the "Close" button at the bottom of the window or the "X" at the top of the window to return to the main form.

5.3 Setting Thresholds: Spatial Units with the Same Threshold

In this scenario the user chooses a spatial unit column, so the parsed column grid includes a "Spatial Unit" column, and the user checked the "Same thresholds across spatial units" box in the "Set Analysis Type" window (see **Figure 5.7**).

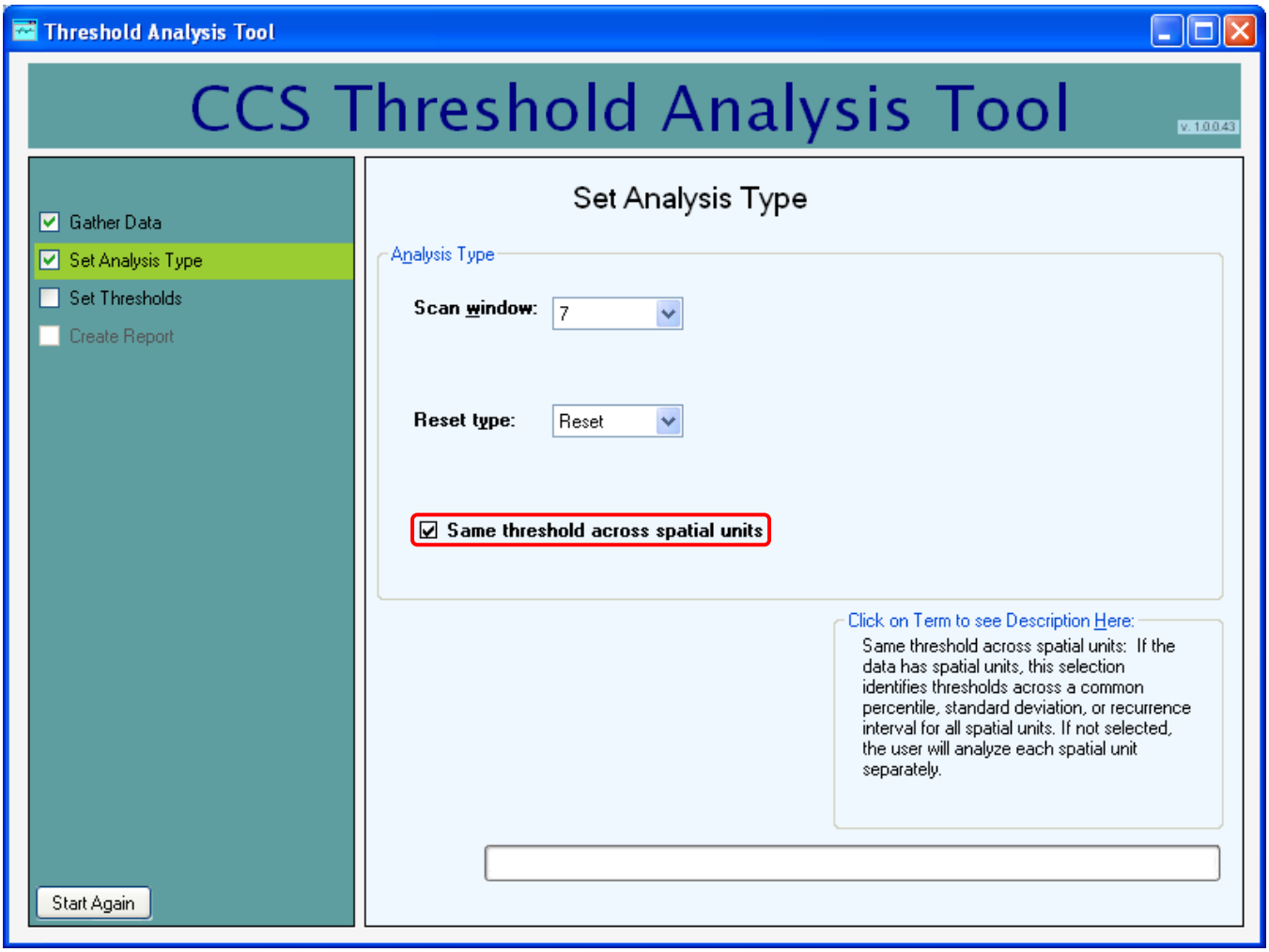

Figure 5.7. Spatial Units with the Same Thresholds

Upon checking "Set Thresholds" on the left of the screen, the user will see a screen for choosing thresholds across spatial/temporal units based on the same percentile, standard deviation or recurrence interval criteria (see **Figure 5.8**).

