TITLE: CProb: A Computational Tool for Conducting Conditional Probability Analysis

Jeffrey W. Hollister - US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive, Narragansett, RI 02882

Henry A. Walker - US Environmental Protection Agency, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division, 27 Tarzwell Drive, Narragansett, RI 02882

John F. Paul - US Environmental Protection Agency, Office of Research and Development, Mail Code B343-06, Research Triangle Park, NC 27711

ABBREVIATION LIST: Conditional Probability Analysis (CPA), National Land Cover Dataset (NLCD), Environmental Monitoring and Assessment Program (EMAP)

CORRESPONDING AUTHOR: Jeffrey W. Hollister, office: (401)782 9655, fax: (401)782 9711, email: hollister.jeff@epa.gov

Abstract:

Conditional probability measures the probability of observing one event given that another event has occurred. In an environmental context, conditional probability helps to assess the association between an environmental contaminant (i.e., the stressor) and the ecological condition of a resource (i.e., the response). These analyses, when combined with controlled experiments and other methodologies, show great promise in evaluating ecological conditions from observational data and in defining water quality and other environmental criteria. Current applications of conditional probability analysis are largely done via scripts or cumbersome spreadsheet routines, which may prove daunting to end users and do not provide access to the underlying scripts. Combining spreadsheets with scripts eases computation through a familiar interface (i.e., Microsoft Excel) and creates a transparent process through full accessibility to the scripts. With this in mind, we developed a software application, CProb, as an Add-in for Microsoft Excel with R, R(D)com Server, and Visual Basic for Applications. CProb calculates and plots scatterplots, empirical cumulative distribution functions and conditional probability. In this short communication, we describe conditional probability analysis, our motivation in developing a conditional probability analysis tool, its implementation as a Microsoft Excel Add-in, and illustrate its use with an example focusing on water quality criterion. CProb is freely available for download from http://www.epa.gov/emap/nca/html/regions/cprob.

Keywords: Conditional probability analysis; Water quality criteria; Stressor-response associations; Causal analysis; The R Project for Statistical Computing; R(D)COM server; RExcel

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

INTRODUCTION

A goal of many applied environmental science efforts is to understand the association between environmental pollutants (i.e., stressors) and ecological condition (i.e., response). Understanding this relationship helps to identify areas of likely ecological impairment, explore potential causes and consequences of those impairments, and set acceptable levels of a given stressor so that the health of an ecosystem may be protected. It is possible to use both field and laboratory methods to explore stressor-response relationships and set criteria that are protective of ecological condition. Inference based on field survey data alone can be problematic because of confounding, uncontrolled variation in field experiments. Conversely, controlled laboratory experiments alone are unable to replicate the full range of factors that affect ecosystems. Causal assessment is an approach that combines field survey information from independent datasets and lab and field experimental results to develop multiple lines of evidence that elucidate these associations (Angradi, 1999; US EPA, 2000). An example of the causal assessment strategy is found in a recent document on developing water quality criteria for suspended and bedded sediments (US EPA, 2006). A key component of this approach is Conditional Probability Analysis (CPA) which was identified as having significant promise and has been used elsewhere to develop possible water quality criteria and explore ecological condition – human health associations (Paul and McDonald, 2005, Paul et al. In Press).

In our experience, causal assessment and CPA are well received by many state environmental managers; however, a common stumbling block for implementation is access to user-friendly software. As such, we developed a computational tool that uses Microsoft Excel® as an interface and the R Language for calculating and plotting conditional probabilities. In this short communication we describe CPA, our motivation for developing a CPA software tool, and demonstrate the use of the tool with two examples: a water quality criterion example from Paul and McDonald (2005) and, building on prior work, a landscape ecological example (Comeleo et al., 1996; Paul et al., 2002; Hollister et al., 2008a; 2008b)

Conditional Probability Analysis

A conditional probability is the probability of an event Y occurring given that some other event X also has occurred. It is denoted P (Y|X). Thus, a conditional probability describes the probability of observing an event of interest in a subset of samples drawn from the original statistical population. These subsets are defined by conditions when X has occurred, in addition to those used to define the entire statistical population. Conditional probability is calculated as the ratio of the joint probability that Y and X occur simultaneously in a given sample from the original statistical population (P(Y, X)), to the probability of X in the original population. The notation for this is written:

$$P(Y \mid X) = \frac{P(Y, X)}{P(X)} \tag{1}$$

The possible applications of conditional probability are very broad. We focus on the more limited context of environmental condition, in which conditional probability may be interpreted as the probability of environmental or ecological impairment given a pollutant, nutrient, or other stressor is larger or smaller than a certain amount (e.g., the probability of benthic community impact given copper concentration in sediments exceed 10 ppm). Refer to Paul and McDonald (2005) for additional information.

Motivation for developing a Conditional Probability Analysis Tool

CPA usually involves informal spreadsheet routines or scripts. The spreadsheet routines require a high degree of user interaction, which increases the potential for transcription errors, incorrect formulas and may lead to an overall decline in the reproducibility of the analysis.

