SampleOptimizer™
and
SampleTracker™

Version 2.0

Reference Manual

1217 Bandana Boulevard North • St. Paul, Minnesota 55108 • Phone (651) 644-8080 • Fax (651) 647-0888
www.summite.com

Table of Contents

Table of Figures

Introduction

Welcome to **SampleOptimizer™** and **SampleTracker™**!

SampleOptimizer™ represents the latest evolution in long term monitoring optimization (**LTMO**) software. For the first time, the power of true mathematical optimization has been applied to LTMO in an easy-to-use desktop software tool to reduce sampling redundancy. *SampleOptimizer™* is the culmination and combination of years of research in mathematical optimization, data analysis, and environmental engineering. Experts in the fields of geology, hydrogeology, computer and environmental engineering, geostatistics, contaminant geochemistry, and remedial system optimization have applied their technical and field expertise to develop the sound approaches implemented in the software.

SampleTracker™ is an additional software module that reviews new monitoring data against historical data. The software identifies cases where current data deviates from expectations that are based on the historical dataset.

ModelBuilder™ is an additional component within the software that is utilized by the *SampleOptimizer™* module and, in some cases, by the *SampleTracker™* module. *ModelBuilder™* has two sections: one for model fitting, visualization, and analysis, and another for visualizing uncertainty.

For your convenience, we have created a **Quick Start Guide** and accompanying tutorial files which can be accessed from the Start Menu along with the rest of the software executables, or from the Help menu in the software.

We hope that you find the software to be highly effective in your task of optimizing the monitoring of environmental data at your site. We welcome all feature suggestions and comments at our website http://www.sampleoptimizer.com.

Environmental Security Technology Certification Program (ESTCP)

As described on its website, ESTCP is a Department of Defense (DoD) program that promotes innovative, cost-effective environmental technologies through demonstration and validation at DoD sites. ESTCP's goal is to demonstrate and validate promising, innovative technologies that target the most urgent environmental needs of the DoD. These technologies provide a return on investment through cost savings and improved efficiency.

In late 2005, ESTCP funded a project to demonstrate and evaluate SampleOptimizer™ and SampleTracker™ on three DoD sites. The objective of this project was to demonstrate and validate the use of SampleOptimizer™ and SampleTracker™ for reducing costs and improving effectiveness of LTM through adaptive assessment while achieving remediation goals.

Substantial benefits are expected from the use of *SampleOptimizer*™ and *SampleTracker*™, including:

1. Realizing significant cost savings by eliminating redundant data;

2. Enabling new data to be assessed for significant deviations and other features of interest without substantial labor, a benefit that will become even greater as more facilities move into long-term monitoring modes and/or emerging sensor technologies produce larger volumes of data to be analyzed; and

3. Providing a framework for effective implementation of an adaptive approach to monitoring, enabling limited resources to be efficiently directed where the greatest benefits are likely to be incurred.

The version number of the software as used in the ESTCP project is 2.0.

Credits and Thanks

Summit Envirosolutions, Inc. would like to extend our gratitude to those who have been instrumental to the development and testing of this software product:

- **Dr. Barbara Minsker**, professor of Civil and Environmental Engineering at the University of Illinois and former president and founder of HMSI, for organizing the software development effort beginning with her research group.
- The **National Center for Supercomputing Applications** (NCSA), especially Michael Welge, David Goldberg, Loretta Auvil, Colleen Bushell, Lisa Gatzke, and Tom Redman, for creating D2K, a data mining and modeling toolkit, as well as EMO, a multi-objective analysis toolkit. Both of these tools were used to create the initial versions of SampleOptimizer™.
- The **U.S. Department of Defense**, for funding the ESTCP demonstration and validation of this product.
- **Dennis Beckmann**, BP Remediation Engineering and Technology, for funding and contributing to a large part of the software development and testing effort.
- **Riverglass, Inc.** for administering the projects to develop and apply the initial version of this software to two BP test sites, and for licensing D2K to be used in this product.
- **Dr. Patrick Reed**, professor at Penn State, for developing the concept of using Genetic Algorithms (GA's) to optimize monitoring network design while he was a member of Dr. Minsker's research group.
- **Dr. Meghna Babbar-Sebens**, former member of Dr. Minsker's research group and former consultant to Riverglass**, Peter Groves**, former Riverglass software developer, and **David Clutter**, former NCSA researcher, for developing the software prototype using D2K.
- **Matthew Zavislak**, former member of Dr. Minsker's research group, former consultant to Riverglass and HMSI, and a lead software engineer at Summit Envirosolutions, for developing the prototype D2K-based design into its current state as a desktop application.
- **John Dustman**, the founder of Summit, for adding his knowledge of industry, environmental applications, data management, and data visualization into the software.
- **Dr. Abhishek Singh**, former member of Dr. Minsker's research group, for his assistance with GA troubleshooting, verification, and tuning.
- **David Tcheng**, NCSA researcher, for his help in utilizing D2K.
- **Dr. Tim Ellsworth**, professor at the University of Illinois, for his contribution to the development of the Ordinary Kriging algorithms used in the software.
- Lastly, many thanks to all the users who have shared their feature ideas and bug reports.

Site Suitability

Before utilizing *SampleOptimizer™* to optimize the sampling on your site, it is important to review these baseline requirements and recommendations to make sure your site is suitable for use with this software suite. Your site must have:

- Sampling analytical results (data) which can be
  - Entered into a spreadsheet, or
  - Exported from a database into a spreadsheet

- A minimum of 15 sampling locations (20 recommended).

- A minimum of 4 sampling events of historical data per sampling location (8 recommended) for temporal or spatio-temporal optimization. Only one event is needed for spatial optimization.

- See the SampleOptimizer™ EDD for more information about the data requirements.

It is recommended that *SampleOptimizer™* be used at sites where potential exists for negotiation of an alternate sampling plan with the relevant regulatory agency. This is not necessary if only *SampleTracker™* will be used, or if the site is not under regulatory guidance or control.

In order to use *SampleTracker™* on your site it must have:

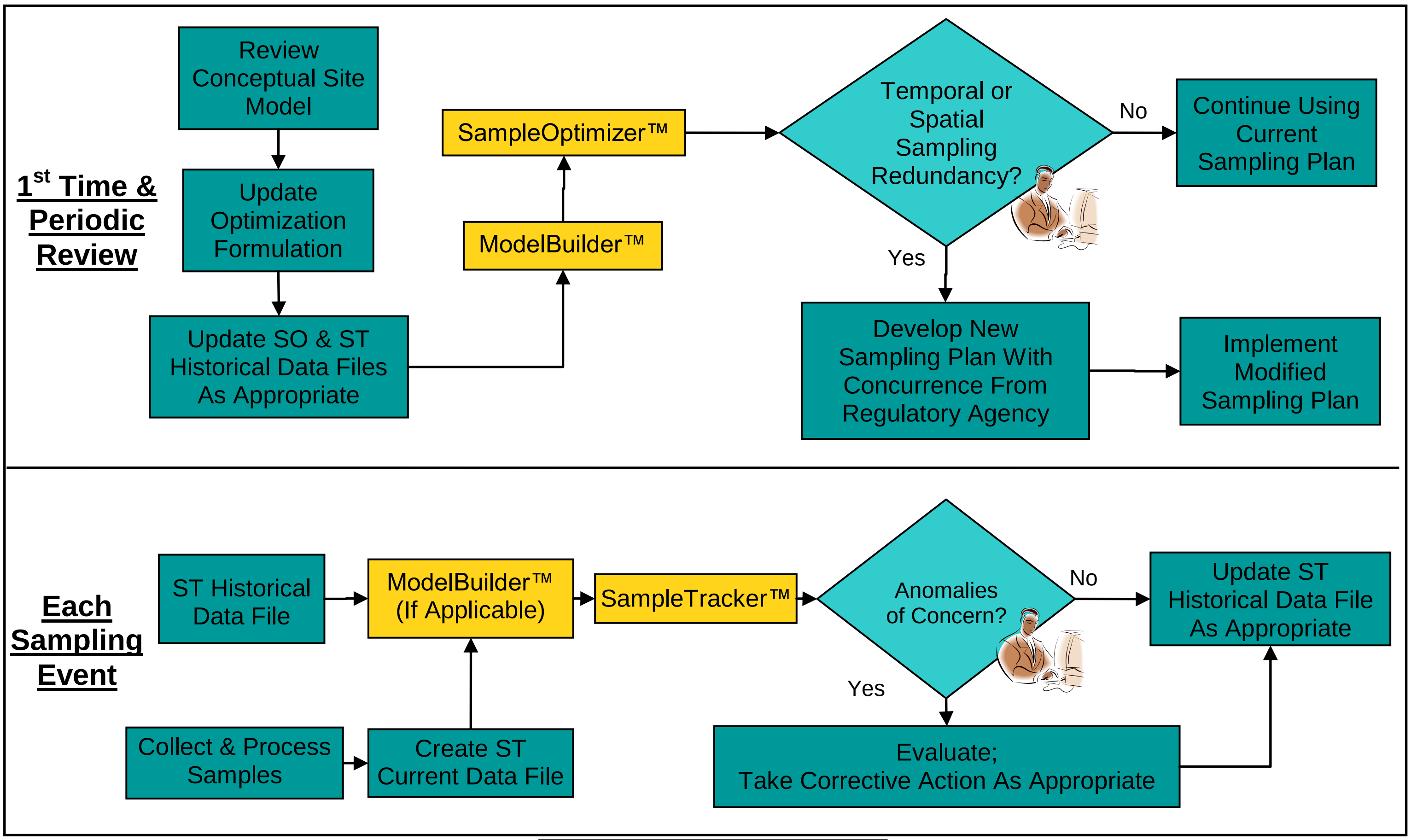
- One or more sampling locations with at least four samples per sampling location.

- In order to enable mass metric tracking with *SampleTracker™* via *ModelBuilder™*, your site must meet the minimum data density requirements required by *SampleOptimizer™*.

- Please see the SampleTracker™ EDD for more information about the data requirements.

LTMO Process Flow

The following is a flow chart describing the steps involved in LTMO using *SampleOptimizer™* and *SampleTracker™*.



Legend
SO = SampleOptimizer™
ST = SampleTracker™

Figure 1. Overview of Optimization and Tracking Process

Many of the above steps will be covered in detail later in this reference; below we provide details for steps not otherwise mentioned in this document.

GA Basics

(Adapted from "Introduction to Genetic Algorithms")

Genetic algorithms (GA's) are inspired by Darwin's theory about evolution.

The GA begins with a randomly-generated set of solutions (in *SampleOptimizer™* these are called **plans**), called a population. Each solution is composed of chromosomes; in *SampleOptimizer™* each chromosome represents a sampling decision for a sampling location. For example, a chromosome could say to "Sample MW-01 Semi-Annually".

Solutions from one population are taken and used to form a new population (also called a new **generation**). This is motivated by a hope that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. After selection, new solutions are created from the selected solutions through operations called **mutation** and **crossover**.

After the fitness of every solution in a generation has been evaluated, *SampleOptimizer™* shows the population on a scatter plot referred to as the **Pareto front** or **tradeoff curve**. This plot shows how well the solutions perform relative to two or more user-selected measures of fitness (objective functions).

This process is automatically repeated until the appropriate number of generations has been completed.

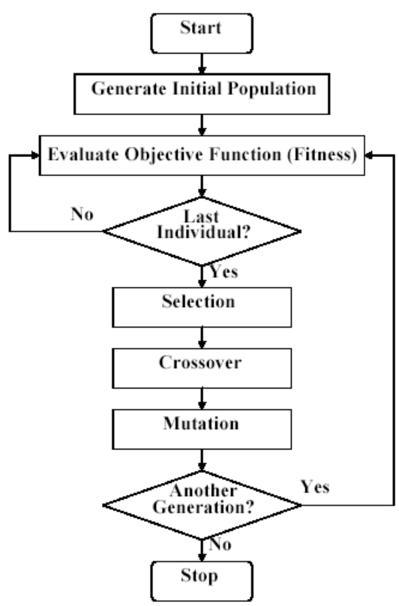


Figure 2. GA Flowchart (created by Professor Chunmiao Zheng, U. of Alabama)

SampleOptimizer™

- **Create / Review Conceptual Site Model (CSM)**
  - Before sending a set of data to the optimizer, it is necessary to check the site data. For example, with groundwater optimization, look for wells which are very close to each other and have similar screened intervals and make sure that they have similar hydraulic head readings and COC concentration values. If there are large differences in such metrics between two or more sampling locations that are very close to each other, there could be a problem with the data.
  - If the CSM shows that your site has multiple hydraulic regions (for example, different aquifers), each region should be optimized separately.
  - Verify to the extent possible that the information is accurate (SiteID, coordinates, values, correct aquifer)
  - Make sure the data file being prepared for import is compliant with the applicable EDD (see SampleOptimizer™ EDD and SampleTracker™ EDD)
- **Identify Spatial or Temporal Redundancy**
  - After running the Optimizer, you should identify several potential new sampling plans from the tradeoff curve results generated by the *SampleOptimizer*™. We recommend that you select plans which provide a level of interpolation accuracy (low error) acceptable to your regulator, at a lower annual cost than the current plan.
  - Similarly, after running the *ModelBuilder*™, you should identify any areas with unusually high uncertainty which may be of potential interest for characterization according to the site's monitoring objectives.
- **Develop New Sampling Plan With Regulatory Agency**
  - Typically a regulator must approve changes to a site's sampling plan. In order to justify a new reduced sampling plan, you will need to provide evidence (such as the plume images generated by the *SampleOptimizer*™) that reduction or elimination of sampling at certain locations, possibly in combination with installation of new sampling locations or increased sampling frequency at other locations, will maintain or improve the fulfillment of the monitoring program objectives.

SampleTracker™

- **Results – Do Anomalies of Concern exist?**
  - The results of running *SampleTracker*™ may indicate that a COC was "out-of-bounds" at a location. In such a case, it will be incumbent upon the analyst using the software to decide if an out-of-bounds COC is an "Anomaly of Concern" needing further study and possible corrective action.
  - For example, a COC being found below the anticipated range will most likely not be a concern; a COC being found slightly above the anticipated range may not be a concern either.
  - Some possible reasons for anomalies:
    - Real emergencies requiring a change in remediation strategy, such as a breakdown in a containment or treatment system.
    - Bad data, due to errors during collection, faulty laboratory analysis, mislabeling of bottles, errant sample preparation, etc.
    - Real differences between actuality and expectation that do not require action, but do require readjustment of one's CSM or expectation, possibly related to a real change in geochemistry or hydrology.
- **Evaluate & Take Corrective Action As Appropriate**
  - After identifying one or more anomalies of concern, the analyst may need to notify additional parties who can decide if further action is appropriate, and what action will be needed.
- **Update ST Historical Data File**
  - Lastly, at the end of each sampling event and the analysis thereof, the analyst needs to update the *SampleTracker*™ historical data file.
  - All <u>appropriate</u> samples should be added to the historical data file. For example, samples which were found to be bad data should not be added.
  - If there are fewer than 8 historical samples for a given COC at a sampling location, it is recommended that new samples be added to the historical data until the requisite 8 historical samples have been gathered.
  - It is important that the analyst understands that adding additional samples to the historical data can cause the upper bound to increase over time if the added samples are of a high concentration.
    - A good rule of thumb to use when deciding whether or not to add a sample to the historical data is if a concentration at a sampling location would almost certainly never be a concern for the foreseeable future, then it can be safely added to the historical data – otherwise it may be advisable to leave it out.
  - During periodic review, it is advisable to check if changes to the CSM should be reflected in the historical data. For example, a value which was once thought to not be appropriate for the historical data could be added into it, and vice-versa, as appropriate for the "new" groundwater understanding.

Installation

Requirements

Computer requirements: Windows XP or Vista, 1 GB RAM (2 GB preferred), 1.5 GHz single-core processor (3.0 GHz dual-core preferred). The installer must be run from a user login with at least power-user permissions.

To use *SampleOptimizer*™ *2.0*, you must have a valid license file (.inst).

Currently the software is only available for Microsoft Windows XP and Vista. Operation under Windows 7 is likely but has not been fully tested. In addition, you must have Sun Java JRE 6.0 or better installed, as well as Microsoft .NET runtime 3.5 or better. Please refer to Appendix A for more information on downloading those tools if you don't already have them installed.

Running the Installer

Simply double-click on the installer file, **SampleOptimizerSetup.msi** when you have it downloaded and have a license file (which has an .inst extension) ready.

Running the Software

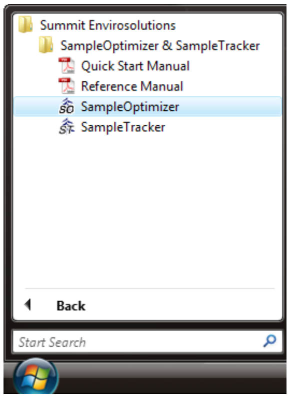


Figure 3. Running the Software (shown in Vista)

Upon running either *SampleOptimizer*™ or *SampleTracker*™ for the first time, you will be prompted to input your license file (ends in **.inst)**. You will not be able to proceed unless your license is validated by Summit's validation server. This procedure sends no personally-identifiable information and is conducted over a secure https connection. Please keep this license file in a safe location, as it will be required to re-install the software.

Please note that per the terms of your retail license you will only be able to install the software on the number of machines for which licenses were purchased. Multiple-boot systems require a license for each OS installation.

Tips and Tricks

Tool Tips

If you see an option in the software and are unsure about it, allow your mouse cursor to hover over it for a few seconds and a "Tool tip" will pop up and give you more information about that option.

Help Menu

At any time, you may access the Help Menu from which the user can open this Reference, the Quick Start, or the product web site.

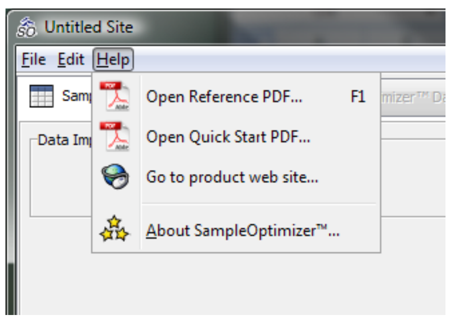


Figure 4. Help Menu

SampleOptimizer™

Getting Started

Before you begin using *SampleOptimizer*™, you will need a set of historical data to work with. The format currently supported for data input & output is .csv. CSV is a universal, standard, non-proprietary, and royalty-free format supported by nearly all spreadsheet and database software including *Microsoft's Excel* ™ & *Access* ™, and *OpenOffice.org's* free *Calc* and *Base*. It is very simple and CSV files can even be easily created by hand using a text editor such as Microsoft's *Notepad*.

Overview

SampleOptimizer™ strives to answer the following question:

How should this site be sampled in the future, in order to best fulfill our goals without spending more than is needed?

In this process, *SampleOptimizer*™ looks at a site's past data and makes decisions based on the assumption that overall trends in the past will continue into the future. For LTMO sites, this is very likely to be a valid assumption.

In order to determine if there is a sufficient amount of information about the site, the site's physical characteristics (e.g., hydrogeology) and regulatory constraints, as well as professional engineering judgment must be used together with the sampling data, plume maps, and other information provided by *SampleOptimizer*™ in order to make an overall determination as to the best course of future action at a site.

Electronic Data Deliverable (EDD)

The EDD for *SampleOptimizer*™ is designed to import site-specific sampling data in a simple cross-tab format. Here is an example.

Date	SiteID	EastCoordinate	NorthCoordinate	Benzene	Chlorobenzene
4/12/1995	BL003	222.5	768	4.10	240.00
4/12/1995	OS004	256	720.75	0.90	20.2
4/12/1995	OS003	441	6	1.20	180.30
4/12/1995	OS005	517.25	449	0.01	
3/13/1996	MWSL001	846	727	2.70	8.20

Figure 5. Example SampleOptimizer™ Data File

The first four columns (Date, SiteID, EastCoordinate, and NorthCoordinate) must have those exact column names in that exact order. Add or remove columns to the right of NorthCoordinate based on the actual number of COC's. COC column names can be anything you wish, but there cannot be more than one column with the exact same name. Also, for ease of display, it's best to keep COC name lengths to a minimum. For example, use "MTBE (ppb)" rather than "Methyl tertiary butyl ether (parts per billion)".

Important! *Each COC's units must be consistent within that COC. Different COC's can have different units, however, since they are analyzed individually by the interpolation model. For example, in* Figure 5, *benzene could be in μg/l, while chlorobenzene could be in mg/L.*

The coordinate system must remain constant throughout the site's data, and only one sampling location is allowed per EastCoordinate, NorthCoordinate pair. The program does not impose a minimum separation distance between locations as long as they are not numerically identical.

If your data contains two or more samples for a sampling location at a given time, you must decide which value you choose to input to SampleOptimizer™. Possible choices include the minimum, maximum, or average of the multiple samples. Here are some general recommendations:

- Average regular and field duplicate data values if neither value is an obvious outlier and the samples are equally within the working range of the analytical method.
- Do not average in quality assurance (QA) samples, usually labeled "Lab Replicate" or "DUP". Often, such values are not comparable and should therefore be ignored.
- If both values are non-detects with different reporting limits, use the result with the lower reporting limit (RL).
- If one value is a non-detect with a "typical" RL and the other a value, average half the RL and the value.
- If one value is a non-detect with a high RL and the other a value, take the value.

Report Dates

Sample dates must be arranged into discrete sampling events and labeled with one report date per sampling event, since the software will internally reference the data in quarters. This convention was chosen because sites are usually sampled relative to the quarters in each year.

Please note that only one report date per quarter is allowed and only one sample per COC per location per quarter is allowed. If the input data do not follow these requirements, the software will display an error message explaining the source of the problem, and the data file must be corrected before proceeding.

For reference, the quarter cutoff dates, which are inclusive of the end points, are as follows:

Q1:	January 1 to March 31	Q2:	April 1 to June 30
Q3:	July 1 to September 30	Q4:	October 1 to December 31

Non-Detects in SampleOptimizer™

- Zero values are not allowed. Instead, non-detect data with typical RLs should be replaced by a numerical substitute value which should be consistent for each COC. While there is no perfect non-detect substitute value, 1/20th RL has been found to work acceptably in many cases.

- If the typical RLs vary from event to event, using a common value such as 1/20th the median of typical RLs across events may be useful.

EDD Notes

- The maximum number of significant digits supported for coordinates and sample values is 15, in the approximate range of -1.79769E308 to 1.79769E308. Additionally, in extreme cases where the site area is astronomically large from a numerical standpoint, the software will show an error message that optimization will not be possible. This should never happen in practical circumstances.
- The supported date formats are as follows:
  - Leading zeros are OK as in 05/05/2007
  - 4-digit and 2-digit years are both supported
  - 5/15/2007
  - 5/15/2007 1:00:00 PM (Note: the time will be ignored)
  - 5/15/2007 13:00 (Note: the time will be ignored)
  - 15-May-07
- If a value for a COC is missing, simply leave the appropriate cell blank as in the example above.
- If you are using *Excel* to create the CSV, all columns & rows outside the valid data should have an "Edit > Clear > All" done to them, or else the exported CSV may contain a potentially large number of blank rows and/or columns which will cause errors during the data import process. You can check for excess columns/rows by opening the CSV in a text editor such as Microsoft Windows' *Notepad*.

Spatial Data, Temporal Data, and Spatio-Temporal Data

There are two categories of data which *SampleOptimizer™* can work with: spatial, and temporal/spatio-temporal. **Spatial** data is a set of COC concentration values for sampling locations at one reporting date. **Temporal** and **spatio-temporal** data contain COC concentration values for sampling locations at multiple reporting dates.

Depending on the available data for the chosen sampling locations at your site, it may be possible to perform either a spatial analysis or a temporal/spatio-temporal analysis.

If a spatial analysis is desired, a representative sampling event must be chosen from the historical data, stored in a properly formatted .csv file, and then input into the optimizer. Alternatively, you may choose to create an artificial sampling event based on the historical data, perhaps averaging locations' historical values or using the last value at each location, in order to create a representative sample value for each sampling location within the newly created artificial sampling event.

If a spatio-temporal or temporal analysis is desired, a properly formatted .csv file containing at least 4 or more representative sampling events must be input into the optimizer. These may be the actual historical data, or, at the discretion of the user, may be artificial sampling events constructed based on the actual historical data.

For further guidance on the construction of artificial sampling events, please contact one of our optimization consultants.

Loading Data

Upon running *SampleOptimizer™*, you will be presented with the following screen.

If you already have a saved **.site** file, you can load it with the File menu, drag it into the window area, or simply double-click on it.

To start a new site analysis, either click the Load button and select the data file, or simply drag the data file onto the window area. Alternatively, you can right-click on a .csv data file you have prepared, select "Open With", and then choose *SampleOptimizer™*.