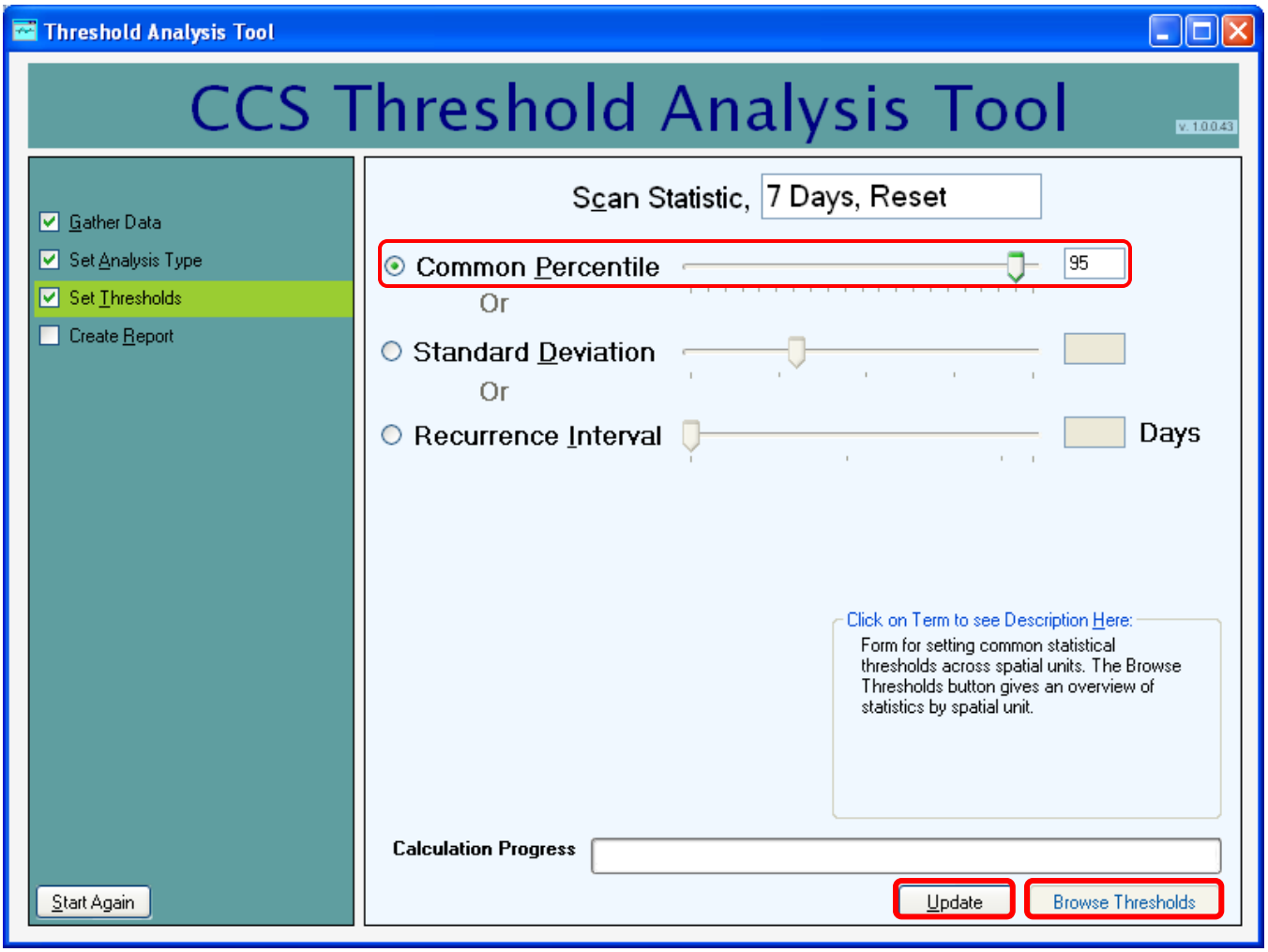

Figure 5.8. Spatial Units with the Same Threshold, Set Thresholds Screen

This screen allows the user to choose thresholds one criterion at a time and applies the choice across spatial units. For example, if the user were to choose a 95th percentile, click on "Update," click on "Browse Thresholds" (see **Figure 5.8**), and view the "Browse Spatial Units" screen (see **Figure 5.9**), the user would see that 95th percentiles are assigned to all spatial units. Further, the tool does some default processing in cases where there is no possible match. For example, if there is only one event for one of the spatial units, the only possible standard deviation is 0. Finally, the tool will set the standard deviation or recurrence interval based on the maximum possible threshold for the given spatial unit that will produce one alert.

Note: When switching between the common percentile, standard deviation and recurrence interval threshold options, the calculations are not updated until the selection slider has been moved or a value is entered in the text box and the "Update" button is selected.