Scripts improve reproducibility and reliability, but are much less accessible as they require statistical programming languages such as S-Plus® or R. While these software applications are quite robust and reliable, they suffer from either being quite expensive, in the case of S-Plus, or require a steep learning curve.

The anticipated end users of CPA are local or state environmental managers who have the technical expertise to understand CPA, but are unlikely to have the time to learn a new application. Additionally, these users often prefer spreadsheets such as Microsoft Excel® for statistical analysis and data management. Although the limitations of using spreadsheets for statistical analysis and for managing data are well documented (Su et al., 2003; Knüsel, 2005; McCullough and Wilson, 2005), spreadsheets do provide a familiar interface for managers and provide a useful front end for other computational resources and new software tools (Su et al., 2003; Baier and Neuwirth, 2007). With this in mind, we use the statistical programming language R to build a script to calculate CPA. We combine the script with the familiar front end of Microsoft Excel 2003®. We accomplish this with a CPA Excel Add-in, CProb, that uses the R(D)COM server and RExcel (Baier and Neuwirth, 2006a; 2006b; 2007; R Development Core Team, 2006).

MATERIALS AND METHODS

Software requirements and installation information

Freely available for download from http://www.epa.gov/emap/nca/html/regions/cprob, CProb version 1.0 is a Microsoft Excel® Add-in developed with Microsoft Office Excel 2003®, Visual Basic for Applications, R version 2.4.0, R(D)Com Server version 2.01 and RExcel version 1.50. The Add-in uses Microsoft Excel® as a front end interface and calculates CPA using R as the statistical processor with R(D)Com providing the connection. Details about acquiring, using and installing these programs are available from Microsoft, The R Project for Statistical

Computing, and the R(D)Com page at the University of Vienna, respectively (Table 1). Detailed installation instructions for CProb are in a README file (CProbREADME.txt) included with the download. Since development, new versions of R, R(D)Com Server and RExcel are available. and CProb is compatible with these.

CProb Data Requirements

Data with paired stressor and response variables are required to calculate CPA. The stressor is either a discrete or continuous variable such as dissolved oxygen concentration, density of suspended sediments, or number of landscape patches. The response is a dichotomous variable or a continuous or discrete variable transformed to a dichotomous variable (e.g., with a threshold) that is thought to be responsive to the stressor. For instance, Paul and McDonald (2005) compared percent fines in bedded sediments to probability of taxa richness of sensitive species less than a threshold. CProb also accepts inclusion probabilities (i.e., from data acquired with a probability sampling design) as an optional parameter. These probabilities are used to calculate unbiased estimators from the data for extrapolation to the statistical population from which the sample was drawn.

CProb does not accept non-numeric values and will return an error message. Non-numeric values must be corrected, removed, or converted to a missing value prior to running CProb. Allowable missing values are permissible with the missing value settings in RExcel (i.e., either blank cells or “#N/A”). CProb will pass these to R and since paired data are required, the records with missing values are omitted from all calculations.

Bootstrapped Confidence Intervals

CProb uses standard bootstrap resampling to estimate the confidence intervals for the conditional probabilities (Manly, 2007). The raw data is resampled with replacement and a conditional probability is calculated for each bootstrap sample. The upper and lower confidence intervals are extracted as percentiles from the distribution of all bootstrapped conditional probabilities. The interval is determined by the "Confidence Interval Alpha" and the default value is 0.05. The default value for the number of bootstrap iterations is 100. This particular default value is provided to allow for an exploratory analysis that balances the interpretability of the analysis while limiting the demand for computational resources. One hundred iterations often results in unstable confidence intervals thus, the default value should only be used on initial, exploratory calculations. As a general rule of thumb, users of CProb should consider at least 1000 iterations with an α of 0.05 and 5000 iterations with an α of 0.01 (Manly, 2007).

Water Quality Criterion Example

To verify the output and test the functionality of CProb, we repeat the analysis conducted by Paul and McDonald (2005) for stream impairment and bedded sediments. In this example stream condition is defined by the total taxa richness of benthic macroinvertebrates in the orders Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies) (i.e., EPT Taxa Richness). Bedded sediments are characterized by percentage fines in the substrate. Data for this example are from the U.S. EPA's Environmental Monitoring and Assessment Program's Mid-Atlantic Highlands Streams Assessment. See Paul and McDonald (2005) for details and additional references. The dataset for our example (included with the software download as "JEQData_wq.csv") includes records for 99 stations and contains four columns: Station Identifier (SITE.ID), Inclusion Probabilities (WGT.FS), EPT Taxa Richness (EPT.RICH), and Percentage of Embedded Fines (PCT.FN). We choose arguments to match the methodology in Paul and McDonald (2005) as closely as possible.

