

The Need for Specialized Statistical Software for Ground-Water Analysis

By: Kirk Cameron, Ph.D, Statistician*

August 20, 2004

Peer-reviewed guidance documents for the statistical analysis of ground-water monitoring data have been available for years, from the US EPA and from other bodies such as state Departments of Environmental Quality and ASTM. The truth, however, is that many of the statistical requirements are not being enforced by state regulators, with the result that statistics — where performed at all — may be not applied correctly. The long-term consequences of this inconsistent regulatory environment are impossible to predict, but this paper will argue that the inconsistency is unnecessary: using specialized ground-water statistical software and meeting the accepted statistical requirements is both practical and affordable.

In today's competitive market and amidst the phenomenal growth of both personal computers and software, the array of choices when it comes to statistical and analytical packages may seem staggering. In the arena of ground-water monitoring statistics, several options are available to facility operators, analysts, and regulatory personnel. On one hand, ubiquitous spreadsheet packages such as Microsoft Excel claim to offer an increasing suite of statistical functions and graphics. On another, many dedicated statistical programs — usually costing hundreds to thousands of dollars — provide a dizzying array of statistical tools, most of which are not clearly related to ground-water statistical analysis. In between, there are well-known programs like Sanitas that are specifically tailored to the ground-water arena and are marketed to that audience. But how does one decide if a specialty software program is worth the cost and investment of human resources?

This white paper attempts to answer that question, and to illustrate why such a package can be a very wise investment when it comes to ground water data analyses. Yes, Sanitas costs more and is less widely available than Excel (Sanitas is approximately \$1400 — although free to regulators — for the first license). No, Sanitas is not a general-purpose statistical program such as Minitab, Statistica, SAS, SYSTAT, or JMP (an advantage, as we will see). Nevertheless, Sanitas occupies a unique and valuable position within the ground-water statistics market.

The single most important reason for this is that the RCRA ground-water regulatory environment is, and by its nature must be, quite complex. For nearly twenty-five years, ever since the first RCRA statistical regulations for ground water were proposed, the rules and guidance connected with ground-water analyses have become substantially more intricate and sometimes difficult to either decipher or reconcile. This regulatory framework, while depending on a variety of statistical methods, is an amalgamation of hydrogeologic practice and desired environmental objectives. Statistical procedures have been fit to the framework and vice-versa in a somewhat patchwork fashion.

Sanitas was originally developed to navigate through the maze of ground-water regulation and guidance by attempting to automate the complicated series of statistical steps and analyses either required or recommended by EPA. Because ground-water data have their own typical characteristics, Sanitas is intentionally tailored to the ground-water monitoring environment and to the analysis of such data. In fact, over the years, environmental regulators around the country have developed special reporting requirements for statistical analyses of ground-water measurements. In general, Sanitas meets these requirements by offering standard reports and results that are specifically designed to mimic the EPA ground-water statistical program, as well as options to perform a variety of customizations.

Compared to Sanitas, both spreadsheets like Excel and general purpose statistical packages leave it up to the individual user to not only:

- Develop the correct decision logic and chain of steps in the statistical analysis to satisfy EPA regulations, but also to;
- Set up the ground-water measurements properly for each statistical test, and;
- Format the results in an EPA- and regulator-preferred manner.

Each of these general steps is not only important, but can be quite daunting to users with more modest

statistical backgrounds. The primary reason is that to do ground-water statistics correctly with either a spreadsheet or a general purpose statistics program, one must first understand how statistical methods have been melded into the complex RCRA regulatory environment. The RCRA statistical requirements have been adapted over the years to the unique characteristics of landfills and hazardous waste facilities. They have also been tailored to the nature of ground-water data. Ground-water measurements are frequently non-detect, and often highly skewed. Non-normal distributions within such data sets are quite common, as are historical data series with seasonal patterns or other auto-correlated trends.

