TreeScan™ User Guide

for version 1.2

By Martin Kulldorff

November, 2015

TABLE OF CONTENTS

Introduction

TreeScan Software

Purpose

TreeScan is a free data mining software that allows users to analyze large data sets using different versions of the tree-based scan statistic.[1,2,3] The software was originally designed for pharmacovigilance, with the purpose to detect unsuspected drug and vaccine adverse reactions using large electronic health plan databases. The software may also be used for similar problems in other medical as well as non-medical fields, whenever the data can be classified into a hierarchical tree-like structure.

Tree Structure and Cluster Detection

To perform an analysis, the user must provide a pre-determined hierarchical tree structure of their data. For example, the tree may consist of (i) ICD-10 medical diagnosis codes with related diagnoses located on the same branch of the tree, (ii) pharmaceutical drugs with similar drugs on the same branch of the tree, or (iii) occupational codes with similar occupations on the same branch of the tree. The tree can have two or more hierarchical levels, representing the number of increasingly smaller sized 'branches' as one moves further away from the trunk towards the 'leaves'. On each leaf and/or branch of the tree, there are observed and expected counts of some outcome. By considering cuts on different branches of the tree, closer or further away from the trunk, the tree-based scan statistic scans the tree for clusters where there are significantly more cases than expected, evaluating both very specific outcome definitions represented by a leaf as well as large groups of related outcomes represented by a big branch.

Data Granularity

A key feature of the tree-based scan statistic is that the granularity of the data does not have to be pre-specified. For example, when we are looking for potential adverse reactions to a pharmaceutical drug, it is impossible to know a priori if the drug may cause a very specific health outcome such as cardiac dysrhythmia, a more general set of related outcomes, such as a variety of heart problems, or an even wider set such as different types of cardiovascular issues. To cast the data mining net as wide as possible, the tree-based scan statistic simultaneously tests many overlapping groups of related outcomes.

Temporal Data

The tree-temporal scan statistic is an extension of the tree-based scan statistic, where scanning is also done over time, to detect temporal clusters on one or more branches of the tree. The

TreeScan software can also perform a purely temporal scan statistic, in which case the tree is non-existent or ignored.

Multiple Testing

Multiple testing is by nature present in all forms of data mining. The tree-based scan statistic automatically adjusts for multiple testing. This is critical in order to conduct proper statistical analysis without generating a large number of false positives.

Developers and Funders

The TreeScan™ software was developed by Martin Kulldorff together with Information Management Services Inc. Financial support for TreeScan has been received from the National Institutes of Health and the Food and Drug Administration. Their financial support is greatly appreciated. The contents of TreeScan are the responsibility of the developer and do not necessarily reflect the official views of the funders.

Related Topics: *Statistical Methodology*, *TreeScan Bibliography*, *Acknowledgments*

Download and Installation

To install TreeScan, go to the TreeScan web site at: http://www.treescan.org/ and select the TreeScan download link. Choose between the TreeScan versions for Linux, Mac or Windows. After downloading the TreeScan installation executable to your computer, click on its icon and install the software by following the step-wise instructions.

Related Topics: *Test Run*, *New Versions*.

Test Run

Before using your own data, we recommend trying one of the sample data sets provided with the software. Use these to get an idea of how to run TreeScan. To perform a test run:

1. Click on the TreeScan application icon.

2. Click on 'Open Saved Session'.

3. Select one of the parameter files, which has a *.prm ending.

4. Click on 'Open'.

5. Click on the Execute ⏩ button. A new window will open with the program running in the top section and a Warnings/Errors section below. When the program finishes running, the results will be displayed.

Note: The sample files should not produce warnings or errors.

Related Topics: *Sample Data Sets*.

Help System

The TreeScan help system consists of three parts:

i. The TreeScan User Guide in PDF format, which you are currently reading, located in the same folder as the TreeScan executable. It can also be obtained from the TreeScan web site (www.treescan.org/techdoc.html) or directly within the TreeScan software by selecting Help > User Guide. You may print this as a single document for easy reference.

ii. Methodology papers describe the details about the statistical methods available in the TreeScan software[1,2,3] These papers are listed in the TreeScan bibliography, which can be found both at the end of this User Guide and on the web (http://www.treescan.org/references.html).

iii. The sample data sets described below makes it easy to familiarize oneself with the software.

Related Topics: *Sample Data Sets, TreeScan Bibliography.*

Sample Data Sets

Three sample data sets are provided with the software. They are automatically downloaded to your computer together with the software itself. The numbers are completely made up, and do not reflect any real data. Their purpose is to illustrate the input file content and format and to conduct a simple test run.

Tree Scan Statistic, Poisson Model

Parameter file: Poisson.prm

Count file: Poisson.cas

    Format: <nodeID> , <#cases> , <population>

Tree file: tree.tre

    Format: <nodeID> , <parent nodeID>

Study period: n/a

Tree Scan Statistic, Bernoulli Model

Parameter file: Bernoulli.prm

Count file: Bernoulli.cas

> **Format:** <nodeID> , <#cases> , <#controls>

Tree file: tree.tre

> **Format:** <nodeID> , <parent nodeID>

Study period: n/a

Tree-Temporal Scan Statistic

Parameter file: TreeTemporal.prm

Count file: TreeTemporal.cas

> **Format:** <nodeID> , <#cases> , <time of cases>

Tree file: tree.tre

> **Format:** <nodeID> , <parent nodeID>

Study period: 1 to 28

Purely Temporal Scan Statistic

Parameter file: TemporalOnly.prm

Count file: TemporalOnly.cas

> **Format:** <#cases> , <time of cases>

Tree file: none

Study period: 1 to 28

Related Topics: *Test Run*, *Help System*.

Statistical Methodology

For the best description of the statistical methodology underlying the TreeScan software, we recommend reading the scientific papers describing the various versions of the tree-based scan statistic.[1,2,3] These are listed in the TreeScan bibliography at the end of the User Guide. Here we only provide a brief summary.

Related Topics: *TreeScan Software*, *Basic TreeScan Features*, *Analysis Tab*, *Methodology Papers*.

Tree-Structured Variable

The foundation of the tree-based scan statistic is a hierarchical tree structured variable. The tree must be pre-specified by the user, and it is not something that is created by the method. The tree should represent some relational/hierarchical aspect of the data. Here are a few examples of potential trees:

Disease Diagnosis Tree: In vaccine safety surveillance, the interest may be in detecting any unanticipated adverse reaction to a new vaccine.[2] The tree will then consist of all medical diagnoses that could conceivably be caused by the vaccine. For example, non-febrile seizures and febrile seizures may be close to each other on one part of the tree while heart murmurs and cardiac dysrhythmias may be close to each other on another part of the tree.
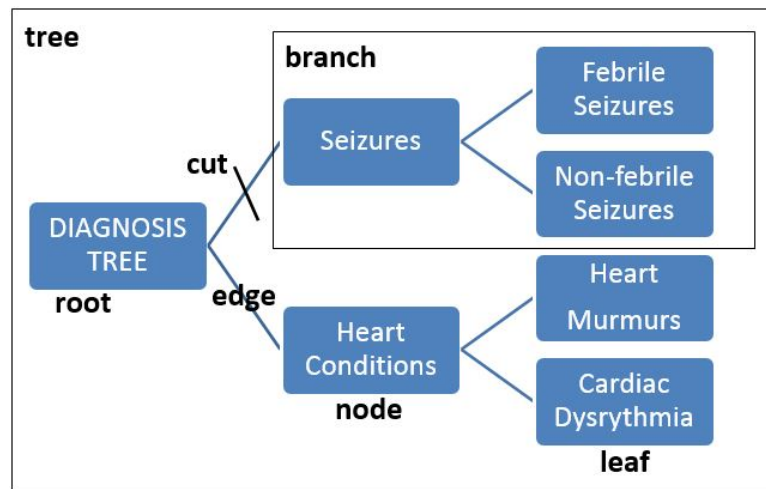
Figure 1: Example of a very small diagnosis tree.

Pharmaceutical Drug Tree: In pharmacovigilance, the interest may be detecting drugs or a class of related drugs that cause drug induced liver injury.[4] The tree will then consist of all pharmaceutical drugs with drugs in the same class close to each other on the tree. For example, rofecoxib (vioxx) and celecoxib (celebrex) may be very close to each other on the tree since they are both cox-2 inhibiting nonsteroidal anti-inflammatory drugs, while ibuprofen is a little further away on the tree but on the same big branch since it is also a nonsteroidal anti-inflammatory drug but not a cox-2 inhibitor.

Occupational Tree: In occupational disease surveillance, the interest may be to determine if people with a particular occupation or group of related occupations are at higher risk for some disease.[1] The tree will then consist of all occupations of interest with closely related occupations close to each other on the tree. For example, elementary school teachers and high school teacher

may be close to each other on one branch of the tree, while coal miner and iron ore miners may be close to each other on another branch of the tree.

Related Topics: *TreeScan Software, Tree Terminology, Tree File, Methodology Papers.*

Tree Terminology

The terminology used for the tree-based scan statistic is derived from graph theory and computer science. For a graphical depiction, see Figure 1.

Tree A hierarchical variable. In mathematical terminology, it is a directed acyclic simple graph consisting of nodes/vertices and edges.

Node A vertex on the tree. Each node contains a number of cases, a population at risk and/or a number of controls. The population must be non-negative and may for example represent expected counts. The cases must be non-negative integers. A node may have zero population and zero cases, and there will typically be many such nodes. If the population is zero, the number of cases must also be zero. Each node must have a unique name identifier, called the node identifier or the node ID.

Edge A line that connects two nodes on the tree.

Parent The parent of a node is the node that is immediately above it on the tree, connected through an edge. A node can only have one parent.

Child The child of a node is a node that is immediately below it on the tree, connected through an edge. A node may have zero, one or multiple children. Note that a node can be a parent to one node while it is a child to another node.

Leaf A node that does not have any children. On most trees, the majority of nodes are leaves.

Root A node that does not have a parent. A typical tree has only one root, but there can be more than one.

Siblings Two or more nodes that share the same parent.

Descendants The descendants of a node are all its children, its children's children (grandchildren), its children's children's children (great grandchildren), and so on.

Ancestors The ancestors of a node are its parent, its parent's parent (grandparent), its parent's parent's parent (great grandparent), and so on

Cut A cut may be done on any one of the edges, just above a node. A simple cut defines a branch, the collection of nodes that include the node just below the cut together with all its descendants. The cut is identified by the name of the node that is just below the cut.

Branch   The collection of nodes defined by a cut, including the node just below the cut and all its descendants.

Cluster   A branch that has more observed cases than expected. Some clusters are statistically significant, and hence unlikely to have occurred by chance, but most clusters are not statistically significant. For the tree-temporal scan statistic, the cluster is defined by the collection of nodes together with a time interval.

