SUSCEPTIBILITY OF SURFACE WATER IN TEXAS TO NONPOINT SOURCE PESTICIDE CONTAMINATION

William A. Battaglin, Randy Ulery, Toby Welborn, and Thomas Winterstein*

ABSTRACT: Surface-water (streams and reservoirs) sources of public drinking-water supply (PWS) are susceptible to contamination from point and nonpoint sources. The Texas Commission on Environmental Quality (TCEQ) and the U.S. Geological Survey (USGS) developed a system for assessing the susceptibility of PWS source waters to the occurrence of 227 contaminants. To accomplish these assessments for surface-water PWS's, water-quality data at 323 USGS and TCEQ sampling sites are matched with GIS-derived basin characteristics data for the watersheds upstream from these sites. Logistic regression models are then used to estimate the probability that a particular contaminant will occur above a given threshold concentration. Once developed, these equations are used to estimate the probability of contaminant occurrence at PWS's. For many of the 227 contaminants, either monitoring data or contaminant detections were insufficient to produce viable logistic regression models. Significant models were developed for 20 (32%) of the 64 contaminants classified as pesticides.
KEY TERMS: pesticides, surface water, drinking water, GIS, logistic regression

INTRODUCTION

In Texas, surface water and ground water are used as sources of public drinking-water supply (PWS). The 1996 Amendment to the Safe Drinking Water Act (U.S. Environmental Protection Agency, 2003) requires that each State prepare a source water assessment for each PWS. States are required to identify the drinking-water source, the potential contaminants of concern, and the intrinsic susceptibility of the source water. To help protect sources of drinking water and to develop protective and economically feasible water-supply monitoring strategies, the Texas Commission on Environmental Quality (TCEQ) and the U.S. Geological Survey (USGS) developed an automated system for assessing the susceptibility of PWS source waters to the occurrence of 227 contaminants. One component of this system is the assessment of susceptibility of surface-water PWS's to contamination from nonpoint sources. This report presents susceptibility model results for a subset (64) of the 227 contaminants, those classified as pesticides. The results for the remaining contaminants will be available on-line at http://co.water.usgs.gov/midconherb/html/texas.html.

Texas Source-Water Assessment Project

The automated system for assessing susceptibility of PWS source water to contamination enables TCEQ staff to (1) focus source-water protection efforts on the most susceptible PWS's, (2) reduce monitoring costs, (3) help educate the public, and (4) identify land-management practices needed to protect source water. A PWS susceptibility assessment is composed of eight components (Ulery, 2000; also see http://www.tnrcc.state.tx.us/permitting/waterperm/pdw/swap/swavap.html):

1. **Identification Component** - determine hydrologic attributes used to delineate the source area(s) for the PWS.
2. **Delineation Component** - delineate PWS source area(s) from hydrologic attributes.
3. **Intrinsic Susceptibility Component** – determine the intrinsic attributes (for example, soil erodibility) within PWS source area(s) that have the potential to make a PWS more or less susceptible to contamination.
4. **Nonpoint-Source Susceptibility Component** – determine statistical relations between measured contaminants in surface water and environmental variables (for example, land use) within associated watersheds. In cases of limited or no data, susceptibility is assigned based on the frequency of contaminant detection.
5. **Point-Source Susceptibility Component** – determine potential point sources of contaminants to the PWS.
6. **Area of Primary Influence Component -** determine the density and proximity of potentially threatening activities or land uses and associated contaminants within the PWS source area(s).
7. **Contaminant Occurrence Component** – determine contaminant detections (above TCEQ-threshold concentrations) at water-quality monitoring and finished water sites near the PWS.
8. **Susceptibility Summary Component** – combine the first 7 components into a summary assessment for the PWS.

*Respectively, Hydrologist, USGS, Box 25046 MS 415 DFC, Denver, CO 80225, (303) 236-4882 x256, wbattagl@usgs.gov; Hydrologist, USGS, 8027 Exchange Dr., Austin, TX, (512) 927-3558; Cartographic technician, USGS, 8027 exchange Dr., Austin, TX (512) 927-3567; and Hydrologist, USGS, 2280 Woodale Dr, Mounds View, MN, (763) 783-3150

Surface-Water Nonpoint-Source Component - Pesticides

The surface-water nonpoint-source component is addressed by investigating the relations between known pesticide occurrences in surface water and environmental variables, either natural or anthropogenic, within each of the contributing watersheds. In cases where monitoring data are sufficient, logistic regression models are developed to quantify these relations and to provide a means for extrapolation of results from monitoring sites to actual PWS intakes. In cases where monitoring data are not sufficient or pesticide occurrence is infrequent, the monitoring results are used to identify susceptibility of PWS to contamination in all areas of the State and not at specific PWS intakes.