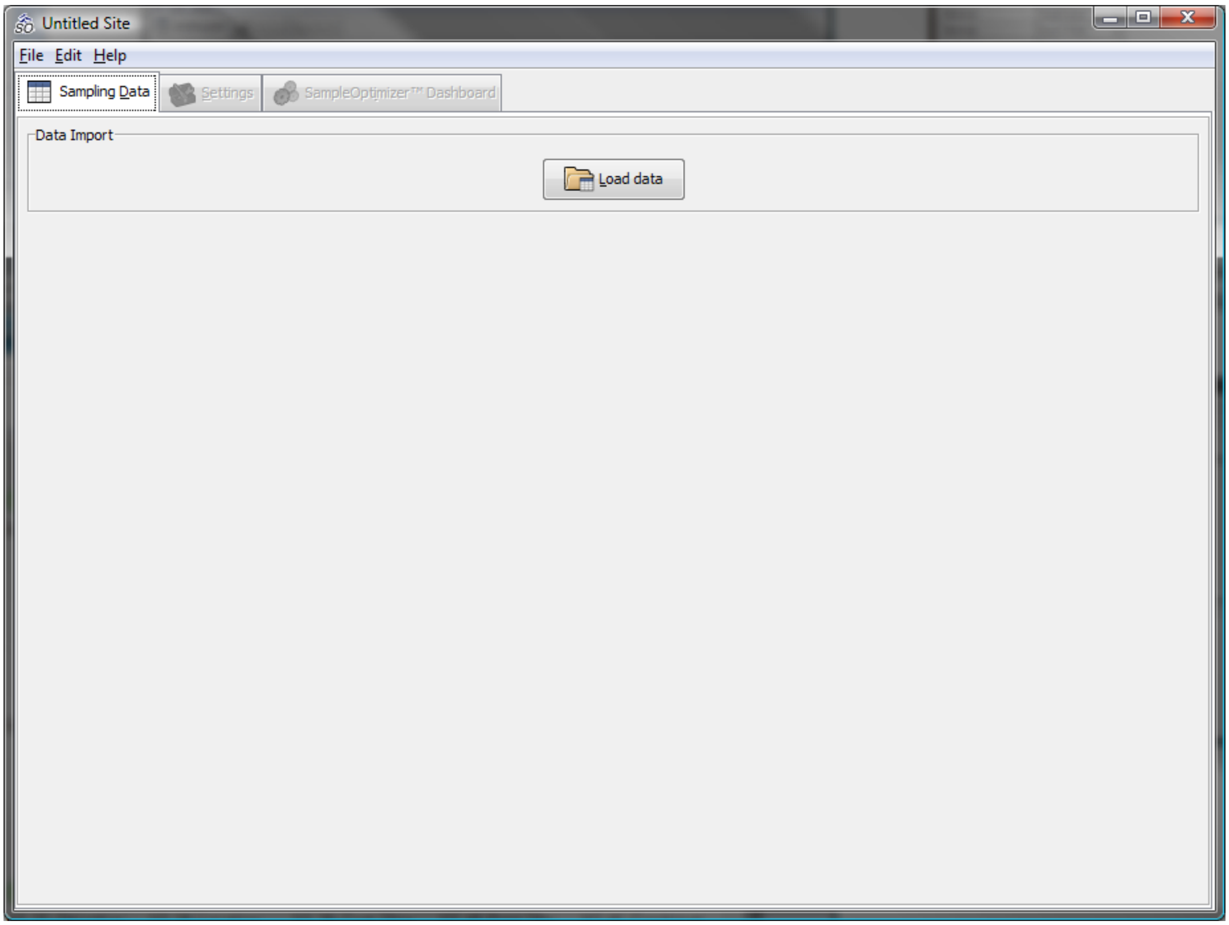


Figure 6. SampleOptimizer™ Initial Screen

If your data file is not in the correct format, you will receive an error dialog explaining the condition. You can then fix the problem in a spreadsheet editor program or text editor, and try to import it again.

If data density checking is enabled in **Application Options** (by default it is disabled), when you attempt to load a new data file which has low data density, you will see a warning screen which explains the source of the warning as well as options on how to proceed. Please note that if there are multiple sampling locations and/or COC's with low data density, you will see multiple confirmation dialogs unless you choose either the "Yes to all" or "No to all" options.

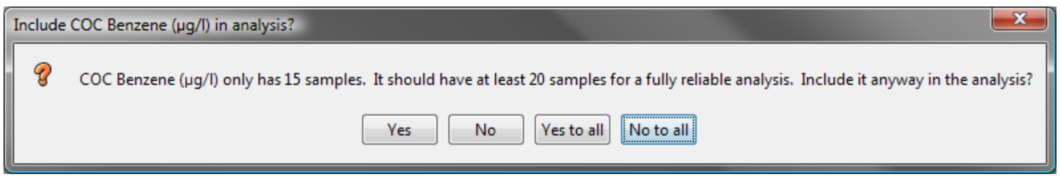


Figure 7. Data Density Warning Dialog

After successfully loading a data file or opening a saved site file, you will see the Data Summary screen. In Figure 8, a spatial dataset has been loaded.

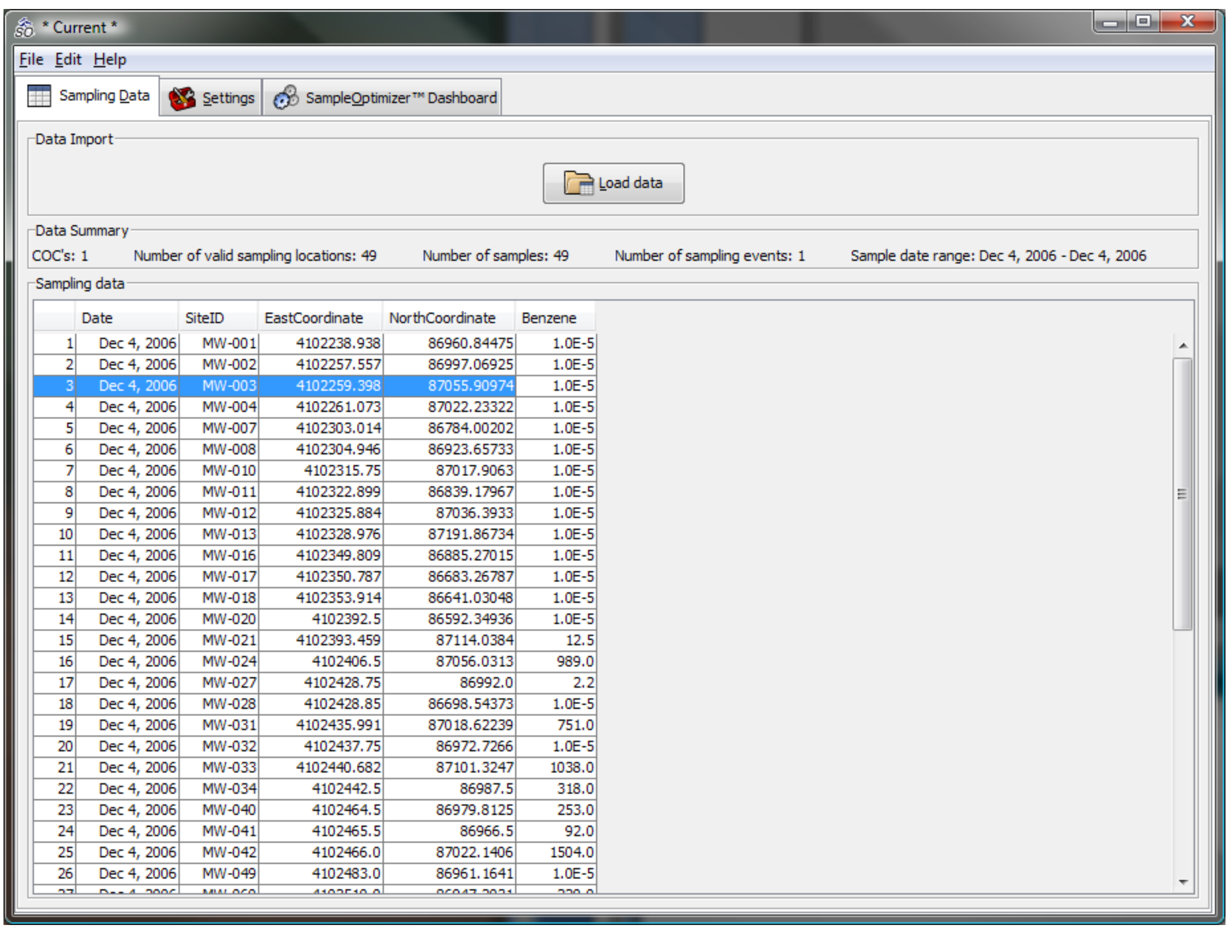


Figure 8. Spatial Data Loaded

If sampling locations, contaminants, or sampling events have been excluded from the analysis because of lack of sufficient data density, their samples will be grayed out. Figure 9 shows an example of a temporal dataset which has been loaded, but with some disabled samples (locations RW-03, RW-08, and RW-11 have been disabled).

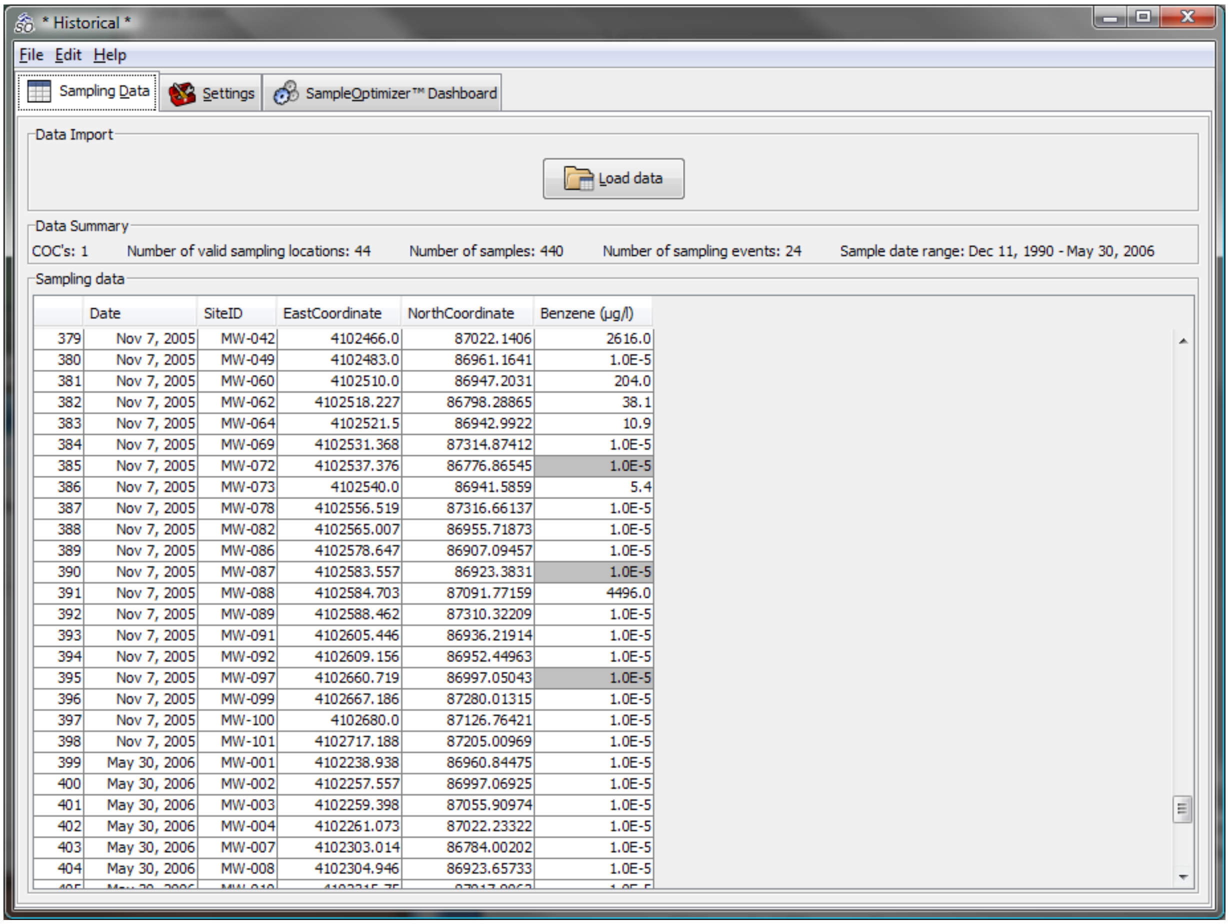


Figure 9. Temporal Dataset Loaded (with exclusions)

Data Density Requirements

For temporal/spatio-temporal analysis, *SampleOptimizer*™ will not include sampling locations in the analysis which have fewer than 4 samples, and for spatial and temporal/spatio-temporal analysis will not include sampling events which have fewer than 15 samples.

Application Options

For advanced users, there are some configuration options (in the **Edit** menu under **Application Options**) which are available to be configured during most phases of the software operation.

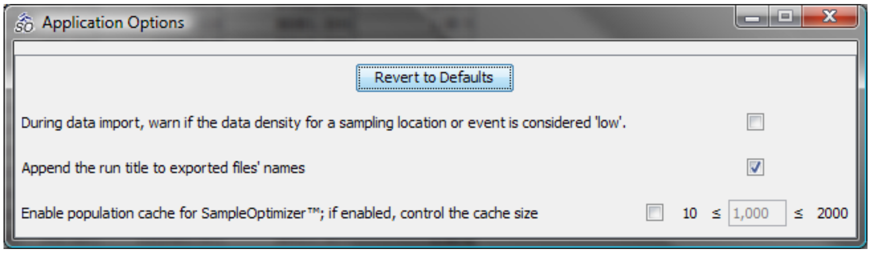


Figure 10. Application Options.

- Data input density warning (disabled by default)
  - If enabled, will warn the user when 15 to 19 (inclusive) valid, non-excluded samples are available in a sampling event, or for temporal/spatio-temporal analysis if between 4 and 7 (inclusive) valid, non-excluded samples are available for a sampling location. Normally the software includes such events or locations without warning, but the user can choose to be notified and have the opportunity to exclude such events or locations from the analysis.
- Append run title (enabled by default)
  - Appends the run title to the current date and site name when auto-generating file names for image and data exporting.
- SampleOptimizer™ population cache (disabled by default)
  - This is an experimental (beta) feature which in some cases has been found to reduce the computational time of the optimization process. However, this will increase memory consumption and may cause other tasks on your computer to slow down severely, especially if your installed memory is low and/or the population size is too large. The program may even crash if it runs out of memory. Therefore, use this feature at your own risk. Note: this feature does not affect model or plume image computation.

Settings

Next, in the **Settings** tab, you will see the following screen when performing a spatial optimization:

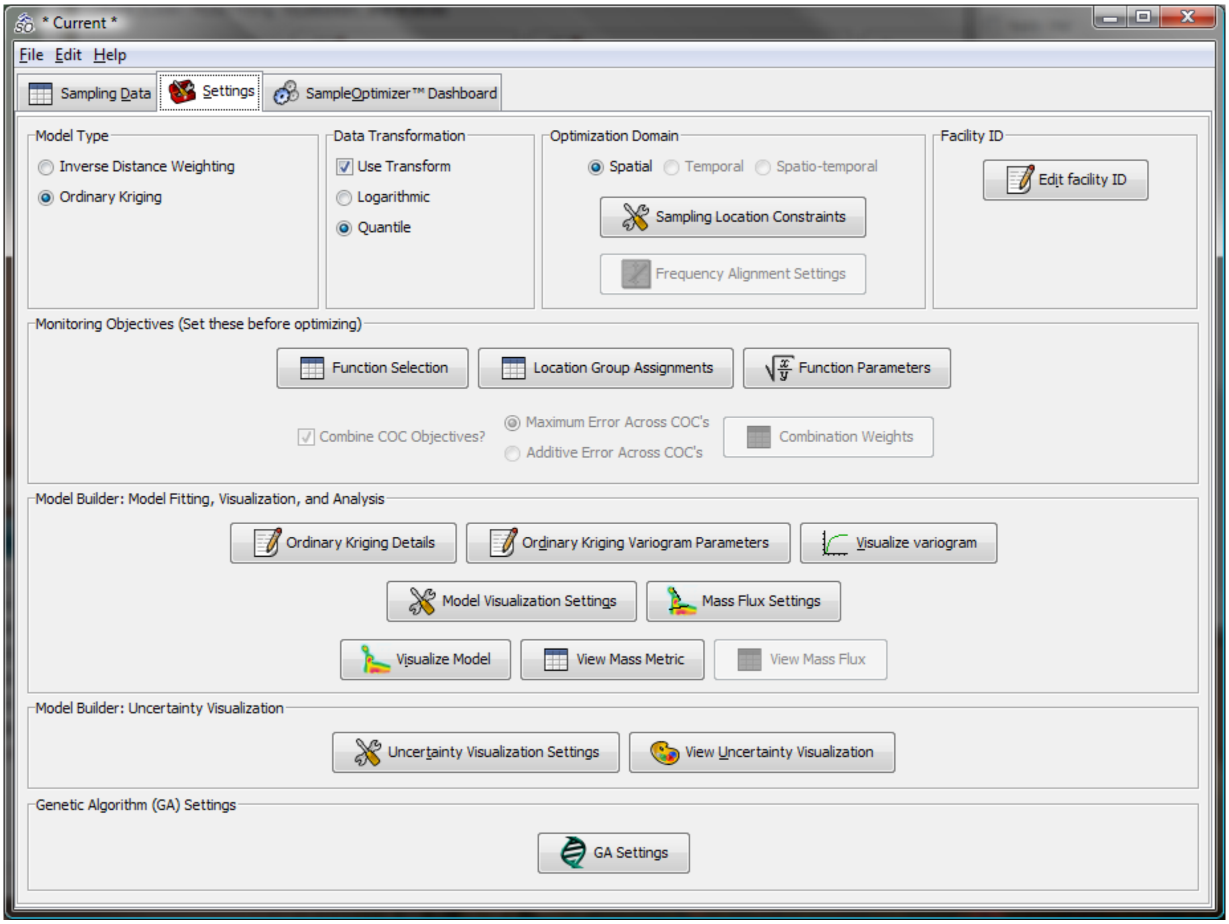


Figure 11. Spatial Data Settings

Please note that when working with a temporal/spatio-temporal dataset, **spatial** domain will be disabled, while **temporal** and **spatio-temporal** domains will be enabled.

This is the screen where you will be able to edit the settings for the optimization, and use **ModelBuilder™** to optimize the parameters for the model. The defaults will usually be a good place to start, and most or all settings will not have to be changed in most cases.

You should use the **Visualize Model** tool to verify that the plume interpolation is acceptable before performing an optimization. If you want to increase the resolution of the images generated, see ModelBuilder™ Settings and increase the setting for **# of vertical slices**.

Model Type

Two types of models are featured in *ModelBuilder™*: **Inverse Distance Weighting** and **Ordinary Kriging**. Details and configurable parameters for each model are given in the “Model Parameters” dialog. If you need additional information on these models, please consult a geostatistical reference book.

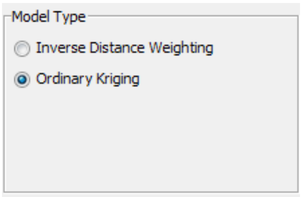


Figure 12. Model Type

In general, while Ordinary Kriging is more computationally-intensive and complex than Inverse Distance Weighting, it is also generally considered to be more statistically sound than Inverse Distance Weighting.

Data Transformation

Inverse Distance Weighting and Ordinary Kriging assume that the input data are normally distributed. Many datasets do not follow this assumption. In such cases, a data transformation should be applied. In *SampleOptimizer™*, transformations are applied to the data automatically as the data goes into and out of the genetic algorithm code. The two transformations offered by *ModelBuilder™* are **quantile transformation** and **logarithmic transformation**.

The quantile transformation creates statistics on the dataset and substitutes a sample’s quantile ranking in the data for its concentration value. The function used is equivalent to the PERCENTRANK function in *Microsoft Excel* ™.

The logarithmic transformation substitutes the natural log of a concentration for its concentration value.

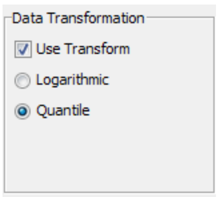


Figure 13. Data Transformation

Optimization Domain

Also see *Spatial Data vs. Temporal Data & Spatio-Temporal Data*

The optimization domain is not configurable when working with a spatial dataset. However, when working with temporal / spatio-temporal data, you can choose whether you would like to perform a temporal or spatio-temporal optimization.

Temporal optimization means that sampling locations' frequencies will be optimized, but no sampling locations can be turned off. Spatio-temporal optimization, on the other hand, allows sampling locations to be turned off. Location sampling recommendations are based on past performance. For example, if turning off a certain location nevertheless results in sufficiently accurate plume interpolation, that location will be recommended to be turned off.

How Spatio-Temporal / Temporal Analysis Works

ModelBuilder™ searches for a parameter set (Inverse Distance Weighting) or theoretical variogram (Ordinary Kriging) for each COC which works well for all historical sampling events. The Optimizer searches for monitoring plans which minimize the maximum errors across all sampling events and COC's. Frequencies are recommended based on their past performance. For example, if reducing sampling from semi-annual to annual at a given location nevertheless results in sufficiently accurate plume interpolation, annual sampling will be recommended for that location.

Since spatio-temporal optimization uses data from multiple sampling events and must cover multiple plume configurations across all the analyzed sampling events, it is typically produces more conservative results than a spatial analysis which only utilizes one sampling event. This effect can be minimized by configuring the objective functions to allow higher error for locations in the plume interior.

Sampling Location Constraints

Additionally, you can configure the **Sampling Location Constraints** for the optimization by clicking **Sampling Location Constraints** in the **Optimization Domain** area of the **Settings** tab.

Below are examples of the dialog box which will pop-up when **Sampling Location Constraints** is clicked, depending on which domain is selected.

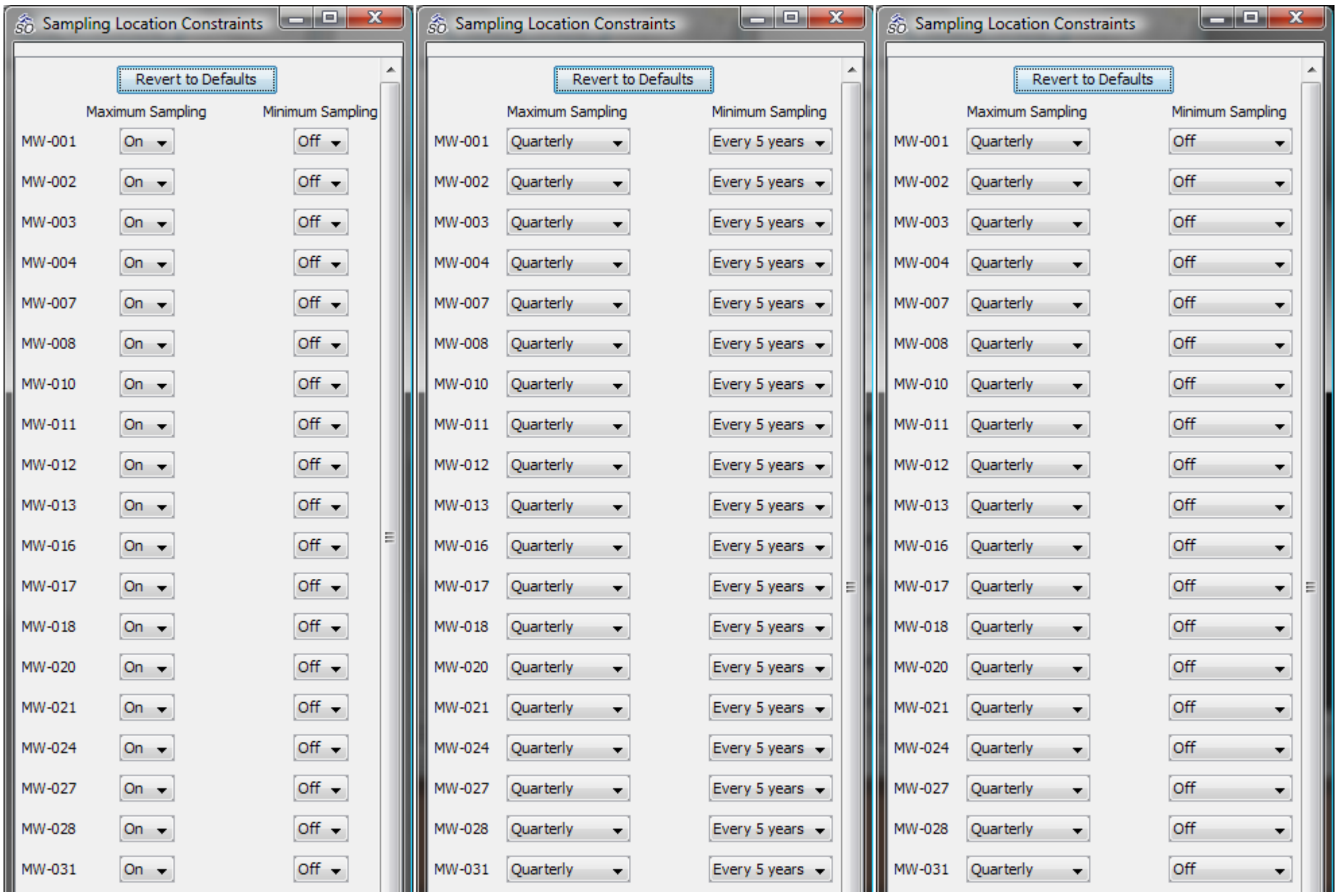


Figure 14. Sampling Location Constraints
(From left to right: Spatial, Temporal, and Spatio-Temporal)

The minimum frequency is the least frequent sampling rate which will be allowed in the optimization, and the maximum frequency is the most frequent sampling rate which will be allowed in the optimization. For example, minimum frequency is useful when sampling locations have been designated as sentinels by a regulator, and maximum frequency is useful to include human assessment that a sampling location should definitively have its sampling frequency capped at a certain rate.

Frequency Alignment Settings

For best optimization results when using temporal or spatio-temporal optimization, the frequency alignment settings must be properly set. The description in the dialog box below explains how to correctly set this parameter.