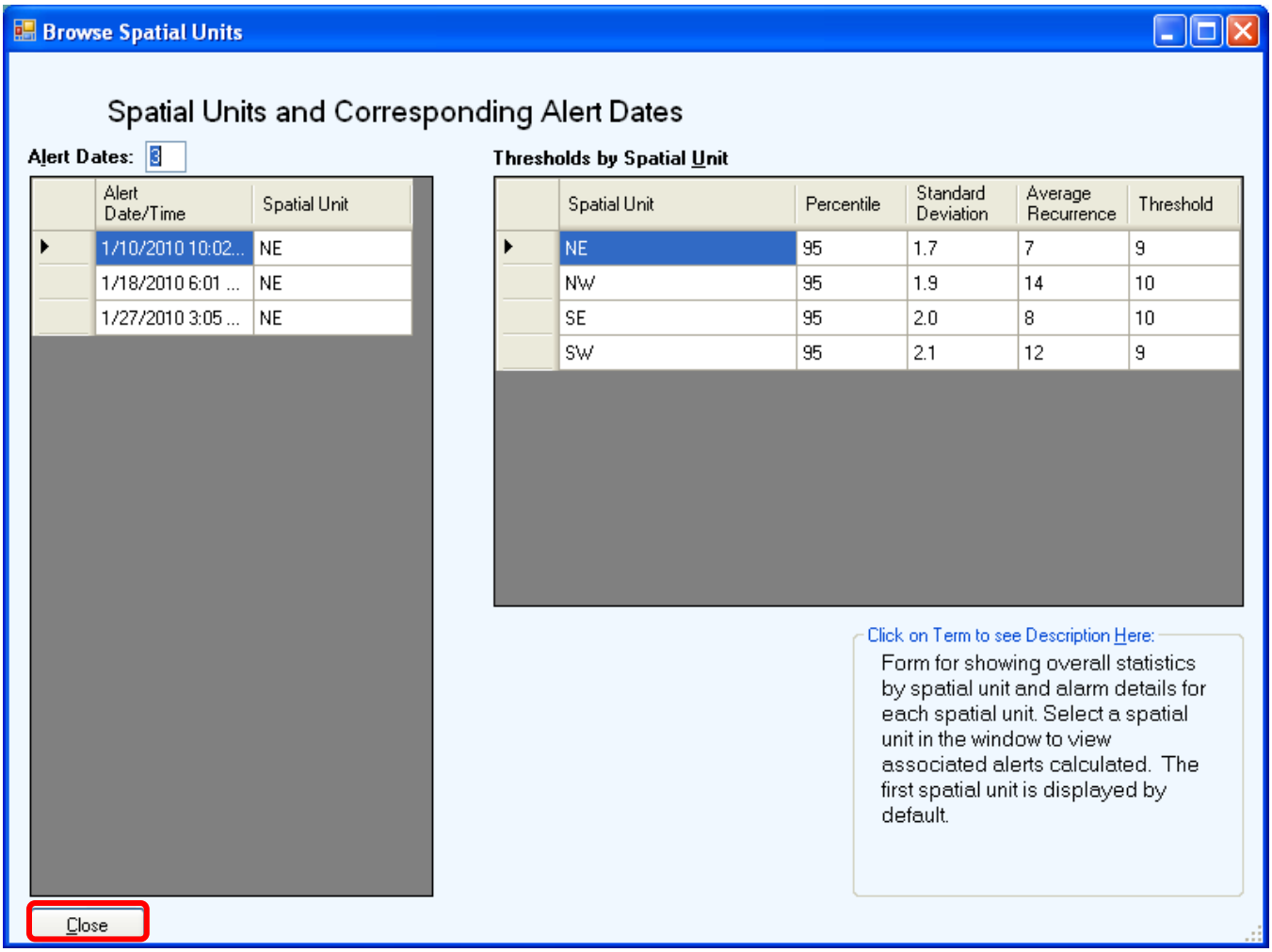

Figure 5.9. Spatial Units with the Same Thresholds, Browse Spatial Units Screen

The same percentile (95^{th}) has been assigned to all spatial units in **Figure 5.9**, but the corresponding standard deviation, average recurrence and thresholds vary. This is normal and is to be expected, as percentiles differ statistically from standard deviation and recurrence. Once the analysis has been completed, the user may close the "Browse Spatial Units" window and export the results by clicking on the "Create Report" option in the left side of the main menu.

Section 6.0: Creating a Report

Once the user has determined the threshold, the first step in creating a report is to select the "Save Files(s) As…" button in the "Threshold Report Output" window (see **Figure 6.1**).

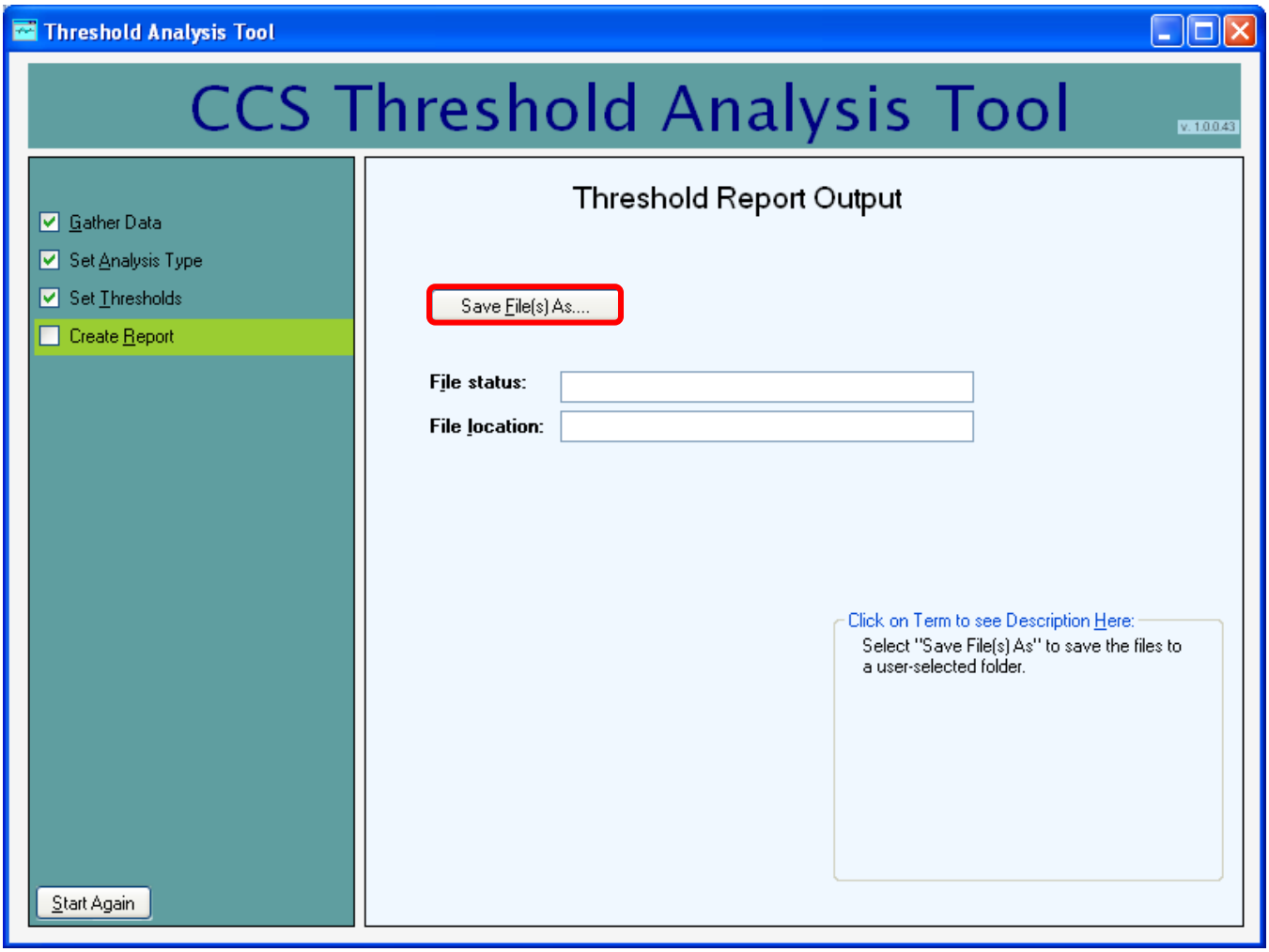

Figure 6.1. Creating a Report: Saving File

This will open a dialog box (see **Figure 6.2**) that will allow the user to browse folders for a location to save a file.