Landscape Example

To independently test CProb, we conduct a CPA between landscape metrics and the condition of estuarine sediments. It is accepted that developed lands impact the ecological condition of sediments in Mid-Atlantic estuaries of the United States (Comeleo et al., 1996; Paul et al., 2002; Hollister et al., 2008a; 2008b) . We use estuarine data (included with software download as JEQData_le.csv) from the U.S. EPA's Environmental Monitoring and Assessment Program Estuaries program (EMAP) and total developed land cover is from the 1992 National Land Cover Dataset (NLCD). We use a total of 112 stations. Hollister et al. (2008a) identified spatial extents with the strongest linear association with different sediment metals. Following suggestions in Hollister et al. (2008a), we define contributing watersheds as all land area within 15 km or less from a station that also drains to that estuarine system. For this example, we define impaired ecological condition by any site that had at least one metal exceeding the Effects Range Median (ERM) value, a benchmark for potential effects on benthic biota (Long et al., 1995). A site with at least one metal exceeding the ERM value has been previously used as a measure of impairment (Paul et al., 1999). We use the metals examined in Hollister et al. (2008a). We expect to see higher amounts of developed lands within 15 km of stations that also have high concentrations of sediment metals. The probability of impairment should increase as the total amount of developed lands in the contributing watershed increases. Conversely, the probability of being impaired would become less as total amount of developed lands decreases.

RESULTS AND DISCUSSION

Water Quality Criterion Example

To ensure that CProb works as expected, we compare the CProb results to those obtained by Paul and McDonald (2005). With the Mid-Atlantic Highlands Data, the resultant CProb scatterplots (Figure 1a) and cumulative distribution function plots (Figure 1b) match Figures 3 and 4 of Paul and McDonald (2005). The conditional probability plot from CProb (Figure 1c) very closely matches Paul and McDonald's (2005) Figure 5. As expected, the only differences are a product of the confidence intervals resulting from the different bootstrapped samples.

Landscape Example

Of the 112 stations in the landscape example, 100 have no metals exceeding an ERM value (Figure 2a). On average, these sites have 6169 ha of developed land with a standard deviation of 8041 ha. The average size of the contributing watersheds (i.e., drainage within 15 km of the station) is 62,749 ha with a standard deviation of 7513 ha. CPA results for the landscape example corroborate past studies and suggest that higher amounts of developed land in contributing watersheds lead to poor condition in estuaries. Although separating developed land into finer classifications (e.g., industrial and residential) might yield stronger relationships, the accuracy of those finer classes is less than the documented accuracy of the coarser developed land classification and would lead to greater uncertainty (Yang et al., 2001; United States Geological Survey, 2003; Hollister et al., 2004). The cumulative distribution function indicates that the amount of developed land can discriminate between unimpaired sites (no metals greater than an ERM) and impaired sites (at least one metal greater than an ERM) as indicated by the separation between the two curves (Figure 2b). In fact 100% of all observed impaired sites have at least 14000 ha of developed land in the contributing watershed (i.e., drainage within 15km of the sampling station). The unimpaired sites have a range of developed land between 51 and 43,000 ha, but no sites with less than 14,000 ha of developed land have any metals exceeding an ERM value. Only 15% of the unimpaired sites have total developed land amounts that overlapped those of the bad sites (Figure 2b).

The conditional probability of impairment increases as total developed area in a watershed increases (Figure 2c). Non-overlapping confidence intervals (i.e., where the conditional confidence interval does not overlap the unconditional confidence interval) may be used as a conservative test for differences. The unconditional probability and confidence interval is represented by the first point in the conditional probability plot. In this case, non-overlapping confidence intervals suggest that the conditional probability of impairment is significantly higher between approximately 7500 ha and 37,500 ha than the unconditional probability of impairment of approximately 10% (i.e., P(Y)). This suggests that if development was limited to approximately 7500 ha or less within 15 km of a sampling station, the probability of metal contamination should not be significantly greater than what would be normally expected for this area. There are few stations with total developed land higher than 37,500 ha. The uncertainty about this part of the curve is large and thus it is not possible to draw any conclusions.

CONCLUSIONS

Our goals for this short communication were to describe a new tool for conducting Conditional Probability Analysis and illustrate its use with two examples: one to validate the results and one to show the versatility of the tool with an independent dataset. Our results validate the calculations and analyses from CProb as we reproduced the results of (Paul and McDonald, 2005). Furthermore, the power of CPA is that it provides a measure of the association between any two variables, environmental or otherwise, that explicitly incorporates uncertainty. Lastly, CProb is proving to be user friendly and is gaining acceptance with a number of state environmental managers in the Northeast. This tool along with the analytical techniques described here and in Paul and McDonald, (2005) are helping set scientifically defensible water quality criteria and more clearly describing stressor-response relationships.

ACKNOWLEDGEMENTS

We would like to thank Sandra Benyi, Anita Morzillo, and Patricia Shaw-Allen for their thoughtful reviews. They greatly enhanced the quality of the manuscript and of CProb. Also, we would like to thank the environmental managers who have used and given feedback on CProb. The research described in this paper has been funded by the US Environmental Protection Agency. This paper has not been subjected to Agency review. Therefore, it does not necessary reflect the views of the Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use. This is contribution number AED-07-095 of the Atlantic Ecology Division, Office of Research and Development, National Health and Environmental Effects Research Laboratory.