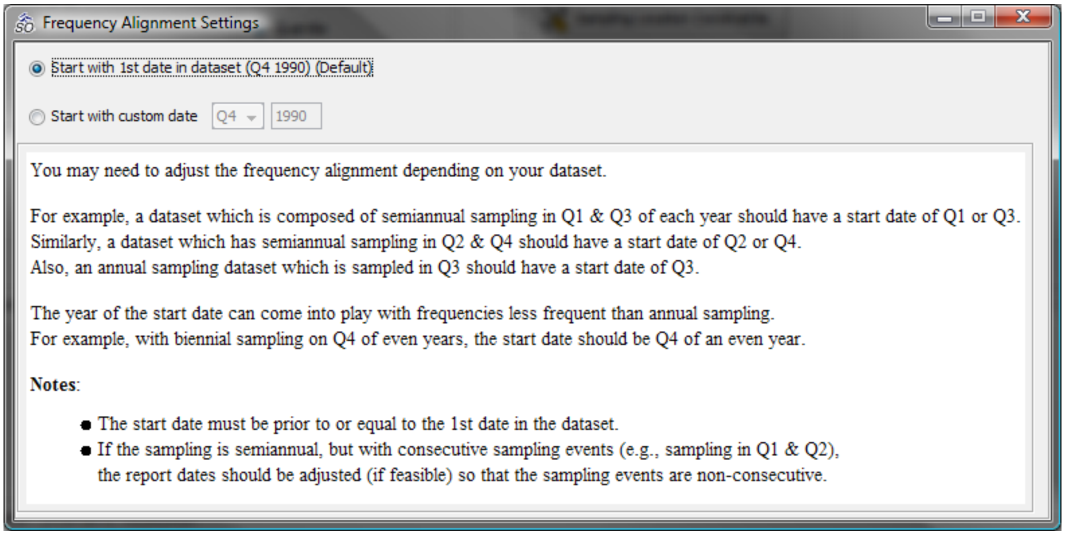


Figure 15. Frequency Alignment Settings

Facility ID

The **Facility ID** dialog allows you to store information about the dataset and the optimization process. This information is stored in the site file.

In future updates to the program, the user will be able to include this information in generated reports.

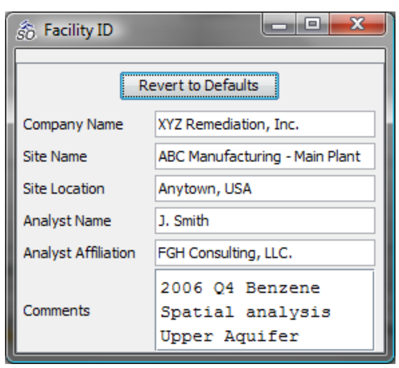


Figure 16. Facility ID

Objective Functions

Objective functions are methods of evaluating a potential sampling plan. In *SampleOptimizer™*, there are two types of objective functions: accuracy (error) objectives, and cost objectives. By default, the x-axis displays the accuracy objective, and the y-axis displays the cost objective.

The cost objective is the way for the user to specify the optimization objective of minimizing monitoring costs through minimizing the number of sampling locations and/or the sampling frequency. The default cost objective is basic, and lets the user specify a cost per sample for purposes of displaying the overall cost of a sampling plan.

The accuracy (error) objective is the way for the user to specify the optimization objective of minimizing the loss of information which can occur when fewer locations are sampled and/or when sampling frequencies are reduced. Accuracy is based on the similarity between an interpolated value for a sampling location, versus the actual value at that location.

Important! *We highly recommended that you configure the settings for the Error Calculator according to the COC's and monitoring objectives of your site. Please see the section* Function Parameters *for additional information.*

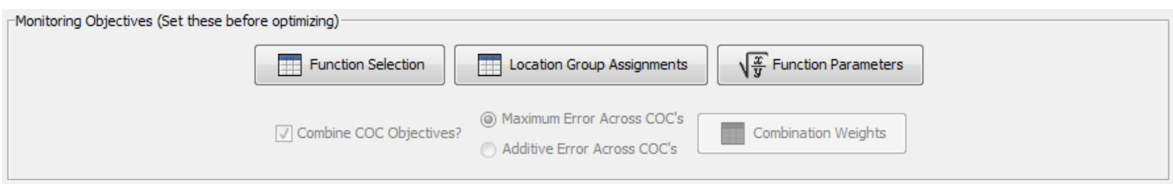


Figure 17. Monitoring Objectives Settings

Location Group Assignments

For additional flexibility and power, SampleOptimizer™ allows the user to assign sampling locations to either of two location groups: exterior and interior locations. This categorization enables the user to choose different objective functions and/or objective function parameters for interior and exterior.

You may find this feature to be useful for allowing less interpolation accuracy inside a plume interior than in the locations considered to be external to the plume. For example, you could use the **Cutoff Error Calculator** (discussed later) and set the percentage error allowed for external locations to be greater than that allowed for internal locations.

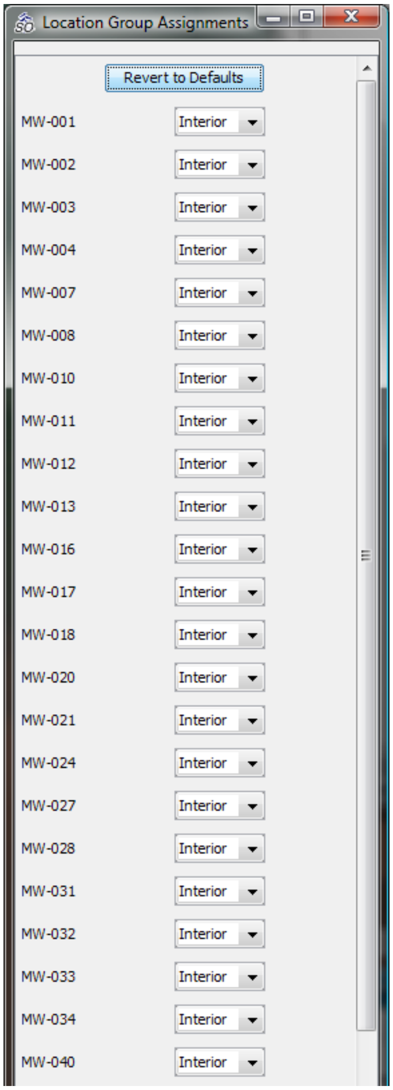


Figure 18. Location Group Assignments

Function Selection

Clicking **Function Selection** opens a window where you can change the objective functions and disable or enable analysis of each COC. By default, all COCs with sufficient data density will be analyzed by the Optimizer. If there is more than one COC with sufficient data density and you wish to exclude one of them from analysis, click the check box below the COC name (see Figure 19 below, showing Chlorobenzene disabled). To change the objective function for a COC or for the cost objective, select an alternate function from the drop-down menu.

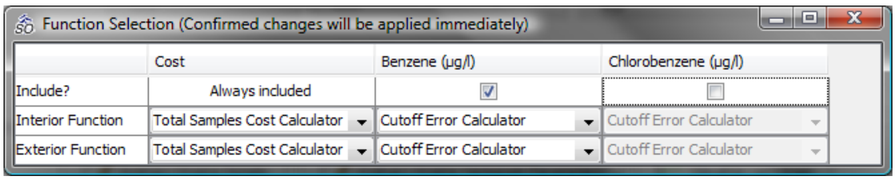


Figure 19. Disabling a COC

If your site requires a new objective which is not in the drop-down list, please contact our Technical Support department for more information on ordering custom objective function(s).

Function Parameters

For each combination of COC and location group (interior and exterior), there are different options for calculation of cost and error. These include "Cutoff Error Calculator" and "Percentage Error Calculator" for error, and "Total Samples Cost" for cost. To set these function values, click the **Function Parameters** button.

Error Calculation

To evaluate a potential sampling plan with fewer samples than the baseline maximum sampling plan, the optimizer uses interpolation algorithms (either Inverse Distance Weighting or Ordinary Kriging) to predict the sample values at the locations with reduced frequency, using the values at the remaining locations, and compares the interpolated values at the locations of the removed samples to the actual values.

For example, let's say we have the following data where Benzene is sampled semi-annually at 15 wells in the baseline sampling plan:

Date	SiteID	Benzene
3/1/2006	W01	26.3
3/1/2006	W02	51.1
3/1/2006	W03	14600
3/1/2006	W04	41
3/1/2006	W05	1.00E-03
3/1/2006	W06	1.00E-03
3/1/2006	W07	1.00E-03
3/1/2006	W08	1.00E-03
3/1/2006	W09	1.00E-03
3/1/2006	W10	1.00E-03
3/1/2006	W11	1.00E-03
3/1/2006	W12	1.00E-03
3/1/2006	W13	1.00E-03
3/1/2006	W14	1.00E-03
3/1/2006	W15	1.00E-03

Date	SiteID	Benzene
8/1/2006	W01	1.00E-03
8/1/2006	W02	1.00E-03
8/1/2006	W03	16000
8/1/2006	W04	0.96
8/1/2006	W05	1.00E-03
8/1/2006	W06	1.00E-03
8/1/2006	W07	1.00E-03
8/1/2006	W08	1.00E-03
8/1/2006	W09	1.00E-03
8/1/2006	W10	1.00E-03
8/1/2006	W11	1.00E-03
8/1/2006	W12	1.00E-03
8/1/2006	W13	1.00E-03
8/1/2006	W14	1.00E-03
8/1/2006	W15	1.00E-03

Date	SiteID	Benzene
3/6/2007	W01	1.00E-03
3/6/2007	W02	1.00E-03
3/6/2007	W03	9280
3/6/2007	W04	1380
3/6/2007	W05	1.00E-03
3/6/2007	W06	1.00E-03
3/6/2007	W07	1.00E-03
3/6/2007	W08	1.00E-03
3/6/2007	W09	1.00E-03
3/6/2007	W10	1.00E-03
3/6/2007	W11	1.00E-03
3/6/2007	W12	1.00E-03
3/6/2007	W13	6.9
3/6/2007	W14	1.00E-03
3/6/2007	W15	1.00E-03

Date	SiteID	Benzene
8/9/2007	W01	1.00E-03
8/9/2007	W02	1.00E-03
8/9/2007	W03	2500
8/9/2007	W04	110
8/9/2007	W05	1.00E-03
8/9/2007	W06	1.00E-03
8/9/2007	W07	1.00E-03
8/9/2007	W08	1.00E-03
8/9/2007	W09	1.00E-03
8/9/2007	W10	1.00E-03
8/9/2007	W11	1.00E-03
8/9/2007	W12	1.00E-03
8/9/2007	W13	4.8
8/9/2007	W14	1.00E-03
8/9/2007	W15	1.00E-03

Figure 20. Example Spatio-Temporal Sampling Data

If we wanted to evaluate a potential plan (let's call it "Plan A") which turned off W06 and W07, and reduced W08 and W09 to annual sampling, the optimizer would perform interpolations to estimate values for the removed samples based on the remaining samples in each sampling period.

Date	SiteID	Benzene
3/1/2006	W01	26.3
3/1/2006	W02	51.1
3/1/2006	W03	14600
3/1/2006	W04	41
3/1/2006	W05	1.00E-03
3/1/2006	W08	1.00E-03
3/1/2006	W09	1.00E-03
3/1/2006	W10	1.00E-03
3/1/2006	W11	1.00E-03
3/1/2006	W12	1.00E-03
3/1/2006	W13	1.00E-03
3/1/2006	W14	1.00E-03
3/1/2006	W15	1.00E-03

Date	SiteID	Benzene
8/1/2006	W01	1.00E-03
8/1/2006	W02	1.00E-03
8/1/2006	W03	16000
8/1/2006	W04	0.96
8/1/2006	W05	1.00E-03
8/1/2006	W10	1.00E-03
8/1/2006	W11	1.00E-03
8/1/2006	W12	1.00E-03
8/1/2006	W13	1.00E-03
8/1/2006	W14	1.00E-03
8/1/2006	W15	1.00E-03

Date	SiteID	Benzene
3/6/2007	W01	1.00E-03
3/6/2007	W02	1.00E-03
3/6/2007	W03	9280
3/6/2007	W04	1380
3/6/2007	W05	1.00E-03
3/6/2007	W08	1.00E-03
3/6/2007	W09	1.00E-03
3/6/2007	W10	1.00E-03
3/6/2007	W11	1.00E-03
3/6/2007	W12	1.00E-03
3/6/2007	W13	6.9
3/6/2007	W14	1.00E-03
3/6/2007	W15	1.00E-03

Date	SiteID	Benzene
8/9/2007	W01	1.00E-03
8/9/2007	W02	1.00E-03
8/9/2007	W03	2500
8/9/2007	W04	110
8/9/2007	W05	1.00E-03
8/9/2007	W10	1.00E-03
8/9/2007	W11	1.00E-03
8/9/2007	W12	1.00E-03
8/9/2007	W13	4.8
8/9/2007	W14	1.00E-03
8/9/2007	W15	1.00E-03

Figure 21. Interpolator input data
(Some samples have been removed compared to Figure 20.)

The interpolator would use the data in Figure 21to predict what the values for the missing samples would have been:

1. Use the 3/1/2006 data in Figure 21 to predict the values for W06 & W07 on 3/1/2006.
2. Use the 8/1/2006 data in Figure 21 to predict the values for W06, W07, W08, and W09 on 8/1/2006.
3. Use the 3/6/2007 data in Figure 21 to predict the values for W06 & W07 in 3/6/2007.
4. Use the 8/9/2007 data in Figure 21 to predict the values for W06, W07, W08, and W09 on 8/9/2007.

The objective calculator then compares the interpolated values to the actual values and assigns an error value (which is the maximum error at any removed point in any of the events, based on the spatial interpolations in each event as described above) which is then used in the GA to compare "Plan A" to the other plans.

The GA seeks to preserve diversity of the population while also selecting plans which are clearly advantageous compared to other plans. For example, if there are two plans which have the same cost but one is more accurate than the other, the more accurate one will be chosen. If there is one plan with less cost and less accuracy and another with more cost and more accuracy, the GA will seek to keep both, to the extent that there is room in the population.

The overall purpose of Error Calculators is to come up with an objective function value that represents the overall similarity of a new sampling plan to the baseline Max Sampling plan. This can be thought of as an accuracy metric, or lack thereof (error metric).

In most cases we recommend that the Cutoff Error Calculator be used. When accuracy inside a plume interior can be less important than the accuracy outside it, we recommend utilizing the **Location Group Assignments** feature and setting a different percentage error allowed for interior vs. exterior points.

The user must be careful when using the Percentage Error Calculator when very small non-detect values are in the dataset, because a very small error can yield a very high percentage error of a non-detect. For example, an error of just 0.001 for a non-detect value of 0.0001 is a 1000% error, which will almost certainly be much higher than any error for any detected sample values.

Figure 22 shows the settings screen for the Cutoff Error Calculator which is displayed after clicking on **Function Parameters**. Note that we recommend that o, p, q values be chosen such that $o = p*q$, to insure that there is continuity between errors calculated above and below the cutoff concentration level.

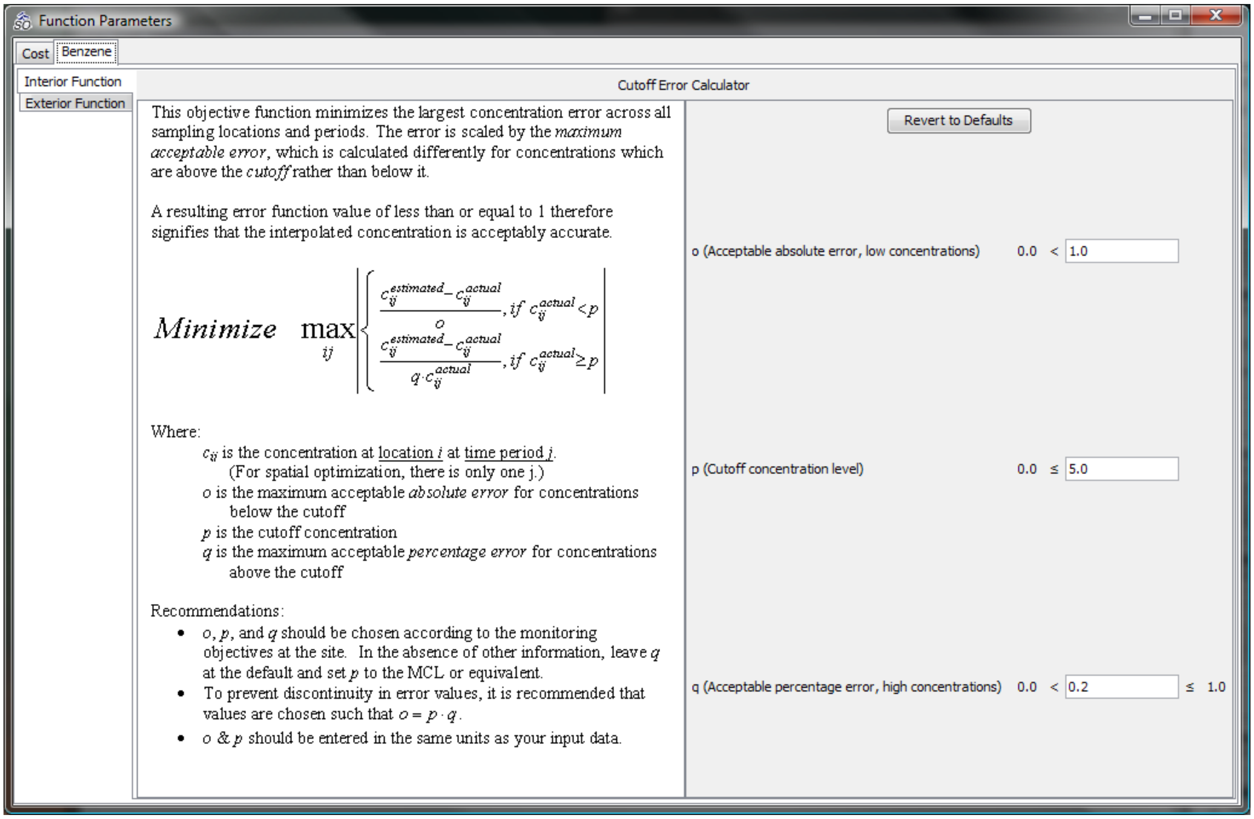


Figure 22. Error Objective Calculator Settings

The **Cutoff Error Calculator** is designed so that error is calculated in a manner which allows more significant percentage deviation, but less significant absolute deviation, between interpolated and actual values in areas of low concentration than in areas of high concentration. This is accomplished as follows:

- The user defines a cutoff concentration (p) for the actual data values that differentiates between low concentrations versus high concentrations, and also defines a value for Acceptable absolute error (o).

- When a low concentration data point is removed (i.e., below the cutoff), error is calculated as the absolute value of the actual value minus the interpolated value, divided by the acceptable absolute error. For example, if the actual value is 5 µg/l (i.e., below the cutoff concentration of 10 µg/l) and acceptable absolute error is 1.0, and the difference between the actual and interpolated value is 5 µg/l, then the error would be |5| / 1 = 5.

- When a high concentration data point is removed (i.e, above the cutoff), error is calculated as the absolute value of the actual value minus the interpolated value, divided by a percentage (q) of the actual value, where q is specified by the user. For example, if the actual value is 100 µg/l (i.e., above the cutoff concentration of 10 µg/l) and the percentage input by the user is 10%, and the difference between the actual and interpolated value is 5 µg/l, then the error would be 5 / (0.10 * 100) = 0.5.

In these examples, the difference between the actual value and the interpolated value was 5 µg/l in both cases, but in the first case the calculated error is 5.0 whereas in the second case it is only 0.5. This illustrates how the Cutoff Error Calculator appropriately reduces the significance of deviations between actual and interpolated values in lower concentration areas of the plume.

If your dataset does not have any very small values relative to the rest of the data (for example, non-detect values), you can use the Percentage Error Calculator. This calculator is a more basic version of the Cutoff Error Calculator that does not have a cutoff or acceptable absolute error value. You can also use the Percentage Error Calculator for external locations which do not have non-detects.

Cost Calculation

Figure 23 shows the configuration window for the Total Samples Cost Calculator. This is used to calculate per event cost (spatial analysis) or per year cost (spatio-temporal or temporal analysis). The default is $100 per sample, but if you specify a more accurate number for your site, the tradeoff curve will be more useful for deciding the best cost/accuracy tradeoff for your site.

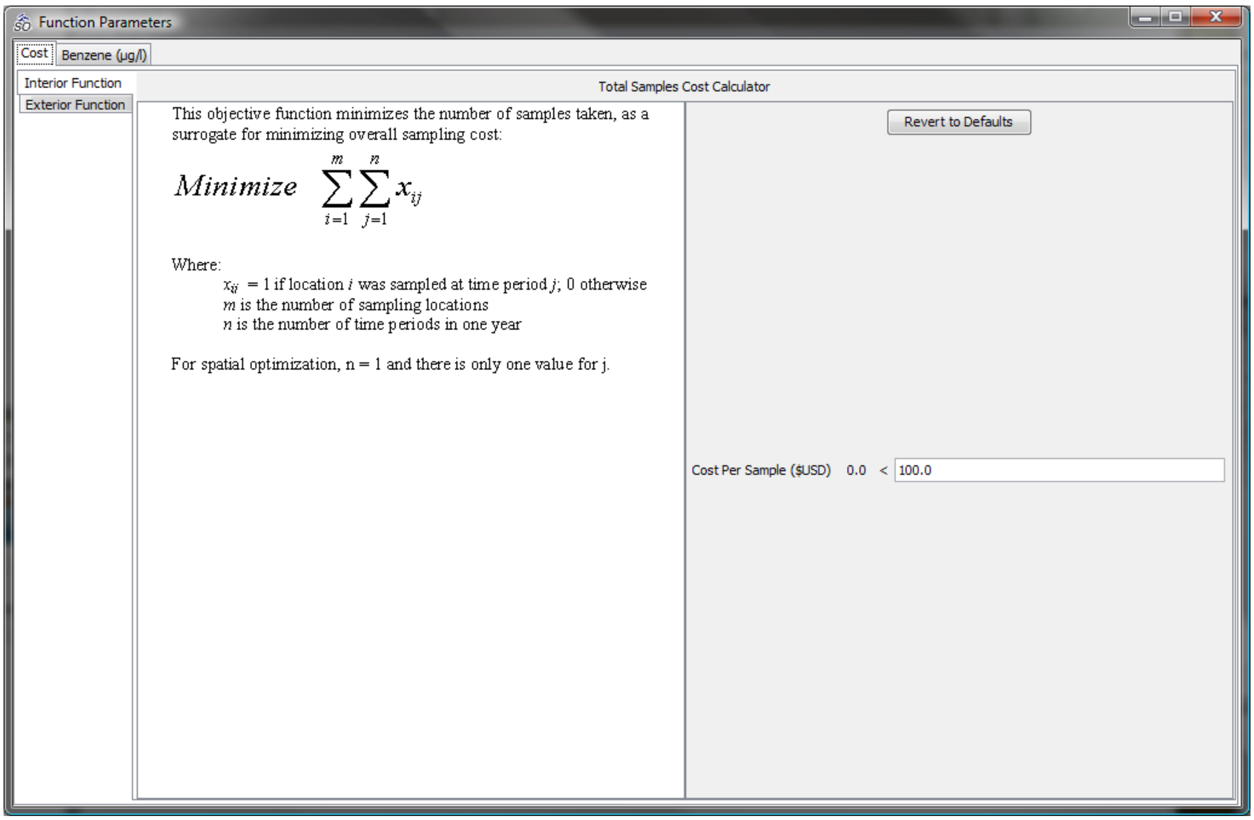


Figure 23. Total Samples Cost Calculator

Combining COC Objectives

If the dataset you have loaded has more than one valid, enabled COC, by default the software enables COC objective combination. This can help make the optimization results easier for the user to analyze successfully, by reducing the number of objectives to analyze. Instead of looking at two tradeoff curves for a two-COC analysis, the user just has to look at one curve.

Some care should be taken in deciding how to combine the objectives, however. The easiest and safest method is to choose the "Maximum Error Across COC's" method, and this is the default. With this approach the error for more than one COC is simply the highest error across all COC's.

Alternatively, you can choose to use the "Additive Error Across COC's" method, and optionally specify the combination weights used (see Figure 24). This method can be useful to specify the relative importance of accuracy for each COC.

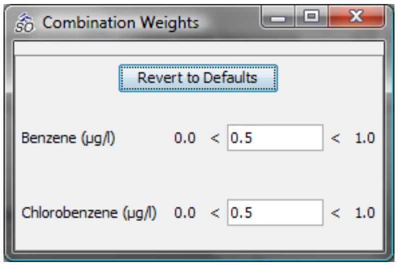


Figure 24. Combination Weights

ModelBuilder™

ModelBuilder™ is used to perform the geo-statistical functions needed during use of *SampleOptimizer™*. It has two sections: one for model fitting, visualization, and analysis, and another for visualizing uncertainty. It has been designed to be powerful, yet easy to use effectively.

Model Fitting, Visualization, and Analysis

This portion represents the bulk of *ModelBuilder™*'s capabilities. Perhaps most importantly, you can fit model parameters to your data and visualize the results. Both automated and manual model parameter fitting are supported for Ordinary Kriging as well as Inverse Distance Weighting.