Related Topics: *Tree Structured Variable, Tree File.*

Tree-Based Scan Statistic

Under the null hypothesis, a case is equally likely to occur anywhere on the tree, in any of the nodes, in such a way that the expected number of cases in any node is proportional to the population of that node. Under the alternative hypothesis, there is one or more branches on the tree where cases have a higher probability of occurring, constituting a cluster. The goal of the tree-based scan statistic is (i) to detect branches that contain a cluster of cases without pre-specifying the branch a priori, and (ii) to determine whether the detected clusters are statistically significant after adjusting for the multiple testing inherent in the many overlapping branches evaluated.

The tree-based scan statistic is a likelihood ratio test. For each branch of the tree, the likelihood is calculated under both the null and alternative hypotheses, using a mathematical formula that depends on the probability model, as described in the tree scan methodology papers.[1,2,3] The branch with the maximum likelihood ratio is the most likely cluster, or the most likely cut, that is, the cluster of cases that is least likely to have occurred by chance. For computational reasons, the maximum log likelihood ratio (LLR) is used as the test statistic rather than the maximum likelihood ratio. Since one is a monotone function of the other, the two give exactly the same statistical test.

The distribution of the test statistic under the null hypothesis is not known, making it impossible to conduct inference using analytical methods, but we do know how to generate data under the null hypothesis. Because of that, we can do inference using computer simulations and Monte Carlo hypothesis testing.[8] First we generate 999 (say) random replicas of the data generated under the null hypothesis. For each of these data sets, we calculate the maximum log likelihood ratio. Note that these maxima are attained for different branches for different random data sets. If the real data set is also generated under the null hypothesis, then there is a 5% probability that the maximum LLR from the real data set is among the 50 highest LLRs from the real and random data sets. If it is, then we can reject the null hypothesis at the alpha=0.05 significance level. This probability is exact, so the hypothesis test is neither conservative nor liberal.

The Monte Carlo p-value is calculated as p=R/(S+1), where R is the rank of the maximum LLR from the real data compared to the random data sets and S is the number of simulated Monte Carlo replications. In order for the p-value to be a 'nice looking' number, the number of simulations is restricted to 999 or some other number ending in 999 such as 1999, 9999 or 99999. That way it is always clear whether to reject or not reject the null hypothesis for typical cut-off values such as 0.05, 0.01 and 0.001. Additional Monte Carlo replications will increase statistical power, but beyond 999, the increase is marginal.

Related Topics: *TreeScan Software*, *Basic TreeScan Features*, *Analysis Tab*, *Methodology Papers*.

Poisson Probability Model

With the Poisson model,[1] the number of cases in each node is Poisson-distributed. For the unconditional version, the expected number of cases under the null hypothesis is provided in the count file. For the conditional version, the expected number of cases under the null hypothesis is proportional to its population, in such a way that the total expected across the whole tree is equal to the total number of observed cases. For the conditional Poisson model, the results will be the same if the population is multiplied by the same constant everywhere on the tree.

Example: For the Poisson model, cases may be health outcomes while exposed to a particular pharmaceutical drug, while the population is the age adjusted expected number of health outcomes.

Related Topics: *Analysis Tab*, *Bernoulli Probability Model*, *Methodology Papers*.

Bernoulli Probability Model

With the Bernoulli model[2], there are cases and non-cases represented by a 0/1 variable. These variables may represent people with or without a disease, people being exposed or unexposed, or two different time periods of a person's life in a self-control type analysis. They may reflect cases and controls from a larger population, or they may together constitute the population as a whole. Whatever the situation may be, these variables will be denoted as cases and controls throughout the user guide, and their total will be denoted as the population. For each node, the population is fixed and non-random.

In the unconditional version, the user must specify the probability of being a case under the null hypothesis. In the conditional version, the analysis is conditioned on the total number of cases observed in the whole tree.

Example 1: In comparative drug safety cohort study, cases may be adverse events among recipients of drug A, while the controls are adverse events among recipients of comparator drug B. If a 1:1 matching is used, the probability of being a case would be ½, using the unconditional Bernoulli model.

Example 2: In a vaccine safety self-control design, cases may be health outcomes occurring in a risk window 1-14 days after vaccination while controls are health outcomes occurring in a control window 29-56 days after vaccination. If the unconditional version is used, the probability of being a case would be 1/3, since the control window is twice as long as the risk window.

If an unconditional Bernoulli model is selected, one should indicate if data comes from a self-control design. This does not affect the statistical analysis per-se, but it uses a different self-control formula to calculate the relative risk, the excess number of cases and the attributable risk.

Related Topics: *Analysis Tab*, *Poisson Probability Model*, *Methodology Papers*.

Tree-Temporal Scan Statistic

The tree-temporal scan statistic is a fusion of the standard tree-based scan statistic[1,2] and the temporal scan statistic.[9,13] Temporal scan statistics are used to detect and evaluate the statistical significance of temporal clusters, without a prior specification of the risk window. The method adjusts for the multiple testing in the many possible cluster times and lengths. For example, with a study time period of 56 days, it may consider all 2002 time intervals with a length of at most 28 days as possible temporal clusters, including [1-4], [3-6], [5], [6-23], [16-43], [31-32] and [29-56]. Alternatively, it may only consider a subset of those time intervals, as requested and specified by the user.

The maximum temporal cluster length should never be more than 50% of the total study time period. The reason for this is that a 'cluster' in a longer time interval would more accurately be thought of as two negative clusters with less cases than expected before and after the time interval. For example, a 'cluster' in the [2-55] time period would more naturally be interpreted as a lack of cases on days 1 and 56 rather than an excess number of cases on days [2-55].

The tree-temporal scan statistic[3] requires a time for each case. Only case data are needed, with no information needed about controls or a background population at risk. In addition to the tree-structured variable, it is necessary to specify a study time period. That could be, for example, the 1 to 56 days following the initial (incident) use of a pharmaceutical drug. For each observed case, one records not only the node to which it belongs but also the time of the case, which should fall within the study time period.

With the tree-temporal scan statistic, we are in essence performing multiple temporal scan statistics, one for each of the many overlapping branches of the tree, adjusting for the multiple testing stemming both from the many branches and the many time intervals evaluated. Each time interval is evaluated on each of the branches, so if there are for example 1000 nodes on the tree and 2002 potential time intervals, there would be 2,002,000 potential clusters to evaluate and for which we need to adjust for multiple testing. If these were 2 million independent tests with independent non-overlapping data, there would be a huge loss in power when adjusting for all that multiple testing. With scan statistics, such a large loss in power does not happen, since

the 2 million potential clusters are highly overlapping with each other. Hence, the penalty for adjusting for the multiple testing is fairly modest and not as bad as one may originally think.

The tree-temporal scan statistic conditions the analysis on the number of cases observed in each node. This means that, unlike the pure tree scan statistic, there is no probability distribution to model the number of cases in each node. That number is now deterministic. What is probabilistic is the time of each case, which under the null hypothesis is assumed to be uniform across the study time period. Under the alternative hypothesis, there is at least one branch for which there is a temporal cluster of cases during a shorter or longer time interval.

Related Topics: *Analysis Tab*, *Methodology Papers,* *Purely Temporal Scan Statistic*.

Purely Temporal Scan Statistic

While the tree and tree-temporal scan statistics are the key features of the TreeScan software, it is also possible to run a purely temporal scan statistic. In fact, the purely temporal scan statistic is a special case of the tree-temporal scan statistic when there is only one node in the 'tree'.

Related Topics: *Analysis Tab*, *Methodology Papers*, *Tree-Temporal Scan Statistic*.

Conditional versus Unconditional Analyses

Scan statistics exist in both unconditional and conditional forms. Unconditional scan statistics need external information about the true probabilities underlying the null hypothesis. Conditional scan statistics are conditioned on the total number of cases observed, either in the whole data set, or in each of a subset of the data. The conditional Poisson and Bernoulli versions of the tree-based scan statistics both condition on the total number of cases observed in the whole tree. In this way, the inference is done as a comparison between the risk in different parts of the tree, and the total number of cases in the tree does not make it either more or less likely to reject the null hypothesis. The tree-temporal scan statistics not only conditions on the total number of cases in the whole tree, but also on the vector of cases in each of the nodes.

Related Topics: *Bernoulli Probability Model*, *Poisson Probability Model*, *Tree-Temporal Scan Statistic*, *Methodology Papers*.

Secondary Clusters

TreeScan also identifies secondary clusters in the data set in addition to the most likely cluster, and orders them according to their likelihood ratio test statistic. There will often be a secondary cluster that overlaps with the most likely cluster, with the cuts being made higher or lower on the same branch of the tree. There may also be secondary clusters that do not overlap with the most likely cluster. The p-values for such clusters should be interpreted in terms of the ability of the secondary cluster to reject the null hypothesis on its own strength, whether or not the more

likely clusters are true clusters or not. Hence, these p-values are not adjusted for the fact that there may be other clusters in the data.

Related Topics: *Results*.

Comparison with Other Methods

Other Scan Statistics

Scan statistics were first studied in detail by Joseph Naus.[13] Scan statistics has been developed for one dimension, such as time, as well as for spatial and spatio-temporal data. Excellent reviews of scan statistics have been provided in books by Glaz and Balakrishnan[9], Glaz, Naus and Wallenstein[10] and Glaz, Pozdnyakov and Wallenstein.[11] For spatial, temporal and spatio-temporal scan statistics, the free SaTScan™ software can be used, which is available at www.satscan.org.

Classification and Regression Trees (CART)

When using the tree-based scan statistic, the tree must be defined by the user, and as such, it is a tree-type variable, different from a categorical, ordinal, multinomial or continuous variable. The purpose of the tree-based scan statistic is not to create the tree. The method is hence very different from statistical data mining methods such as Classification and Regression Trees (CART), where the purpose is to create a tree structure.[7]

Input Data

Data Requirements

Required Files: The input data should be provided in a number of files. A count file with the observed data is always needed, and a tree file defining the structure of the tree is needed except for a purely temporal analysis.

File Format: The data input files must be in TreeScan ASCII file format or you may use the TreeScan import wizard for dBase, Excel, comma delimited or space delimited files. Using such files, the wizard will automatically generate TreeScan file format files. Both options are described below.

Related Topics: *Input Tab*, *Count File*, *Tree File*, *TreeScan Import Wizard*, *TreeScan File Format*.

Tree File

The tree file defines the structure of the tree. Each node on the tree must be assigned a Node ID, which can be any text string. Each line of the file represents one node. The first column contains the node ID. The second column denotes the node ID of the 'parent' of that node, that is, the closest node one level up on the hierarchical tree. If there is no parent node, the second column is left blank. This file is not needed for the purely temporal scan statistic.

Note: All nodes must be included in this file, including those that do not have a parent node.

Related Topics: *Input Tab*, *Tree Structured Variable*, *Tree Terminology*, *TreeScan Import Wizard*, *TreeScan File Format*.

Count File

The count file provides information about observed cases. For the Bernoulli model it should also contain information about controls, and for the Poisson model, it should also contain expected cases or population numbers. The count file is used for all probability models and it should contain the following information:

> **Node ID:** Any numerical value or string of characters.