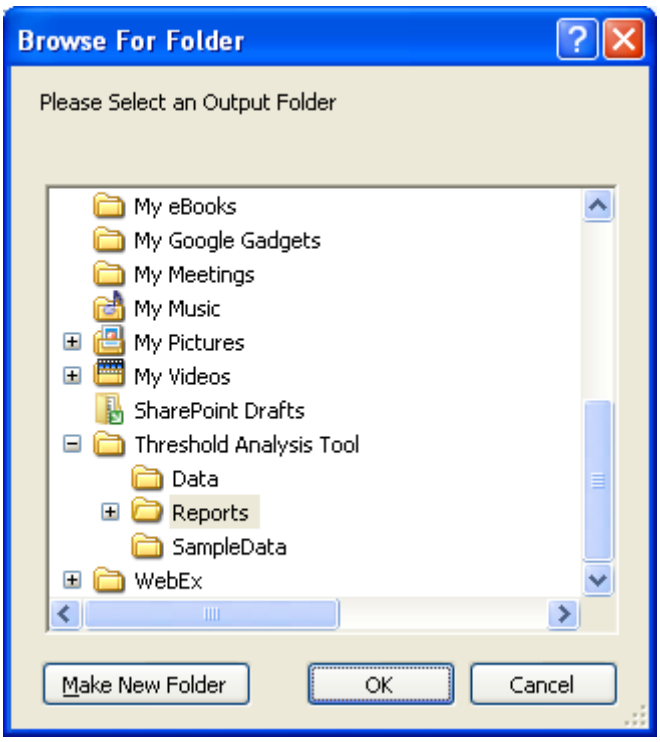

Figure 6.2. Creating a Report: Selecting an Output Folder

Once the user has chosen an output folder, and clicks the "OK" button, the tool will begin the process of creating a report file. When the tool finishes creating a report, a message box (**Figure 6.3**) will inform the user that the process has been completed. The "File Status" text box in the "Threshold Report Output" window now displays the number of report files created. The "File Location" text box provides the location of the output folder to allow the user to copy the location for future use. Once saved, files can be viewed and printed in Excel or Notepad.

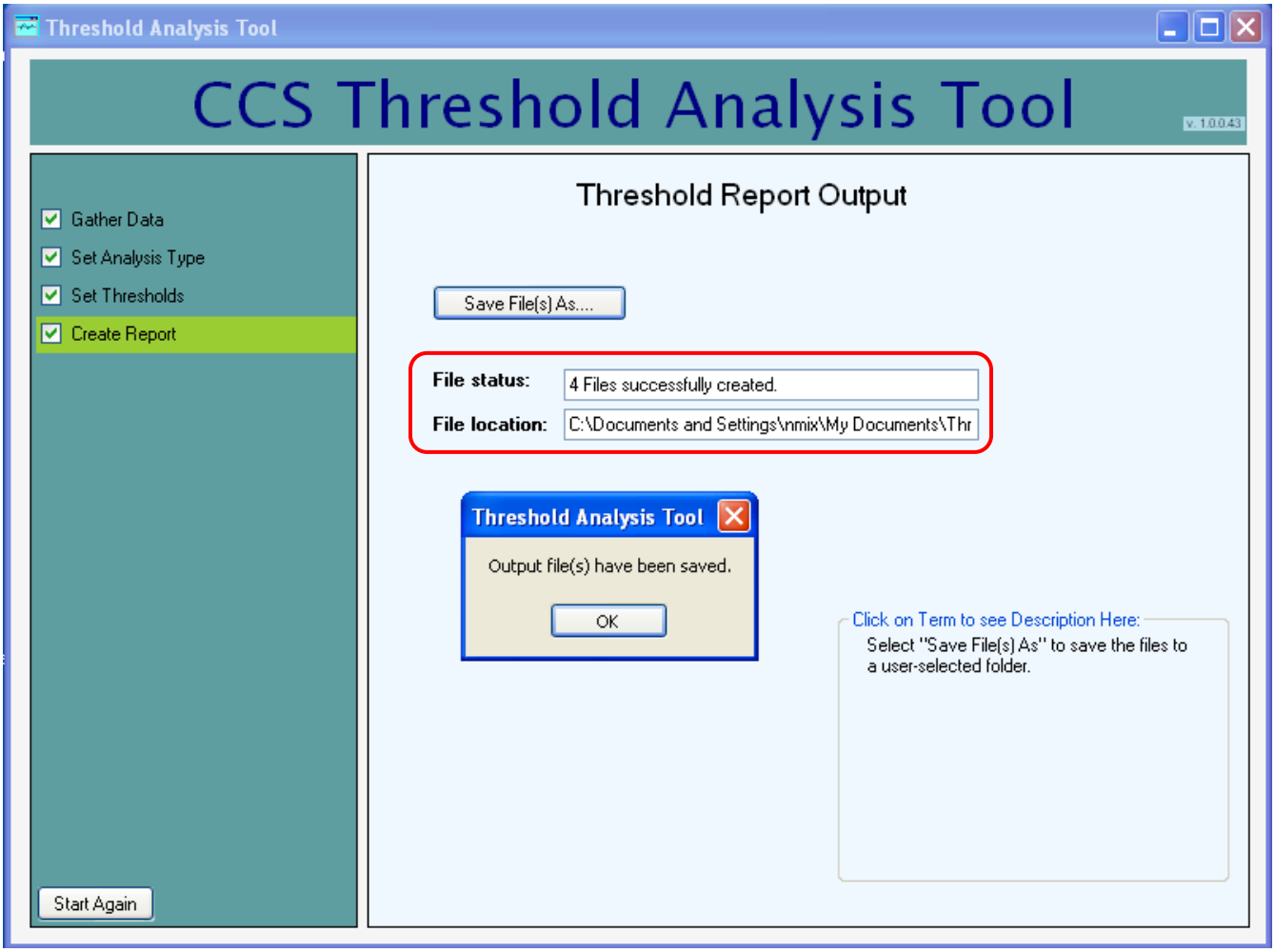

Figure 6.3. Creating a Report: Confirmation Message and File Location

If the user has data without corresponding spatial/temporal units, or has data for only one spatial/temporal unit, the tool will save the report in a file called "TAT Results.{algorithm type}- {scan window}- {date column}- {time column}- {spatial unit column}-{count column}.csv" (e.g. TAT Results.Reset-7-1-2-3-6.csv).