Important! *It is extremely important to visualize and review the interpolation model used by ModelBuilder™ before attempting to analyze the Optimization Results. It is best to review and approve the model before running the Optimizer.*

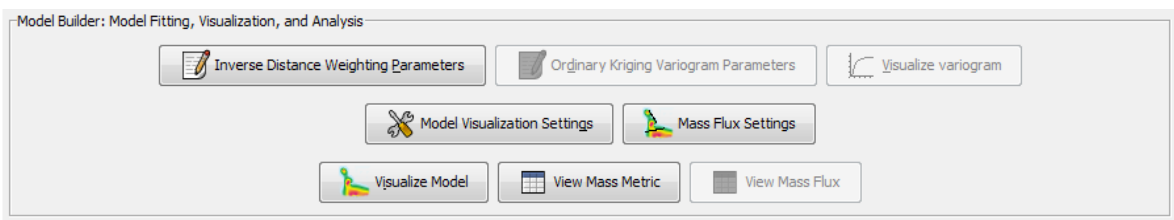


Figure 25. ModelBuilder™ (Inverse Distance Weighting)

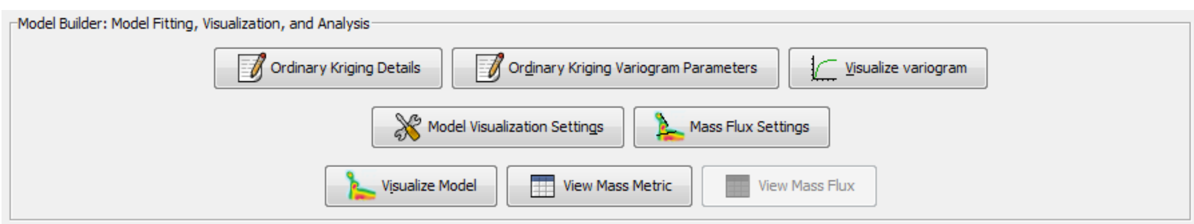


Figure 26. ModelBuilder™ (Ordinary Kriging)

When you load your data, *SampleOptimizer™* chooses Ordinary Kriging for the geo-statistical model by default, and *ModelBuilder™* automatically finds a set of model parameters which should achieve a good visual representation of the data (plume image).

Manual Model Fitting

Inverse Distance Weighting

To manually fit an Inverse Distance Weighting model, adjust the **Distance Weighting Power** in the model parameters and click on **Visualize Model** after closing this input screen to create and view an interpolated image of the COC distribution in a 2-D plan view.

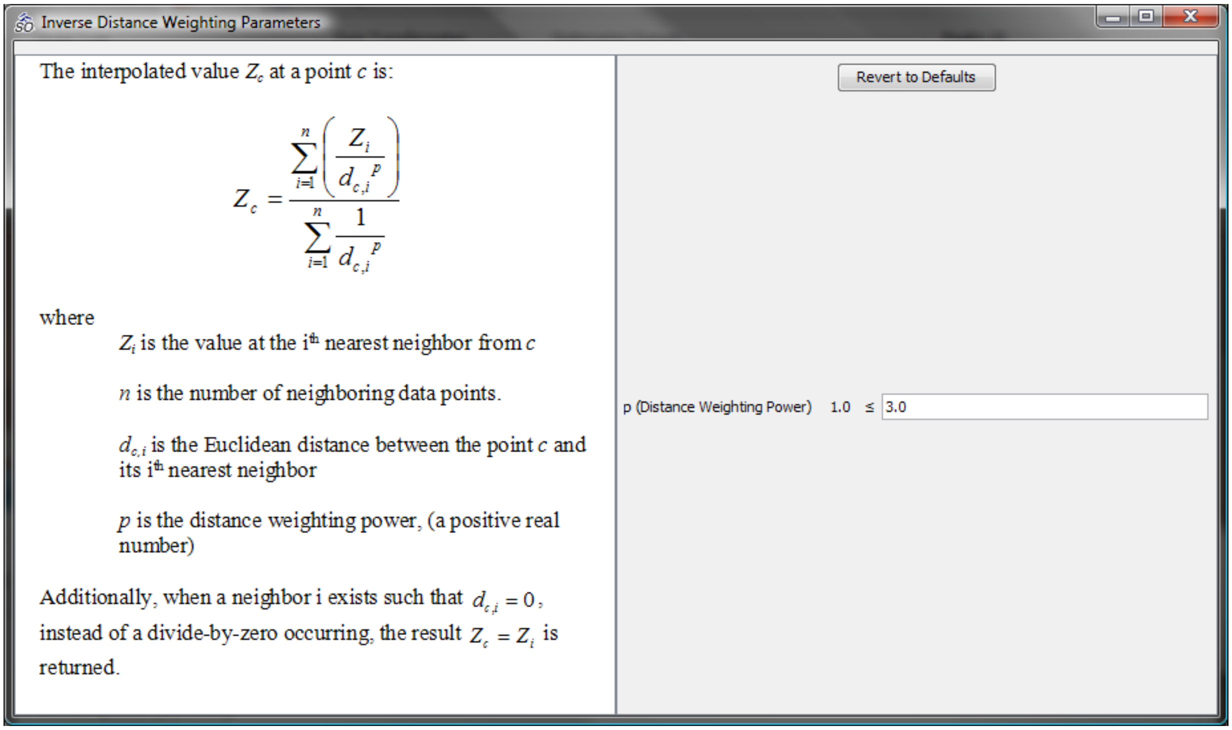


Figure 27. Inverse Distance Weighting Model Parameters

Kriging

Kriging is a more complex model than Inverse Distance Weighting. Its user-configurable parameters are set in **Variogram Parameters**, while additional details about the algorithm can be seen in **Ordinary Kriging Details**.

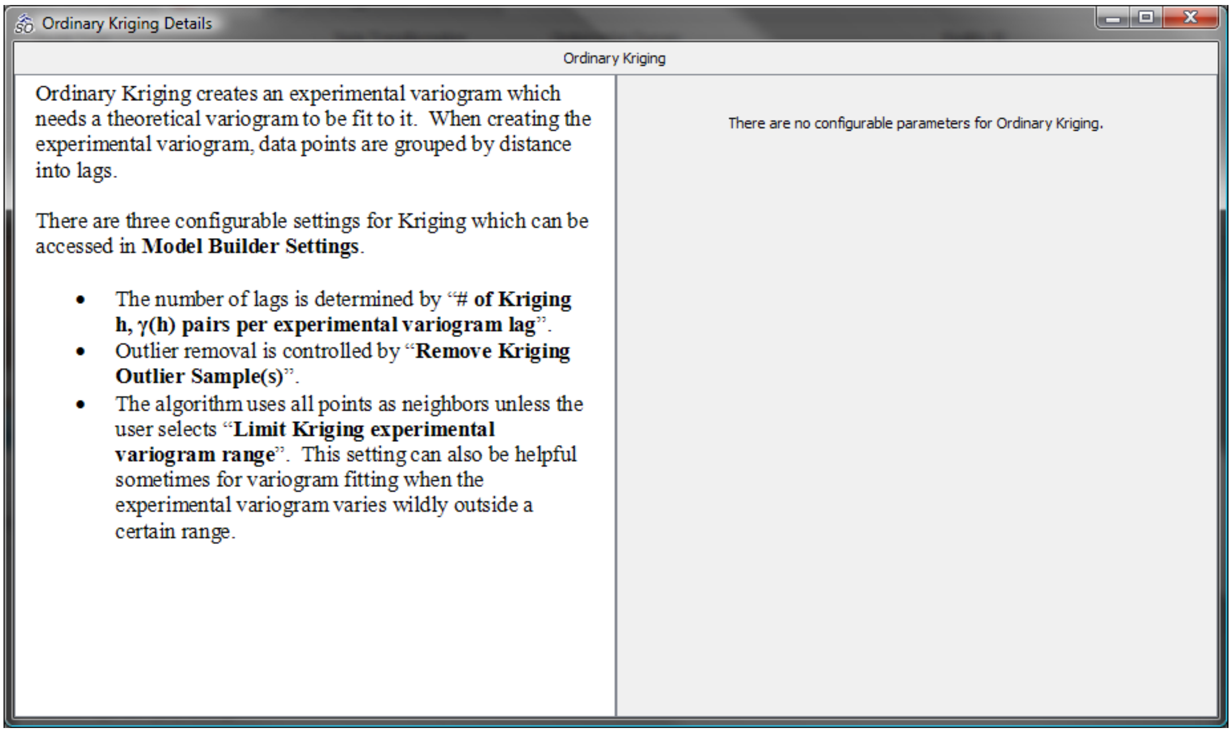


Figure 28. Ordinary Kriging Details

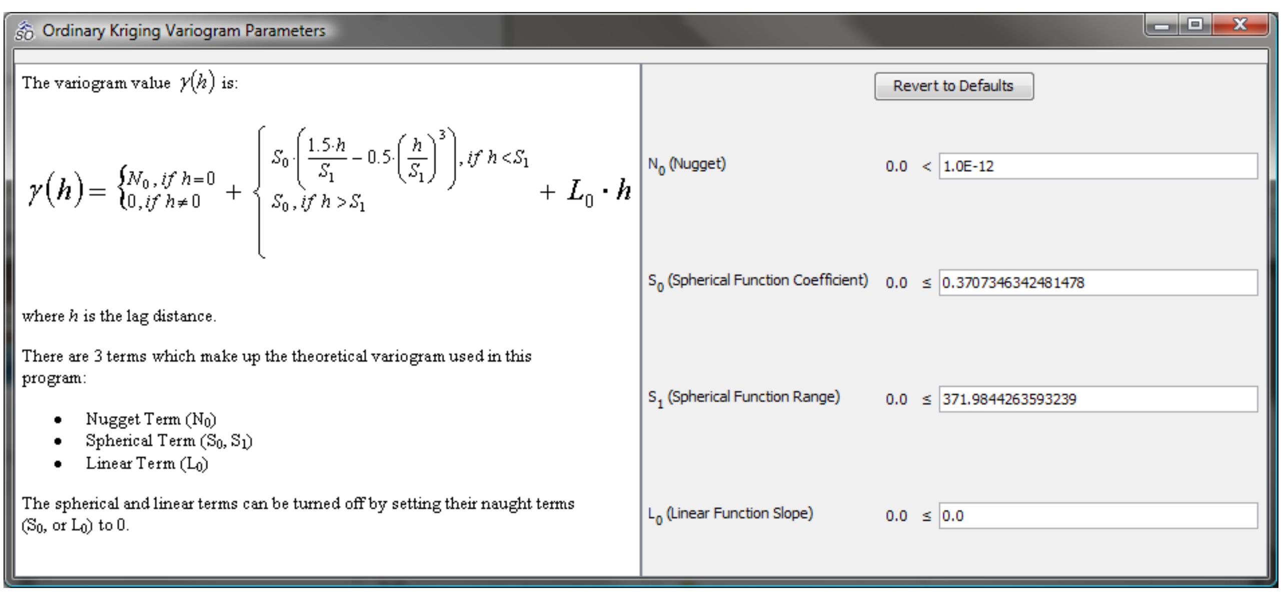


Figure 29. Variogram Parameters

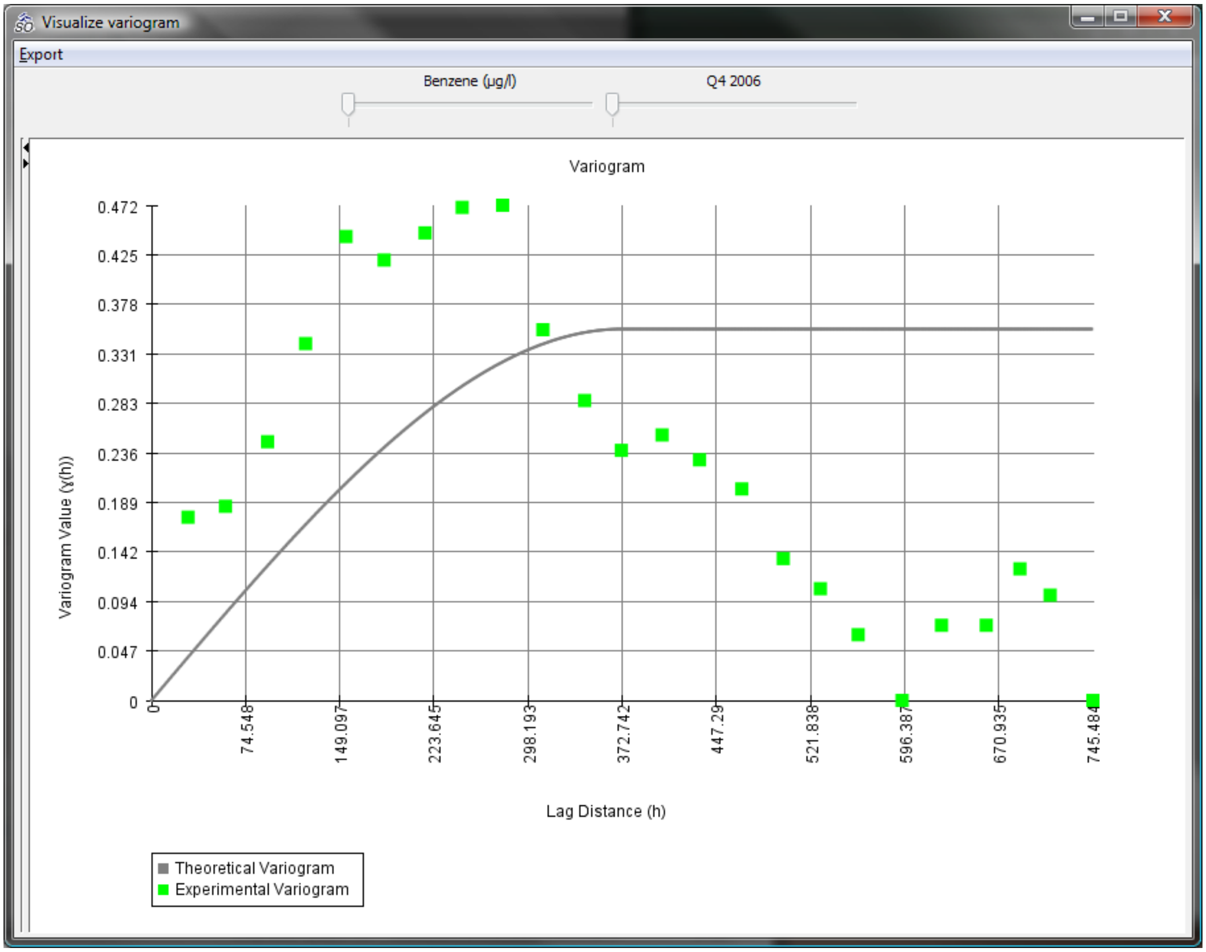


Figure 30. Kriging Variogram

The overall process for manually fitting a Kriging model is similar to that for Inverse Distance Weighting: you should adjust the parameters until the interpolated plumes are acceptable.

It is highly recommended that you verify the appropriateness of the model parameters by looking at some medium to high resolution plume images of each COC plume.

- Increase the number of vertical slices in the ModelBuilder™ settings to at least 200 (possibly more depending on your site) and then click on **Visualize Model**. . (note – this value for # vertical slices only pertains to plume visualization in *ModelBuilder*™ – a separate value for # vertical slices for visualizing plumes for resulting plans selected from the tradeoff curve in *SampleOptimizer*™ is input in the **Dashboard Settings** window)
- Turn on sampling location labels and look at the sample values vs. the values interpolated by the model, and verify that they look correct.
- A good first step for model validation is to adjust the color range to different maximum values to see the area(s) of the plume which are above certain values, and compare them to the reported sample values.

Model Visualization

You can visualize your data just by clicking on **Visualize Model**, but you can also adjust the **Model Visualization Settings**.

For both Inverse Distance Weighting and Ordinary Kriging, the resolution, or # of vertical slices can be chosen. Note that the number of horizontal slices is chosen by the software based on the chosen # of vertical slices in order for every cell to be a square. The # of border slices determines the amount of blank space around the edge of the visualization. Also, the run title can be specified, and this can optionally be appended (see Application Options) to the auto-generated filenames for exported data and images.

For Ordinary Kriging, additional options are available but these are intended for advanced users who have some knowledge of how Ordinary Kriging models operate.

- Remove outlier sample(s) (disabled by default)
    - If enabled, the algorithm will iteratively remove the highest concentration values until no lag changes by more than the fractional value chosen.
- Number of pairs per lag
    - This parameter influences how many lags will be created in experimental variograms. The number of lags is equal to the total number of h, γ (h) pairs which are created by the data divided by the number of pairs per lag.
- Use means to create lags (enabled by default)
    - If selected, when creating lags, the mean of the experimental variogram points within a lag will be used as the substitute data point for that lag. If unselected, the median of the experimental variogram points will be used instead. NOTE: data which contain a very large number of one value (such as a small non-detect value) should always use the mean to help insure that an experimental variogram can be successfully created.
- Limit variogram range (Disabled by default)
    - If enabled, the experimental variogram range will be limited, and the value entered becomes the distance range fractional cutoff used when creating experimental variograms. This fractional number times the distance between the two monitoring points furthest away from each other is used as the largest distance which will be included in experimental variograms.

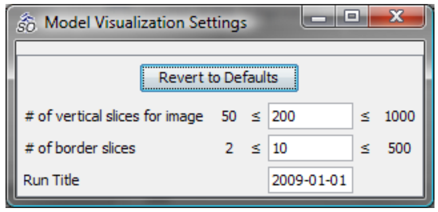


Figure 31. ModelBuilder™ Settings (Inverse Distance Weighting)

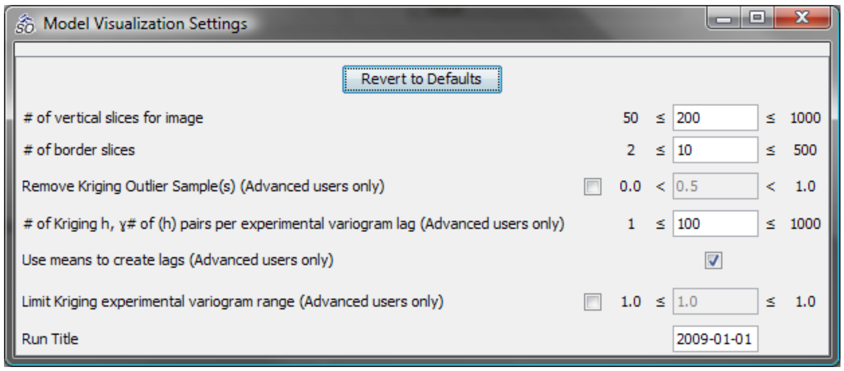


Figure 32. ModelBuilder™ Settings (Ordinary Kriging)

While *ModelBuilder™* is working you will see a window similar to that shown in Figure 33.

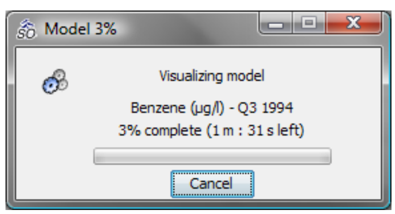


Figure 33. ModelBuilder™ Progress

Note that displaying the image(s) may take a while to complete. Figure 34 shows an example completed model visualization using Ordinary Kriging.

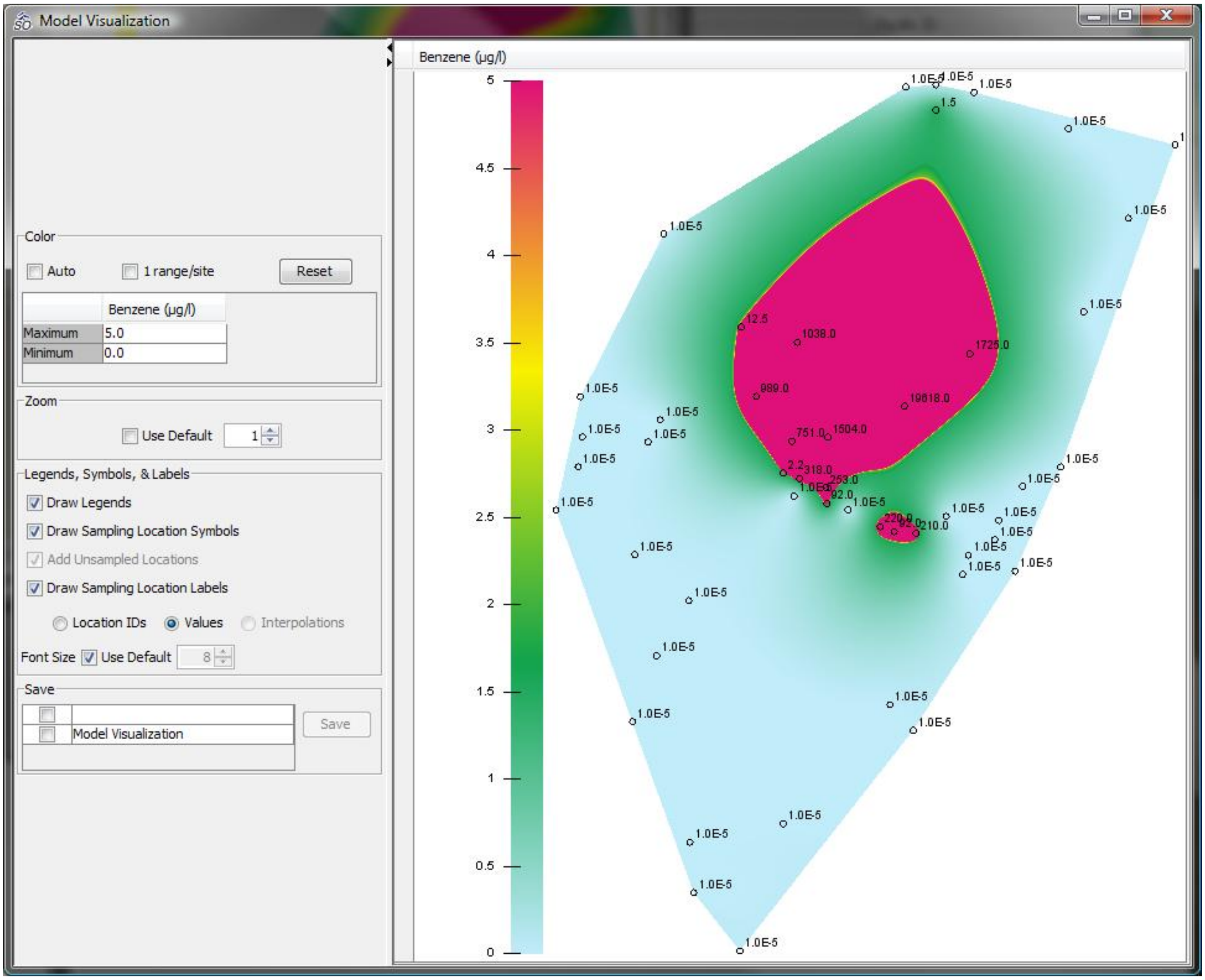
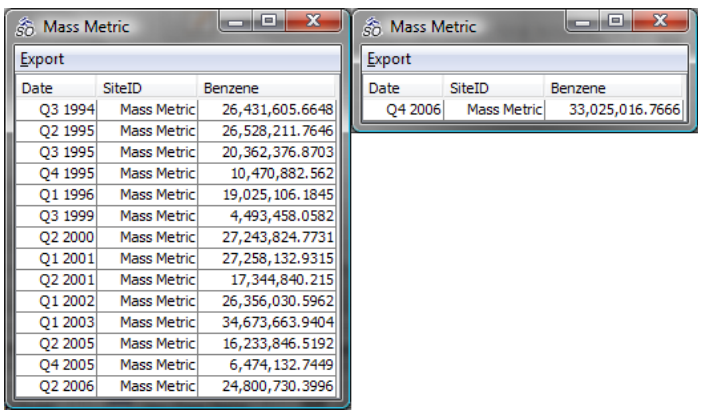


Figure 34. Model Visualization (Ordinary Kriging model)

Mass Metric

This feature is intended for use in tracking data over time (as is done with *SampleTracker™*), in order to get a rough estimate of a mass metric which could be tracked in time so that relative mass of contaminants in different sampling events can be compared. However, this functionality is based on spatial interpolations and therefore is included in the *SampleOptimizer™* menus of the software. To use it, click on Mass Metric to bring up a window similar to that shown in Figure 35. Note that if model visualizations have not yet been generated, they will automatically be generated so that the Mass Metric data can be shown.

Important! *Mass Metric is not designed to authoritatively estimate the absolute amount of mass at a site. Rather, it calculates mass per unit volume of aquifer. Therefore, it is useful for comparisons of relative mass between sampling events, but it is not intended to estimate <u>absolute</u> mass within the plume.*



Date	SiteID	Benzene
Q3 1994	Mass Metric	26,431,605.6648
Q2 1995	Mass Metric	26,528,211.7646
Q3 1995	Mass Metric	20,362,376.8703
Q4 1995	Mass Metric	10,470,882.562
Q1 1996	Mass Metric	19,025,106.1845
Q3 1999	Mass Metric	4,493,458.0582
Q2 2000	Mass Metric	27,243,824.7731
Q1 2001	Mass Metric	27,258,132.9315
Q2 2001	Mass Metric	17,344,840.215
Q1 2002	Mass Metric	26,356,030.5962
Q1 2003	Mass Metric	34,673,663.9404
Q2 2005	Mass Metric	16,233,846.5192
Q4 2005	Mass Metric	6,474,132.7449
Q2 2006	Mass Metric	24,800,730.3996

Date	SiteID	Benzene
Q4 2006	Mass Metric	33,025,016.7666

Figure 35. Mass Metric

To estimate the mass metric for a COC at a given time period at a site, the interpolated concentration value at every cell in the plume map for that COC at that time period is summed, and each cell is approximated to represent the same volume (i.e., incorporating vertical extent and porosity) as every other cell.