> **Number of Cases:** The number of observed cases for the specified node. Must be a non-negative integer.

> **Number of Controls (Bernoulli model):** The number of observed controls for the specified node. Must be a non-negative integer.

> **Population (Poisson model):** The population size for the specified node. For the conditional Poisson model this could be raw population numbers, a covariate adjusted

population at risk, or, the expected number of cases under the null hypothesis. If the unconditional Poisson model is used, it must be the expected counts. If the population

size is zero for a particular node, it may either be left out or included in the count file with a specified value as zero. The population can be specified as integers or decimal numbers. It cannot be a negative number.

Time (Tree-temporal scan statistic): Specified in a generic format, which typically represent days, weeks, months or years, although it could also represent seconds, minutes, decades or centuries. Must be an integer, but could be either negative or positive.

Example: If there were 3 seizures 7 days after receiving the measles-mumps-rubella-varicella vaccine and 5 seizures 8 days after the vaccination, the following information should be provided for the tree-temporal model:

Seizure 3 7
Seizure 5 8

Note: Multiple lines may be used for different cases in the same node/leaf, having the same time attribute. TreeScan will automatically add them. For example, the above information can also be provided as:

Seizure 1 7
Seizure 1 7
Seizure 1 7
Seizure 4 8
Seizure 1 8

Related Topics: *Input Tab*, *TreeScan Import Wizard*, *TreeScan File Format*.

Cut File

The cut file is optional. If used, it should contain the following information:

Node ID: Any numerical value or string of characters.

Cut type: The type of cuts used below this particular node. The options are simple cuts, pairs cuts, triplets cuts and ordinal cuts, which are denoted as follows:

simple cut:	s	*or*	simple
pairs cut:	p	*or*	pairs
triplets cut:	t	*or*	triplets
ordinal cut:	o	*or*	ordinal

The definitions of these cuts are provided under Complex Cuts in the Advanced Input Tab section.

Note: Not all node IDs must be included in the Cut File. The default is simple cuts, so that will be used if a node is not listed in this file, or if this file is not provided.

Related Topics: [Advanced Input Tab](#), *[Complex Cuts](#)*, *[Tree File](#)*, *[TreeScan Import Wizard](#)*, *[TreeScan File Format](#)*.

TreeScan Import Wizard

The TreeScan Import Wizard can be used to import dBase, comma delimited, or space delimited files. It works for all import files. Launch the Import Wizard by clicking on the File Import [...] button furthest to the right of the text field for the file that you want to import. Follow the steps below to import files. Use the **Next** and **Previous** buttons to navigate between the dialogs.

Step 1 – Selecting the Source File

1. At the bottom of the Select Source File dialog, select the file type extension you are looking for. If you are unsure, select the All Files option. Supported file formats are: dBase III/IV, CSV, Excel, Text (*.txt) and TreeScan file formats.

2. Browse the folders and highlight the file you want to open. It will appear in the File Name text field.

3. Click on Open. The TreeScan Import Wizard will now appear.

Step 2: Specifying the File Structure

If you are importing a dBase or an Excel file, this step is automatically skipped. For all other source files, you need to specify the file structure using the File Format dialog box.

1. First specify the delimiting character and grouping indicator of the file.

2. If there are extraneous lines in the beginning of the file, type the number of lines that you would like to ignore in the text field below data sample area.

3. Click on **Next** to proceed to the next dialog box.

Step 3: Matching Source File Variables with TreeScan Variables

The top grid in this dialog box links the TreeScan variables with the input file variables from the source file. The bottom grid displays sample data from the chosen input file.

1. To match the variables, click on one of the places where it says `unassigned'.

2. Select the appropriate variable form the input file to go with the chosen TreeScan variable.

---

3. When all the required and optional variables that you selected have been matched, click on the Execute button to import the file. This will create a temporary file in TreeScan ASCII file format.

Step 4: Saving the Imported File

The imported file, which is in TreeScan ASCII file format, must be saved at least temporarily. The default is to save it to the TEMP directory and after the analysis is completed you may erase the file. You can also save it to some other directory of your choice and use it for future analyses without having to recreate it by using the Import Wizard again.

Related Topics: *Input Tab*, *Count File*, *Tree File*, *Cut File*.

TreeScan File Format

As an alternative to using the TreeScan Import Wizard, it is possible to directly write the name of the input files in the text fields provided on the Input Tab, or to browse the file directories for the desired input files using the button to the right of that box. The files must then be in TreeScan comma delimited file format, which are ASCII files with one row for each node/time combination with columns as defined below. Such files can be created using any text editor and most spreadsheets. The order of the columns in the file is very important, but the rows can be in any order. The optional variables, defined above, are optional columns in the TreeScan file format.

Tree File Format (*.tre):

<node ID> , <parent node ID>

Count File Format (*.cas):

<node ID> , <#cases> , <population>   (Poisson model)

<node ID> , <#cases> , <#controls>   (Bernoulli model)

<node ID> , <#cases> , <time>   (tree-temporal model)

Cut File Format (*.cut):

<node ID> , <cut type>

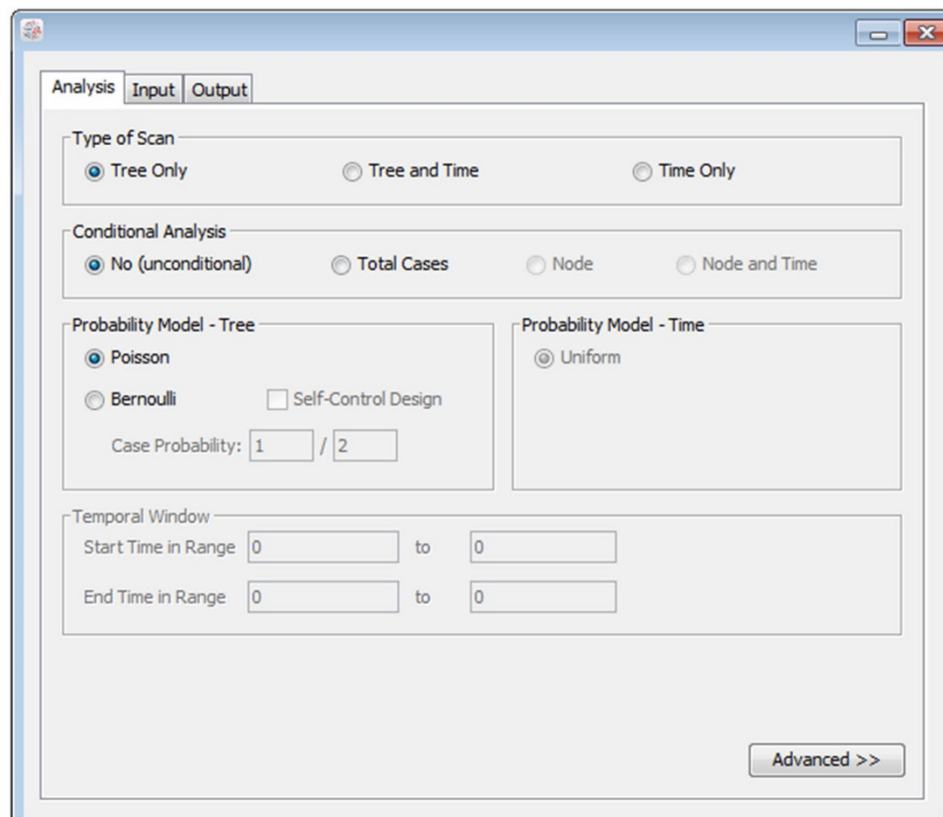Alternative Hypothesis File Format (*.alt):

<node ID> , <relative risk>   (Poisson model)

<node ID> , <probability>   (Bernoulli model)

Related Topics: *Input Tab, Count File, Tree File, Cut File, TreeScan Import Wizard.*

Basic TreeScan Features

Most TreeScan analyses can be performed using a set of basic analysis and data features. The users specify these on three different window tabs for analysis, input and output options respectively. These contain all required specifications for a TreeScan analysis as well as a few optional ones. Additional features, all optional, can be specified on the advanced features tabs.

Related Topics: *Statistical Methodology, Input Tab, Analysis Tab, Output Tab, Advanced Features.*

Analysis Tab

Figure 2 - Analysis Tab Dialog Box

The Analysis Tab is used to specify the fundamental analysis options. Additional optional features are available by clicking on the Advanced button in the lower right corner.

Type of Scan

TreeScan may be used for a purely tree-based analysis, scanning only the tree, as well as for a tree-temporal scan statistic which simultaneously scans both the tree and a temporal time period. It is also possible to run a purely temporal scan statistic, not using a tree at all.

Conditional Analysis

With the TreeScan software, it is possible to conduct either unconditional or conditional analyses, where the analysis is conditioned on the total number of cases in the whole tree or in parts of the tree. In an unconditional analysis, random data is generated either from expected counts (Poisson model) or by using a pre-specified event probability (Bernoulli model). The expected counts and the event probability must be provided by the user.

With a conditional scan statistic, the analysis is conditioned on some aspect of the observed data. The analysis can be conditioned on the total number of cases found in the tree as a whole. This means that each random data set has exactly the same number of total cases as the real data set. For the Poisson model this means that under the null hypothesis, the expected number of cases in a particular node is no longer identical to the expected count provided by the user. Instead, the expected number of cases in that node is the total number of cases times the proportion of the expected count that is in that node. For the Bernoulli model, it means that under the null hypothesis, the probability of being a case is equal the total number of cases in the whole tree divided by the total number of observations (cases and controls).

For the tree-temporal scan statistic, the analysis is always conditioned on the number of cases in each node. That is, a node has exactly the same number of cases in each of the random data sets as in the real data set. This means that the methods does not evaluate whether there are branches on the tree with more cases than expected. Instead, it evaluates if there are branches on the tree that has a temporal cluster of cases. In addition to conditioning on the number of cases in each node, a tree-temporal analysis may also condition on the total number of cases at each time point. The purpose of this is to adjust for any purely temporal variation in the data that is common to the tree as a whole.

For a purely temporal scan statistic, the analysis is always conditioned on the total number of cases observed.

Probability Model - Tree

There are two different probability models that can be used for the number of cases observed in each node: Poisson and Bernoulli.

Poisson Model: The Poisson model should be used when the background population reflects a certain risk density such as total person years, or, some covariate adjusted expected counts.

Bernoulli Model: The Bernoulli model should be used for 0/1 type data, such as individuals who may or may not have a disease. Those who have the disease are 'cases' and those who don't have the disease are 'controls'. For an unconditional analysis, it is necessary to specify the probability of being a case, which must be greater than 0 and smaller than 1. If you are doing a self-control type analysis, that should be indicated. This does not affect the statistical analysis, but ensures that the self-control formulas are used when calculating the relative risk, the excess number of cases and the attributable risk.

When the tree-temporal scan statistic is selected, the analysis is conditioned on the number of cases in each node, so there is no probability model for the number of cases on the nodes. That is, the number of cases in each node is deterministic rather than probabilistic.