References

Angradi, T.R. 1999. Fine sediment and macroinvertebrate assemblages in Applachian streams: A field experiment with biomonitoring applications. Journal of the North American Benthological Society 18:49-66.

Baier, T., and E. Neuwirth. 2006a. R(D)com server v2.01.

Baier, T., and E. Neuwirth. 2006b. RExcel v 1.50.

Baier, T., and E. Neuwirth. 2007. Excel::Com::R. Computational Statistics 22:91-108.

Comeleo, R.L., J.F. Paul, P.V. August, J.L. Copeland, C. Baker, S.S. Hale, and R.W. Latimer. 1996. Relationships between watershed stressors and sediment contamination in Chesapeake Bay estuaries. Landscape Ecology 11 307-319.

Hollister, J.W., P.V. August, and J.F. Paul. 2008a. Effects of spatial extent on landscape structure and sediment metal concentration relationships in small estuarine systems of the United States' mid-Atlantic coast. Landscape Ecology.

Hollister, J.W., P.V. August, J.F. Paul, and H.A. Walker. 2008b. Predicting estuarine sediment metal concentrations and inferred ecological conditions: An information theoretic approach. J. Environ. Qual. 37.

Hollister, J.W., M.L. Gonzalez, J.F. Paul, P.V. August, and J.L. Copeland. 2004. Assessing the accuracy of the National Land Cover Dataset area estimates at multiple spatial extents. Photogramm. Eng. Remote Sens. 70:405-414.

Knüsel, L. 2005. On the accuracy of statistical distributions in Microsoft Excel 2003. Computational Statistics and Data Analysis 48:445-449.

Long, E.R., D.D. MacDonald, S.L. Smith, and F.D. Calder. 1995. Incidence of adverse biological effect within ranges of chemical concentration in marine and estuarine sediment Environmental Management 19 81-97.

Manly, B.F.J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. 3rd ed. Chapman Hall/CRC, Boca Raton.

McCullough, B.D., and B. Wilson. 2005. On the accuracy of statistical procedures in Microsoft Excel 2003. Computational Statistics and Data Analysis 49:1244-1252.

Paul, J. F., M. E. McDonald, S.F. Hedtke. In Press. Stream condition and infant mortality in U.S. mid-Atlantic states. Human and Ecological Risk Assessment.

Paul, J.F., and M.E. McDonald. 2005. Development of empricial, geographically specific water quality criteria: A conditional probability analysis approach. Journal of the American Water Resources Association. 41:1211-1223.

Paul, J.F., R.L. Comeleo, and J.L. Copeland. 2002. Landscape metrics and estuarine sediment contamination in the mid-Atlantic and southern New England regions J. Environ. Qual. 31 836-845.

Paul, J.F., Gentile, J.H., Scott, K.J., Campbell, D.E., & Latimer, R.W. . 1999. EMAP-Virginian province four-year assessment report (1990-1993), pp. 119 pp. U.S. Environmental Protection Agency, Atlantic Ecology Division, Narragansett, RI.

R Development Core Team. 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Su, Y., C.H. Langmuir, and P.D. Asimow. 2003. Petroplot: A plotting and data management tool set for Microsoft Excel. Geochemistry, Geophysics, Geosystems 4:1-14.

US EPA. 2000. Stressor identification guidance document. Office of Water, Office of Research and Development, Washington, DC.

US EPA. 2006. Framework for developing suspended and bedded sediments (sabs) water quality criteria EPA-822-R-06-001. United States Environmental Protection Agency, Washington, DC.

United States Geological Survey. 2003. Accuracy assessment of 1992 National Land Cover Data.

Yang, L., S.V. Stehman, J.H. Smith, and J.D. Wickham. 2001. Thematic accuracy of MRLC land cover for the eastern United States. Remote Sens. Environ. 76:418-422.

Table 1. Software sources and links.

Software	**Source**	**Link**
Excel 2003	Microsoft	http://office.microsoft.com/excel
R	R Project for Statistical Computing	http://www.r-project.org
R(D)Com Server and RExcel	Sunsite	http://rcom.univie.ac.at

Figure 1a. Scatterplot showing relationship between EPT Taxa and Bedded Sediments.

Figure 1b. Cumulative Distribution Function showing percentage of stations with a given amount, or more, of percent fines.

Figure 1c. Conditional probability plot showing probability of impairment in streams with a given percentage of fine material in the sediment. Solid lines indicate the 95% confidence intervals.