If data corresponds to multiple spatial/temporal units, the tool will create a separate file for each spatial/temporal unit, using the following file naming format: "TAT Results.Spatial Unit [spatial unit]{algorithm type}- {scan window}- {date column}- {time column}- {spatial unit column}-{count column}.csv" (see **Figure 6.4**).

Currently there are no options for customizing the naming convention when saving these files, or for establishing a naming scheme for created reports. The user can rename the files at later time (within Windows Explorer), or make new output folders by clicking "Make New Folder" on the "Browse for Folder" Screen (**Figure 6.2**)

Figure 6.5 shows the data contained in an example report generated by the tool.

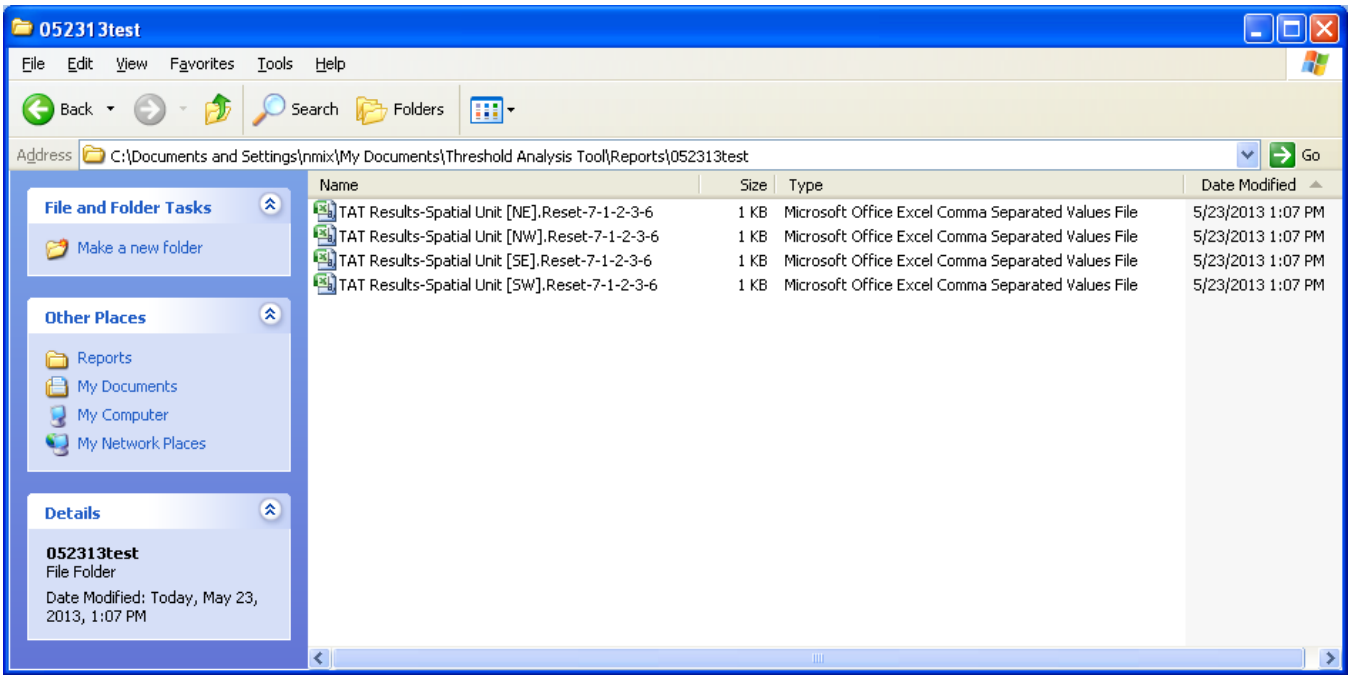

Figure 6.4. Examples of Saved Reports with Spatial Units

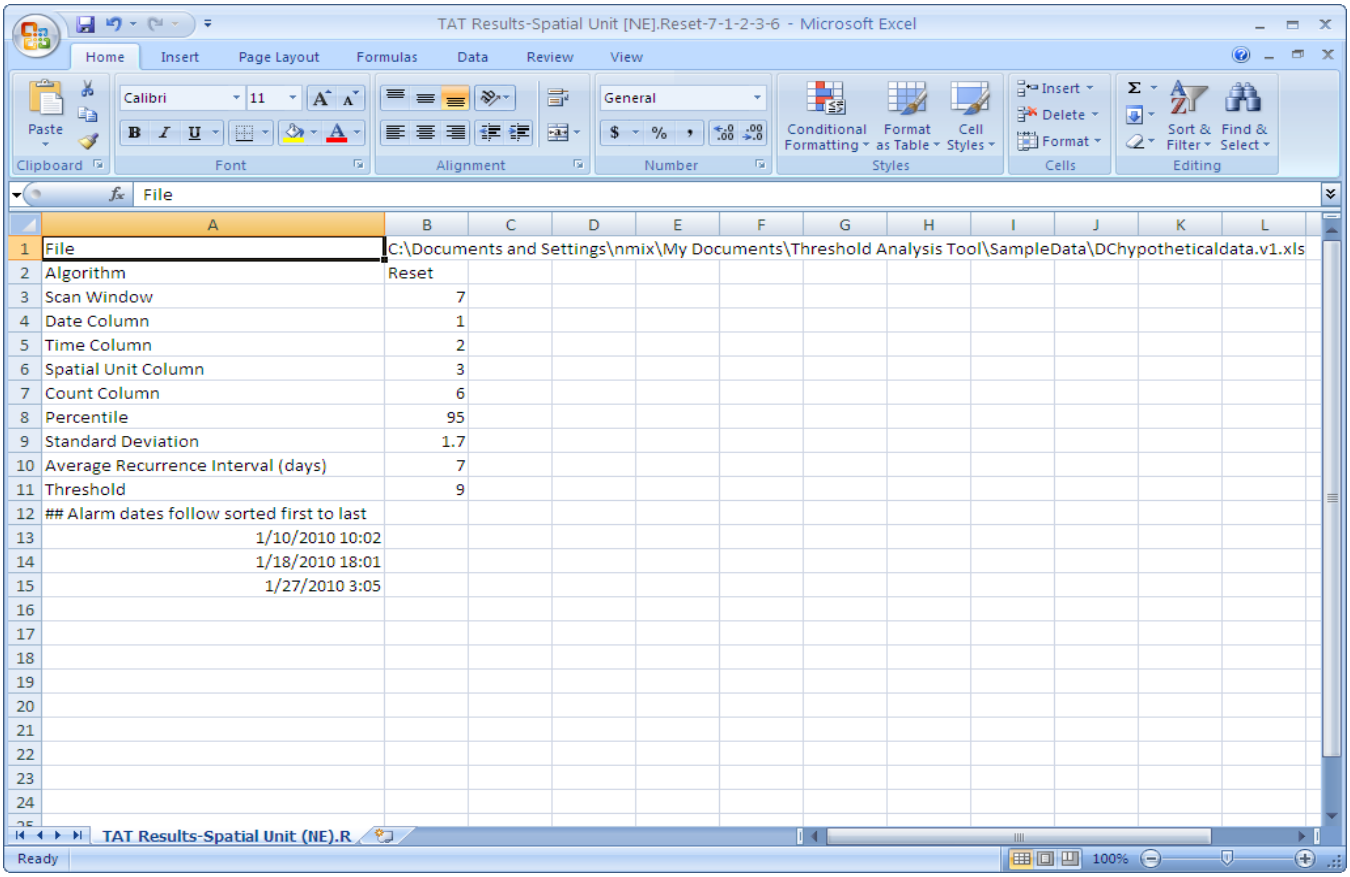

Figure 6.5. Example Report Data

Appendix A: Glossary of Terms

- **Contamination incident:** in the context of water security, a contamination incident is defined as the intentional or unintentional introduction of a chemical, radiochemical, pathogen or biotoxin into the drinking water supply which may result in negative public health consequences and/or infrastructure damage to a utility's drinking water distribution system.
- **Surveillance and Response System:** an early warning system which uses multiple monitoring and surveillance components for timely detection of and effective response to drinking water contamination in the distribution system.
- **Continuous:** the "Continuous" algorithm option after each Alert will stop issuing any new "Alert" until the number of customer complaints in the scan window dips below the threshold. Then the number of customer complaints in the scan window will be reset to zero and the whole operation will be repeated.
- **Customer Complaint Surveillance:** in the context of a drinking water SRS, a CCS component enhances and automates the collection and analysis of calls by customers reporting unusual water quality concerns and compares trends against an established baseline to detect possible contamination incidents.
- **Event(s):** customer complaints (e.g., phone calls, work orders) received and documented by the drinking water utility.
- **File Delimiter:** a delimiter is a character that marks a boundary between data items. For example, a comma –delimited file places commas between each complaint date-time. This is sometimes synonymous with the file type. If the system does not automatically select the correct delimiter, the user can change the delimiter with this option. The File Delimiter choices are displayed on a drop-down menu and are limited to comma, tab, single line and Excel.
- **Percentile:** the value on a scale of one to one hundred that indicates the percent of a distribution that is equal to or below the value.
- **Recurrence interval:** the average time between alerts. The TAT uses the average recurrence interval.