Probability Model – Time

For the tree-temporal and purely temporal scan statistics, the only probability model that is available for the time dimension is uniform. That is, a case is equally likely to occur during any of the days (or other unit) in the data time range.

Note: The Bernoulli model is a special case of the tree-temporal scan statistic when there is only two time periods in the data time range, such as [1,2], and when the start and end ranges of the temporal window are [1,1] and [1,1].

Temporal Window Range

The maximum temporal window size is always set to be at most 50% of the data time period. That is, the time period inside the cluster can never be larger than the sum of the time periods before and after the cluster. This is to ensure that we are evaluating an excess number of cases inside the window, rather than a deficit of cases during a very small time period at the very beginning and/or end of the data time range.

As an addition to this firm restriction set by the software, it is necessary to specify the collection of windows to be evaluated in terms of a range of start times for the temporal clusters as well as a range of end times. For example, if [1,2] is selected as the start range and [4,5] as the end range, only the following temporal clusters are evaluated: [1,4], [1,5], [2,4] and [2,5]. Most commonly, a much larger collection of windows are used. By specifying both the start and end range to be identical to the data time range on the Input Tab, all possible windows with a length less than 50% are evaluated.

This option is only relevant and available for the tree-temporal scan statistic.

Related Topics: *Basic TreeScan Features, Statistical Methodology, Poisson Probability Model, Bernoulli Probability Model, Tree-Temporal Scan Statistic, Conditional Versus Unconditional Analyses.*
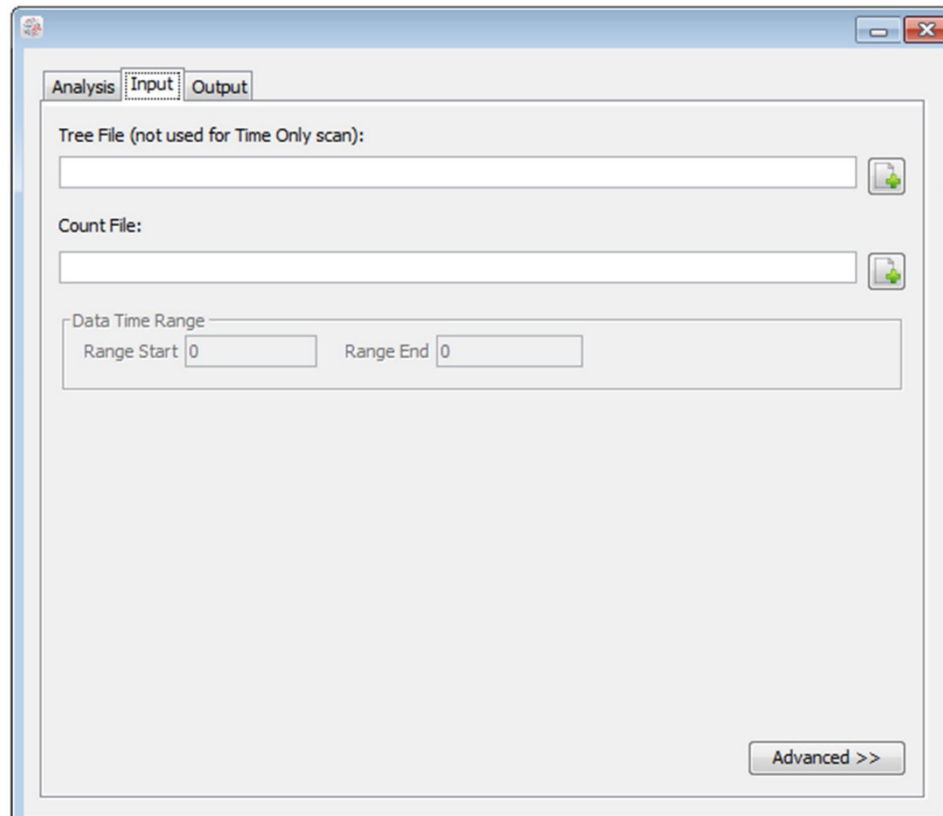
Input Tab

Figure 3 - Input Tab Dialog Box

The Input Tab is used to specify the names of the input data files as well as the nature of the data in these files. If the files are in TreeScan comma delimited file format, they may be specified either by writing the name in the text box or by using the Open File [icon] button. If they are not in TreeScan comma delimited file format, they must be specified using the TreeScan import wizard, by clicking on the File Import [icon] button. Both the TreeScan comma delimited file format and the TreeScan import wizard are described in the Input Data section.

Tree File Name

Specify the name of the input file that defines the tree structure. This file is required for all analyses except the time only scan statistic, irrespectively of the probability model used.

Count File Name

Specify the name of the input file with the observed case data. This file is required for all scan statistics, irrespectively of the probability model used.

Data Time Range

Only used for the tree-temporal and purely temporal scan statistics. Specify the start and end of the time range for which temporal data was collected. All times in the count file should fall on or between the start and end times.

Data Time Range Start: The earliest time included in the study period.

Data Time Range End: The latest time included in the study period.

The times may be negative or positive. Provided in a generic format, the times can represent seconds, minutes, hours, days, weeks, months, years or any other unit of time.

Related Topics: *Basic TreeScan Features*, *Count File*, *Tree File*.
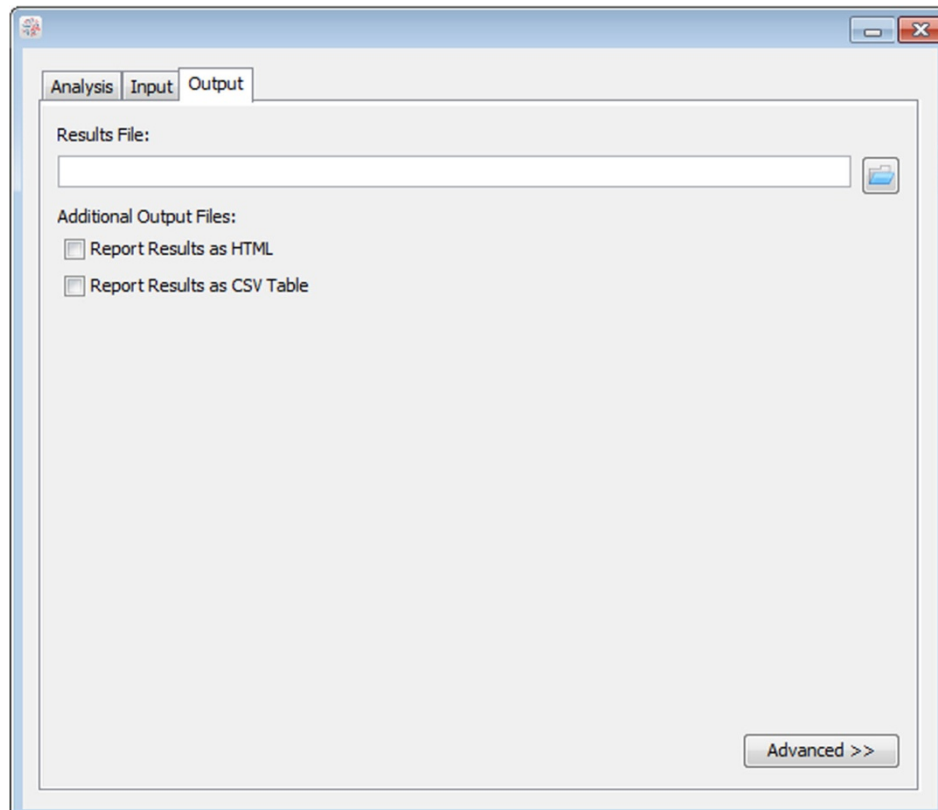
Output Tab

Figure 4 - Output Tab Dialog Box

The Output Tab is used to set parameters defining the output information provided by TreeScan. These parameters do not influence the actual analysis; just the way that the results are reported.

Text Output Format

A standard text based results file is automatically shown after the completion of the calculations. It contains information about the clusters detected, summary information about the data, computing time and the analysis parameters chosen. Specify the name of this file. Other optional output files will have the same name, but with different filename extensions.

Warning: If you specify the name of a file that already exists, the old file will be overwritten and lost.

HTML Output Format

If requested, TreeScan will create a HTML file that shows the results in a table, using the web browser. It will automatically launch after the analysis is complete. The name of this output file is the same as the name of the text output format file, but with a different filename extension (*.html).

CVS Table Output Format

If requested, TreeScan will create a comma delimited (CSV) output file, that will show the detected clusters in table format. This file can easily be imported into Excel, SAS, R or other software for further formatting, depiction or analyses. The name of this output file is the same as the name of the text output format file, but with a different filename extension (*.csv). You must manually open this file after the TreeScan run is completed.

Related Topics: *Basic TreeScan Features*, *Results*.

Advanced Features

While most TreeScan analyses can be performed using the features on the three basic tabs for analysis, input and output, a few additional options are available as advanced features. These features are reached through the Advanced button on the lower right corner of each of the three main tabs. 'Advanced' should be interpreted as 'additional' or 'uncommon' rather than 'complex', 'difficult' or 'better'.

Since many of the advanced options depend on the selections made on the Analysis and Input Tabs, it is recommended that those two tabs be filled in first.

Related Topics: *Basic TreeScan Features*, *Inference Tab*, *Advanced Output Tab*.

Temporal Window Tab
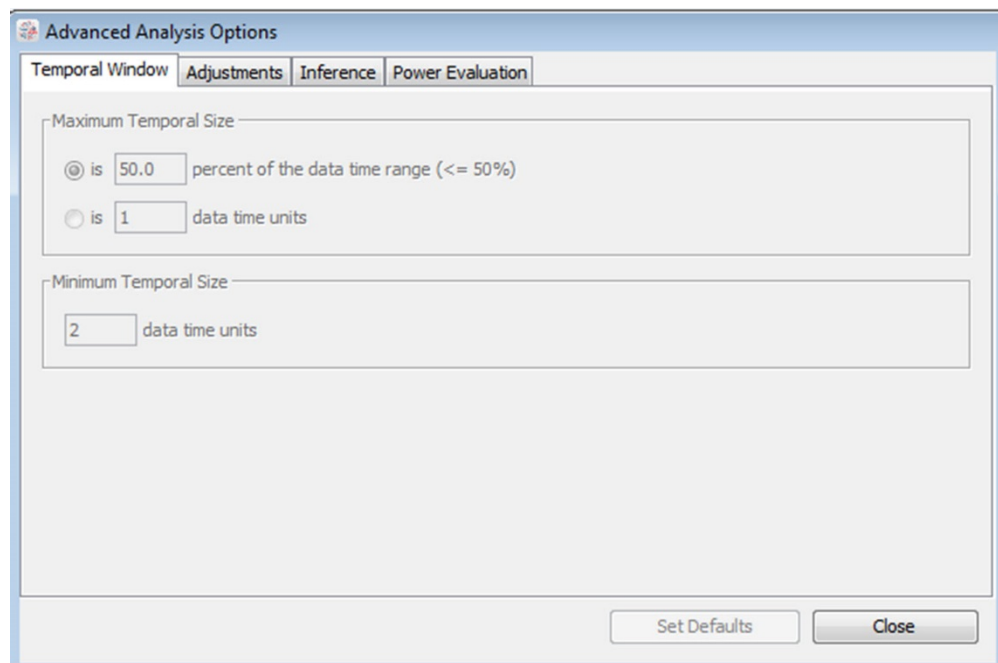
Figure 5 - Temporal Window Tab Dialog Box

This tab is reached by clicking the Advanced button in the lower right corner of the Analysis Tab. It is only relevant when using the tree-temporal scan statistic.

