Pacific Northwest
NATIONAL LABORATORY

PNNL-29015

Regression for Assessing Contaminant Plume Trends

August 2019

Ingrid A. Jennings
Christian D. Johnson

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: mb-reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-29015

Regression for Assessing Contaminant Plume Trends

August 2019

Ingrid A. Jennings
Christian D. Johnson

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

The Department of Energy's Hanford site has significant dissolved-phase groundwater contaminant plumes, for which remediation activities are ongoing. Assessing concentration trends is critical for making decisions regarding remedy implementation or performance. The SOCRATES (Suite of Comprehensive Rapid Analysis Tools for Environmental Sites) single-page web application provides a quality-assured framework for data access and consistent data analytics using standard methods. To provide trend information for remedial decisions, statistical analyses based on ordinary least squares (OLS) are already supported by SOCRATES. The presence of censored data, however, can skew the result of an OLS regression. Other methods, such as Akritas-Theil-Sen (ATS) or Tobit regressions can be more powerful. These two regression methods are used by Hanford site personnel to analyze groundwater contaminant concentration data, which often include nondetects (censored data). The ATS method uses comparisons of observation-pair ranks and the Tobit method uses a maximum likelihood estimation method. The Tobit method can also accommodate covariates, such as river stage or groundwater level fluctuations. Both ATS and Tobit regressions can be performed with the statistics software R, but to incorporate these methods into SOCRATES, their R code was ported into JavaScript. For quality assurance verification, the new code was tested with R code test data for trichloroethene concentrations in groundwater and compared to regression analyses that were previously conducted on this data using R.

Acknowledgements

This work was performed at the Pacific Northwest National Laboratory under the Deep Vadose Zone − Applied Field Research Initiative, which is funded by the U.S. Department of Energy Richland Operations Office. This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships (SULI) Program.

Acronyms and Abbreviations

ATS	Akritas-Theil-Sen
cenken	function in the NADA package for R
censReg	Censored regression (Tobit) models package for R
CH2M	CH2M Hill Plateau Remediation Company
DOE	Department of Energy
jStat	JavaScript statistical library
logsigma	natural log of the estimated standard deviation of the residuals
maxLik	maximum likelihood estimation package for R
NADA	Nondetects and Data Analysis for environmental data package for R
OLS	ordinary least squares
plm	linear models for panel data package for R
PLATO	PLume Analysis TOol
PNNL	Pacific Northwest National Laboratory
SOCRATES	Suite Of Comprehensive Rapid Analysis Tools for Environmental Sites
SULI	Science Undergraduate Laboratory Internships
TCE	trichloroethene
TCEReg	TCE concentration data set for Long Island, NY groundwater from the NADA package for R
WDTS	Workforce Development for Teachers and Scientists

Contents

Figures

Tables

1.0 Introduction

The Hanford Site was established in 1943 as part of the Manhattan project. During World War II, plutonium was produced at the site using the first large-scale plutonium production reactor in the world, the B Reactor. Plutonium production operations continued through the Cold War until the late 1980s. Historical disposal practices during plutonium separation operations included disposal of liquid waste to the soil column in cribs and trenches. This resulted in the distribution of contaminants in the vadose zone and groundwater, much of which still remains at the Hanford Site more than seventy years after plutonium production began. Environmental restoration is ongoing in various stages of characterization, feasibility study, and remedy implementation. An understanding of groundwater contaminant plume dynamics is required to determine the effectiveness of remediation efforts and to guide further remedial decisions.

Trend analysis provides information on plume dynamics, but is often implemented as an ordinary least squares (OLS) regression, or "best-fit" line. However, regression methods based on OLS can lead to misleading results when applied to data with censored values (e.g., values below analytical detection limits). The OLS method treats all values as known measurements, and does not account for the different nature of censored values. That is, a quantitative value cannot be assigned to measurements below a certain threshold (e.g., a method detection limit) in the case of left censored data or above a threshold if the data are right censored. The true value remains unknown. In this case, regression methods intended for censored data are more appropriate. Several such methods, notably the Akritas-Theil-Sen (ATS) and Tobit regressions, can account for the presence of censored data and are already used at the Hanford site [e.g., CH2M, 2015, 2016, 2018].