Surface-Water Sites and Water-Quality Data

The USGS operates an extensive surface-water data-collection network in Texas, with more than 500 sites on rivers and reservoirs (Gandara and Barbie, 2001). For this investigation, 323 sites were selected to represent surface-water quality in the State of Texas. These 323 sites were selected because there was information about the occurrence of a least one of the 227 target contaminants in water-quality data from the site. Most sites are at USGS gages, but some are PWS reservoirs, and others are sites sampled by TCEQ.

Sixty-four pesticide compounds were evaluated in this study. All water-quality data associated with the 323 selected sites from the USGS and TCEQ databases were accumulated and merged into a single database of approximately 2 million values. A contaminant occurrence code was assigned to each site for each contaminant as follows: -1, the contaminant was never analyzed for at the site; 0, the contaminant was analyzed for but never detected at the site; 1, the contaminant was analyzed for and detected but always at a concentration less than the TCEQ threshold; and 2, the contaminant was analyzed for and detected one or more times at a concentration equal to or greater than the TCEQ threshold. This coding scheme transforms the raw concentration data into a format suitable for analysis by logistic regression.

Watershed Characteristics Data

Watershed characteristics data were quantified within the watersheds of the 323 sampling sites by using a geographic information system (GIS). Land-cover characteristics (U.S. Geological Survey, 2000), soils characteristics (U.S. Department of Agriculture, 1995), land and stream physiography (Brown et al., 2000), population density (U.S. Department of Commerce, 1990), oil and gas well density, agricultural chemical use, fertilizer use, and manure production estimates (Battaglin and Goolsby, 1994; Goolsby and others, 1999) were calculated for individual watersheds by overlaying basin boundaries on source datasets and summarizing the information. Twenty-five land-use, 6 soils, 2 population (people and oil wells), 44 agricultural chemical use, 2 manure production, and 5 land and stream physiography variables were tested for potential use in the logistic regression models. Using logistic regression, variables not significantly related to pesticide occurrence can be identified and excluded from the models. The variables listed in table 1 are those that were statistically significant in one or more of the models of pesticide occurrence.

Many of the watershed characteristics are related to or correlated with each other. For example, the percentage of high-intensity residential land in the 323 watersheds is correlated ($r^2 = 0.86$) with population density in the watersheds. Including strongly correlated explanatory variables in multivariate statistical models can have undesirable consequences in a regression equation such as coefficients with unrealistic signs or unstable values (Helsel and Hirsch, 1992). For the results presented here, explanatory variables that were highly correlated ($r^2 = 0.7$ or greater) were not allowed to appear in the same regression equation.

Logistic Regression Models

Logistic regression is a statistical technique that uses one or more explanatory variables to predict the probability of a categorical response. Logistic regression is well suited for modeling the probability or likelihood of the occurrence or non-occurrence of some event (Hosmer and Lemeshow, 2000). Used primarily in medical research, logistic regression has been gaining popularity in environmental investigations (Battaglin and Goolsby, 1998; Nolan, 2001; Smith et al., 2001).

The SAS statistical software system (use of trade names in this publication does not connote product endorsement by the USGS) was used to develop logistic regression models. Both the “Best” and the “Stepwise” model selection procedures were used to screen potential models. Neither the “Best” nor the “Stepwise” model selection procedures will prevent correlated explanatory variables from entering the models, so supervision of the process is required. Logistic regression equations for individual pesticides were compared based on model-fit statistics such as the Akaike information criterion, and the predictive abilities of the models were compared based on the Somers’ D measure of rank correlation between the observed and predicted responses, and by contingency tables analysis of predicted versus observed responses. The final logistic regression models should not be affected by multicollinearity among the explanatory variables because variables that were correlated ($r^2 = 0.7$ or greater) were not permitted to appear in the same model. A logistic regression model was developed for a pesticide only if there were more than 50 sample analyses from a total of 32 or more sites, and the pesticide was detected in 15% or more of the samples.

Table 2. Statistically significant explanatory variables in pesticide occurrence logistic regression models.