Maximum Temporal Cluster Size

The maximum temporal cluster size can be specified in terms of a percentage of the study period as a whole or as a certain number of time units. If specified as a percent, it can be at most 50 percent. If specified in time units, it can be set to at most half the length of the data time range specified on the Input Tab.

Minimum Temporal Cluster Size

A minimum temporal cluster size can be specified in terms of a certain number of time units. The default value is one time unit, which means that there is no minimum restriction on the temporal cluster size.

Related Topics: *Analysis Tab*, *Input Tab*, *Tree-Temporal Scan Statistic*.
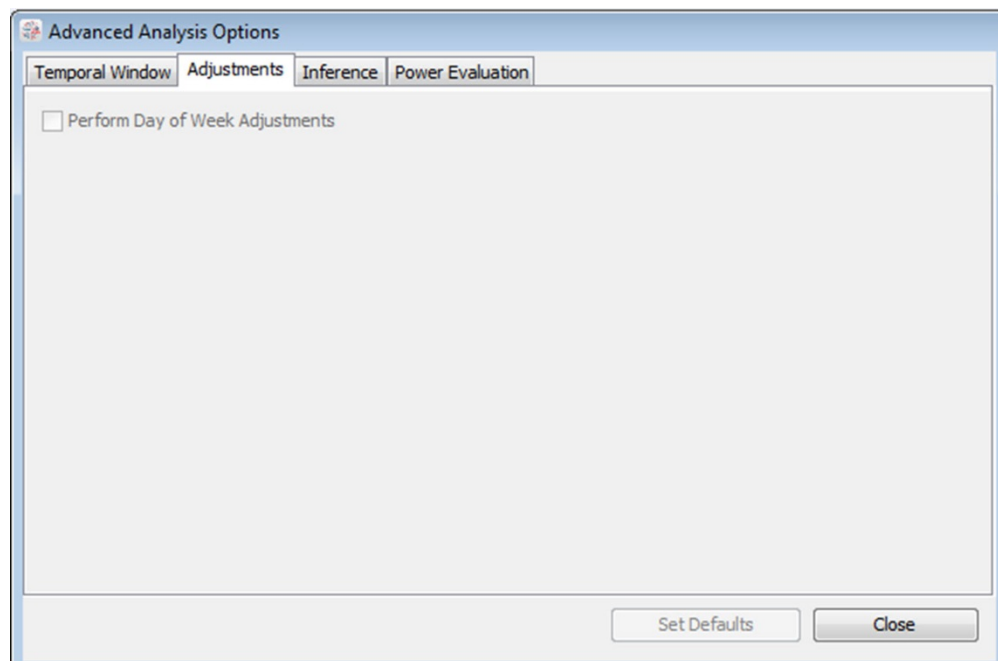
Adjustments Tab

Figure 6 - Adjustments Tab Dialog Box

This tab is reached by clicking the Advanced button in the lower right corner of the Analysis Tab.

Day-of-Week Adjustment

By checking this box, TreeScan will automatically perform day-of-week adjustment. This option is only relevant for the tree-temporal and purely temporal scan statistics.

Suppose that we are looking for adverse events after vaccination, and the time is the number of days between the vaccination and the adverse event. If *both* the vaccinations and the outcome have a day-of-week pattern, such as being more common on weekdays than weekends, the assumption of a uniform probability over time is no longer valid. The day-of-week adjustment takes care of this, using stratified randomization. If only the exposure, or only the outcome, has a day-of-week effect, the uniform assumption holds, and there is no need to do a day-of-week adjustment.

Related Topics: *Analysis Tab*, *Tree-Temporal Scan Statistic*.
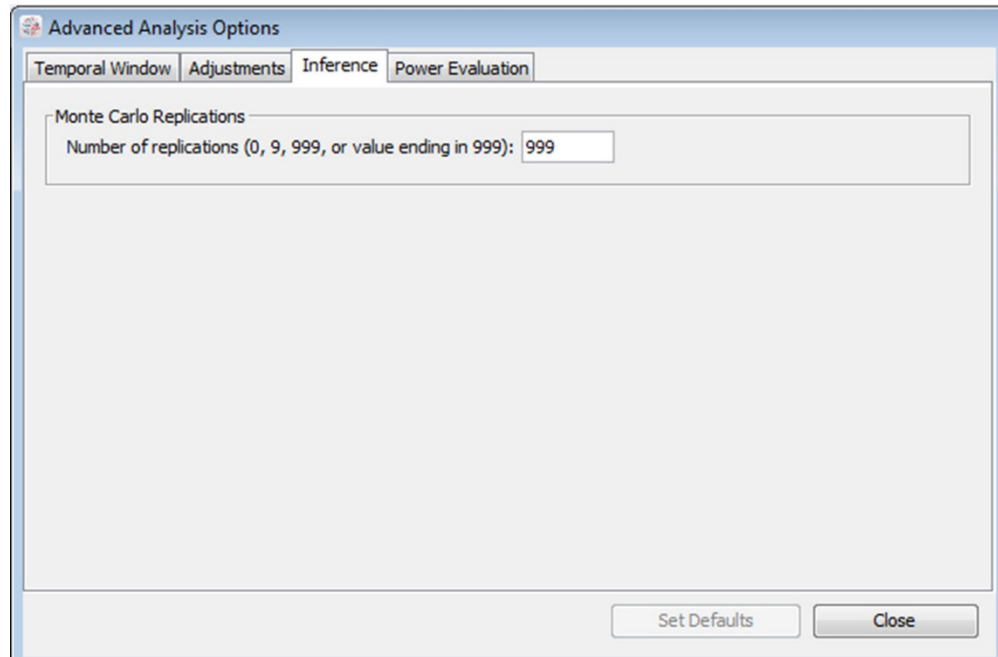
Inference Tab

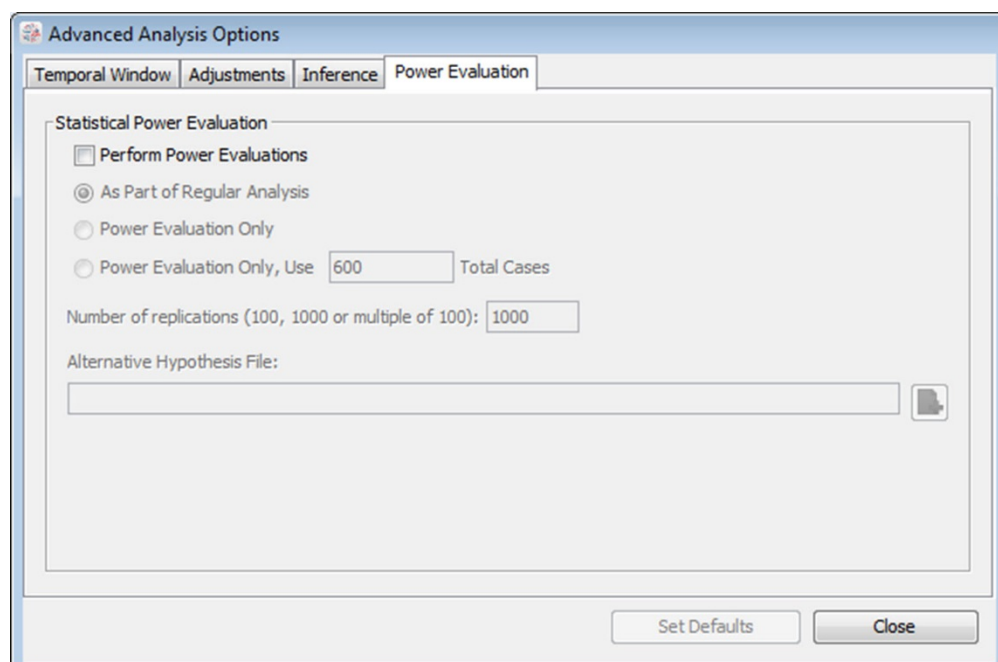*Figure 8 - Inference Tab Dialog Box*

This tab is reached by clicking the Advanced button in the lower right corner of the Analysis Tab.

Monte Carlo Replications

To calculate p-values for detected clusters, TreeScan uses computer simulations to generate a number of random replications of the data set under the null hypothesis. If the maximum likelihood ratio calculated for the most likely cluster in the real data set is high compared to the maximum likelihood ratios calculated for the most likely clusters in the random data sets, that is evidence against the null hypothesis and for the existence of clusters.

The test statistic is calculated for each random replication as well as for the real data set, and if the latter is among the 5 percent highest, then the test is significant at the 0.05 level. If it is among the 1 percent highest, the test is significant at the 0.01 level, and so on. This is called Monte Carlo hypothesis testing, and was first proposed by Dwass.[8] Irrespective of the number of Monte Carlo replications chosen, the hypothesis test is unbiased, generating a correct significance level that is neither conservative nor liberal nor an estimate. The number of replications does affect the power of the test though, with more replications giving slightly higher power. A higher number also increases the computing time.

In TreeScan, the number of replications must be at least 999 to ensure excellent power for all types of data sets. For small to medium size data sets, where computing time is not an issue, at least 9999 replications are recommended.

Note: The number of Monte Carlo replications can be specified to be 0 or 9 in order to do a test run, but that will not provide p-values for the analysis.

Related Topics: *Analysis Tab*, *Tree-Based Scan Statistic*, *Random Number Generator*, *Results*.

Power Evaluation Tab

Figure 9 - Power Evaluation Tab Dialog Box

This tab is reached by clicking the Advanced button in the lower right corner of the Analysis Tab.

Statistical Power Evaluation

The power of the tree-based scan statistic varies for different branches and leaves on the tree. The higher the expected counts in a Poisson analysis, and the higher the number of cases and controls in a Bernoulli or tree-temporal analysis, the higher the power for that branch or leaf. Of course, power also increases with increased relative risk, and it also depends on the size and nature of the tree. With the TreeScan software, it is possible to estimate the power either as part of a regular analysis or as a stand-alone exercise. The feature is available for the conditional and unconditional Poisson models as well as for the unconditional Bernoulli model.

If the power evaluation is done as part of a regular analysis, the estimated power will be added to the results file. If only a power evaluation is requested, the software will not evaluate any cuts on the tree and not produce any likelihood ratios, relative risks, p-values, etc. If only a power evaluation is performed for the conditional Poisson model, the user has the option to either condition the analysis on the total number of cases in the count file or to specify a different total number of cases on the Power Evaluation Tab.