2.0 Context for Application to Hanford Groundwater Data

Web-based software tools are a fast-growing approach for deploying software that is readily accessible, provides seamless integration of multiple data sources, and provides standardized, quality analyses. SOCRATES (Suite Of Comprehensive Rapid Analysis Tools for Environmental Sites) is one such single-page web application for data access and analysis that is available to Hanford personnel to help access and analyze data to support remedial decisions. SOCRATES includes the PLATO (PLume Analysis TOol) module for analysis of contaminant plume trends. PLATO currently includes tools for performing linear and exponential OLS regressions, as well as a nonparametric Mann-Kendall trend analysis, but lacks trend analysis methods designed for use with censored data. Current Hanford applications of the ATS and Tobit regression methods [e.g., CH2M, 2015, 2016, 2018] are performed in the R software for statistical computing [R Core Team, 2019].

3.0 Description of Regression Methods for Censored Data

The ATS method is a nonparametric regression method for censored data based on Kendall's Tau coefficient. Slopes are calculated between each coordinate pair defined by an independent variable value, x, and a dependent variable value, y. If the slope is increasing, it is said to be concordant, whereas a decreasing slope is discordant. In cases where x-values are equal or the slope is zero, the slope is considered a tie. Kendall's Tau is calculated based on the concordant and discordant slopes between the residuals (defined as $y - b \cdot x$, where b is the slope) and the associated x values. The median slope between the residuals and the x values that produces a Kendall's Tau of zero is chosen for the slope of the regression line. This regression method can be performed in the R software using the cenken function from the NADA (Nondetects and Data Analysis for Environmental Data) package [Lee, 2017]. The NADA package accepts only one dependent variable, but accommodates both x and y censored data. Censored values are input as lists of true or false values, indicating whether a value is left censored or not, with multiple threshold values being allowed for the censored data. The cenken function calculates the slope, intercept, Kendall's Tau, and p-value for the ATS regression.

The Tobit regression method also accommodates censored data. The Tobit regression is a maximum likelihood method developed by economist James Tobin. To accommodate censored data, the method introduces a latent variable, a variable that is inferred rather than directly observed, for censored y values (Figure 1). These y values can be censored either above or below a specified limit (i.e., left or right censored, respectively). Only one threshold value is permitted for each of the left and right censoring limits. To estimate the regression line, the Tobit method uses a maximum likelihood estimation with two parts: a Probit part and a linear part. The Probit part is used for censored observations and the linear part is used for uncensored observations. Unlike ATS, the Tobit method can be used for multivariate regressions and can accommodate covariates. The corresponding R package to do Tobit regressions can be performed in the R software using the censReg package [Henningsen, 2017], which uses functions from the maxLik [Henningsen and Toomet, 2011] and plm [Croissant and Millo, 2008] packages to provide a maximum likelihood algorithm and initial linear model estimate, respectively. The censReg function accepts values for left- and right-censoring limits (–infinity or infinity where data is not left or right censored, respectively). The function calculates the slope(s), intercept, and the natural log of the estimated standard deviation of the residuals (logsigma) for the regression line, as well as the p-value associated with the t-test for each coefficient.