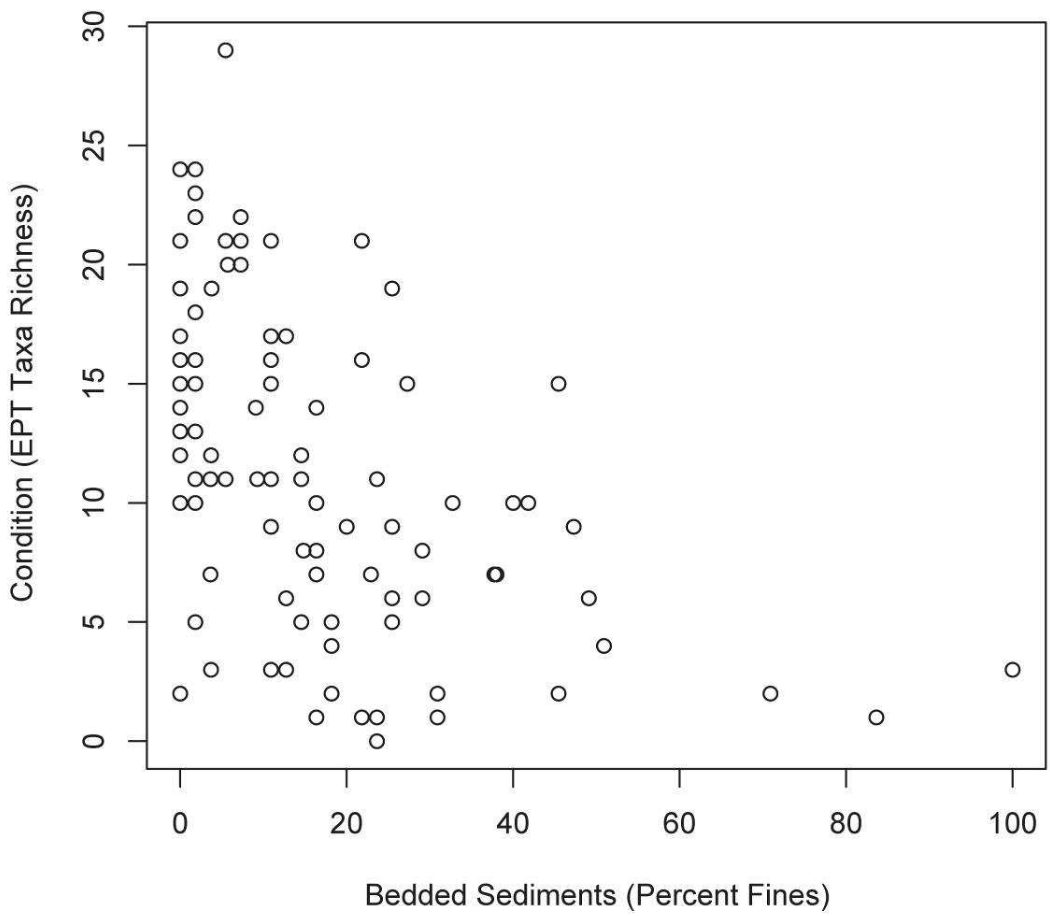
Figure 2a. Scatterplot showing relationship between number of metals exceed the Effects Range Median (ERM) and total developed land (ha) withing 15 km of EMAP stations.

Figure 2b. Cumulative distribution function showing percentage of stations with a given amount, or more, of total developed land (ha) within 15 km of EMAP stations.

Figure 2c. Conditional probability plot showing probability of at least one metal exceeding the ERM given the total developed land (ha) is greater than or equal to X_c.