The statistical power is estimated using data that are simulated under the alternative hypothesis. The number of simulated replications of the data needs to be specified. It is recommended to use at least 1000 replications, but more is always better in order to increase the precision of the power estimates.

The alternative hypothesis is defined in the Alternative Hypothesis File, which is a plain text file that consists of two columns. In the first column, specify the node ID for which there should be an excess risk. In the second column, specify the excess risk. For the Poisson model, specify it in terms of the relative risk. For the Bernoulli model, specify it in terms of a probability. For the alternative hypothesis, it is possible to specify an excess risk for multiple nodes. All of them will then have an excess risk. This is done by having multiple rows in the Alternative Hypothesis File, without an empty row in between. Note that the excess risk can be different for different nodes within the same alternative hypothesis.

Multiple alternative hypotheses can be evaluated within the same TreeScan run. This is done by adding additional rows to the Alternative Hypothesis File, keeping an empty line between the different alternative hypotheses. It is faster to do it this way than to conduct a separate run for each alternative, since the different alternative can use the same set of data that is generated under the hypothesis.

Note: For the Poisson model, it is typical to specify an alternative hypothesis with a relative risk greater than one, and for the Bernoulli model, a probability that is greater than the probability specified on the main Analysis Tab. This ensures that the alternative hypothesis contain an excess risk somewhere on the tree. Technically, it is also possible to specify smaller numbers.

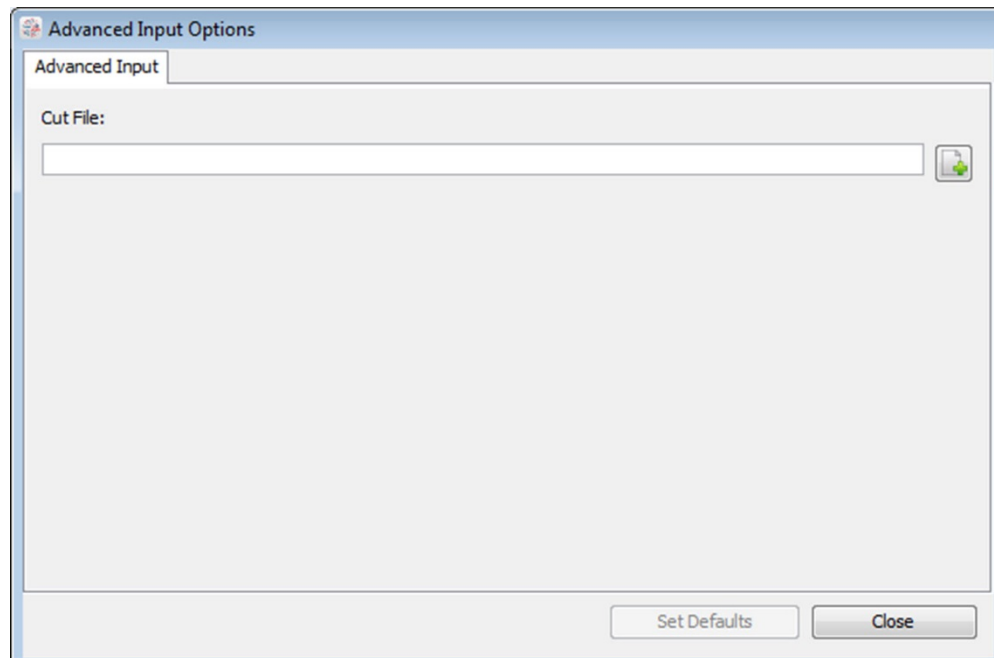**Related Topics:** *Analysis Tab*.

Advanced Input Tab

Figure 10 - Advanced Input Tab Dialog Box

Complex Cuts

As the default, TreeScan uses simple cuts on the tree, where the cut is on a single edge, defining the branch that is a potential cluster. When a node has more than two children, it is also possible to perform pair cuts, triplet cuts or ordinal cuts, as described below. The type of cuts requested is specified in the Cut File, and the name of that file is specified on the Advanced Input Tab.

Suppose a node has five children. With simple cuts below the node, each child and all of its descendants is a cut defining a branch with potential cluster. With a simple cut above the node, all five children and all their descendants is also a branch defining a potential cluster. Two of the children together, but without the other three siblings, cannot define a potential cluster though, when using simple cuts. When using pair cuts, pairs of two siblings with their descendants will also define a potential cluster, consisting of two of the five branches emanating from that node. A group of three siblings with their descendants will not be a candidate for a cluster, when a pair cut is used, but such trios will be considered if triplet cuts are used. When a triplet cut is specified, pair and simple cuts are also utilized. When a pair cut is specified, simple cuts are also used.

With an ordinal cut, there needs to be an order on the children, and we only consider pairs, triplets or larger groups of children that are next to each other in the ordering. For example, suppose the five children are A, B, C, D and E. As potential clusters, we would then consider for example [A,B], [B,C], [C,D,E], and [A,B,C,D], but not [A,C] or [B,E]. The order used for the children is alphabetical. If you want to order the children in some other way, the NodeID's need to be renamed. For example, if the original names were VeryLow, Low, Medium, High and Very High,

that can be renamed as 1VeryLow, 2Low, 3Medium, 4High and 5Very High. Whether the order is on one direction or the opposite does not matter, generating the same analysis and the same results.

 Related Topics: *Advanced Features*, *Input Tab*, *Cut File*.
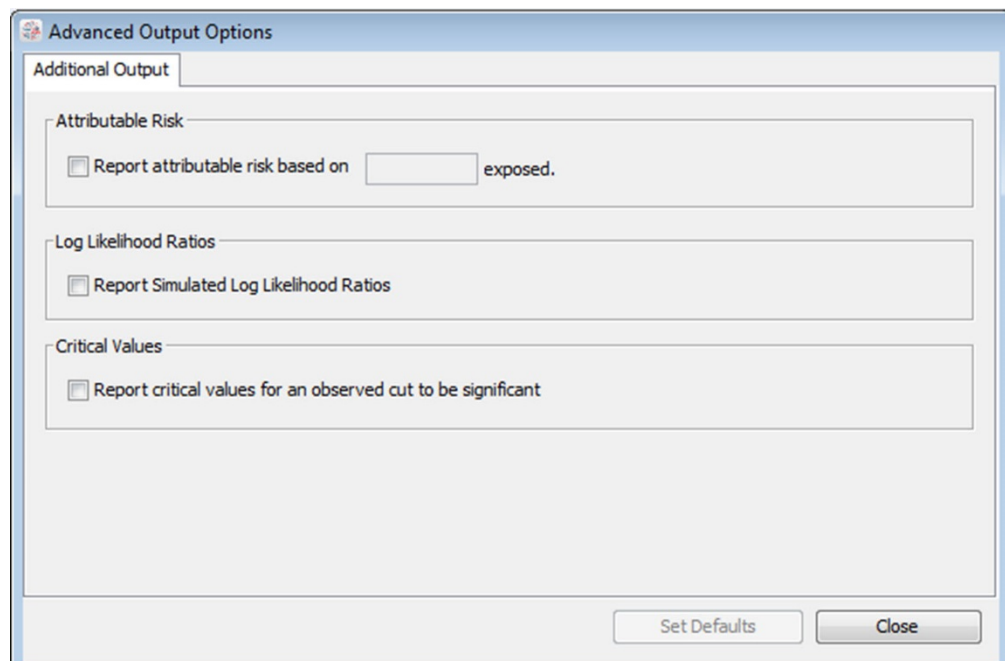
Additional Output Tab

Figure 11 - Additional Output Tab Dialog Box

Attributable Risk

For some analysis, it is of interest to calculate the attributable risk in addition to the relative risk. For this to work, it is necessary to specify the total number of exposed individuals, including those that may not have contributed any cases to the tree.

Related Topics: *Output Tab*.

Simulated Log Likelihood Ratios

If requested, TreeScan will print all the simulated log likelihood ratios to a comma delimited output file. This output is not needed for running or interpreting a TreeScan analysis. The option is available for statistical researchers who want to study the distributional properties of the tree-

based scan statistic. The name of this output file is the same as the name of the text output format file, but with a *.llr.csv filename extension.

Related Topics: *Output Tab*.

Critical Values

If requested, TreeScan will provide the critical values for rejecting the null hypothesis at the alpha = 0.05, 0.01 and 0.001 levels. This output is not needed for running or interpreting a TreeScan analysis. The option is available for statistical researchers who want to study the distributional properties of the tree-based scan statistic.

Related Topics: *Output Tab*.

Running TreeScan

Specifying Analysis and Data Options

The TreeScan program requires that you specify parameters defining analysis, input and output options for the analysis you wish to conduct. A tabbed dialog is provided for this purpose. To access the parameter tab dialog, either press the ➕ button or select the File/New menu item. Once the tabbed dialog has been opened, specify the parameters for your session on the following three main tabs:

- Analysis Tab
- Input Tab
- Output Tab

See the section on Basic TreeScan Features for instructions on how to fill in these tabs.

Most analyses can be performed using only these three tabs. For each tab, there are additional features that can be selected by first clicking on the Advanced button in the lower right corner of the tab. These additional features may be useful in special circumstances.

The available choices for some features may depend on what was selected in other places, and they will then be deactivated accordingly.

Related Topics: *Basic TreeScan Features*, *Sample Data Sets*, *Test Run*, *Input Tab*, *Analysis Tab*, *Output Tab*, *Advanced Features*, *Launching the Analysis*.

Launching the Analysis

Once the data input files have been created, and the parameters defining the analysis input and output options have been specified, select the Start Analysis ➡ button to launch the analysis and

produce the results file. Once the analysis has been completed, the standard results file will appear in the job status window.

Multiple parameter session windows may be opened simultaneously for data entry, and multiple analyses may be run concurrently. If you are running multiple analyses concurrently, please make sure that the output files have different names, or they will overwrite each other.

 Related Topics: *Input Data*, *Test Run*, *Specifying Analysis and Data Options*, *Status Messages*, *Warnings and Errors*, *Computing Time*, *Batch Mode*.

Status Messages

Status messages are displayed as the program executes the analysis, as the data is read, and at each step of the analysis. Normal status messages are displayed in the top box of the job status window. Warnings and error messages are displayed in the bottom box of the job status window. Upon successful completion of the calculations, the standard results file will be shown in the job status window.

Related Topics: *Launching the Analysis*, *Warnings and Errors*.