- **Reset:** the "Reset" algorithm option after each Alert will reset the number of customer complaints in the scan window to zero and begin the operation again.
- **Scan Statistic Variables:** include the number of days "Scan Window" (in Days), and the algorithm "Reset Type" (choose either "Continuous" or "Reset").
- **Scan Window:** the interval, in number of days, from which all the data is analyzed. The Scan Window may be set at 1, 2, 3, 4, 5, 6 or 7 days. The Scan Window should correspond with the type of algorithm the utility uses or will use for the event detection system.
- **Spatial/Temporal Unit:** a unit designated by the utility. Examples of 'spatial units' may include the names of service areas, pressure zones, treatment sources, zip codes, or some other community name. Examples of 'temporal units' may be the names of the day of the week, month or some other descriptive text to distinguish weekday calls from weekend calls.
- **Standard deviation:** a measure of dispersion, or how much the individual measurements vary across all of the data. The TAT displays the deviation above and below the mean in positive and negative values.
- **Threshold:** the number of customer complaints (e.g., calls received by the utility) within the scan window preset by the utility. When the number of customer complaints is equal to or greater than the "**threshold**," an alert is issued.

- **Threshold Analysis Tool (TAT):** a tool developed by EPA for water utilities which retrospectively applies a configurable scan algorithm to call and work management data input by the user and outputs the date and times when an alert would have been issued.
- **.csv:** A .csv, or comma-separated values file format is a set of file formats used to store tabular data in which numbers and text are stored in a plain-text form that can be easily written and read in a text editor. Because the goal of reading and writing the format take precedence over consistency, there is no .csv standard, only the understanding that plain text is delimited by a symbol. Traditionally, lines in the text file represent rows in a table, and commas separate the columns.
- **.tab:** A .tab, or tab-separated values, file is a simple text format for a database table. Each record in the table is one line of the text file and each field value of a record is separated from the next by a tab stop character.
- **.txt:** A .txt, or text, file is a type of computer file that is structured as a sequence of lines. A text file exists within a computer file system. The end of a text file is often denoted by placing one or more special characters, known as an end-of-file marker, after the last line in a text file.
- **.xls, .xlsx:** Excel files are those which are produced by Microsoft Excel, a commercial spreadsheet application written and distributed by Microsoft for Microsoft Windows and Mac OS X. .xls files are produced by older versions of Excel, whereas .xlsx files are produced by Microsoft Excel 2007 and newer versions.