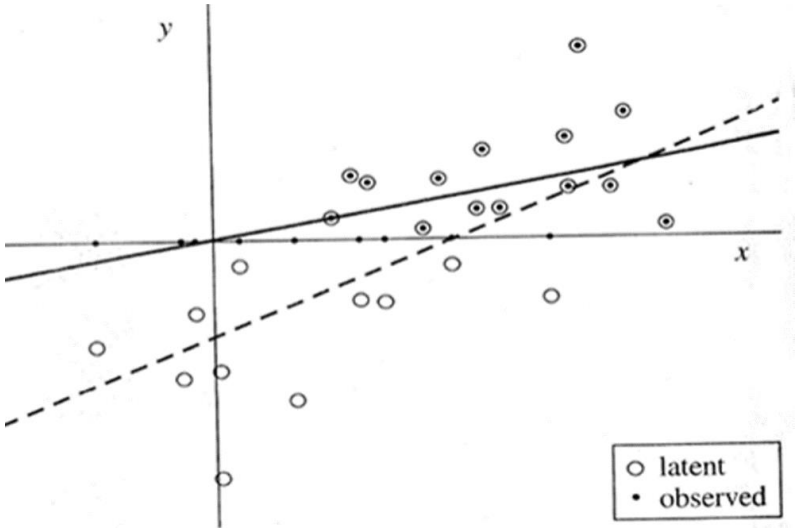

Figure 1. *A dashed Tobit regression line and an OLS regression line with y values censored at zero are shown. The Tobit regression infers latent values for censored data, while OLS does not distinguish between censored and uncensored data. Figure from Ruud [2000].*

4.0 Approach for Implementing Censored Regression Methods in PLATO

Implementation of ATS and Tobit censored regression methods in PLATO would improve the suite of analysis tools available in that web-based software. For web applications, calculational engines can be either implemented in the backend as part of database processing or server calculations, or can be implemented in the frontend in the form of client-side calculations. The existing OLS regression and Mann-Kendall analysis methods are implemented as JavaScript modules in the PLATO frontend. Thus, it made sense to port the ATS and Tobit regression functionality from R codes to JavaScript modules that could be added to PLATO.

The porting process began by analyzing each R package and its dependencies, to determine which code must be rewritten and which functionality was unnecessary in JavaScript. R has many built-in functions for data analysis and matrix manipulation. Some of these functions, such as a function for finding the Cartesian product of two arrays, were rewritten. For statistical methods and distributions, the statistical library jStat [Norris, 2019] was used to provide needed functionality. The JavaScript code was structured to provide results with the necessary precision to avoid artifacts from rounding.

5.0 Code Verification

Quality verification of the final ported JavaScript code for the ATS regression was done by comparison with results obtained using R. The Tobit code is still in the process of being ported and therefore cannot yet be verified against R calculations. The NADA R package [Lee, 2017] provides sample environmental data, some of which are discussed in the book, *Statistics for Censored Environmental Data Using Minitab and R* [Helsel, 2012]. Data for trichloroethene

(TCE) concentrations in groundwater were taken from the NADA package (i.e., the TCEReg data set) and analyzed with R and JavaScript. The natural log of TCE concentrations were plotted versus population density (Figure 2). An ATS regression was then performed in the completed JavaScript code and in R. The regression values calculated by the JavaScript code were equivalent to the values from the R calculations (and the values reported in Helsel [2012] for the same data set), as shown in Table 1. To show the effect of using censored regression methods, the R-calculated OLS regression line is also shown in Figure 2.