Warnings and Errors
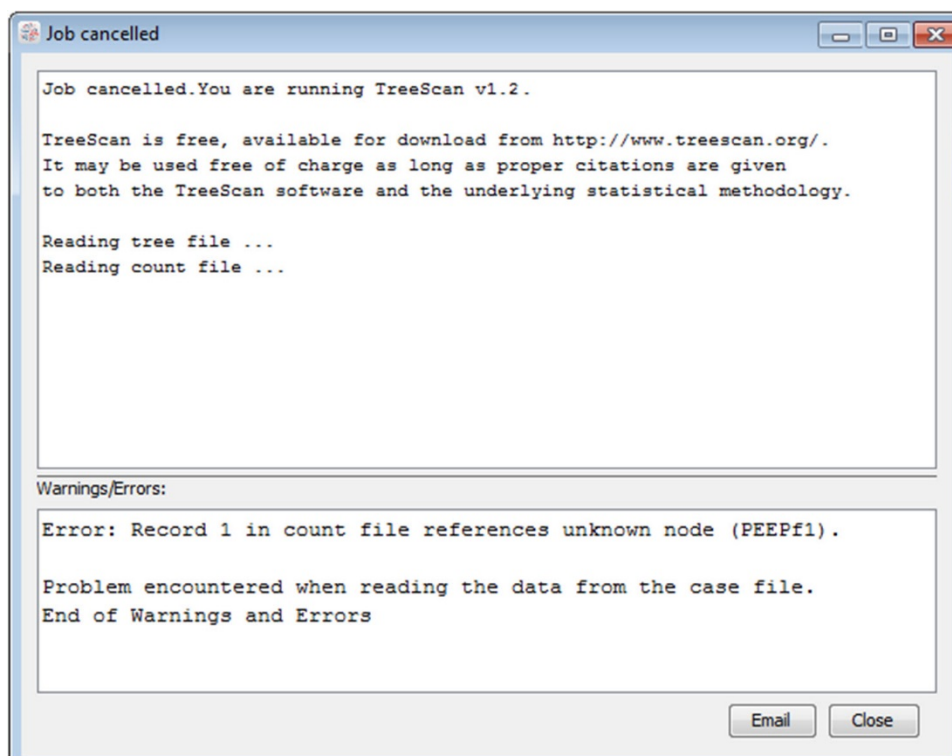
Figure 12 - TreeScan Status Messages and Warnings/Errors Dialog Box

Warning Messages

TreeScan may produce warnings as the job is executing. A message is then displayed in the Warnings/Errors box on the bottom of the job status window. A warning will not stop the execution of the analysis. If a warning occurs, please review the message and access the help system if further information is required.

Error Messages

If a serious problem occurs during the run, an error message will be displayed in the Warnings/Errors box on the bottom of the job status window and the job will be terminated. The user may resolve most errors by reviewing the message and using the help system.

One of the most common errors is that the input files are not in the required format, or that the file contents are incompatible with each other. When this occurs, an error message will be shown specifying the nature and location of the problem. Such error messages are designed to help with data cleaning.

If the error message cannot be resolved, you may press the email button on the job status window. This will generate an automatic email message to TreeScan technical support. The contents of the "Warnings/Errors" box will be automatically placed in the e-mail message. All a user needs to do is press their e-mail Send key. Users may also print the contents of the Warnings/Errors box and even select, copy (ctrl c) and paste (ctrl v) the contents if necessary.

Related Topics: *Input Data, Data Requirements, TreeScan Support.*

Saving Analysis Parameters

Analysis parameters, specified on the Parameter tab dialog, can be saved and reused for future analyses. It is recommended that you save the parameters with a ".prm" file extension. The parameter file is stored in an ASCII text file format.

Save Analysis Parameters

1. If the parameters have not previously been saved, select Save As from the File menu. A 'Save Parameter File As' dialog will open.

2. Select a directory location from the 'Look In' drop-down menu at the top of the dialog box.

3. Enter a name for your parameter file in the 'File Name' text box. It is recommended that the 'Save As Type' selection remain as Parameter Files (*.prm).

4. Press the Save button.

Once the parameter file is initially saved, save changes to the file by selecting `Save' on the File menu. The file will save without opening the `Save Parameter File As' dialog.

Open a Saved Parameter File

1. Select 'Open' from the File menu or click on the ![folder icon] button in the toolbar. A Select Parameter File dialog will open.

2. Locate the desired file using the Look in drop-down menu.

3. Once the file is located, highlight the file name by clicking on it.

4. Press the Open button.

A Parameter tab dialog will open containing the saved parameter settings. The location and name of the parameter file is listed in the title bar of this dialog.

Related Topics: *Specifying Analysis and Data Options*, *Basic TreeScan Features*, *Advanced Features*, *Batch Mode*.

Parallel Processors

If you have parallel processors on your computer, TreeScan will take advantage of this by running different Monte Carlo simulations using different processors, thereby increasing the speed of the calculations. The default is that TreeScan will use all processors that the computer has. If you want to restrict the number, you can do that by clicking on Session > Execute Options, and selecting the maximum number of processors that TreeScan is allowed to use.

Related Topics: *Results*.

Batch Mode

TreeScan is most easily run by clicking the Start Analysis ![icon] button at the top of the TreeScan window, after filling out the various parameter fields in the Windows interface.

An alternative approach is to skip the windows interface and launch the TreeScan calculation engine directly by either:

1. Dragging a parameter file onto the 'treescan32.exe' executable.

2. Writing 'treescan32.exe *.prm' in a batch file or at the command prompt, where *.prm is the name of the parameter file.

By using the batch mode version, it is possible to write special software that incorporates the TreeScan calculation engine with other applications. To use TreeScan in this manner requires a reasonable amount of computer skill and sophistication.

When running TreeScan in batch mode, it is easiest to first create the parameter file using the TreeScan windows interface, and then save that file. Changes can be made in the same way, but it is also possible to change the parameter manually using any text editor or automatically by using some other software product. If only a few parameters should change compared to what is in an existing parameter file, name the parameter file on the command prompt together with instruction on which parameters should change. The command line parameter values will then over-ride the parameter values specified in the parameter file.

When the batch mode version of TreeScan is run, the standard results file does not automatically pop up on the screen, but must be opened manually using any available text editor such as Notepad.

Related Topics: *Launching the Analysis*, *Basic TreeScan Features*, *Advanced Features*, *Saving Analysis Parameters*.

Computing Time

The tree-based scan statistic can be computer intensive to calculate. The computing time depends on a wide variety of variables, and depending on the data set and the analytical options chosen; it could range from a few seconds to several days or weeks. The main variables that increase the computing time are the number of nodes and the total number of observed cases. For the tree-temporal scan statistic, the number of time periods and the range of the window start and end times will also influence the computing time. Unconditional analyses are typically faster than conditional analyses.

Related Topics: *Memory Requirements*.

Memory Requirements

TreeScan uses dynamic memory allocation. If there is insufficient memory available on the computer to run the analysis, there are several options available for working around the limitation:

- Close other applications.
- Aggregate the data into fewer nodes.
- Aggregate the temporal data into fewer time periods.
- Run the program on a computer with more memory.

It is highly desirable that there is sufficient RAM to cover all the memory needs, as TreeScan runs considerable slower when the swap file is used, so these techniques may also be used to avoid the swap file.

Related Topics: *Computing Time*, *Warnings and Errors*.

Results

As output, TreeScan always creates a standard plain text based results file in ASCII format. This file is automatically shown after the analysis is complete. It can also be opened using any text editor, such as NotePad or Word. Two additional optional output files with the same information can be generated in HTML and comma delimited CSV formats respectively. If the former is requested, it is automatically launched when the analysis is done.

Related Topics: *Output Tab.*

Standard Results File (*.txt)

The standard results file is automatically shown after the calculations are completed. It is fairly self-explanatory, but for proper interpretation it is recommended to read the section on statistical methodology, or even better, one of the methodology papers listed in the bibliography.

> **SUMMARY OF DATA**
>
> Provides data concerning the tree as a whole. For example, it includes information about the size and depth of the tree and the total number of cases. Use this to check that the input data files contain the data that you expect it to contain.
>
> **MOST LIKELY CUTS**
>
> Summary information about the most likely cuts, that is, the tree branches with the cluster of cases that are least likely to be due to chance. They are presented in order by their likelihood, from larger to smaller; that is, from stronger to weaker clusters.
>
> **Node Identifier:** This is the node below where the cut was made.
>
> **Node Cases:** The total number of cases in the node regardless of whether they are inside or outside the time window of the cluster. Only reported for the tree-temporal scan statistic, as it is otherwise identical to the number of cases.
>
> **Time Window:** The temporal window of the cluster. Only reported for the tree-temporal and purely temporal scan statistics.
>
> **Cases:** The number of observed cases in the cluster. For a purely tree-based scan statistic, this is the number of observed cases in the node where the cut was made. For the tree-temporal scan statistic, it is the number of observed cases in the node that are also within the time window of the cluster.
>
> **Expected:** The number of expected cases in the cluster, under the assumption that the null hypothesis is true, and conditioned on whatever the analysis is conditioned on. The mathematical formulas are provided below.
>
> **Relative Risk:** The risk in the cluster divided by the risk outside the cluster. For the Poisson and Bernoulli models, outside means for other parts of the tree. For the tree-temporal

and purely temporal scan statistics, outside means outside the cluster time window but inside the same node. The formula is different for different probability models. See below.

Excess Cases: The excess number of cases. The formula is different for different probability models. See below.

Attributable Risk (optional): The attributable risk is the excess number of cases divided by the total number of people exposed that were specified on the Additional Output Tab.

Log Likelihood Ratio: The natural logarithm of the likelihood ration for the cut. This is the test statistic, and a larger value is evidence for a cluster and against the null hypothesis.

P-value: The p-values are adjusted for the multiple testing stemming from the multitude of cuts evaluated. This means that under the null-hypothesis of complete randomness there is a 5% chance that the p-value for the most likely cut will be smaller than 0.05 and a 95% chance that it will be bigger. Under the null hypothesis there will always be some area with a rate higher than expected just by chance alone. Hence, even though the most likely cut always has an excess rate, the p-value may actually be very close or identical to one.

PARAMETER SETTINGS

A list of the parameter settings used for the analysis.

COMPUTATIONAL INFORMATION

Information about when the program was run, how long it took to execute, and the number of processors used.

Related Topics: *Output Tab*, *Additional Output Tab*, *Mathematical Formulas*, *HTML Results File*, *Comma Delimited Results File*.

Mathematical Formulas

Here we give the mathematical formulas for various result variables. First we define the following notation:

c = Number of cases in the cluster

C = Total number of cases, in the whole tree

n = Population in the node

N = Total population in the whole tree

p = Case probability (for the unconditional Bernoulli model)

w = Length of the temporal cluster window (for tree-temporal and temporal scan statistics)

T = Length of the data time range (for tree-temporal and temporal scan statistics)

z = Number of cases in the temporal cluster window, summed over the whole tree

Depending on the probability model used, the mathematical formulas are:

Type of Scan	Conditional	Expected	Relative Risk	Excess Cases
Poisson	No	n	$\dfrac{c}{n}$	$c - n$
Poisson	Total cases	$n\dfrac{C}{N}$	$\dfrac{c/n}{(C-c)/(N-n)}$	$c - n\dfrac{(C-c)}{(N-n)}$
Bernoulli	No, standard	np	$\dfrac{c}{np}$	$c - np$
Bernoulli	No, self-control	np	$\dfrac{c/p}{(n-c)/(1-p)}$	$c - p\dfrac{(n-c)}{(1-p)}$
Bernoulli	Total cases	$\dfrac{nC}{N}$	$\dfrac{c/n}{(C-c)/(N-n)}$	$c - n\dfrac{(C-c)}{(N-n)}$
Tree-Temporal	Node	$\dfrac{wn}{T}$	$\dfrac{c/w}{(n-c)/(T-w)}$	$c - w\dfrac{(n-c)}{(T-w)}$
Tree-Temporal	Node and time	$\dfrac{nz}{C}$	$\dfrac{c(C-n-z+c)}{(n-c)(z-c)}$	$c - \dfrac{(n-c)(z-c)}{(C-n-z+c)}$
Purely Temporal	Total cases	$\dfrac{wC}{T}$	$\dfrac{c/w}{(C-c)/(T-w)}$	$c - w\dfrac{(C-c)}{(T-w)}$

Related Topics: *Output Tab*, *Results*, *Standard Results File*.