Appendix B: Frequently Asked Questions

Do I need to "clean" my data before importing it into the tool?

Depending on the condition of the data exported from a customer complaint management system, the user may have to "clean" data. The tool accepts a variety of formats including .txt, .tab, .xls, .xlsx and .csv. Straight data imports or downloads from a customer complaint system can be filtered through the data selection process to select only those data necessary for analysis. However, the filtering criteria may not be sufficient to recreate the conditions of the event detection system. For example, if the data set contains both rusty water complaints and taste complaints, but the event detection system only analyzes taste complaints, the user will have to filter the rusty water complaints from the dataset prior to importing it into the tool.

Ultimately, the goal is to analyze the same data that are used (or will be used) by the event detection system in the same way the event detection system analyzes data: the same complaints, the same date-time combination, the same spatial units (if any), the same event counts (if any), the same scan window and the same algorithm type. The selections in the tool should recreate the event detection system for historical data analysis.

Which statistical method is better to use when setting thresholds?

Percentile, standard deviation and recurrence interval methods are related, with neither being preferred over the other for use in establishing thresholds. The choice of method is dependent on the user's comfort with the analysis, and criteria valued.

R*ecurrence interval* refers to the average time between alerts, or how often alerts occur. This method is useful for establishing thresholds based on the level of effort required to investigate each alert. However, this does not have statistical grounding in the unusualness of the complaint volume. Users should be cautious not to set the threshold at too large a recurrence interval, as the utility may miss an actual water contamination incident.

Percentiles are based on the distribution of the number of alerts received at each threshold. For example, utilities may set their threshold to the 95th percentile, meaning that only 5% of higher thresholds generate fewer alerts. This statistical analysis ensures that the utility does not set thresholds beyond a reasonable expectation of complaints.

Utilities may also use *standard deviation (from the mean)* to establish thresholds. Standard deviation is a measure of dispersion, or how much individual measurements vary across all of the data. This method provides a statistical basis for establishing a threshold, beyond the number of complaints received within the time period. However, the utility should be cognizant not to set thresholds at a number beyond the reasonable expectation of the number of complaints.

Why do I get different alerts for the same algorithm, if I use "Dates" compared to "Dates" and "Times"?

Because the tool applies the same algorithm for analysis of all data sets, a consistent date-time format must be used. When a user provides dates and does not provide corresponding times for each complaint, the tool assigns a default time of 12:00 AM to each complaint.

For example, if there are five complaints on 1/2/12, the tool will read each of those complaints as having a date-time of 1/2/12 12:00 AM. When the 1-day reset algorithm checks whether the complaint volume has crossed the threshold at 1/2/12 12:00 AM, it sees five complaints but issues only one alert. This is the case even if the

threshold is set at one, since it has only executed once. In other words, the complaints occur in the same analysis cycle.

If the complaints have varying "Times" throughout the same "Date," and these "Times" are included with the "Dates," an alert will be issued for each complaint at a threshold of one, as the complaints would occur at different analysis cycles. For example, consider a data set with 31 "Dates," and 115 customer complaints occurring during those 31 "Dates." When applying a 1-day reset algorithm using just "Dates," there will be 31 alerts; when applying the same 1-day reset algorithm using both "Dates" and "Times," there could be up to 115 alerts, depending on the times that the complaints occurred. Thus, data sets that include both dates and times allow for more accurate thresholds to be calculated than data sets with dates and the default time of 12:00 AM.

Can I still use the TAT if I only have "Dates"?

Yes. In this case, the tool will assign a time value equal to 12:00 AM to all records.

Will I get different numbers for the same algorithm if I use "Dates" compared to "Dates" and "Event Counts"?

Yes, unless each day has exactly one complaint in the "Event Counts" column. If the "Event Counts" value for any entry is greater than one, then the number of alerts will differ. If only "Dates" are used, each date is treated as a single unique complaint. If "Event Counts" are used for the analysis, the tool will analyze all "Event Counts" records.

Why do I get different values when I run the same data set with "Event Counts" and without "Event counts"?

The tool is designed to process all data with a date and a time regardless of whether an "Event Count" exists. If there is no time, it assigns a value of 12:00 AM. If there is an "Event Count" and a corresponding time the tool uses the time provided. Because of this, if there are multiple entries for a single date, each with a different time on that date, the number of alerts may be different with the time selected.

"Event Counts" are not intended to be used if the user has an actual time value associated with each date value. "Event Counts" are intended to be used in instances where the user has data that does not include any times but the user knows that there were multiple complaints on that given date.

Note: If the user alters the data so the first time is used for the "Event Count," the results may not be indicative of the data (i.e., assigning four records a time of 11:15 AM versus assigning each a different time).

Running the same data using "Dates" and "Times," I get a different answer if I add "Event Count" as a User option. Why?

The tool defaults the time value in the column to 12:00 AM (i.e. "0") for the quantity identified in the "Event Count" column. If the user has four event counts for a given date, it means that the tool has defaulted the time to 12:00 AM, and created four separate records for that date. Thus, the data set the tool is analyzing is different than the data set input into the tool (the times are different, and more records have been created for the "Event Counts").

"Event Counts" are not intended to be used if the user has an actual "Time" values associated with each "Dates" value. "Event Counts" are intended to be used in instances where the user has data that does not include any "Times", but knows that there were multiple complaints on that given "Date".