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

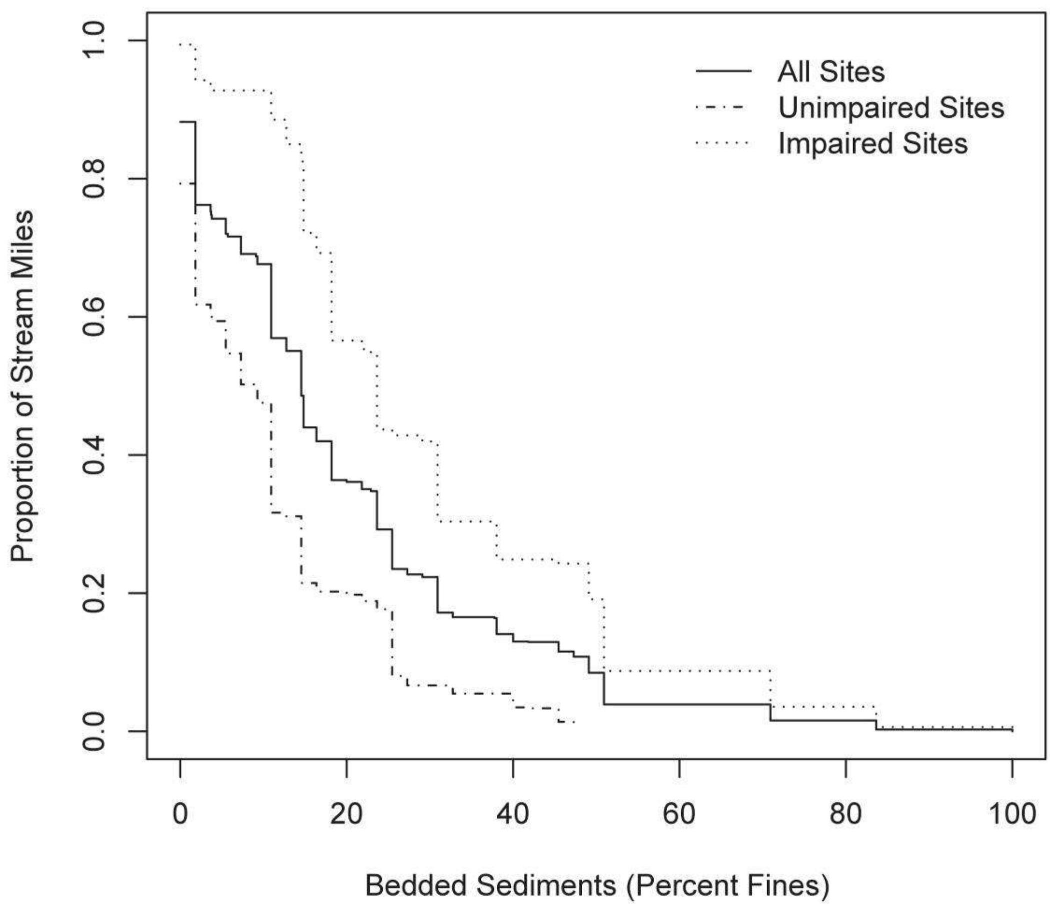
Figure 1A; Q07−0536; J.W. Hollister



1

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

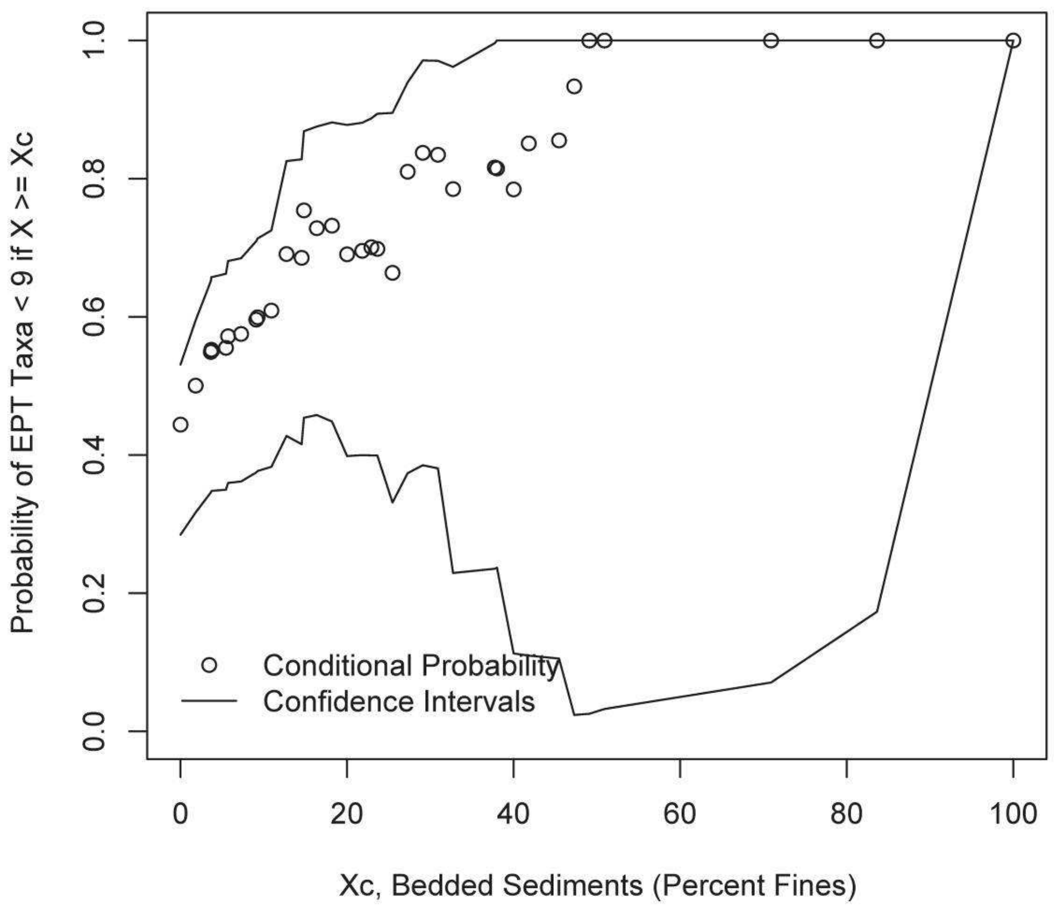
Figure 1B; Q07−0536; J.W. Hollister



1