HTML Results File (*.html)

The HTML results file has the same information as the standard ASCII results file, but it is presented as an HTML table. If selected, it is automatically launched when the analysis is completed. The name of the file is the same as the standard results file, except for the *.html extension.

Related Topics: *Output Tab*, *Results*, *Standard Results File*, *Comma Delimited Results File*.

Comma Delimited Results File (*.csv)

The Comma Delimited results file takes the information from each of the most likely cuts, presented as a comma delimited CSV table. This file is useful for importing the TreeScan output into other software. It needs to be opened manually. The columns are:

Cut: The index number of the most likely cuts, in decreasing order with respect to log likelihood ratios.

Node Identifier: The node identifier.

Node Cases: The total number of cases in the node. This is only reported for the tree-temporal scan statistic as it is identical to Cases for the other scan types.

Time Window Start: The start time of the detected cluster window. Only reported for the tree-temporal and purely temporal scan statistics.

Time Window End: The end time of the detected cluster window. Only reported for the tree-temporal and purely temporal scan statistics.

Cases: The observed number of cases in the cluster.

Expected Cases: The expected number of cases in the cluster, when the null hypothesis is true, and conditioned on whatever the analysis was conditioned on.

Relative Risk: Relative risk.

Excess Cases: The excess number of cases.

Attributable Risk: Attributable risk.

Log Likelihood Ratio: The log likelihood ratio.

P-value: The p-value for the cluster.

The name of the file is the same as the standard results file, except for the *.csv extension.

Related Topics: *Output Tab*, *Results*, *Standard Results File*, *HTML Results File*.

Simulated Log Likelihood Ratios File (*_llr.csv)

The log likelihood ratio test statistics that were calculated from each of the random data sets are not provided as part of the standard output. As an advanced option, they can be printed to a special file. There is no need for this file when doing a regular analysis, but it can be interesting for statistical researchers who are studying the distributional properties of the tree-based scan statistic under various scenarios. By default it has the same name as the output file but with the extension *_llr.csv

Related Topics: *Output Tab*, *Results*.

Miscellaneous

New Versions

To check whether there is a later version of TreeScan than the one you are currently using, simply click on the update button ![update button] on the tool bar. If a newer version exists, you will be asked whether you want to download and install it. You can request that TreeScan automatically checks for new versions once a week, once a month or every time TreeScan is used. Alternatively, you can set TreeScan to only check for new versions manually, when you decide to do so.

At any given time, it is also possible to download the latest version of TreeScan at 'www.treescan.org'.

Related Topics: *Download and Installation*.

Random Number Generator

The choice of random number generator is critical for any software creating simulated data. TreeScan uses a Lehmer random number generator[12] with modulus $2^{31}-1 = 2147483647$ and multiplier 48271, which is known to perform well.[14]

Related Topics: *Monte Carlo Replications*.

Contact Us

Please direct technical questions about installation and running the program, as well as the web site, to:

support@treescan.org

Please direct substantive questions about the statistical methods and suggestions about new features to:

Martin Kulldorff
Department of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics
Harvard Medical School and Brigham and Women's Hospital
1620 Tremont Street, suite 3030, Boston, MA 02120, USA
Email: kulldorff@treescan.org

Acknowledgements

Comments and Suggestions

Feedback from users is greatly appreciated. Very valuable suggestions concerning the TreeScan software have been received from various individuals, including:

Judith Maro, Harvard Medical School and Harvard Pilgrim Health Care Institute

Michael Nguyen, Food and Drug Administration

Logo

The TreeScan logo was created by Bulkhead Design (www.bulkheaddesign.com).

Frequently Asked Questions

Input Data

1. **I tried running TreeScan using one of the sample data sets and all went well, but when I try it on my own data there is an error. What should I do?**

   TreeScan makes sure that the input data is compatible with each other, and with the options specified on the windows interface. For example, it complains if a node ID in the count file is not present in the tree file, because it must know where to assign those cases. For most data sets there is some need for data cleaning and TreeScan is designed to help with this process by spotting and pointing out any inconsistencies found.

2. **I have constructed the ASCII input files exactly according to the description in the TreeScan User Guide, but TreeScan complains that they are not in the correct format. What is wrong?**

   The most likely explanation is that the files are in UNICODE rather than ASCII format. Just convert to ASCII and it should work.

3. **In my data, there is zero or only one case in many nodes. Can I use TreeScan for such sparse data?**

   Yes, you certainly can. One of the main reasons for using TreeScan is to avoid arbitrary aggregation of the data, letting the scan statistic consider different smaller or larger aggregations by considering different cuts on the tree. With finer resolution of the input data, TreeScan can evaluate more different cluster locations and sizes without restrictions imposed by higher level groupings.

4. **What is the minimum size of the tree needed to run TreeScan?**

   Technically, the tree-based scan statistic can be run using only two nodes, providing correct inference. Unless there is a time component, there is no point using scan statistic for such data though, for which a regular chi-square statistic can be used instead, as there is no multiple testing to adjust for. With three nodes or more, the fundamental scan statistic concept of including different combinations of locations into the potential clusters is being utilized. In most practical applications though, the tree-based scan statistic is used for data sets with hundreds or thousands of nodes.

   Regarding the tree-temporal scan statistic, it can make sense to run it with only two nodes on the tree, in which case it will look for temporal clusters in either or both of the nodes.

5. **When should I use the Bernoulli versus the Poisson model?**

   Use the Bernoulli model when you have binary data, such as cases and controls in a matched design or exposed and unexposed time periods in a self-control design. Use the Poisson model when you have cases and expected counts from a historical or concurrent comparison population at risk.

6. **I have memory problems when running TreeScan. What should I do?**

   Make sure you are running TreeScan in 64-bit mode. For this you must (i) have a 64 bit computer, and (ii) have 64-bit Java installed on your computer.

Results

7. **I get an error stating that the output file could not be created. Why?**

   In Windows, permission to write to the "Program Files" folder is given only to administrators and power users of that machine. If the output file path includes the "Program Files" folder and you do not have administrative or power user privileges on your computer, Windows prevents TreeScan from creating the output file in the designated location. The solution is to specify a different directory for the output file.

8. **Since the TreeScan results are based on Monte Carlo simulated random data, why are the p-values the same when I run the analysis twice?**

   All computer-based simulations are based on pseudo-random number generators. When the same seed is used, exactly the same sequence of pseudo-random numbers will be generated. Since TreeScan uses the same seed for every run, you obtain the same result for two runs when the input data is the same.

9. **In the results for a Poisson analysis, the population is smaller than the number of cases. How can that be?**

   When using the Poisson model, the 'population' could be a variety of things. If it is the total number of people exposed to a drug, then the population must be greater or equal to the number of cases. Population could also be expected counts, calculated using some other statistical software; it could be person-years; or it could be the number of million people exposed. In these three scenarios, the observed cases can be less than the population. Note that, with the conditional Poisson model, results will be the same if you multiply all population counts with the same constant.

Operating Systems

10. **Is TreeScan available for Windows/Mac/Linux?**

    The TreeScan software is available for Linux, Mac and Windows, and all three versions can be downloaded from the www.treescan.org web site.

TreeScan Bibliography

Suggested Citations

The TreeScan software may be used freely, with the requirement that proper references are provided to the scientific papers describing the statistical methods. Depending on the application, the suggested citations can be found among the methodological papers below.

Methodology Papers

General Statistical Theory, Poisson Model

1. Kulldorff M, Fang Z, Walsh S. A tree-based scan statistic for database disease surveillance. Biometrics, 2003,59:323-331.

Bernoulli Model, Conditional and Unconditional

2. Kulldorff M, et al. Manuscript in preparation.

Tree-Temporal Scan Statistic

3. Kulldorff M, et al A Tree-Temporal Scan Statistic for Drug and Vaccine Safety Data Mining. Manuscript in preparation.

Applications by Field of Study

Drug Safety Surveillance

4. Kulldorff M, Dashevsky I, Avery TR, Chan KA, Davis RL, Graham D, Platt R, Andrade SE, Boudreau D, Gunter MJ, Herrinton LJ, Pawloski P, Raebel MA, Roblin D, Brown JS. Drug safety data mining with a tree-based scan statistic. Pharmacoepidemiology and Drug Safety, 2013, 22:517-523.
5. Brown JS, Petronis KR, Bate A, Zhang F, Dashevsky I, Kulldorff M, Avery TR, Davis RL, Chan KA, Andrade SE, Boudreau D, Gunter MJ, Herrinton L, Pawllowski PA, Raebel MA, Roblin D, Smith D, Reynolds R. Drug adverse event detection in health plan data using the gamma Poisson shrinker and comparison to the tree-based scan statistic. Pharmaceutics, 2013,5:179-200.

Vaccine Safety Surveillance

6. Kulldorff M, et al A Tree-Temporal Scan Statistic for Drug and Vaccine Safety Data Mining. Manuscript in preparation.

Occupational Disease Surveillance

7. Kulldorff M, Fang Z, Walsh S. A tree-based scan statistic for database disease surveillance. Biometrics, 2003,59:323-331.

Other References Mentioned in this User Guide

8. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press, 1984.

9. Dwass M. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics, 1957; 28:181-187

10. Glaz J, Balakrishnan N (editors). Scan Statistics and Applications. Birkhäuser: Boston, 1999.

11. Glaz J, Naus JI, Wallenstein S. Scan Statistics. Springer Verlag: New York, 2001.

12. Glaz J, Pozdnyakov V, Wallenstein S (editors). Scan Statistics: Theory and Applications. Birkhäuser: Boston, 2009.

13. Lehmer DH. Mathematical methods in large-scale computing units. In Proceedings of the second symposium on large scale digital computing machinery. Cambridge, USA: Harvard Univ. Press, 1951.

14. Naus J. The distribution of the size of maximum cluster of points on the line. Journal of the American Statistical Association, 60:532-538, 1965.

15. Park SK, Miller KW. Random number generators: Good ones are hard to find. Communications of the ACM, 31:1192-1201, 1988.

Index