Variable	Definition
P21	Percent land use 21 – low intensity residential
P22	Percent land use 22 – high intensity residential
P23	Percent land use 23 – commercial/industrial/transportation
P41	Percent land use 41 – deciduous forest
P42	Percent land use 42 – evergreen forest
P43	Percent land use 43 – mixed forest
P51	Percent land use 51 – shrubland
P61	Percent land use 61 – orchards/vineyards
P71	Percent land use 71 – grasslands/herbaceous
P81	Percent land use 81 – pasture/hay
P82	Percent land use 82 – row crops
P83	Percent land use 83 – small grains
P84	Percent land use 84 – fallow
P85	Percent land use 85 – urban/recreational grasses
P91	Percent land use 91 – woody wetlands
P92	Percent land use 92 – emergent herbaceous wetlands
P_ag	Percent agricultural land – sum of p81, p82, p83, p84, p85, and p61
P_clay	Total clay content of soil expressed as percent clay in material less than 2 mm in size.
P_drain	Percent artificially drained agricultural land
P_forest	Percent forest land – sum of p41, p42, and p43
P_org	Total soil organic matter in percent by weight.
P_urban	Percent urban land – sum of p21, p22, and p23
P_wet	Percent wetlands – sum of p91 and p92
Ala_use	Alachlor use rate in kilograms per square kilometer
Atr_use	Atrazine use rate in kilograms per square kilometer
BS	Average basin slope in feet per mile
DCPA_use	DCPA use rate in kilograms per square kilometer
Diuron_use	Diuron use rate in kilograms per square kilometer
Hydgrp	Average soil hydrologic group
Kfact	Average soil total K factor
MCS	Main channel slope in feet per mile
Perm	Average soil permeability in inches per hour
Pop_den	Basin population density in people per square mile
TDA	Total drainage area in square miles
Well_den	Density of oil and gas wells in number per square kilometer

RESULTS

Two statistically significant logistic regression models were developed for 20 of the 64 pesticide contaminants. Pesticide occurrence was not modeled due to a lack of detections in the otherwise sufficient monitoring data (28 of 44 not modeled), the lack of monitoring data at 10% or more of the 323 sites (10 of 44), or the complete lack of monitoring data (5 of 44). Logistic regression equations return an estimate of the probability (p) of pesticide detection above a particular threshold value. This probability value is used by the TCEQ to assign a PWS susceptibility rating as follows:

If	$p_{Pesticide} > 0.67$	Susceptibility Rating = High
If	$p_{Pesticide} \geq 0.33$ but ≤ 0.67	Susceptibility Rating = Medium
If	$p_{Pesticide} < 0.33$	Susceptibility Rating = Low

An example of the modeling results is given in equation 1 and figure 1. Equation 1 is the model used to predict the probability of simazine occurrence. This model suggests that the presence of residential land and pasture land is positively related to simazine occurrence, while evergreen forest land and woody wetlands are inversely related to simazine occurrence.

$$P = [\exp(-0.719 + 0.098*p21 + 0.052*p81 - 0.028*p42 - 0.463*p91) / (1 + (\exp(-0.719 + 0.098*p21 + 0.052*p81 - 0.028*p42 - 0.463*p91)] \quad (1)$$

In figure 1, the top panel shows simazine monitoring data were available for model development; the bottom panel shows the simazine model results. On both the top and bottom panels, the pink areas are the parts of the State of Texas that were not covered by one of the 323 watersheds. The top panel shows where dissolved simazine occurred at least once at a concentration of 0.01 µg/L or greater (red), occurred but always at a concentration less than 0.01 µg/L (yellow), was analyzed for but not detected (blue), and was not analyzed for (gray). The bottom panel shows the modeled probability that simazine will occur at a concentration of 0.01 µg/L or greater, with red areas having a probability of greater than 0.67, yellow areas

having a probability between 0.33 and 0.67, and blue areas having a probability of less than 0.33. There was monitoring data for simazine at 129 of the 323 sites with about 1,300 analyses.

The complete table of model results is too large for this publication but can be accessed via the Internet at (http://co.water.usgs.gov/midconherb/html/texas.html). Figures like figure 1 for the other pesticides for which logistic regression models were developed are also available at this site.