To export the mass metric data (which can then be imported into *SampleTracker™* if desired), select "Export Table to .csv" from the "Export" menu. Then, add the data in this new file to your site's historical and current data files as appropriate.

Mass Flux

This feature is also intended for use with SampleTracker™, in order to get a rough estimate of the mass which might be flowing through a cross-section of a site. To use it, one must first define a cross-section for the software to analyze. Click **Mass Flux Settings** to bring up a window similar to the one shown in Figure 36.

Important! *Mass Flux is not designed to authoritatively estimate the absolute amount of mass flux at a site, because plume thickness and porosity are not accounted for. Similar to Mass Metric, it is useful for comparisons of relative mass flux across a boundary between sampling events, but it is not intended to estimate* ***absolute*** *mass flux in each sampling event.*

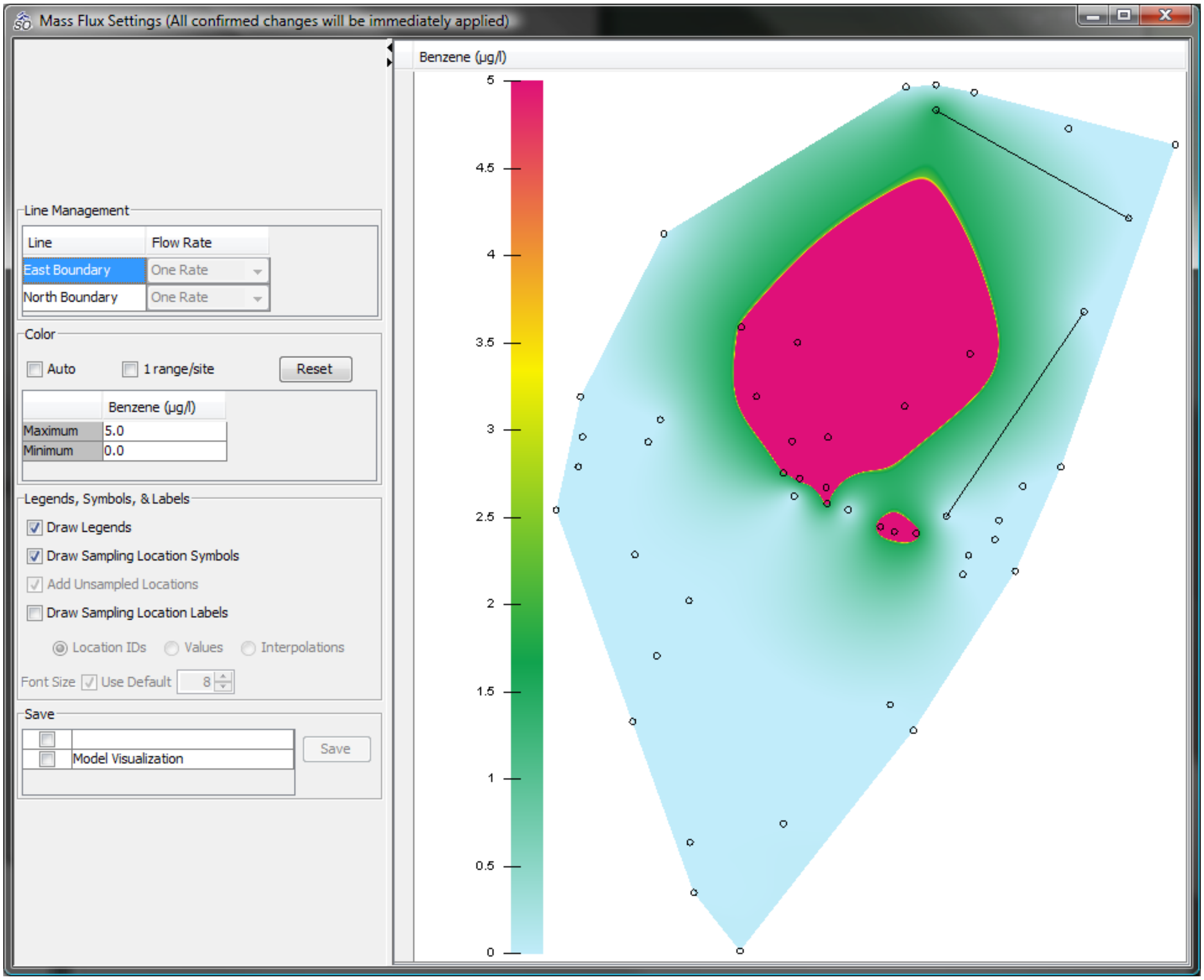


Figure 36. Mass Flux Settings

The user can create a boundary consisting of several line segments, each of which can have a unique groundwater flow rate. To create a new line, first click in a plume image to bring up the crosshairs. Then, click again on the place where you would like the cross-section to begin, and click a third time on the place where you would like the cross-section to end. If you would like to cancel the creation of a line when it is already in progress, press the right mouse button to cancel.

After you have created a line, right click on it in the Line Management table and choose "Adjust Flow Rate Settings" to bring up a window similar to the one shown in Figure 37.

Important! *Make sure that the flow rates you enter are in the same volumetric units as the site concentration data.*

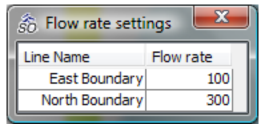


Figure 37. Mass Flux Flow Rate Settings

Mass flux uses similar mathematical and physical principles as Mass Metric. The interpolated value at every location in the cross section is multiplied by the flow rate to get an estimate of the mass flux at each point in the cross section, and the flux at each point is then added together to create the estimate. To view the results, click **View Mass Flux** in the main *ModelBuilder*™ section of the Settings tab. A brief pause may be required to perform the necessary calculations; afterwards, a window similar to the one in Figure 38 will appear.

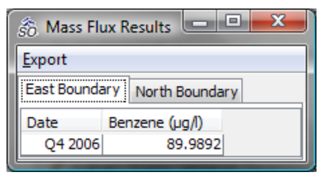


Figure 38. Mass Flux Results

Uncertainty Visualization

By clicking on **View Uncertainty Visualization** in *ModelBuilder*™, you can visualize a metric of uncertainty for your site. This metric shows the user an image which is a combination of variability in the model as well as the data. It can be used along with site-specific knowledge to assist in the siting of new sampling locations.

The uncertainty calculation process uses multiple iterations of a cross validation and visualization generation process, and creates the uncertainty map based on the root-mean-squared (RMS) value of each pixel in the generated visualizations.

Figure 39 shows the Uncertainty Visualization Settings screen. On the left are some more details about how the maps are calculated, and on the right are a set of parameters which can be set by the user to control how the maps are calculated. Hovering the cursor over any of these parameters will bring up a pop-up description.

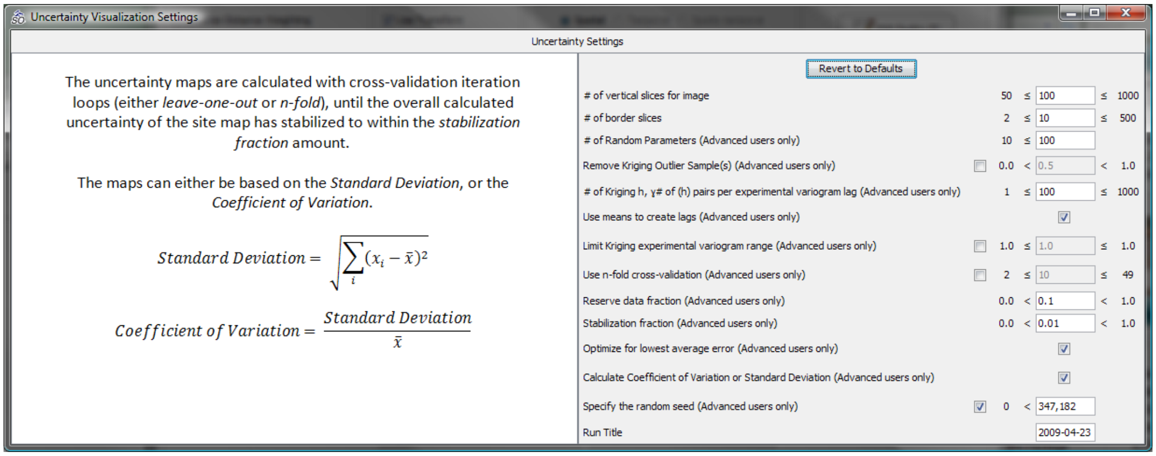


Figure 39. Uncertainty Visualization Settings

Figure 40 shows an example of uncertainty visualization for a spatial dataset using an Ordinary Kriging model. Note that by default, the coloring range goes from the minimum uncertainty to the 75th percentile of uncertainty. As with plume visualization, this coloration can be adjusted by utilizing the "Color" control in the window.

The uncertainty visualization shows areas of the plume map with the highest root-mean-square error in the estimated concentrations. Note that large differences in concentrations in two nearby wells can lead to high cross-validation errors that may not be a major concern if the plume is thought to be stable and well characterized. If there are locations with high errors where accurate concentration estimates are particularly important (e.g., close to down-gradient exposure locations), these locations may be candidates for additional sampling activities.

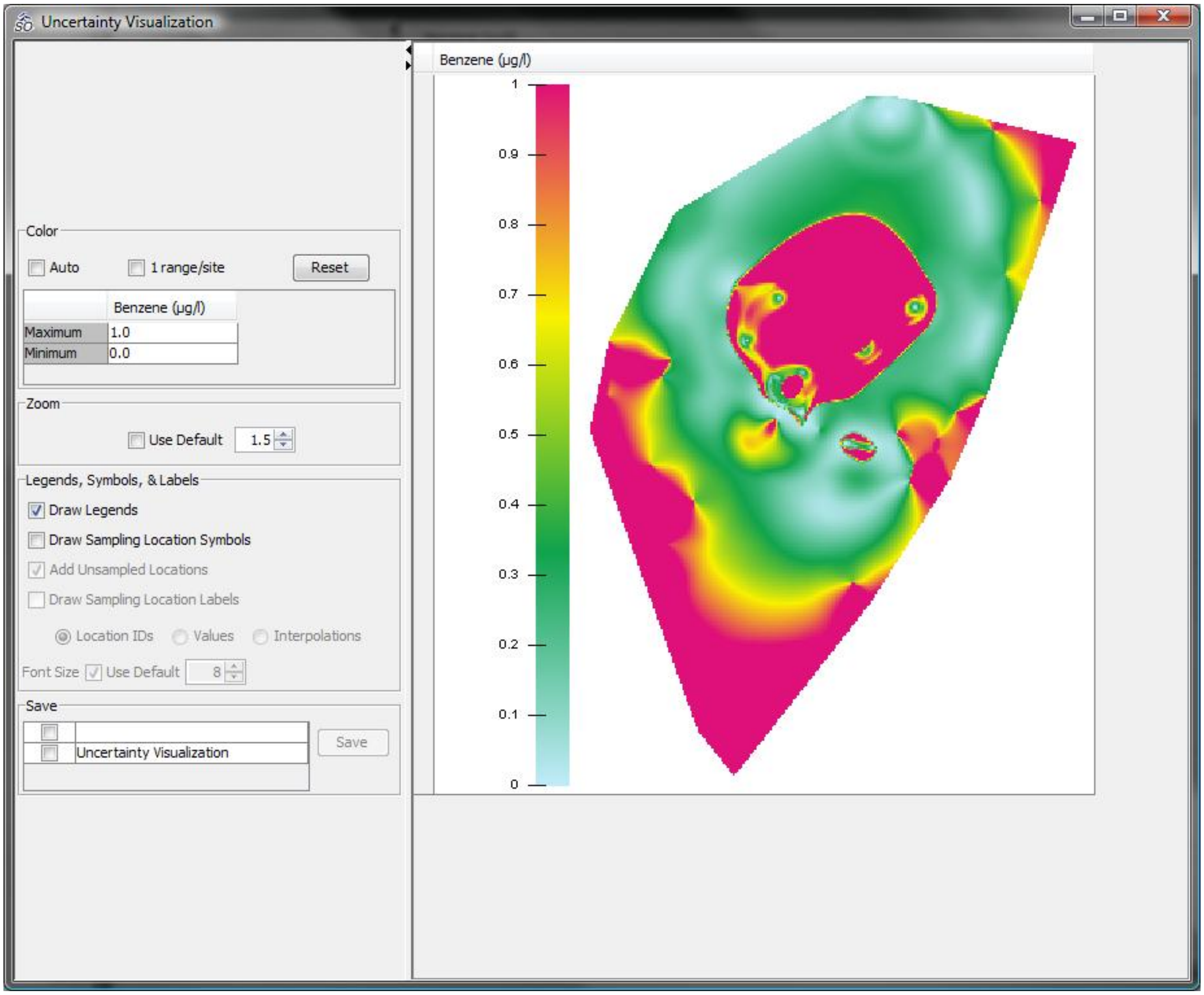


Figure 40. Uncertainty Visualization (Ordinary Kriging)

The Uncertainty calculation process for Ordinary Kriging may take from 15 minutes up to several hours to complete depending on the number of sampling locations, number of COC's, number of time periods (spatio-temporal or temporal optimization only), and the speed of your computer's processor. Inverse Distance Weighting is usually much faster.

Please note that the time listed to completion is the *worst-case* completion time for the calculations to complete. Calculation continues until the change in standard deviation of the uncertainty maps is within the **Stabilization %** for two consecutive iterations, which usually occurs before the maximum number of parameter optimization iterations is completed.

Lastly, keep in mind that when the number of sampling locations is large (over 40 or so), Uncertainty Visualization can take an extremely long amount of time to complete. In such situations, we recommend checking the option for "use n-fold cross validation" with n = 3. This is a less rigorous approach but usually yields appropriate results.

GA Settings

One of the powerful features of the software is that it automatically configures the GA settings to values which should give good performance in many cases. If you wish to manually adjust change the settings, you can do so under **GA Settings** (see Figure 41).

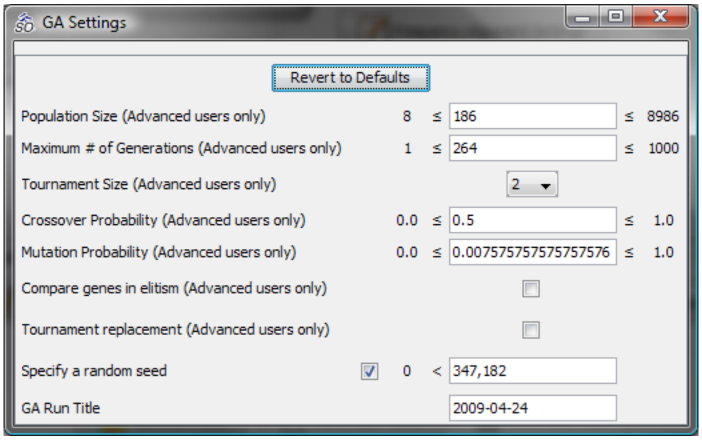


Figure 41. GA Settings.

- Population size
  - The number of candidate solutions that will be held in memory at any given point in the optimization process, as well as the number of initial randomly created solutions from which the GA evolutionary process begins. This must be an even number. Additionally, it must be a multiple of the tournament size if tournament replacement is off.
- Maximum # of generations
  - Maximum number of permutations of the overall GA evolutionary process that will be performed.
- Tournament size
  - The number of individuals that will compete for mating at any given reproduction instance.
- Crossover probability
  - The probability that the Uniform Crossover process will be performed to combine two parent plans into two new child plans. This should be kept at 0.5.
- Mutation probability
  - The probability that each gene will be modified in a random fashion.
- Elitism gene comparison
  - If enabled, the GA will take into account the differences between plans' genes when computing the crowding distances. While this increases the diversity of the population and ultimately can result in better solutions, unfortunately, it takes $O(n^2)$ computational time as the population size increases. Therefore, this option is only practical for populations of less than a few hundred.
- Tournament replacement
  - If enabled, return individuals to the competition pool after they have competed in a tournament.

SampleOptimizer™ Dashboard

Now that the settings have been configured, you are ready to start the main optimizer.

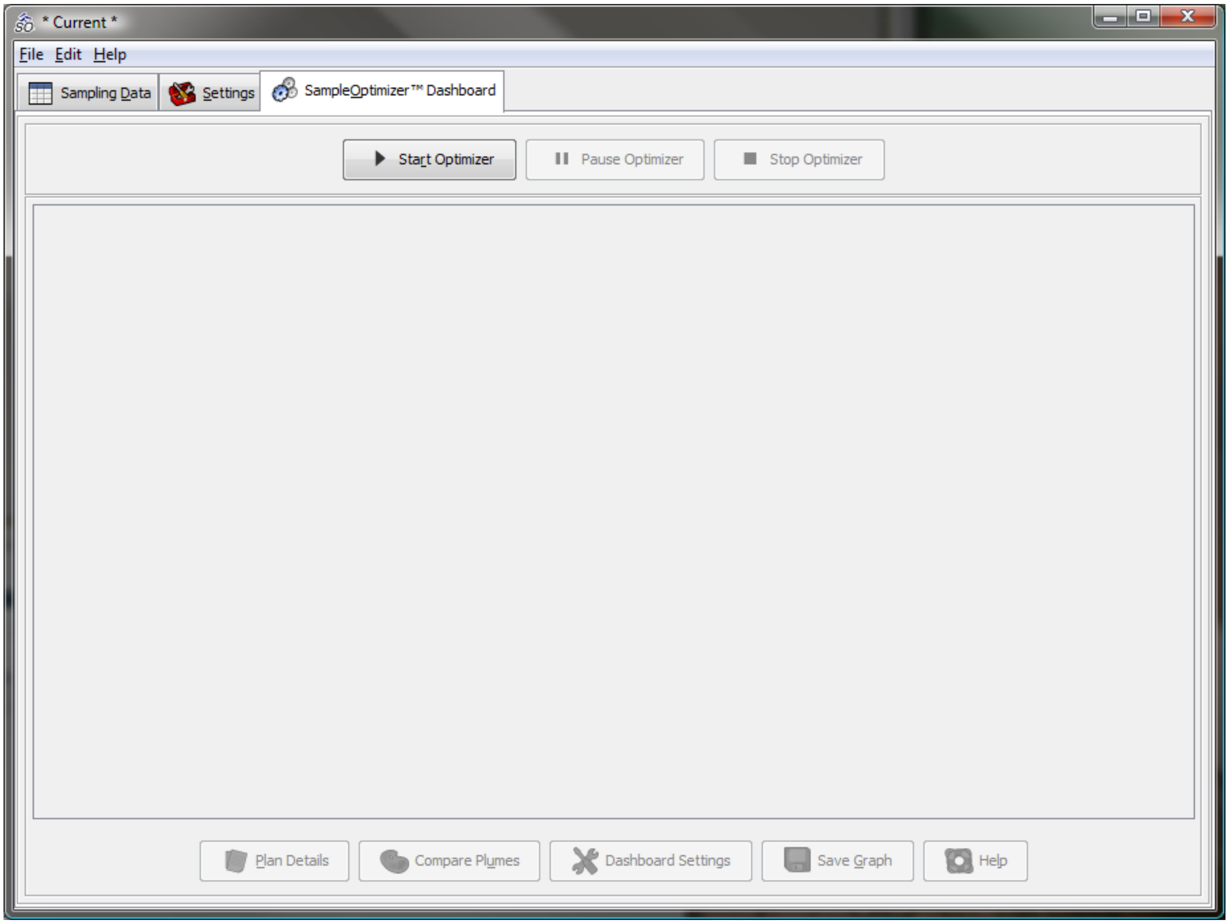


Figure 42. SampleOptimizer™ Dashboard

To begin, click on **Start Optimizer**.

While the optimizer is running, you can click on pause to momentarily halt the progress of the optimization. While you can click on pause at any time while the optimization is running, the optimizer will continue to run until it reaches a checkpoint. Checkpoints occur after every **generation** of the GA.

Figure 43 shows the optimizer paused at a checkpoint. The graph, called a *Pareto Front*, is a visual representation of the current sampling plans under consideration by the optimizer. Each plan is represented by a dot on the graph. By default, the y-axis is the average sampling cost (per event for spatial optimization, and per year for temporal / spatio-temporal optimization), while the x-axis is the modeling error. The axes are user-configurable with a drop-down menu accessible by right clicking on the axis title and using the resulting drop-down menu.

Important! *Error, as used by the "Cutoff" and "Percentage" error objective calculators, does not have a direct physical unit because it is a ratio of the actual error to the acceptable error. When using these Objective Calculators, errors of less than 1 are in compliance with the error objective function.*

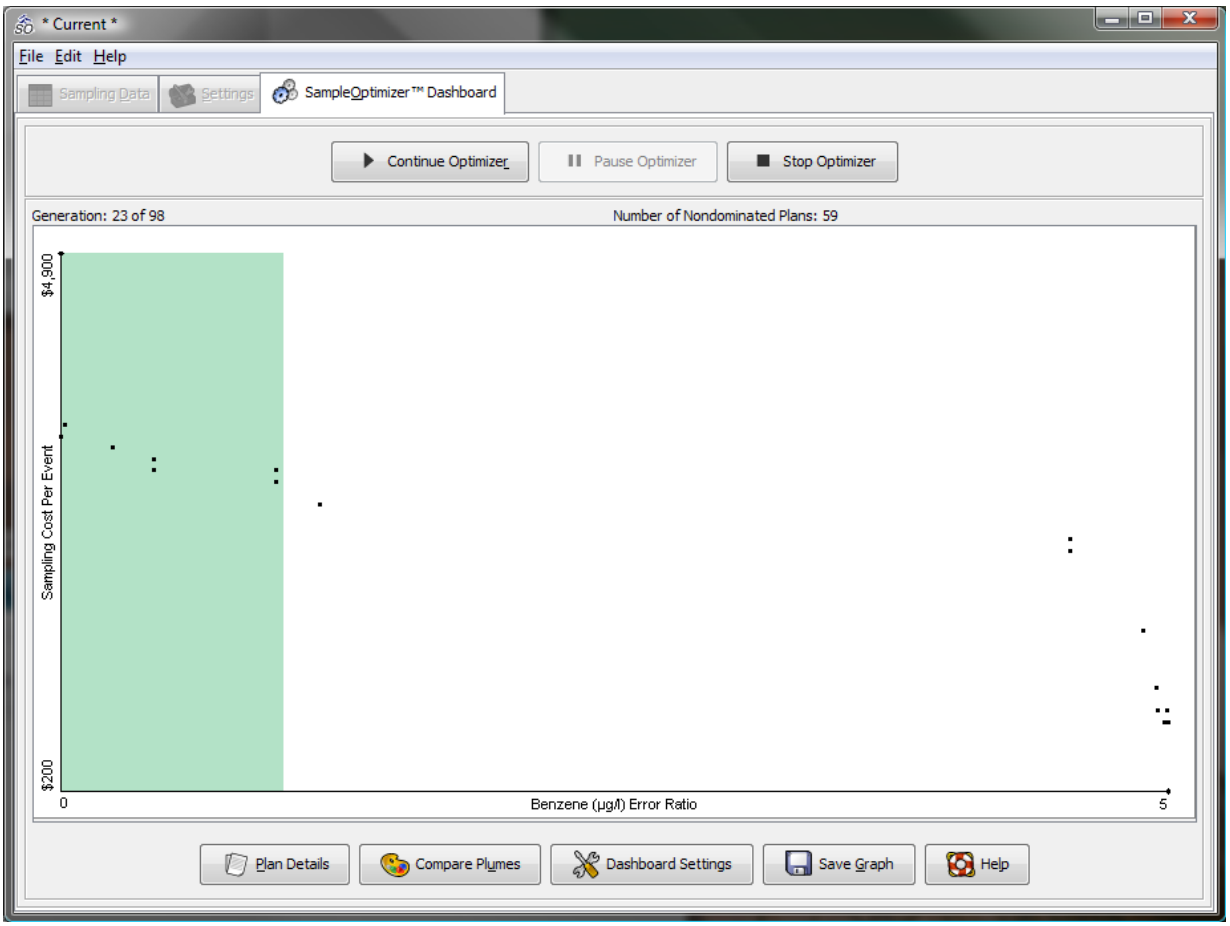


Figure 43. Optimization Paused

The area of the graph highlighted in green encompasses plans which have an error objective less than 1.

Pausing

When the Optimizer is in the paused state (such as in Figure 43), you may interact with the intermediate results.

- Click inside the Pareto Front graph area to bring up the crosshairs.
- Click and drag a selection box around a plan or plans you wish to compare to the maximum sampling plan.
- When the mouse cursor is hovered over the Pareto front a few seconds, information will pop up about that location on the graph, and any plan(s) which may be located therein.
- Hold down the Control key while dragging to add to the current selection.
- Hold down the Alt key to subtract from the current selection.
- Hold down the Shift key to zoom in.
- Right click in the graph to reset the zoom.

There are two options when comparing plans. You can analyze the details of the plan (click **Plan Details**), or look at the plume maps built using them (click **Plumes**).

Dashboard Settings

There are several configurable options which affect the **Plan Details** and **Compare Plumes** functions; these are accessed through the Dashboard Settings button.