The RCRA regulations list specific tests and sampling requirements that must be met or used in the analysis of ground-water monitoring data. But in applying these methods, the measured data must be set up properly so that an appropriate background is designated. Such set-up can entail significant effort when using a spreadsheet or even a general-purpose statistical package. Why? For one, the background data may change depending on what set of wells is under consideration. Should an interwell or intrawell comparison be made? Does the comparison account for non-detected measurements, and possible seasonality or autocorrelation? Does the test method appropriately handle what are typically small sample sizes? And does the selection of background account for changes in flow gradients, site hydrostratigraphy, and other site-specific factors?

Sanitas is designed with a built-in decision logic framework, tailored to the ground-water monitoring environment. Unlike more general statistical programs or spreadsheets, it accommodates all the factors listed above in a thoughtful way, making it easy for the user to step through the required set-up and analysis. And Sanitas provides several statistical methods not found in *any* spreadsheet *or* basic statistical package. In fact, Sanitas has been at the forefront of incorporating the latest EPA statistical guidance into the structure of its decision logic.

This continued development of Sanitas — even as the EPA regulatory environment has evolved — is quite relevant to what makes it important as a product. After many years, EPA is again revising its guidance on recommended statistical strategies for ground-water monitoring. In particular, EPA is continuing its shift of emphasis toward resampling and retesting strategies as methods to ensure that statistically-based ground-water tests are as accurate as possible. None of the general-purpose statistical packages are designed to handle resampling and retesting procedures without significant amounts of additional programming and set-up. None of them properly calculate site-wide false positive rates in accord with EPA recommendations. None of them use EPA's "needle-in-the-haystack" approach in computing statistical power and false negative rates.

In addition, even dedicated statistical packages — since they are not expressly designed for the ground-water statistical framework — generally do not include EPA's recommended version of the Shewhart-CUSUM control chart, nor can they perform without additional programming either Cohen's or Aitchison's adjustment methods for non-detect measurements.

All of these limitations also apply of course to spreadsheet programs like Microsoft Excel. Despite the statistical tests and graphics that Excel advertises, almost all of the statistical framework and decision logic dictated within the RCRA ground-water guidance would have to be programmed by the user in order to get results comparable to Sanitas. Furthermore, some procedures would be very difficult to build on a spreadsheet. Sanitas overcomes these limitations by the very fact that it offers complete solutions proven by years of real-world use.

In particular, for users wanting to follow EPA methodology, Sanitas offers the following specific advantages over either spreadsheet applications or general-purpose statistical programs:

- Not only is Sanitas tailored to perform the required EPA statistical analyses, but it also features reporting of data and results that is consistent with EPA's statistical requirements and regulatory language.
- Sanitas is specifically designed to follow step-by-step the recommended EPA decision logic. This decision logic is encoded in EPA guidance as a lengthy series of flowcharts of statistical steps and testing algorithms.
- Because of its adherence to EPA's statistical framework, Sanitas makes decisions at each step in its analyses that account for well network size and configuration, number and types of constituents (COCs), whether or not different tests are required for different COCs, whether interwell, intrawell, or some combination of tests is needed, and what kinds of retesting and resampling protocols are appropriate.
- Sanitas also is designed to adhere to EPA guidelines and regulatory requirements regarding decision accuracy — specifically the proper calculation of false positive and negative rates, and the computation of statistical power.

All of these benefits of specialized ground-water monitoring software cannot be found in any off-the-shelf general-purpose statistical package, nor in any spreadsheet. Given that any statistical decision comes with a risk of being wrong, it is safe to say that the odds of the less-statistically savvy analyst or facility operator arriving at an incorrect decision using one of these other programs are much greater than if they used Sanitas.

The technology exists to meet RCRA statistical requirements in a cost-effective and efficient manner, and the risks associated with taking short cuts can and should be avoided.

*Author of the EPA 1992 Addendum to the Interim Final Guidance document on the statistical analysis of ground water monitoring data, and currently authoring a comprehensive update to the ground water guidance entitled "Unified Guidance on the Statistical Analysis of Ground-Water Monitoring Data." (kcmacstat@qwest.com)