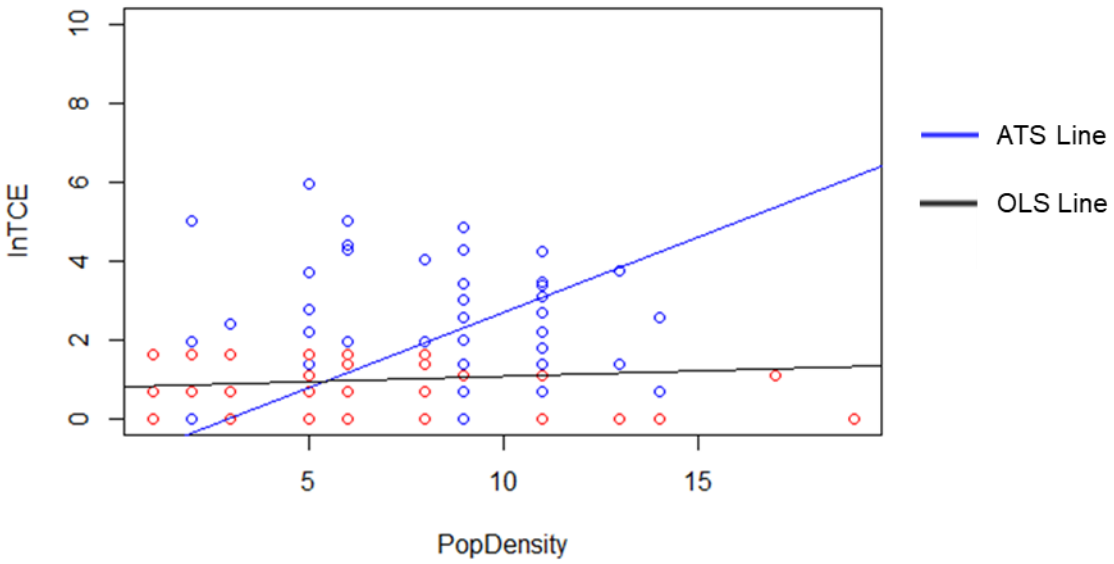

Figure 2. *The natural log of TCE concentrations from Helsel [2012] is plotted versus population density. Nondetects are shown in red, while detects are shown in blue. The calculated OLS and ATS regression lines are also shown for this data.*

Table 1. *Comparison of R and JavaScript ATS calculation results for the TCEReg verification data set [Helsel, 2012].*

Code	Slope	Intercept	Tau	P-value
R (and Helsel [2012])	0.3835066	-1.15052	0.1458477	0.0003007718
JavaScript	0.3835066	-1.1505198	0.1458477	0.0003006800

6.0 Example Application

Several examples are presented to illustrate the implications of using a censored regression approach versus standard least squares regression. In the first example, arbitrary data sets are used, and in the second example actual Hanford data is examined. The JavaScript ATS routine and the R software routines for OLS and Tobit regression were used for these examples.

Where there are no censored values, OLS, ATS, and Tobit regressions produce the same regression line, although each regression uses a different approach. When censored data is present, however, the resulting regression lines of each method will vary. To compare the results of OLS, ATS, and Tobit, regressions were done using arbitrary y-censored data for structured testing. The test data sets were produced to evaluate a range of potential scenarios and assess the differences between censored regression methods and OLS. Several test data sets were designed with different quantities of censored data; the analysis results for two of these data sets are shown in Figure 3. For the data set with 16% censored y values, the regression lines for each method are almost identical. In contrast, the data set with 50% censored values shows a distinct difference between the OLS regression line and the ATS and Tobit regression lines, which is emphasized by the clustering of censored values at later times. Significant differences between OLS and censored regression models can also occur in cases where censored values are more evenly distributed over time (e.g., Figure 2).