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

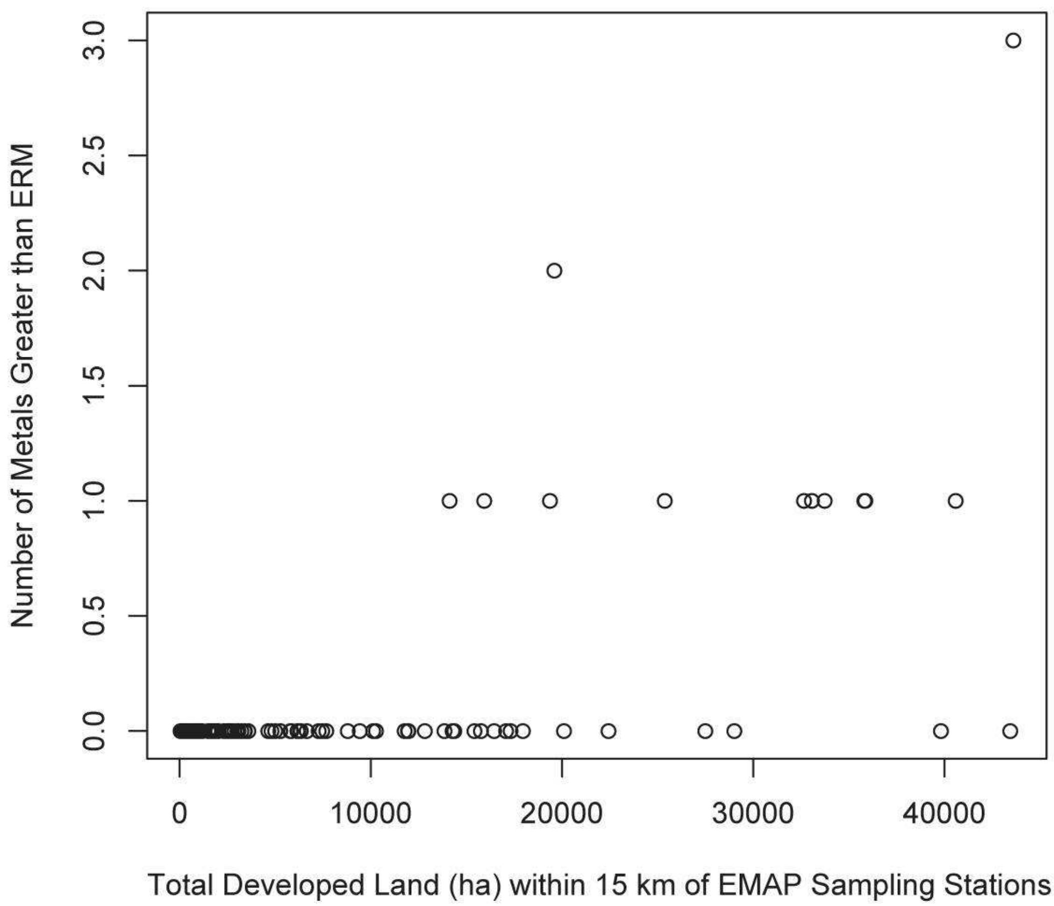
Figure 1C; Q07−0536; J.W. Hollister



1

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

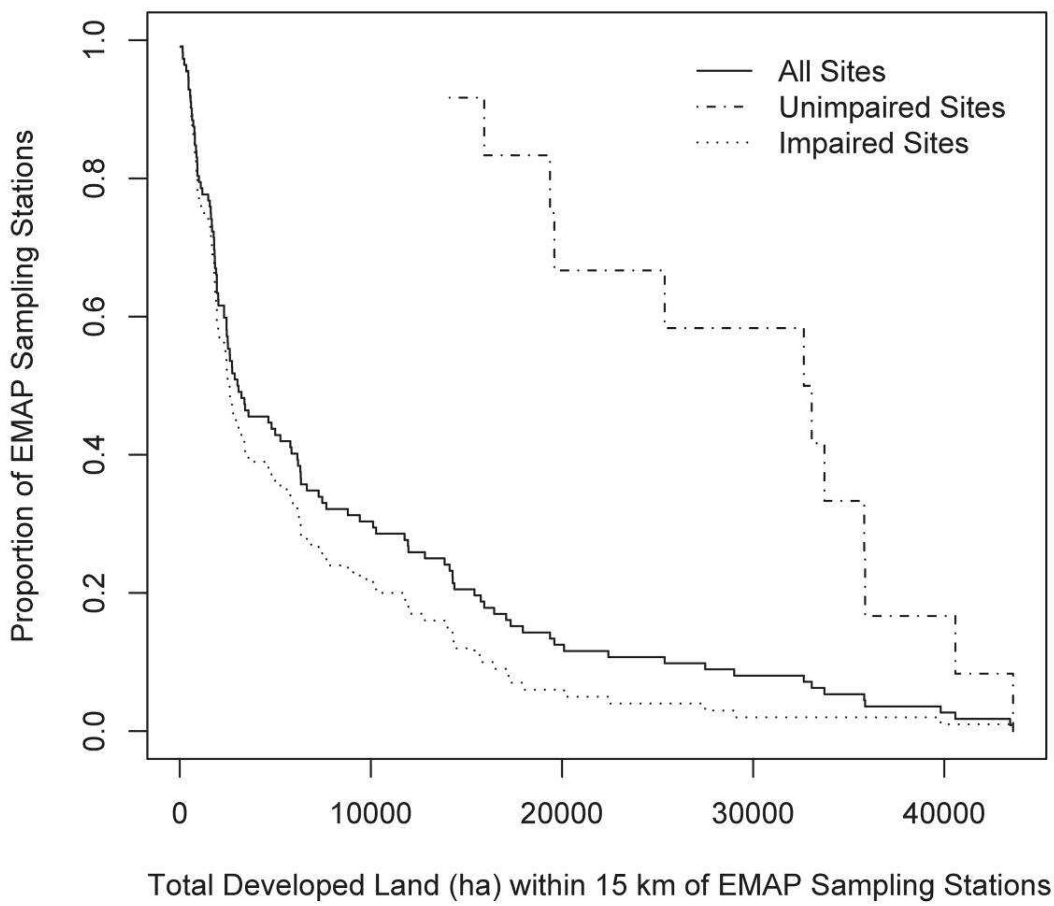
Figure 2A; Q07−0536; J.W. Hollister



1