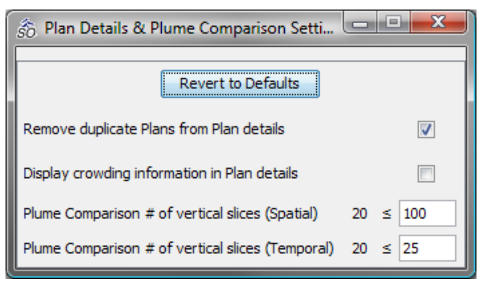


Figure 44. Dashboard Settings

- Show plan details
  - Remove duplicate Plans (enabled by default)
    - By default the software only shows unique plans; if disabled, duplicate plans will be shown in the Pareto front and in the plan details window.
  - Display crowding information (disabled by default)
    - This is extra information typically of use only to a GA expert.
- Default Plume Comparison # of vertical slices
  - Sets the default vertical resolution <u>when generating plume images from the Pareto front</u>, for both spatial and temporal analyses. (note – this value does not apply to the "model visualization settings" used within *ModelBuilder™, which has a separate input value for # of vertical slices)*

Plan Details

Figure 45 and Figure 46 show the plan details comparison for a spatial optimization. Note: rank is a GA term showing which front a solution belongs to. The 0th rank consists of the front of non-dominated solutions.

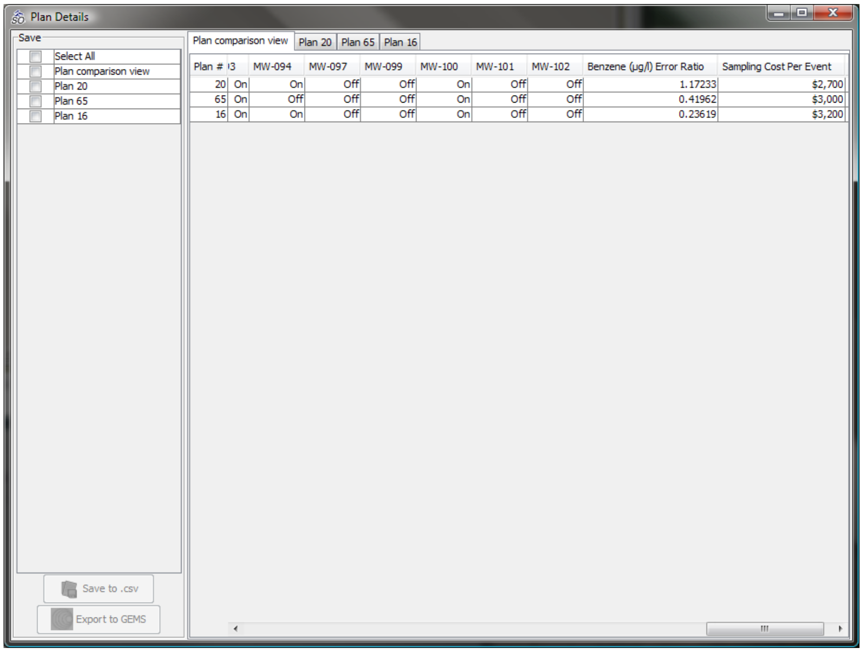


Figure 45. Plan Details - Comparison (Spatial)

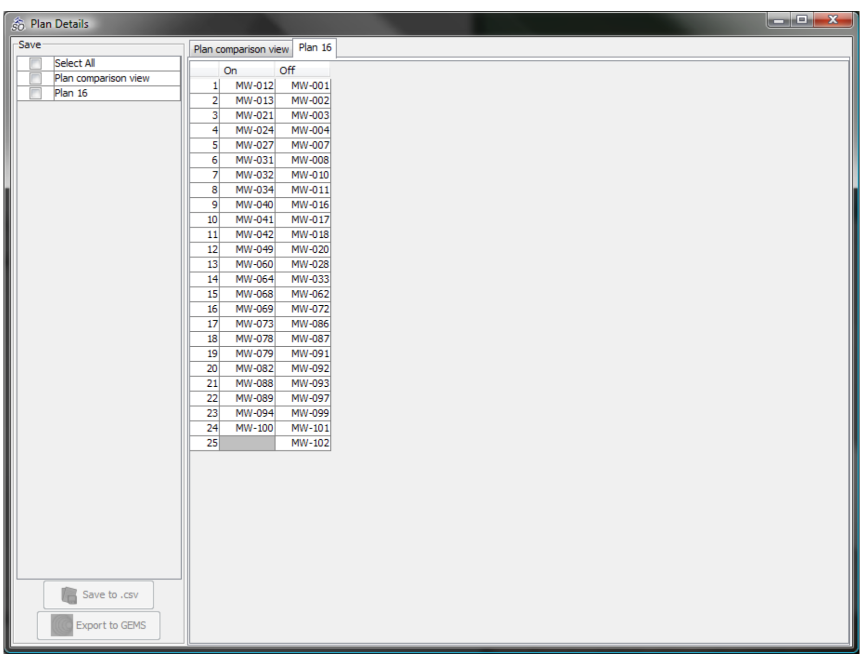


Figure 46. Plan Details - Individual (Spatial)

Figure 47 and Figure 48 show the plan details comparison for a temporal optimization.

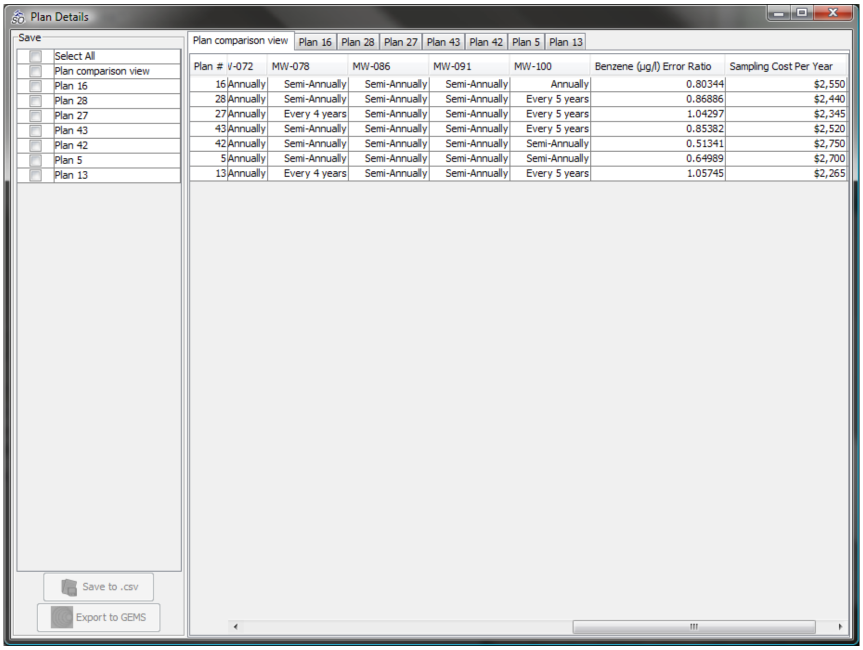


Figure 47. Plan Details - Comparison (Temporal)

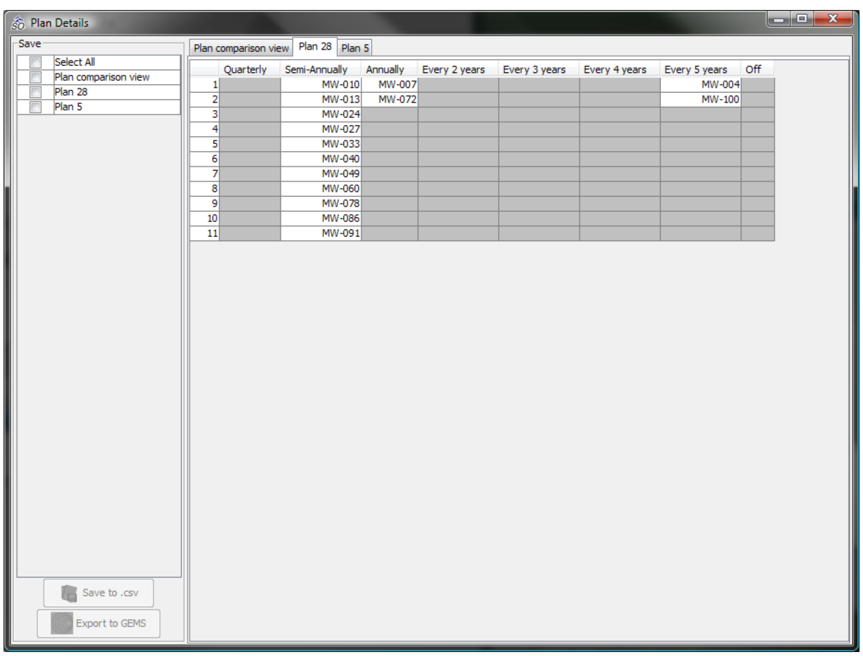


Figure 48. Plan Details - Individual (Temporal)

To export the plan details to .csv format data files, select the details you would like to save in the **Save** portion of the window and click **Save to .csv**.

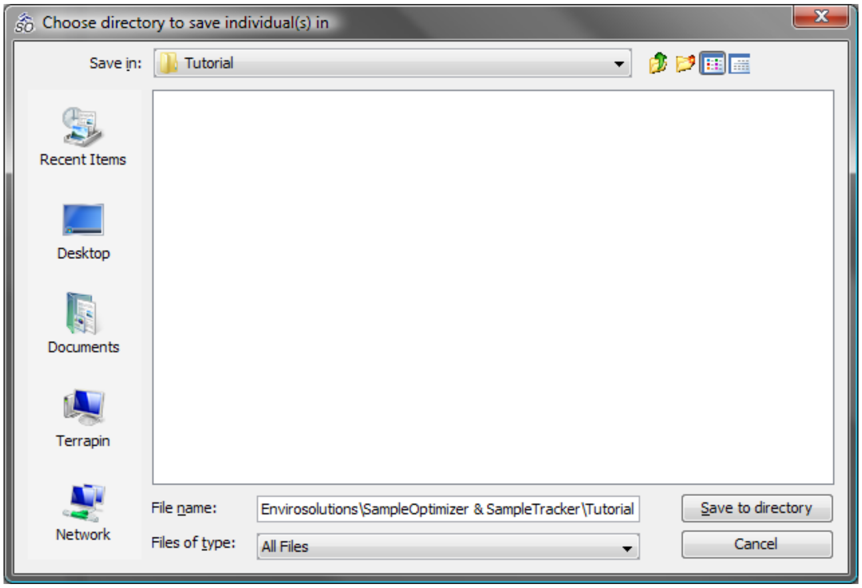


Figure 49. Save Plan Details

Here is what the saved files look like when opened in *Microsoft Excel*:

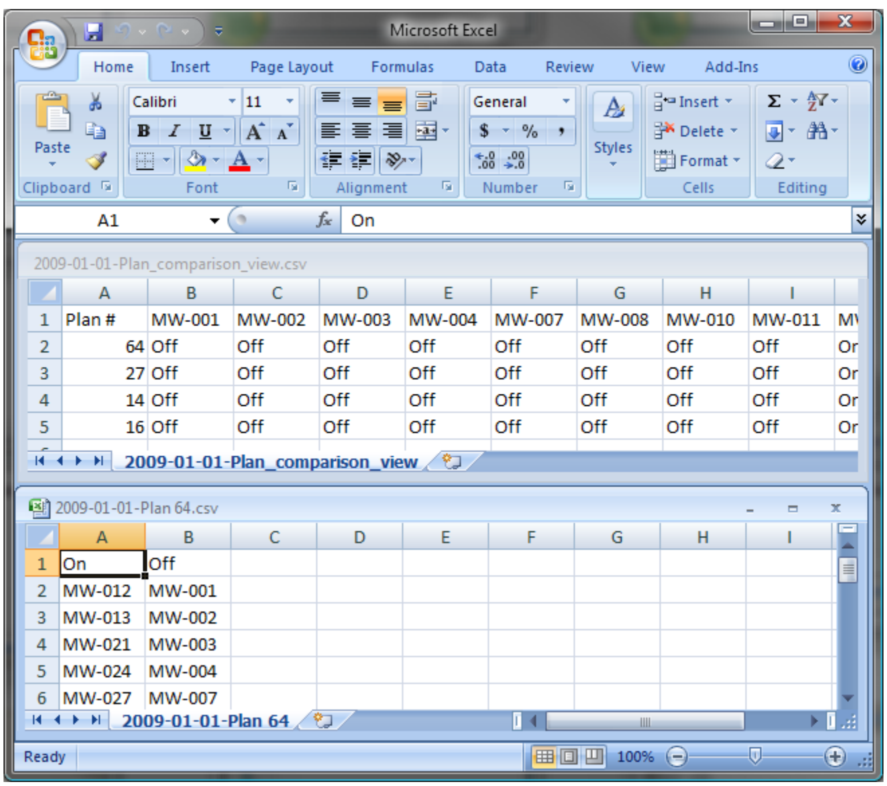


Figure 50. Saved Plan Viewing

In addition to saving a plan to a .csv, you can also export it into a Summit *GEMS* database location group. Once in *GEMS*, the location group can be exported to a shapefile for viewing in ESRI *ArcView*.

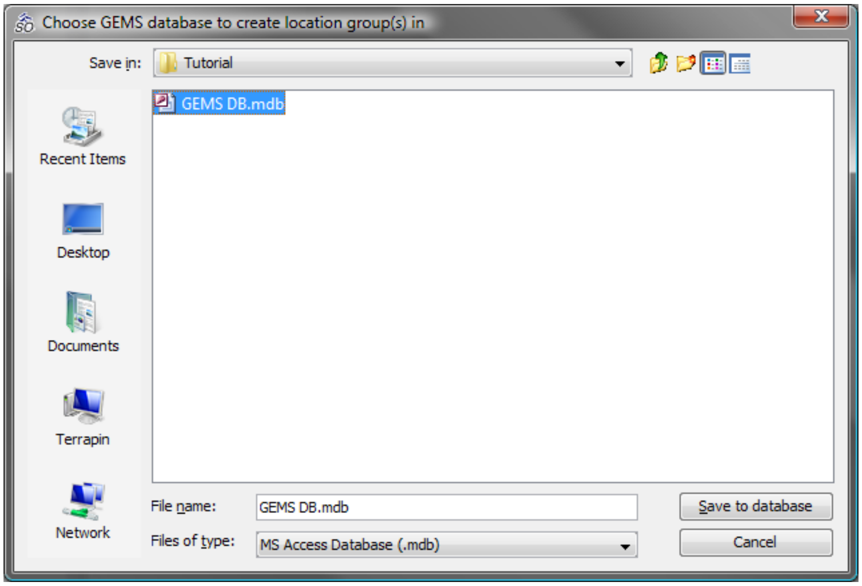


Figure 51. GEMS Export

Plume Maps

Clicking on the Plumes button will instruct the software to generate plume maps for viewing. They will automatically pop up when they are ready. Please note that while you are viewing plumes, you cannot interact with the **SampleOptimizer™ Dashboard**.

Here is what plume comparison looks like with a spatial optimization, showing a 50% sampling cost reduction (note that a "+" symbol in the figure indicates a location that was proposed for removal in the sampling plan):

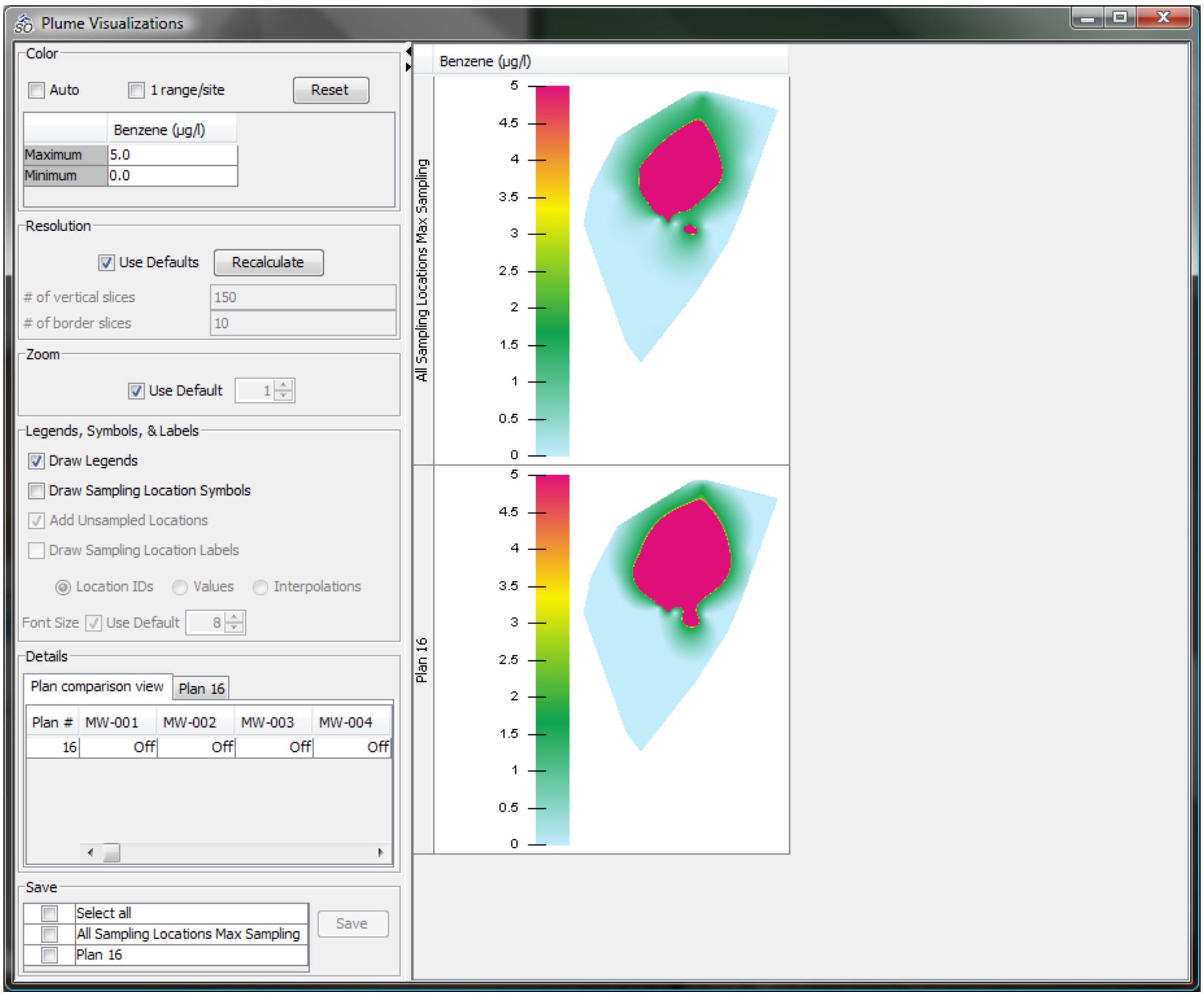


Figure 52. Plume Map Comparison (Spatial)

Here is what plume comparison looks like with a temporal optimization (the actual plume map illustrated is for one sampling event, but all sampling events can be accessed and illustrated):

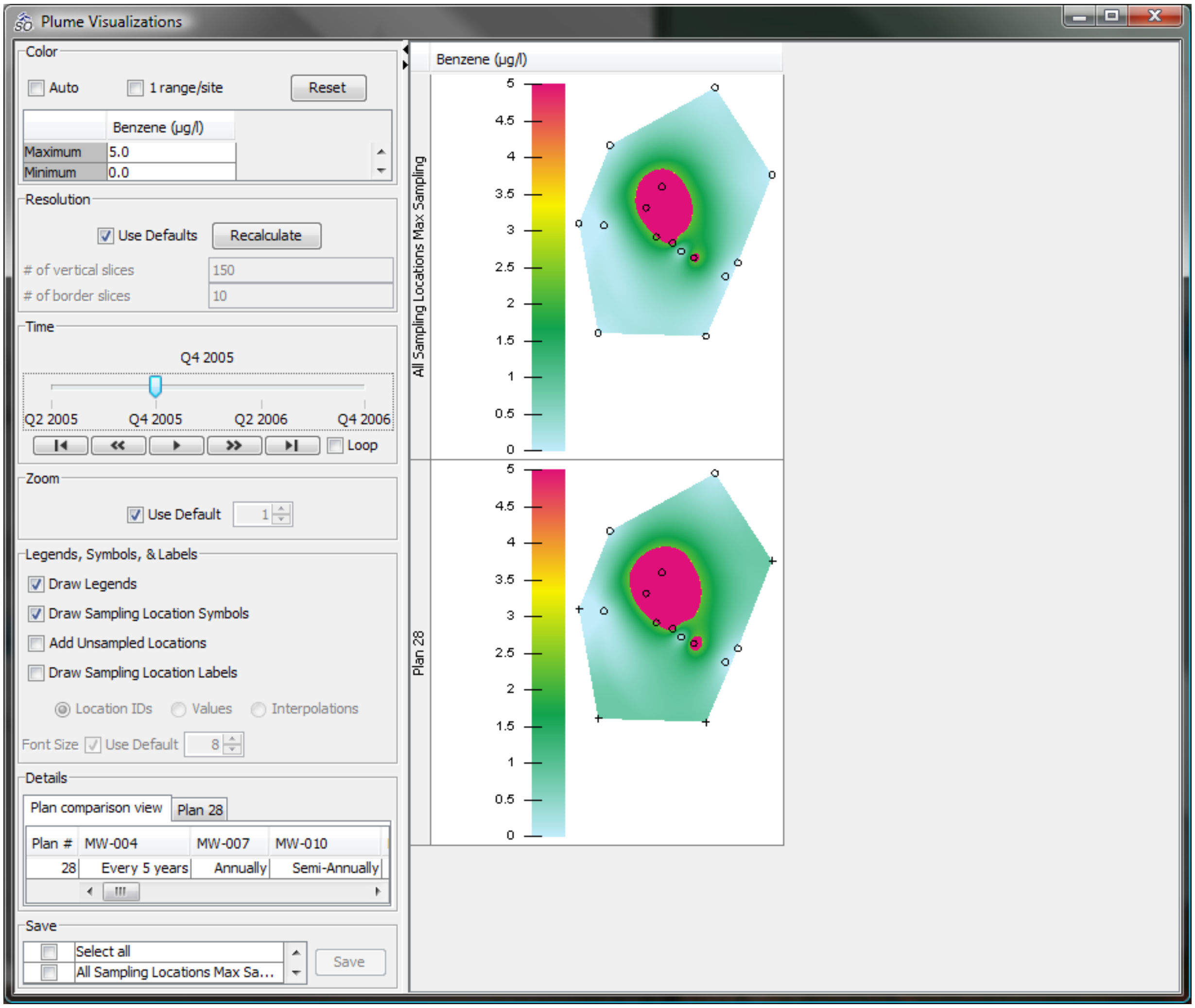


Figure 53. Plume Map Comparison (Temporal, Ordinary Kriging)

To export the plan details to .csv format data files, select the details you would like to save in the **Save** portion of the window and click **Save to .csv**.

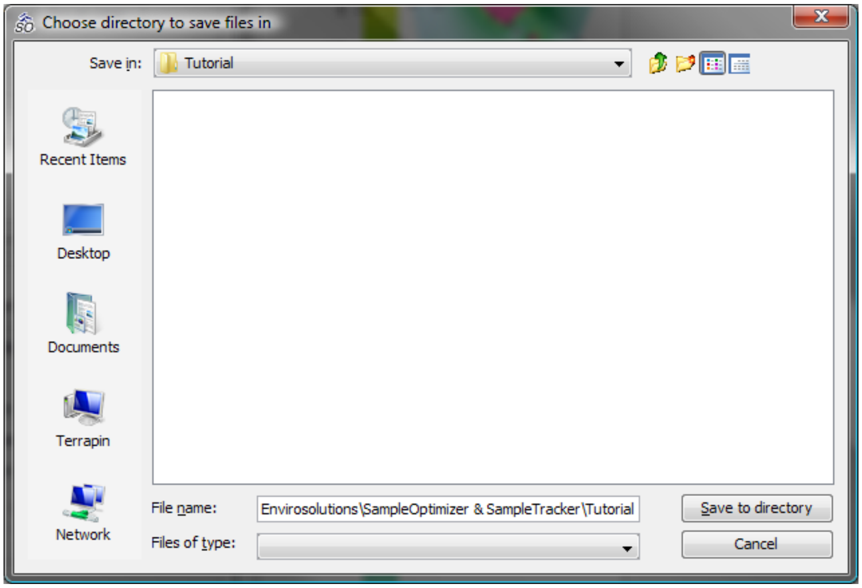


Figure 54. Plume Map Saving

Here is what the saved files look like when opened in *Microsoft Windows Vista's Windows Photo Gallery*:

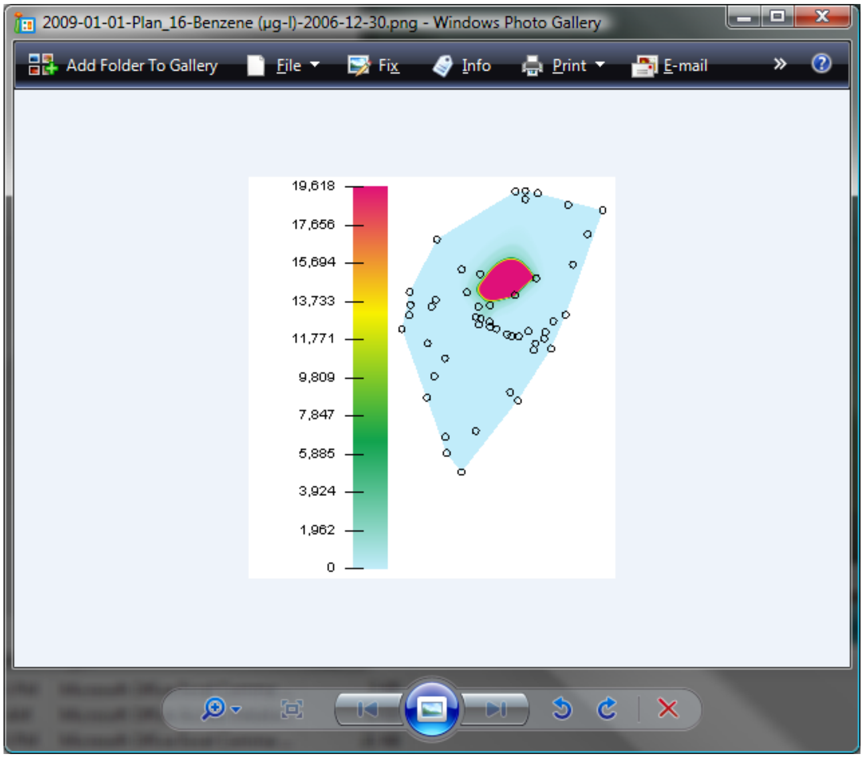


Figure 55. Saved Plume Map Viewing

SampleTracker™

Getting Started

Before you begin using *SampleTracker™* (abbreviated as ST), you will need both a historical dataset and current dataset to work with. If you have generated output files from *ModelBuilder™* for Mass Metric and/or Mass Flux, you need to combine them into your historical and current datasets if you wish to track those metrics. In this case, please pay attention to the report dates used in *ModelBuilder™* to insure they agree with the other data you input to *SampleTracker™*.