For each algorithm, does the TAT produce the same alert results as EPA's Alarm Estimation Tool (AET)?

EPA's AET is a spreadsheet that, like the TAT, was designed to assist drinking water utilities with establishing alert thresholds based on user-supplied data and user preferences. The AET uses a simple tabular and graphical output display, and is limited to recurrence interval analysis. The TAT was designed to build on the AET by offering additional statistical analyses (percentile and standard deviation) and a more refined user interface.

The TAT will produce the same results as the AET, but only if date and time are used and the data are identical. The AET can be accessed at: http://water.epa.gov/infrastructure/watersecurity/techtools/index.cfm.

What can be said about the range of thresholds when "Spatial Units" are selected?

When there is only one spatial unit identified, the results will be identical whether "Same thresholds across spatial units" is check or not. This is because the tool will treat all the data as if the data are in one spatial unit if there are no other "Spatial Units" in the data set selected by the user. When there are two or more spatial units identified, the tool will calculate thresholds for each of the spatial units, regardless of whether the data for each spatial unit are the same or different.

For any algorithm, does the TAT produce the same number of alerts if the "Same threshold across spatial units" box is checked?

No, unless the threshold is one. When "Spatial Units" are selected, the tool parses the data into separate data sets for each spatial unit before running the analysis. For example, imagine three complaints occur in three different spatial units within an hour with an alert threshold of three. When there are no "Spatial Units" there will be an alert, but not when there are "Spatial Units" and "Same threshold across spatial units" is selected. When the "Same threshold across spatial units" is selected, the tool will initially calculate different thresholds for each spatial unit, based on the 95th percentile. It's a good idea to be familiar with the data set before it is input into the tool and analyzed, to have a better understanding of the results.

What are the main things to remember about "Reset" and "Continuous"?

The "Reset" algorithm option sets the total number of alerts to zero when the threshold is reached and begins counting again when new complaints are added. The TAT will look back to previous data records within the scan window, add any previous complaints to new complaints, and compare the value to the threshold, alert and reset.

The "Continuous" algorithm option stops adding alerts until the number of alerts needed to exceed the threshold dips below the threshold and then surpasses the threshold again. The tool will look back to previous data records within the scan window, add any previous complaints to new complaints, and compare the value to the threshold and alert. It stays in an alert mode until the number of complaints within the scan window is less than the threshold, and then alerts again once additional complaints exceed the threshold. So, the TAT is in a "Continuous" alert condition.

As noted in Section 3.1.4, inclusion of the event time is not required for analysis, because the tool will automatically use a default time of 12:00 AM if an event time is not provided. However, results from data sets that default to 12:00 AM may be less precise than results from data sets that include specific event times. For example, when there is no event time specified (such as event count data), the tool may produce additional alerts when using the "Continuous" algorithm setting. While the tool will still produce the correct alerts, there may be additional alerts issued when the count is above the threshold.

Appendix C: Formulas in the Threshold Analysis Tool

The Threshold Analysis Tool is intended for use by water utility management interested in managing the number of alerts requiring a response and the ability to detect possible water contamination. For this reason, the tool calculations (percentile, standard deviation, and recurrence interval) incorporate a utility-centric approach, targeting the frequency of alerts at different thresholds. It does not focus on the number of complaints, or number of days without a complaint.

Percentile

$$\text{Percentile} = 100 * \frac{Rank\ (k)}{a}$$

The percentile is a value on a scale of one to one hundred that indicates the percent of a distribution that is equal to or below the value. In the equation above, *k* is the threshold and *a* is the highest threshold that generates an alert. The tool sorts the thresholds by the number of alerts generated at that threshold. $Rank\ (k)$ is the rank (position) of the threshold in that order list. While the lowest minimum percentile can be different, given the underlying data, the slider for selecting the percentile in the tool, shows the percentile as a scale from 0 – 100. This user interface choice was incorporated to provide the user with a consistent experience regardless of the underlying data. For percentile values lower than the minimum percentile at the lowest possible threshold (1), the other sliders adjust to the respective values associated with the minimum threshold.

Standard Deviation (from the mean)

$$\mu = \frac{\sum(xf)}{n} \qquad S = \sqrt{\frac{\left(\sum x^2 f - \frac{\sum(xf)^2}{n}\right)}{n}}$$

Standard deviation is a measure of dispersion, or how much individual measurements vary across all of the data. The tool displays the deviation above and below the mean as positive and negative values. In the equation above, *x* is the threshold, *f* is the number of alerts at the threshold, and *n* is the total number of alerts at all thresholds. "*u*" is the mean, and "S" is the standard deviation (from the mean). The tool reports the standard deviation from the mean, also known as the z-score or standard score. For values below the threshold, the value displayed will be negative, indicating the number of standard deviations below the mean for the threshold. While a z-score is continuous, thresholds are whole numbers. Consequently, the tool reports the minimum standard deviation from the mean at each threshold on the slider. This results in the user interface slider on the "Set Thresholds" Screen "jumping" as it is moves from one threshold to another as other interface sliders are moved.

While the label associated with the Standard Deviation selection slider says "Standard Deviation," the slider actually represents the standard deviation from the mean. The Help context in the tool states "The TAT displays the deviation above and below the mean in positive and negative values. Adjust the slider to identify thresholds at different standard deviations." Standard deviation was chosen for the label over z-score since this is a term utility managers may be more likely to recognize. The context is that the slider represents standard deviations from the mean.

Recurrence Interval

$$\mathrm{RI} = \frac{d}{(f+1)}$$

Recurrence interval (RI) refers to the average time between alerts. In the equation above, *f* is the number of alerts at the threshold, *d* is the total number of days in the data set, and the equation equals the average recurrence interval or the average number of days between alerts. This calculation incorporates the date boundaries of the data by adding 1 to the denominator; if there is one alert over a 40 day range of data, for example, the average recurrence interval would be 20 days.