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

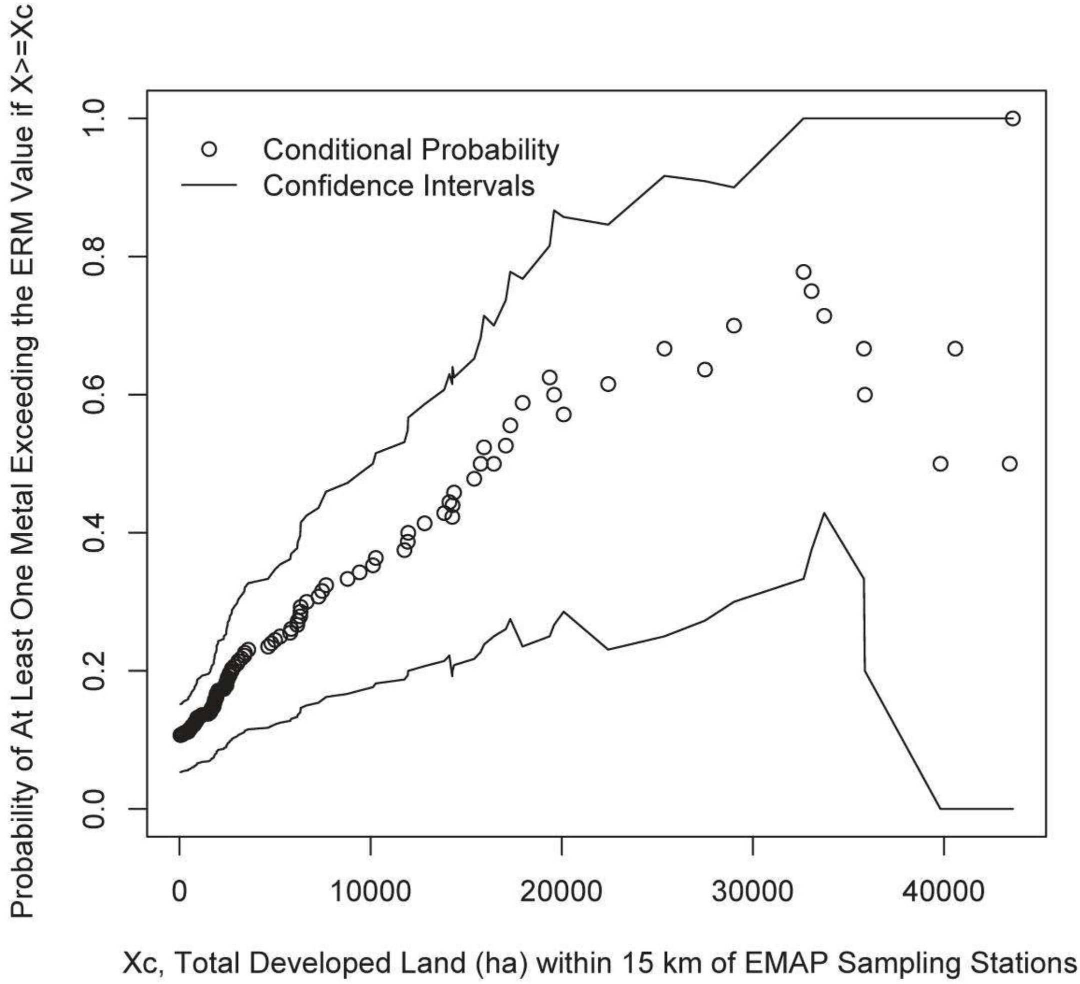
Figure 2B; Q07−0536; J.W. Hollister



1

This is a Green Open Acess copy.
Please see http://dx.doi.org/10.2134/jeq2007.0536 for published version.

Figure 2C; Q07−0536; J.W. Hollister



1