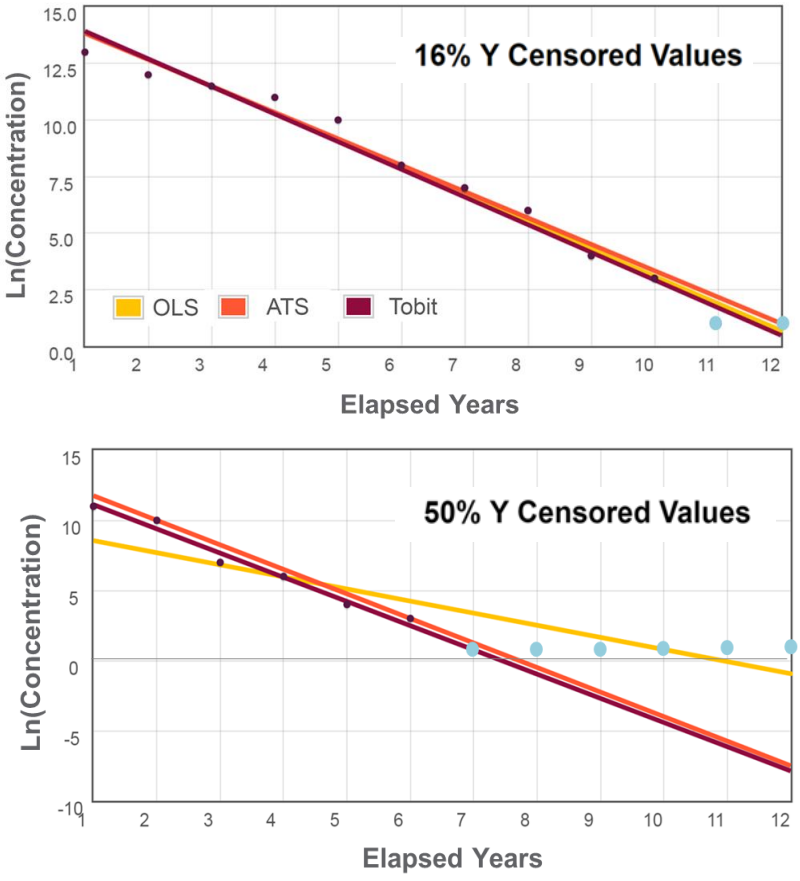

Figure 3. *The natural log of arbitrary concentration data is plotted vs elapsed years for sample data generated with varying levels of censored y values. The regression lines for ATS, Tobit, and ATS are shown.*

As an example of how the ATS methods are used for contaminant data analysis for the Hanford site, hexavalent chromium concentration data were analyzed for two wells (199-K-108A and 199-K-166) from the Hanford 100-K Area. Regressions for both of these wells were done previously using R, and the results were reported by CH2M [2016]. The regression lines for each well were calculated with the ATS JavaScript code and are plotted in Figure 4.