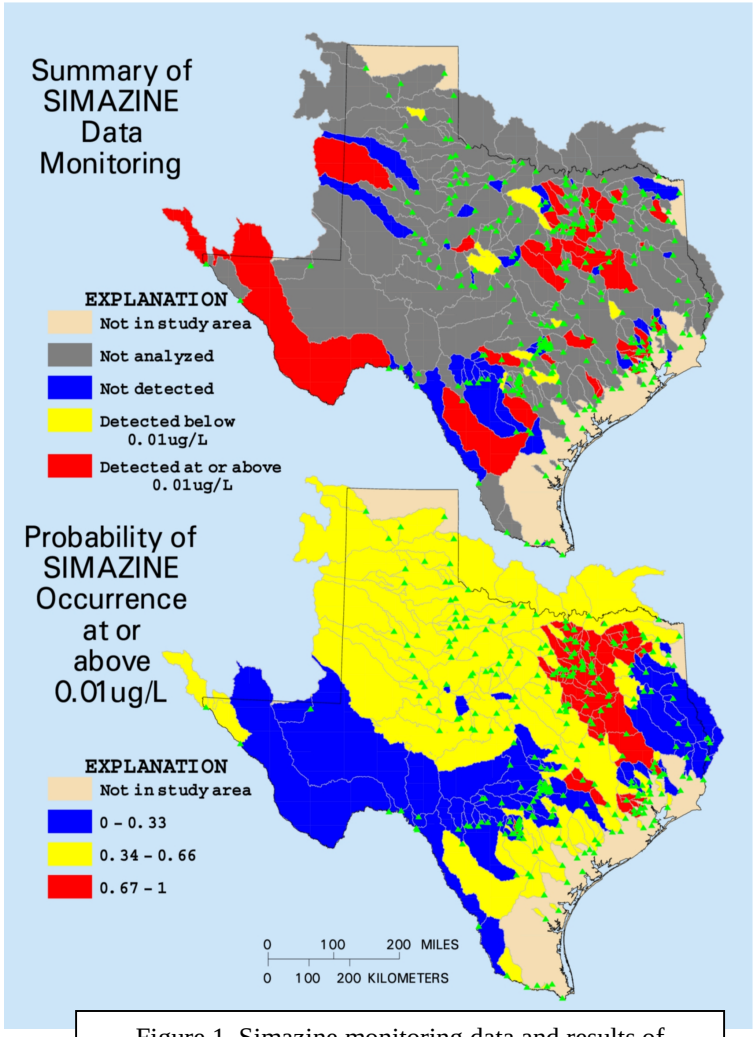

Figure 1. Simazine monitoring data and results of logistic regression model estimating the probability of simazine occurrence at or above 0.01 µg/L in Texas

REFERENCES

Battaglin, W.A. and D.A. Goolsby, 1994. Spatial Data in Geographic Information System Format on Agricultural Chemical Use, Land Use, and Cropping Practices in the United States: U.S. Geological Survey Water-Resources Investigations Report 94-4176, 87 p.

Battaglin, W.A. and D.A. Goolsby, 1998. Regression Models of Herbicide Concentrations in Outflow from Reservoirs in the Midwestern USA, 1992-1993: Journal of the American Water Resources Association 34(6): 1369-1390.

Brown, J.R., R.L. Ulery, and J.W. Parcher, 2000. Creating a Standardized Watersheds Database for the Lower Rio Grande/Rio Bravo, Texas: U.S. Geological Survey Open-File Report 00-065, 18 p. available at URL http://water.usgs.gov/pubs/of/ofr00-065/

Gandara, S.C. and D.L. Barbie, 2001. Index of Stations—Surface-Water Data-Collection Network of Texas, September 1999: U.S. Geological Survey Open-File Report 01-103, 17 p. available at: http://tx.usgs.gov/biblio/abstracts.asp?seq=1105

Goolsby, D.A., W.A. Battaglin, G.B. Lawrence, R.S. Atrz, B.T. Aulenback, and R.P. Hooper, 1999. Flux and Sources of Nutrients in the Mississippi-Atchafalaya River Basin: Topic 3 Report for the Integrated Assessment on Hypoxia in the Gulf of Mexico: NOAA Coastal Decision Analysis Series No. 17, NOAA Coastal Ocean Program, Silver Spring, MD, 130 p. available at: http://www.nos.noaa.gov/pdflibrary/hypox_t3final.pdf

Helsel, D.R. and R.M. Hirsch, 1992. Statistical Methods in Water Resources: Elsevier, Amsterdam, The Netherlands, 522 p.

Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression: John Wiley & Sons, Inc., New York, 375 p.

Nolan, B.T., 2001. Relating Nitrogen Sources and Aquifer Susceptibility to Nitrate in Shallow Ground Waters of the United States. Ground Water 39(2): 290-299.

Smith, J.H., J.D. Wickham, D. Norton, T.G. Wade, and K.B. Jones, 2001. Utilization of Landscape Indicators to Model Potential Pathogen Impaired Waters. Journal of the American Water Resources Association 37(4): 805-814.

Ulery, R.L., 2000. Overview of the Texas Source Water Assessment Project. U.S. Geological Survey Fact-Sheet 101-00, 6 p.

U.S. Department of Agriculture, 1995. STATSGO – U.S. coverage by states including Puerto Rico (excludes Alaska): U.S. Department of Agriculture, Natural Resources Conservation Service, National Cooperative Soil Survey, State Soil Survey Geographic Data Base, CD-ROM.

U.S. Department of Commerce, 1990. 1990 Census of Population and Housing: U.S. Department of Commerce, U.S. Census Bureau, accessed June 15, 2000, at URL http://www.census.gov/geo/www/tiger/index.html

U.S. Environmental Protection Agency, 2003. The Safe Drinking Water Act, Public Law 104-182 104th Congress. Accessed 2/2003 at http://www.epa.gov/safewater/sdwa/text.html.

U.S. Geological Survey, 2000. National Land Cover Dataset: U.S. Geological Survey Fact Sheet 108-00, 2 p. data available at: http://landcover.usgs.gov/nationallandcover.html