As with *SampleOptimizer™*, the currently supported data standard used by *SampleTracker™* for data input & output is .csv.

Tips and Tricks

If you see an option in the software and are unsure about it, allow your mouse cursor to hover over it for a few seconds and a "Tool tip" will pop up and give you more information about that option.

Overview

SampleTracker™ reviews current monitoring data against historical data and identifies cases where current data deviates from expectations that are based on the historical dataset. These expectations are formulated in the form of **bounds**, which represent a statistical estimate of the range of values that might be expected in a specific sampling event, based on the historical data at that location. Any current data value which is outside the bounds range for that location is reported to the user as being out-of-bounds.

ST includes two types of bounds calculators: static and time-dependent. The software automatically makes a recommendation for the bounds type to be used based on the historical data. Advanced users can change the bounds type from the default based on the expectations for a given sampling location.

Electronic Data Deliverable (EDD)

The EDD for *SampleTracker™* is similar to that used by *SampleOptimizer™*; the main difference is that in SampleTracker™, historical data and current data must be separated into two files: one for the historical dataset, and another for the new (current) dataset to be evaluated.

Mass Metric and Mass Flux Tracking

SampleTracker™ can track Mass Metric and Mass Flux data via the *ModelBuilder™* component of *SampleOptimizer™*. Please see Mass Metric and Mass Flux for more information on how to use these features.

Non-Detects in SampleTracker™

- For best performance of the bounds calculators, zero values should not be used. Instead, non-detect data with typical RLs should be replaced by a numerical substitute value. ½ RL is commonly used, but keep in mind that there is no optimal substitute value.
    - In particular, zero values cannot be used with time-dependent decreasing bounds, since those use log-transformed data, and very small artificial or substitute values can distort any actual trends that may be present.
- If the typical RLs vary from event to event, using a common value such as ½ the median of typical RLs across events may be useful in order to avoid introducing trends where there aren't any.

EDD Notes

- ND values with "high" reporting limits should be screened by the user because they can potentially cause the upper bounds calculated by the software to be higher than may be desired. When 4 or more historical detects exist for a COC at a location, any current or historical ND for that COC at that location with reporting limit (RL) above the 25th percentile of the detected historical values for that COC at that location should be eliminated from the analysis.

- Follow the duplicate sample guidelines in the SampleOptimizer™ EDD for any duplicate samples you find in your historical dataset. If your current dataset has duplicates, you can screen them separately by adding them into the current dataset one at a time.

Date	SiteID	EastCoordinate	NorthCoordinate	Benzene	Chlorobenzene
4/9/1995	BL003	110	502	4.1	240.0
4/9/1995	OS004	233	438	0.9	50.0
4/17/1995	OS003	220	224	100.0	1800.0
4/16/1995	OS005	157	525	0.5	
3/11/1996	MWSL001	197	386	2.7	8.2

Figure 56. Example SampleTracker™ Historical Data File

Date	SiteID	EastCoordinate	NorthCoordinate	Benzene	Chlorobenzene
4/12/1997	BL003	110	502	1.0	180.0
4/13/1997	OS004	233	438		50.5
4/11/1997	OS005	220	224	0.5	0.5
4/14/1997	MWSL001	157	525	1.5	0.5

Figure 57. Example SampleTracker™ Current Data File

Managing Historical Data

A minimum number of historical samples are required for a sampling location / COC pair to be analyzed; the user may specify what the minimum number is, but it must be at least four samples (see Figure 56). In particular, for formal RCRA detection monitoring programs a minimum of 8 historical samples is recommended [1].

In general, we recommend that the user add new samples to the historical set for each sampling location / COC pair until it contains at least 8 historical samples present per sampling location / COC pair, keeping in mind that the background data should be representative of values that can be expected in the future.

However, in some monitoring situations the interested parties agree that data from a certain period is representative of what should be expected from that location. This comes up particularly in regulatory settings, where the agency reviews data and approves, or when by regulation the first two years of data forms the “official” historical dataset. In these cases you would keep the historical dataset the same until there is good reason to update it, such as if you start getting repeated exceedences and the interested parties agree that the situation is understood and the old historical data are no longer relevant.

Please note that regularly adding new samples to the historical set may cause the software to not catch any mild trend(s) which may be occurring. If one takes the position that such data are essentially consistent with an in-control process, this is acceptable. Otherwise, the historical set should be kept the same until an evaluation is triggered, and then it should be updated (keeping the last 8, say, observations) after the interested parties agree.

Loading Data

Upon running *SampleTracker™*, you will be presented with the following screen.

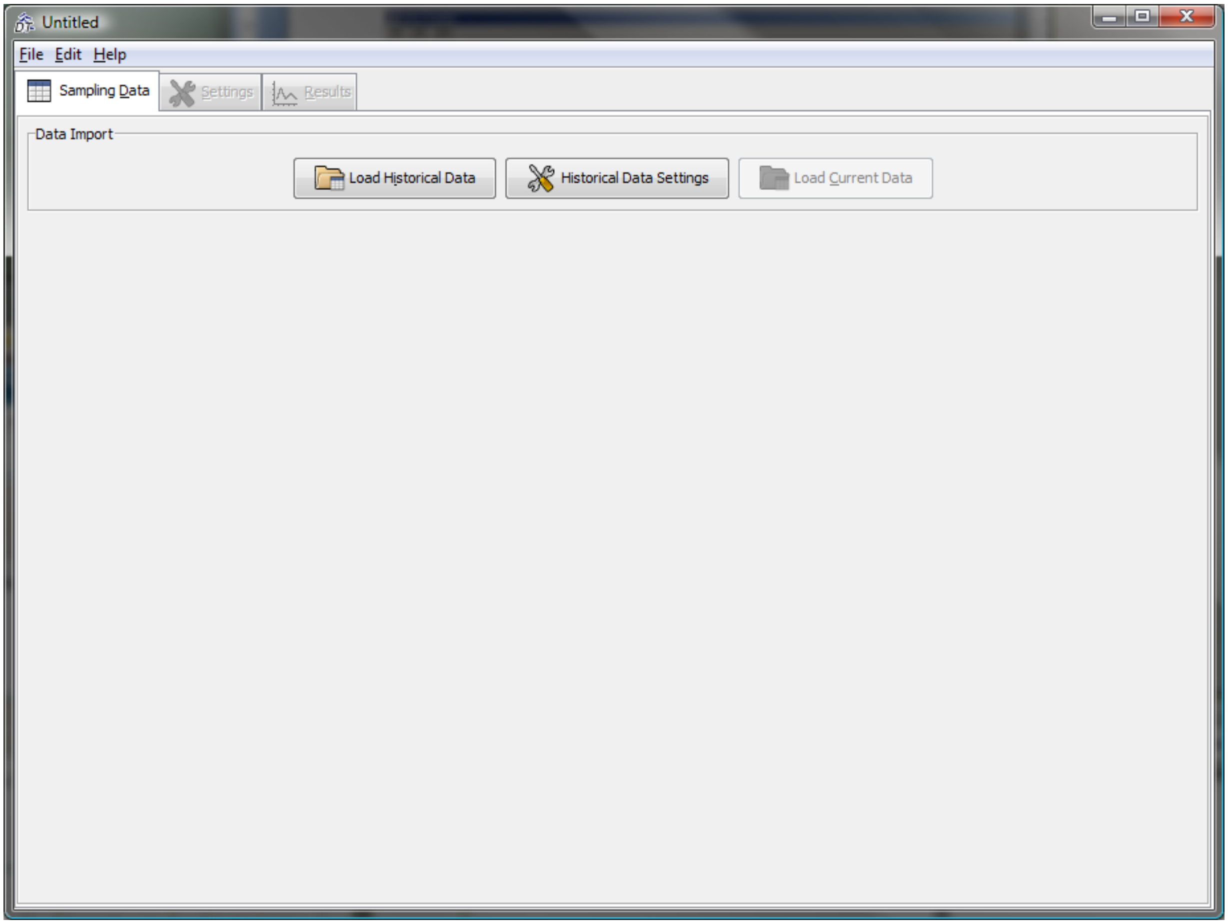


Figure 58. SampleTracker™ Initial Screen

Application Options

At any time, you can adjust the applications options. There are four settings which may be changed here.

- Program Path
  - This is for technical support use only.
- Zero substitute value
  - Time-dependent bounds use natural log scale, and since the log of zero is undefined, a small value must be substituted for zero values in the dataset. That value is specified with this parameter. Note that we do not recommend using zero values at all; as stated earlier, non detects in SampleTracker™ should be replaced by a fraction of the RL.
- Slope test upper tail %
  - This is the confidence value for the slope test done when recommending the bounds type for each location / COC pair.
- Number of digits in result tables
  - The number of digits displayed in the bounds result tables can be adjusted here.

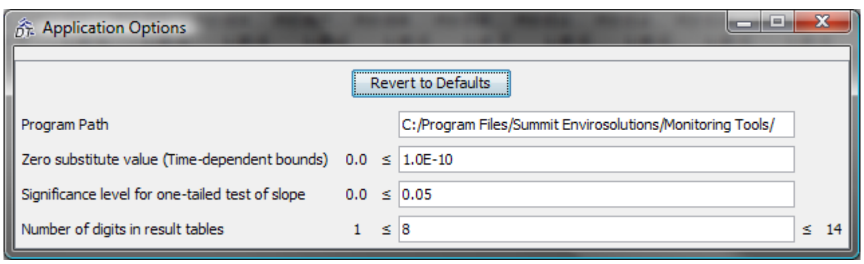


Figure 59. Application Options

Loading Historical Data

Before loading the historical data, you can adjust the historical data settings if desired (Figure 60). This parameter controls the minimum number of historical samples which are required for SampleTracker™ to analyze a COC at a given location.

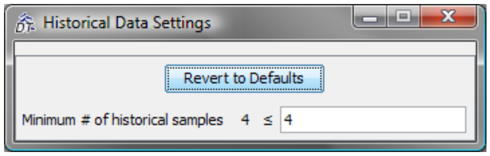


Figure 60. Historical Data Settings

To start a new analysis, click **Load Historical Data** and locate the file you wish to load, or just drag the historical data file onto the application window.

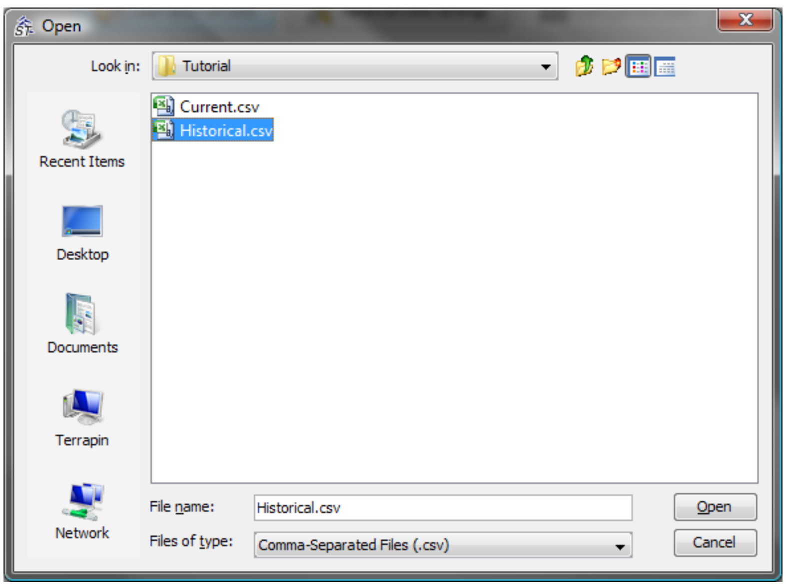


Figure 61. Loading Historical Data

Figure 62 shows an example screen after a historical data file has been input into the software.

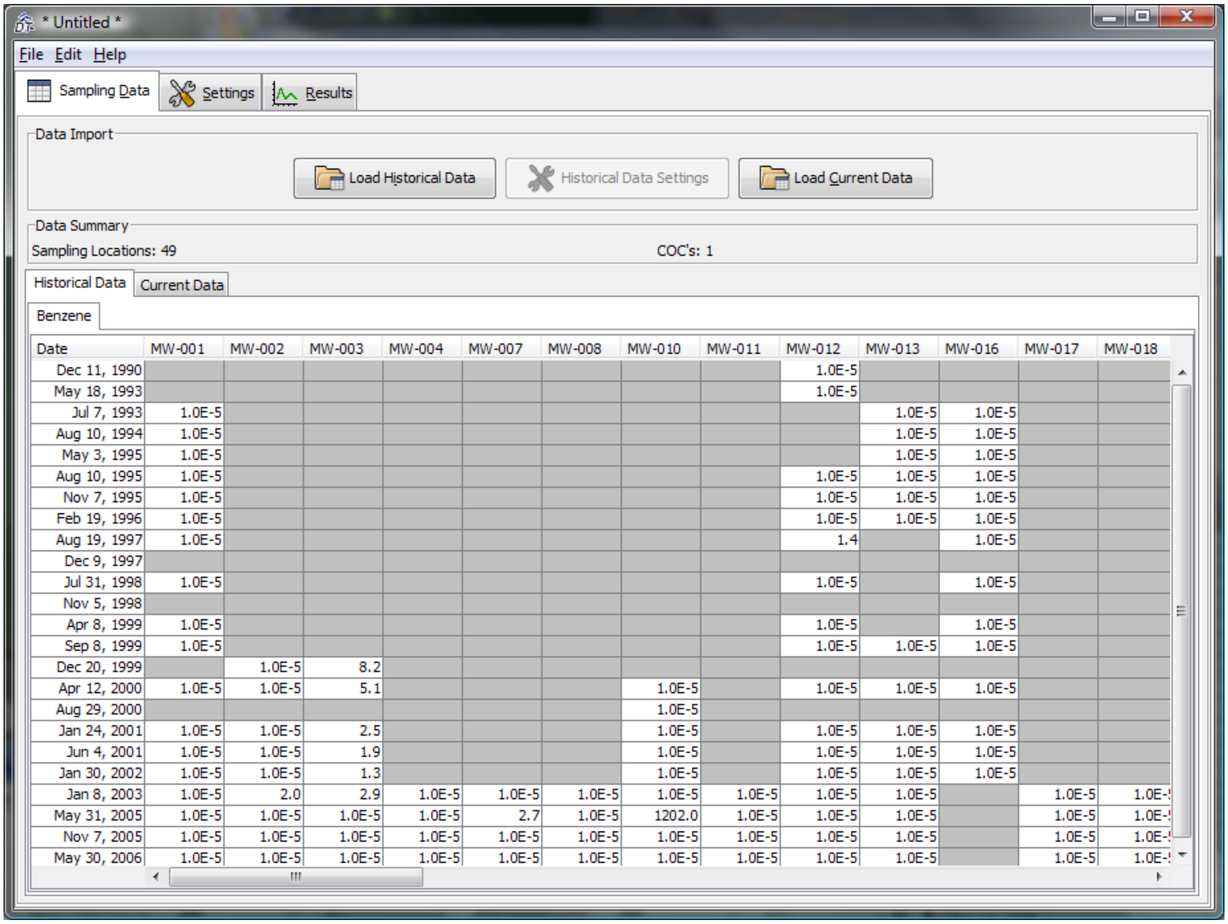


Figure 62. Historical Data Loaded

Next, click on **Load Current Data**.

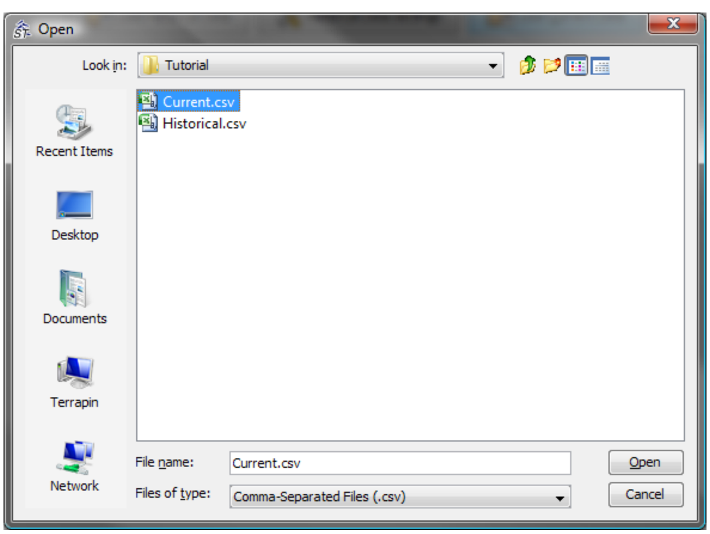


Figure 63. Loading Current Data

Here is an example showing the screen after both a historical data file and a current data file have been loaded into the software.

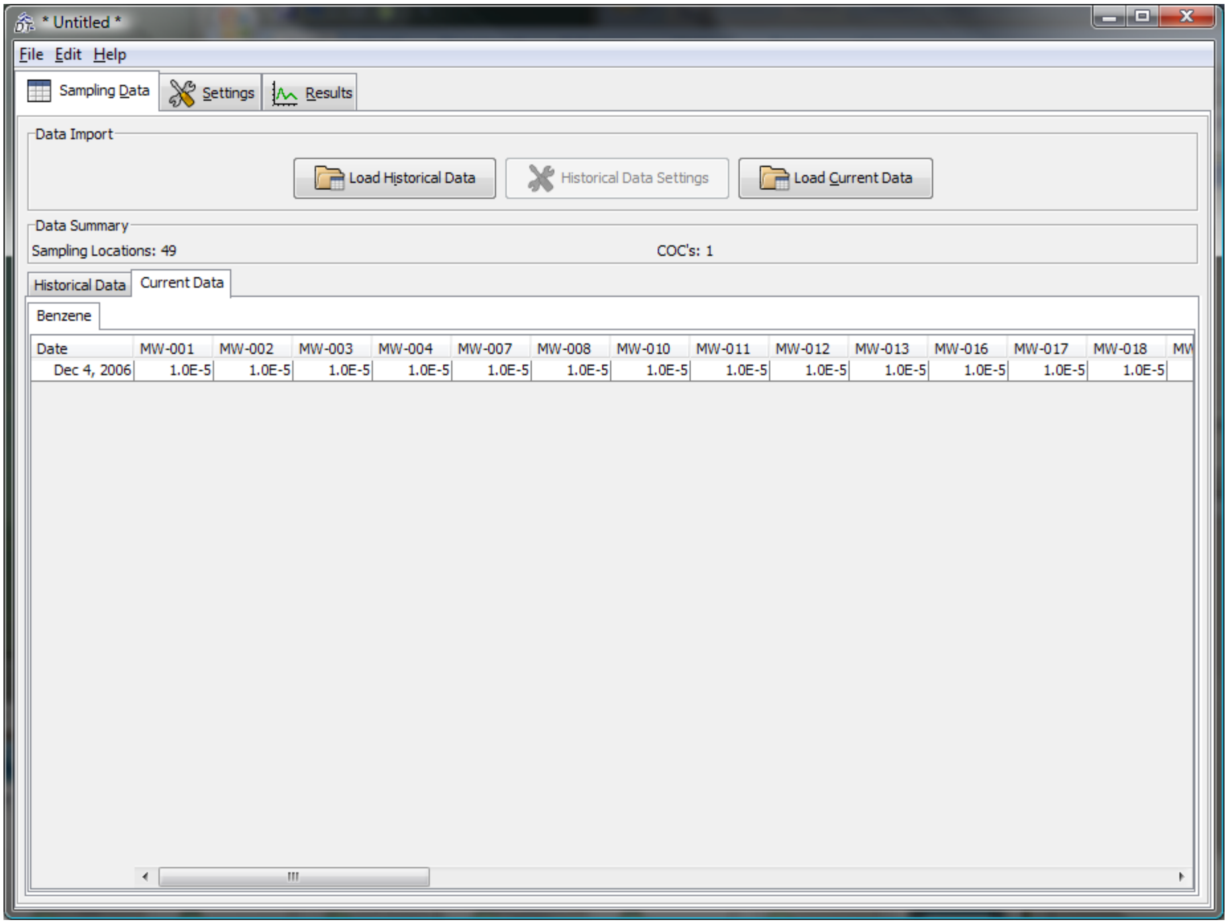


Figure 64. Current Data Loaded

Settings

After loading the historical and current data, you can view and/or change the settings for the bounds calculation in the Settings tab (Figure 65). SampleTracker™ automatically recommends static or time-dependent bounds based on the data history at each location. You can also specify the parameters used in the bounds calculators themselves, by clicking **Bounds Calculation Settings** (Figure 66).

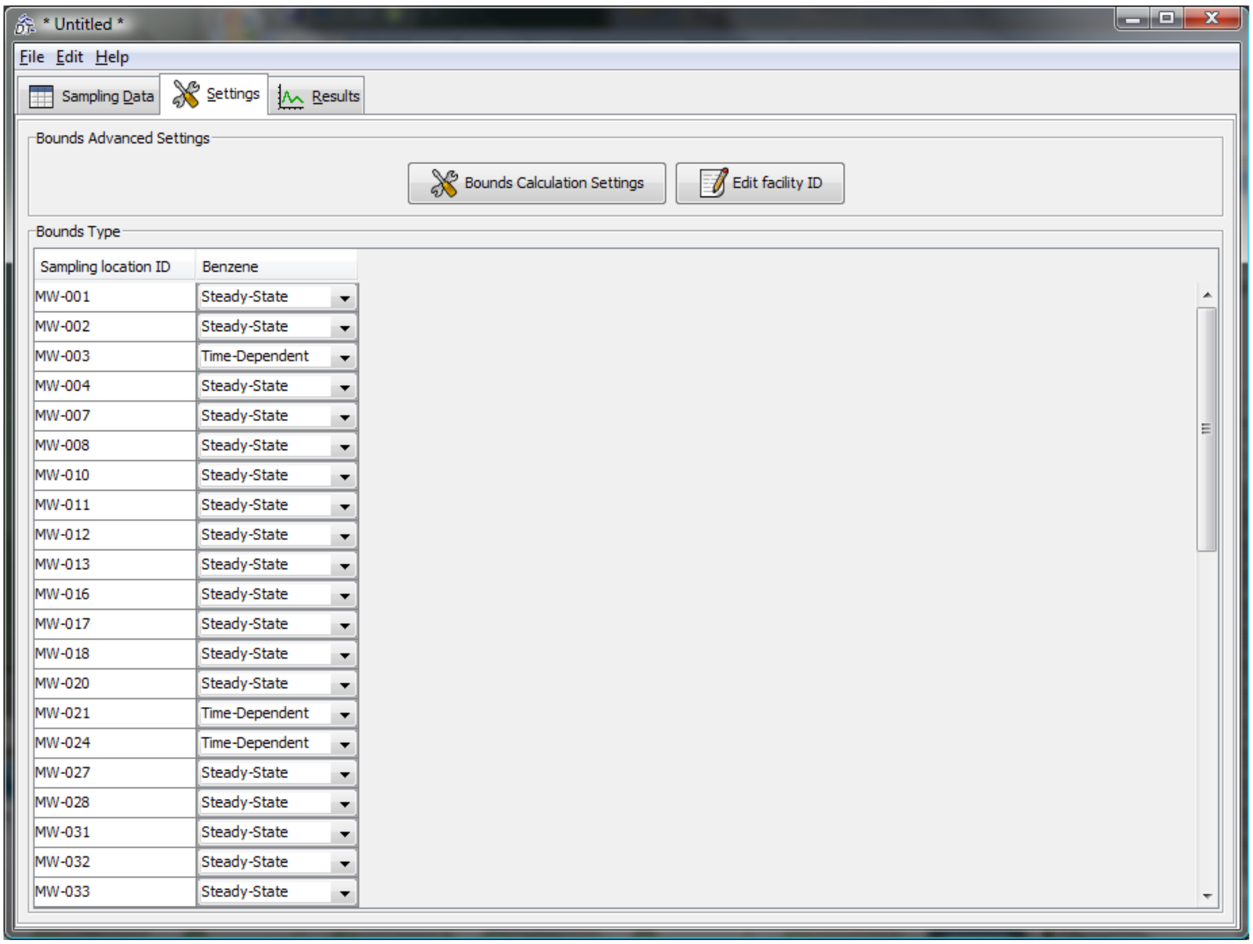


Figure 65. Bounds Type

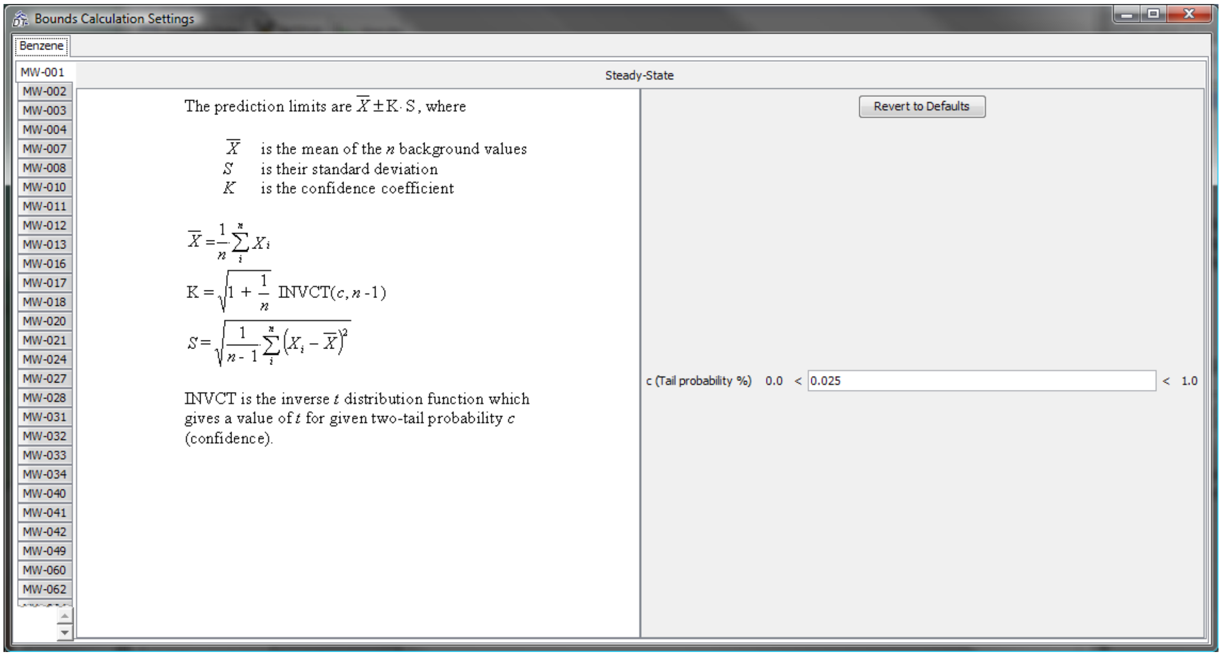


Figure 66. Bounds Calculation Settings

Results

After you have input the data and adjusted any necessary settings, you are ready to have the software compute the results.

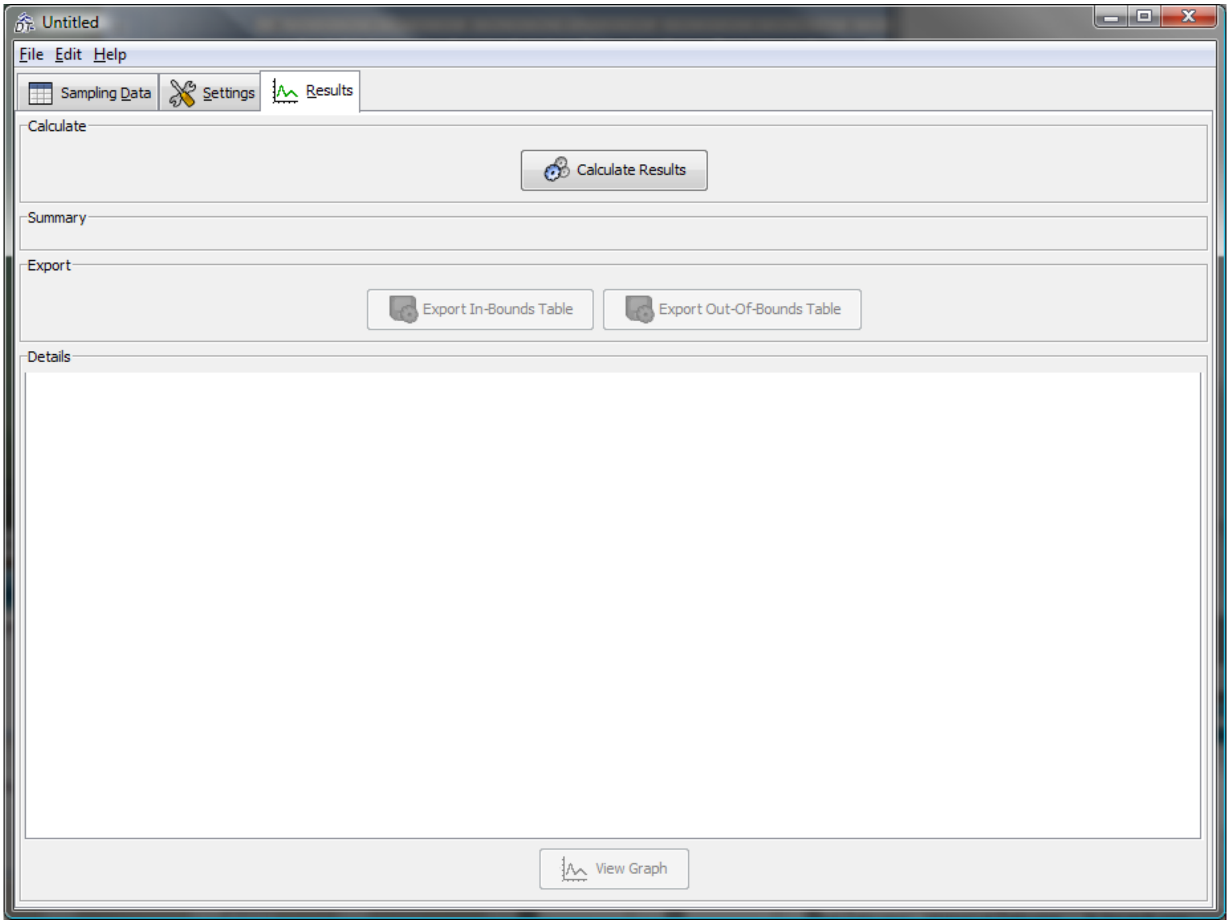


Figure 67. SampleTracker™ (Ready to Compute Results)

After you click on **Calculate Results**, you will see a screen similar to this one:

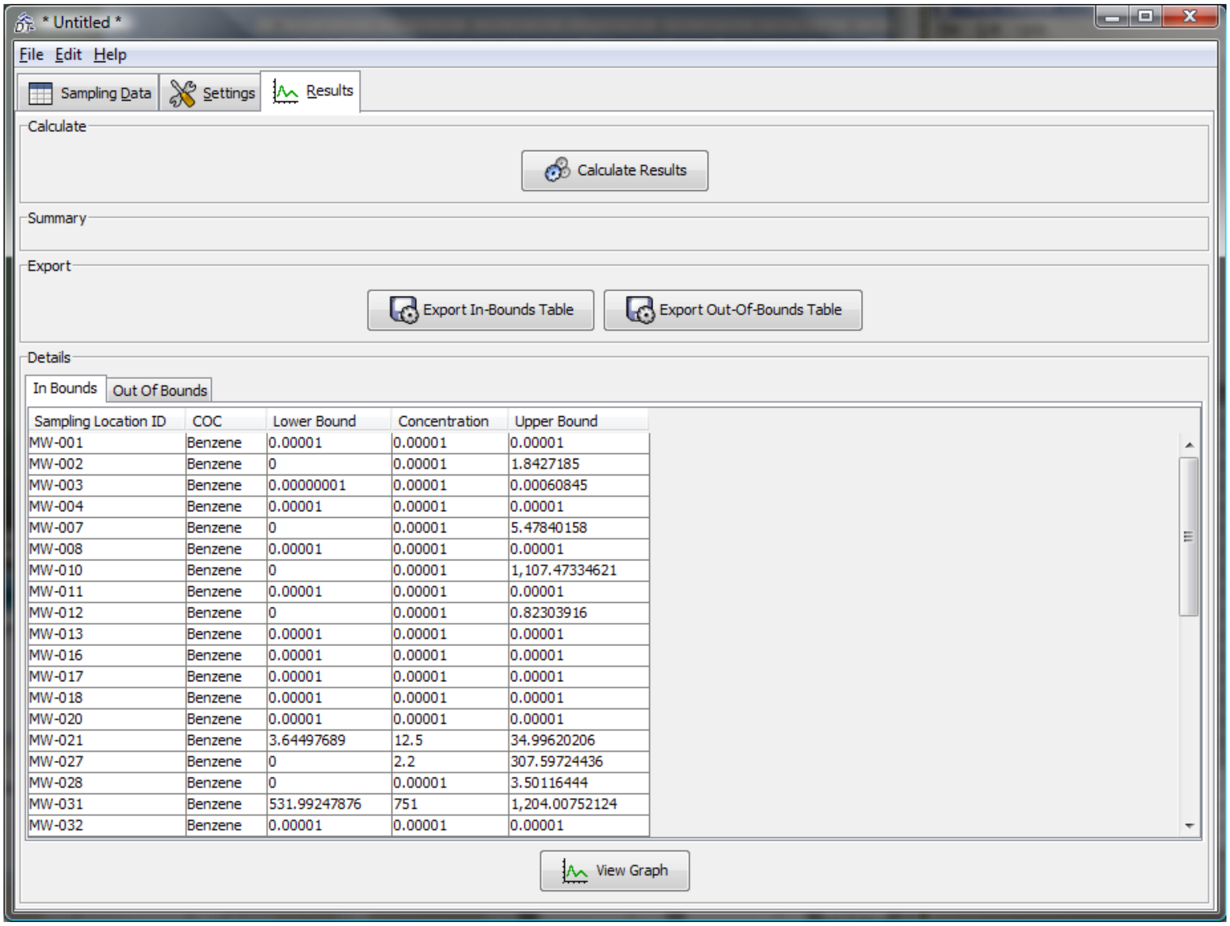


Figure 68. In Bounds Results

If any COC's were out-of-bounds at any sampling locations, they will be visible in the **Out-of-Bounds** tab.

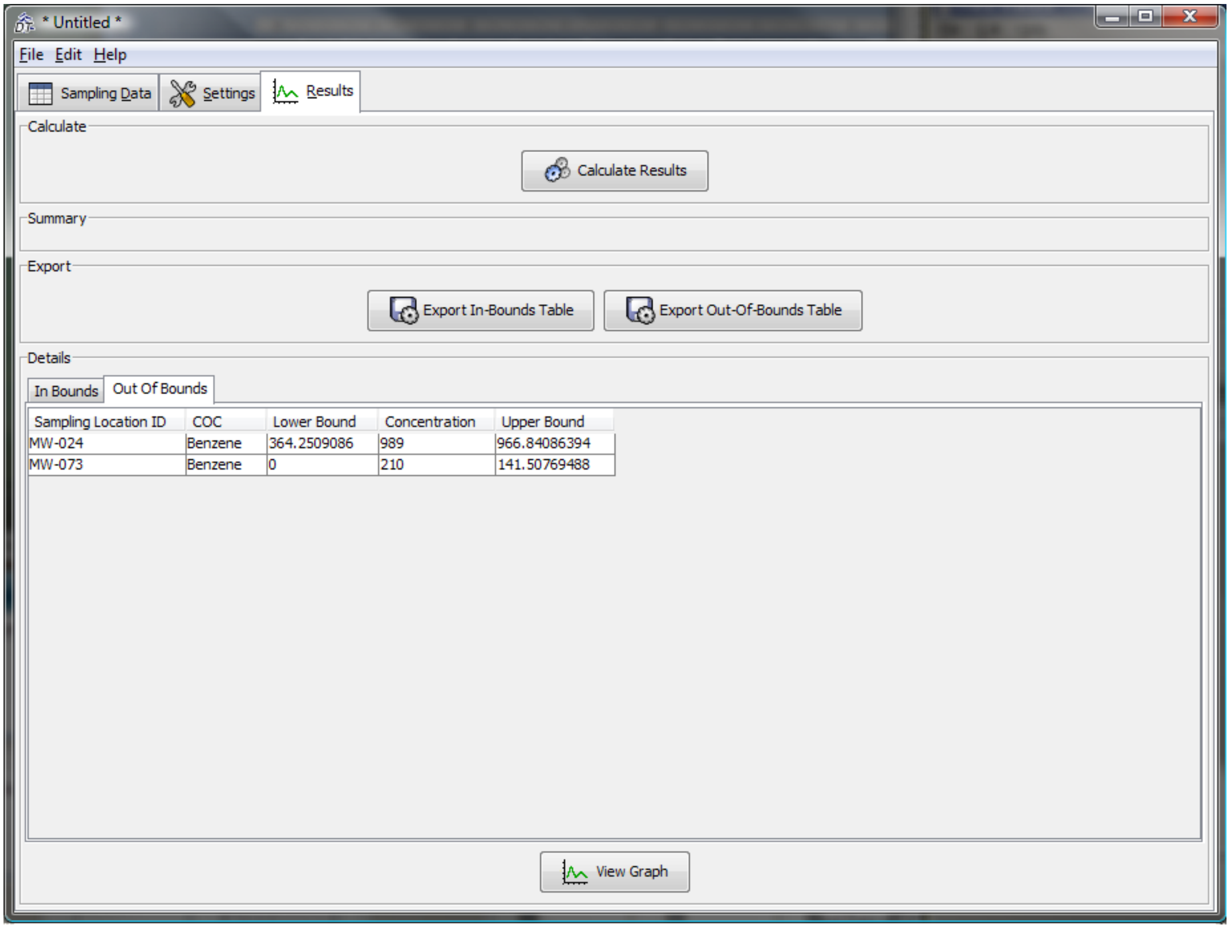


Figure 69. Out-Of-Bounds Results

To save the result tables, hit either the **Export In-Bounds Table** button, or the **Export Out-Of-Bounds Table** buttons.

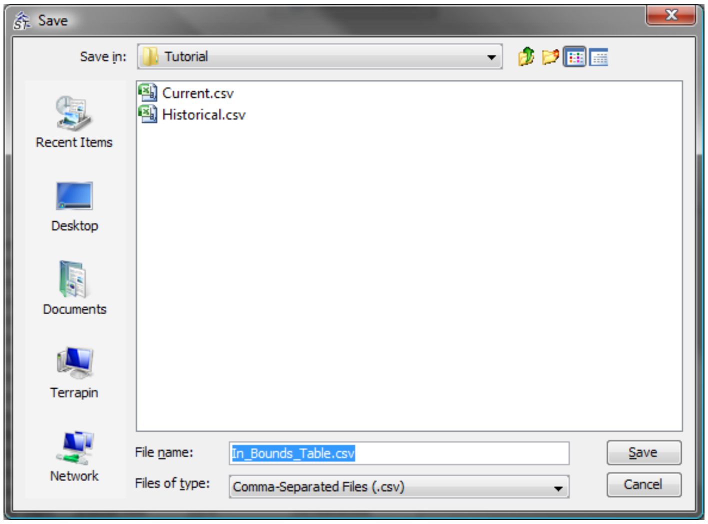


Figure 70. Exporting In-Bounds Table

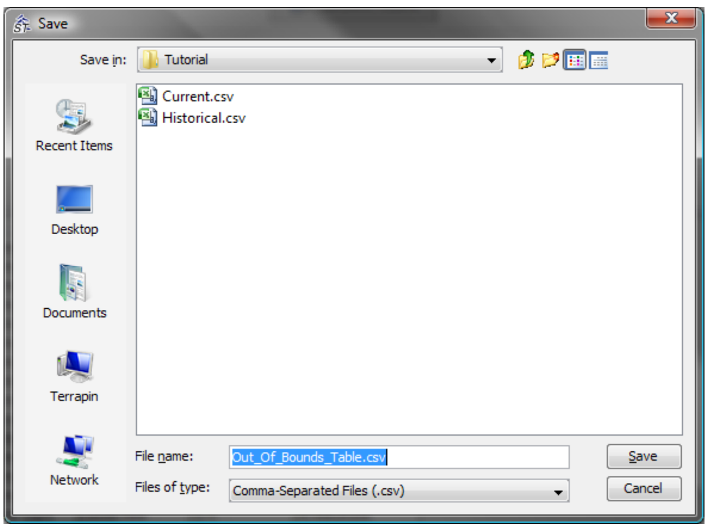


Figure 71. Exporting Out-Of-Bounds Table

To view a graph, simply select a row in the table and hit **View Graph**.

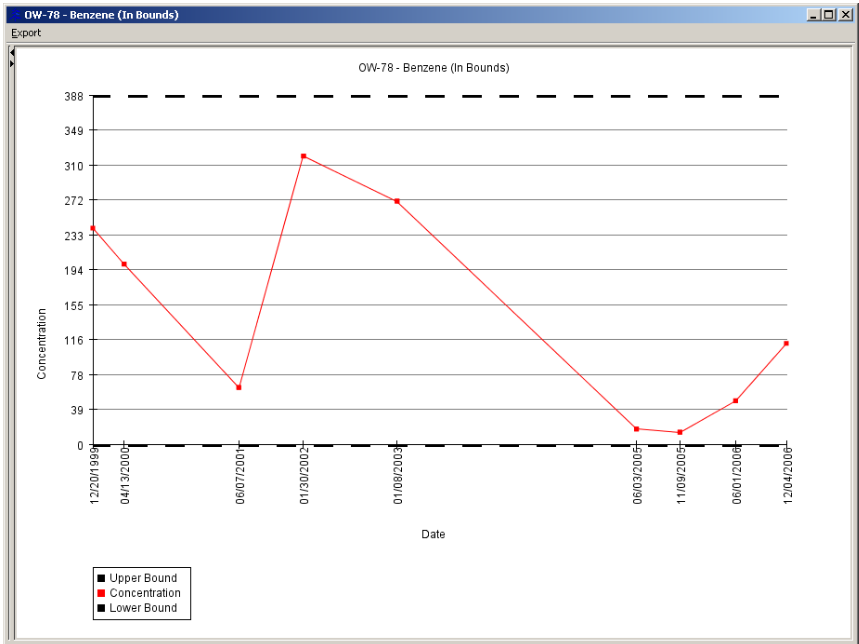


Figure 72. Viewing Graph

To export a graph to a .png graphics file, click on **Export graph to .png** from the **Export** menu.

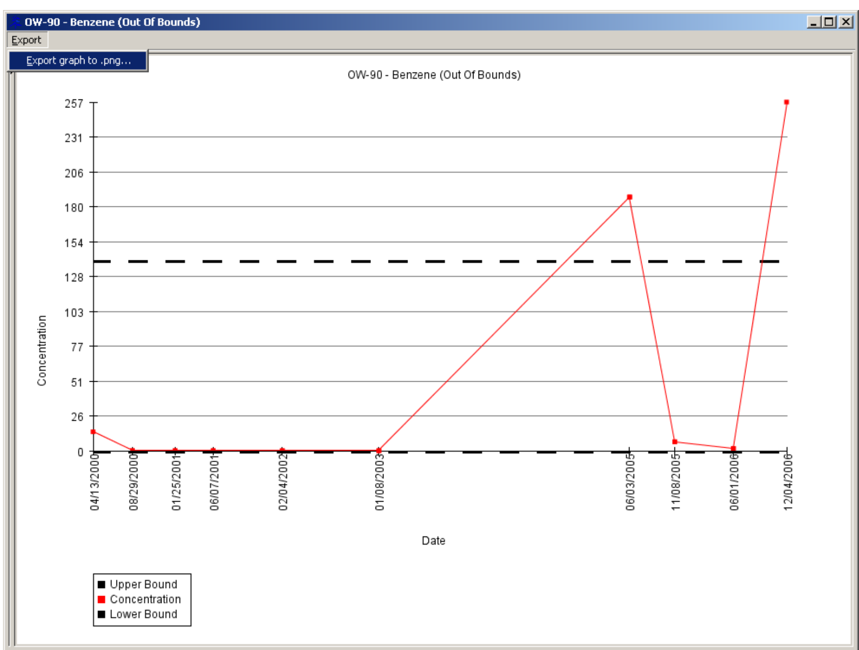


Figure 73. Exporting Graph (1 of 2)

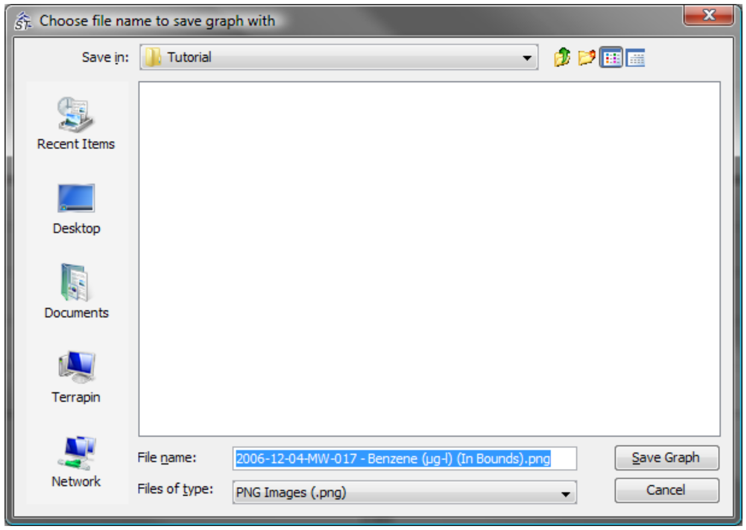


Figure 74. Exporting Graph (2 of 2)

Index

Appendices

Appendix A – Installation prerequisites

Microsoft .net Framework Runtime 2.0:

http://www.microsoft.com/downloads/details.aspx?familyid=0856EACB-4362-4B0D-8EDD-AAB15C5E04F5&displaylang=en

Sun Microsystems Java JRE 6.0:

http://java.sun.com/javase/downloads/index.jsp

Appendix B – Additional References

Websites

1. LTMO
   a. Federal Remediation Technologies Roundtable
      http://www.frtr.gov/optimization/monitoring/ltm.htm

2. Conceptual Site Model Development
   a. EPA Triad
      http://www.triadcentral.org/mgmt/splan/sitemodel/index.cfm
   b. Lawrence Livermore ERD:
      http://www-erd.llnl.gov/rescue/Topics/SCModel/outline.html

3. CSV
   a. Wikipedia
      http://en.wikipedia.org/wiki/Comma-separated_values

Books

1. EWRI Task Committee on the State of the Art in Long-Term Groundwater Monitoring Design, ***Long-Term Groundwater Monitoring: The State of the Art***, ed. by B.S. Minsker, American Society of Civil Engineers, Reston, VA, 2003.

2. Isaaks, E.H. and Srivastava, R.M. ***An Introduction to Applied Geostatistics.*** Oxford University Press, New York, 1990.

3. Goldberg, D.E., ***Genetic Algorithms in Search, Optimization, and Machine Learning***, Addison-Wesley, 1989.

Journal Papers

1. Davis, C. B. and McNichols, R. J., 1994. Ground Water Monitoring Statistics Update: Part I: Progress Since 1988. Ground Water Monitoring and Remediation, Vol. 14, No. 4, pp.148-158; this is cited also in the MAROS manual in the references at the end of Appendix VII.

Documents

1. Roadmap to Long-Term Monitoring Optimization, USEPA, May 2005.
   http://www.cluin.org/download/char/542-r-05-003.pdf

Appendix C – Troubleshooting

- If you get the error dialog shown in Figure 75 ("Could not find the main class."), there may be a problem with the installation of Sun's Java Runtime Environment (JRE) on your computer.
    - Please uninstall the JRE and reinstall it with the latest version available from (http://java.sun.com/).
    - If you have trouble uninstalling the JRE, try using Microsoft's Windows Installer CleanUp (available from http://support.microsoft.com/kb/290301) to uninstall it.

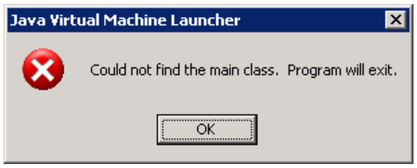


Figure 75. Java Error Dialog – Could not find the main class.

- If you have difficulty finding solutions with sufficiently high accuracy (aka low error), try increasing the population size in the GA settings.

- If automated model building is taking too long, try decreasing the resolution.

Appendix D –Technical Support, Training, Sales, and Consulting Information

User Forum

A free web forum is available for users to communicate with each other about *SampleOptimizer™* related issues, including long-term monitoring optimization in general. To sign up, all you need is an e-mail account; go to http://www.sampleoptimizer.com/forum.

Individualized Support

- If you have a retail license, it may include phone and/or e-mail technical support. For more details, please see the sales receipt that came with your license purchase. If it is not available, please contact the license support & billing department for assistance.

- If you have a non-commercial license (such as an educational, governmental, or demo license), or your retail license support contract is expired, support is available on a paid basis. For more information, please contact Matt Zavislak (see below).

Group Training

Summit plans to offer training seminars available to the general public. Additionally, Summit can come to your company or organization to offer a customized group training event.

Consulting

Summit Envirosolutions, Inc. is a full-service environmental engineering and cultural resources consulting firm which has been helping governments, businesses, and individuals find cost-effective and safe solutions to their environmental concerns for over 15 years. Please contact John Dustman to find out how Summit can assist your organization.

Contact Information

Sales & Consulting

John Dustman jdustman@summite.com 651-842-4203

Technical Support

Matt Zavislak mzavislak@summite.com 651-842-4248

License Support & Billing

Stacey Scarcella sscarcella@summite.com 651-842-4201