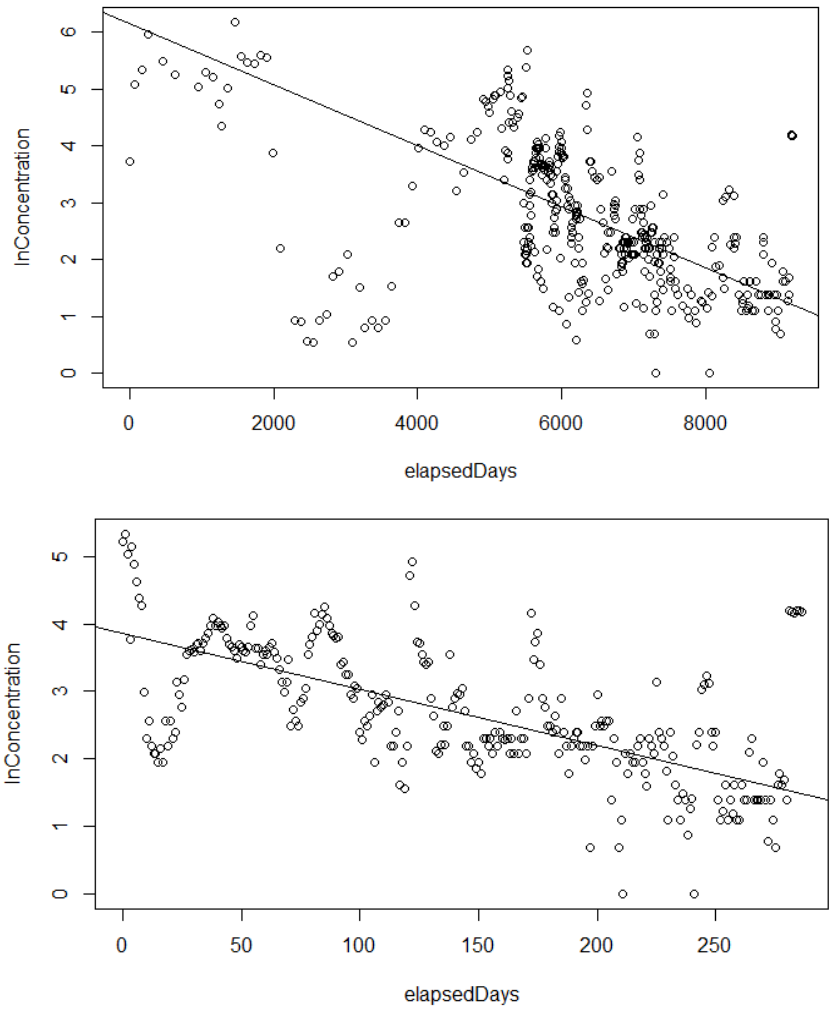

Figure 4. *The natural log of hexavalent chromium concentrations are plotted for wells 199-K-166 (top) and 199-K-108A (bottom). The ATS regression line was calculated using the developed JavaScript code and is shown for each well.*

7.0 Summary

Regression analyses that can handle censored data are important for contaminant concentration trend analysis at the Hanford site. Implementation of the ATS and Tobit regressions in PLATO will aid in plume analysis and remedial decisions. The ATS JavaScript code was verified by comparison against the results from R analyses for TCE data from the NADA code. The resulting

slope, intercept, tau, and p-value are equivalent to the results from NADA in R. The Tobit code is still in the process of being ported and therefore cannot yet be verified against R calculations.

The ATS JavaScript code provides correct outputs, but further improvements could be made to streamline and optimize the code. To further this project, porting the Tobit code can be finished and both regressions can be implemented in PLATO. Additionally, PLATO should allow user-specified censored data treatment options (i.e., multiplying censored values by ½ or setting a specific detection limit). These changes would allow for a better user experience and help provide the necessary tools to further plume analysis and remedial decisions.

8.0 References

Croissant, Y., and G. Millo. 2008. *Panel Data Econometrics in R: The plm Package*. R package version 2.1-0. Available at: https://CRAN.R-project.org/package=plm (accessed on 08/05/2019).

CH2M. 2015. *Calculation of Concentration Trends, Means, and Confidence Limits for cis-1,2-Dichloroethene, Gross Alpha, Nitrate, Trichloroethene, Tritium, and Uranium in the 300-FF-5 Operable Unit*. ECF-300FF5-15-0017, Rev. 0, CH2M Hill Plateau Remediation Company, Richland, WA.

CH2M. 2016. *An Assessment of Concentration Trends in Groundwater at 100 K-West*. ECF-100KR4-16-0074, Rev. 0, CH2M Hill Plateau Remediation Company, Richland, WA.

CH2M. 2018. *Calculation of Concentration Trends, Means, and Confidence Limits for the 200-ZP-1 Operable Unit Contaminants of Concern Before and After 200 West Pump and Treat Startup*. ECF-200ZP1-17-0124, Rev. 0, CH2M Hill Plateau Remediation Company, Richland, WA.

Helsel, D.R. 2012. *Statistics for Censored Environmental Data Using Minitab and R*. John Wiley & Sons, Hoboken, New Jersey.

Henningsen, A. 2017. *censReg: Censored Regression (Tobit) Models*. R package version 0.5-26. Available at: https://CRAN.R-project.org/package=censReg (accessed on 06/28/2019).

Henningsen, A., and O. Toomet. 2011. *maxLik: A package for maximum likelihood estimation in R*. R package version 1.3-6. Available at: https://CRAN.R-project.org/package=maxLik (accessed on 07/10/2019).

Lee, L. 2017. *NADA: Nondetects and Data Analysis for Environmental Data*. R package version 1.6-1. Available at: https://CRAN.R-project.org/package=NADA (accessed on 06/28/2019).

Norris, T. 2019. jStat: Statistical Library for Javascript. Version 1.8.3. Available at: https://jstat.github.io/all.html or https://github.com/jstat/jstat (accessed on 07/05/2019).

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org.

Ruud, P. 2000. *An Introduction to Classical Econometric Theory*. Oxford University Press, Inc.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov