Statistical Approaches to Interpretation of Local, Regional, and National Highway-Runoff and Urban-Stormwater Data

A Contribution to the
NATIONAL HIGHWAY RUNOFF DATA AND METHODOLOGY SYNTHESIS

U.S. Department
of Transportation

USGS

science for a changing world

Statistical Approaches to Interpretation of Local, Regional, and National Highway-Runoff and Urban-Stormwater Data

By GARY D. TASKER and GREGORY E. GRANATO

A Contribution to the
NATIONAL HIGHWAY RUNOFF DATA AND METHODOLOGY SYNTHESIS

Northborough, Massachusetts
2000

U.S. DEPARTMENT OF THE INTERIOR
BRUCE BABBITT, Secretary

U.S. GEOLOGICAL SURVEY
Charles G. Groat, Director

PREFACE

Knowledge of the characteristics of highway runoff (concentrations and loads of constituents and the physical and chemical processes which produce this runoff) is important for decision makers, planners, and highway engineers to assess and mitigate possible adverse impacts of highway runoff on the Nation's receiving waters. In October 1996, the Federal Highway Administration and the U.S. Geological Survey began the National Highway Runoff Data and Methodology Synthesis to provide a catalog of the pertinent information available; to define the necessary documentation to determine if data are valid (useful for intended purposes), current, and technically supportable; and to evaluate available sources in terms of current and foreseeable information needs. This paper is one contribution to the National Highway Runoff Data and Methodology Synthesis and is being made available as a U.S. Geological Survey Open-File Report pending its inclusion in a volume or series to be published by the Federal Highway Administration. More information about this project is available on the World Wide Web at http://ma.water.usgs.gov/fhwa/

Fred G. Bank
Team Leader
Office of Natural Environment
Federal Highway Administration

Patricia A. Cazenas, P.E., L.S.
Highway Engineer
Office of Natural Environment
Federal Highway Administration

Gregory E. Granato
Hydrologist
U.S. Geological Survey

CONTENTS

FIGURES

TABLES

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
LENGTH				
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
AREA				
in²	square inches	645.2	square millimeters	mm²
ft²	square feet	0.093	square meters	m²
yd²	square yards	0.836	square meters	m²
ac	acres	0.405	hectares	ha
mi²	square miles	2.59	square kilometers	km²
VOLUME				
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft³	cubic feet	0.028	cubic meters	m³
yd³	cubic yards	0.765	cubic meters	m³

NOTE: Volumes greater than 1000 l shall be shown in m³.

Symbol	When You Know	Multiply By	To Find	Symbol
MASS				
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
TEMPERATURE (exact)				
°F	Fahrenheit temperature	5(F-32)/9 or (F-32)/1.8	Celcius temperature	°C
ILLUMINATION				
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m²	cd/m²
FORCE and PRESSURE or STRESS				
lbf	poundforce	4.45	newtons	N
lbf/in²	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
LENGTH				
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
AREA				
mm²	square millimeters	0.0016	square inches	in²
m²	square meters	10.764	square feet	ft²
m²	square meters	1.195	square yards	yd²
ha	hectares	2.47	acres	ac
km²	square kilometers	0.386	square miles	mi²
VOLUME				
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m³	cubic meters	35.71	cubic feet	ft³
m³	cubic meters	1.307	cubic yards	yd³
MASS				
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
TEMPERATURE (exact)				
°C	Celcius temperature	1.8C + 32	Fahrenheit temperature	°F
ILLUMINATION				
lx	lux	0.0929	foot-candles	fc
cd/m²	candela/m²	0.2919	foot-Lamberts	fl
FORCE and PRESSURE or STRESS				
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in²

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.

(Revised September 1993)

Statistical Approaches to Interpretation of Local, Regional, and National Highway-Runoff and Urban-Stormwater Data

By Gary D. Tasker *and* Gregory E. Granato

Abstract

Decision makers need viable methods for the interpretation of local, regional, and national-highway runoff and urban-stormwater data including flows, concentrations and loads of chemical constituents and sediment, potential effects on receiving waters, and the potential effectiveness of various best management practices (BMPs). Valid (useful for intended purposes), current, and technically defensible stormwater-runoff models are needed to interpret data collected in field studies, to support existing highway and urban-runoff-planning processes, to meet National Pollutant Discharge Elimination System (NPDES) requirements, and to provide methods for computation of Total Maximum Daily Loads (TMDLs) systematically and economically.

Historically, conceptual, simulation, empirical, and statistical models of varying levels of detail, complexity, and uncertainty have been used to meet various data-quality objectives in the decision-making processes necessary for the planning, design, construction, and maintenance of highways and for other land-use applications. Water-quality simulation models attempt a detailed representation of the physical processes and mechanisms at a given site. Empirical and statistical regional water-quality assessment models provide a more general picture of water quality or changes in water quality over a region. All these modeling techniques share one common aspect—their predictive ability is poor without suitable site-specific data for calibration.

To properly apply the correct model, one must understand the classification of variables, the unique characteristics of water-resources data, and the concept of population structure and analysis. Classifying variables being used to analyze data may determine which statistical methods are appropriate for data analysis. An understanding of the characteristics of water-resources data is necessary to evaluate the applicability of different statistical methods, to interpret the results of these techniques, and to use tools and techniques that account for the unique nature of water-resources data sets. Populations of data on stormwater-runoff quantity and quality are often best modeled as logarithmic transformations. Therefore, these factors need to be considered to form valid, current, and technically defensible stormwater-runoff models.

Regression analysis is an accepted method for interpretation of water-resources data and for prediction of current or future conditions at sites that fit the input data model. Regression analysis is designed to provide an estimate of the average response of a system as it relates to variation in one or more known variables. To produce valid models, however, regression analysis should include visual analysis of scatterplots, an examination of the regression equation, evaluation of the method design assumptions, and regression diagnostics. A number of statistical techniques are described in the text and in the appendixes to provide information necessary to interpret data by use of appropriate methods.

Uncertainty is an important part of any decision-making process. In order to deal with uncertainty problems, the analyst needs to know the severity of the statistical uncertainty of the methods used to predict water quality. Statistical models need to be based on information that is meaningful, representative, complete, precise, accurate, and comparable to be deemed valid, up to date, and technically supportable. To assess uncertainty in the analytical tools, the modeling methods, and the underlying data set, all of these components need be documented and communicated in an accessible format within project publications.

INTRODUCTION

Engineers, planners, economists, regulators, and others concerned with stormwater runoff from highways and urban areas often develop alternative plans to meet demands for desired quantity and quality of water at particular locations and times. Predictive procedures are needed for planning, permitting, and design of highway structures, as well as design of best management practice (BMP) structures (such as swales or retention ponds) and BMP maintenance procedures (such as street-sweeping programs). Predictive procedures include the development and application of conceptual models, simulation models, empirical models, and statistical models.

Any model, to be considered viable, must stem from an initial conceptualization of the system of concern. A conceptual model is a qualitative understanding of the sources and processes relevant to a problem. The model, which can commonly be represented by a diagram, is supported by common sense, scientific principles, data, and research to support the ideas. For example, expecting total pollutants in runoff to increase with increasing average daily traffic at a highway site—because vehicles are considered to be one of the primary sources of highway pollutants—is an example of a conceptual model. Analysis of available information, however, must substantiate the conceptual model to support development of more quantitative models through the application of empirical, deterministic, and (or) statistical techniques.

Stormwater-quality models have historically been used to characterize stormwater flow and quality, predict pollutant runoff loads, assess impacts on receiving waters, and determine the effectiveness of various BMPs to mitigate possible impairment of designated

beneficial uses of receiving waters (Kobriger and others, 1981; Shelley and Gaboury, 1986; Driscoll and others 1990). Now, and in the future, valid, current, and technically defensible stormwater runoff models are needed to

- help in interpreting data collected by field studies,
- support existing highway- and urban-runoff planning processes (Kobriger and others, 1981, Driscoll and others 1990; Young and others, 1996; Granato and others, 1998),
- meet National Pollutant Discharge Elimination System (NPDES) requirements (Young and others, 1996), and
- provide systematic and economical methods for calculation of Total Maximum Daily Loads (TMDLs) (Shoemaker and others, 1997).

Sound statistical planning models that use data collected at monitoring stations representing the region of interest can be used to assist policy makers and public officials in the evaluations of alternative plans. Statistical planning models represent a structured, ordered, and quantitative approach. These models can provide information for debates over proper choices for water-management alternatives and to evaluate competitive alternatives. Statistical planning models are needed for quantitative local, regional, or national water-quality assessments that can be demonstrated to be valid (useful for intended purposes), current, and scientifically/technically defensible when based on readily available monitoring data, land use information, and information about water-quality management practices.

The general objective of this paper is to examine methods for water-quality modeling and to identify and describe statistical methods and procedures that may be used to predict concentrations and loads of chemical constituents and sediment in highway and urban runoff and the potential for effects in receiving waters in terms of the unique characteristics of water-resources data. Although the concentrations, loads and potential effects of runoff constituents are often examined on a site-specific basis, these issues are also examined within the context of regional and (or) national interpretations. This report is designed to provide general (and sometimes very basic) information in terms of the analysis of water-resources data for stormwater analysis, but it is not designed to be a textbook for statistical methods. Appendixes are provided to expand on certain methods discussed in the text, and references to

suggested reading are provided where appropriate. Uncertainties in modeling are discussed, as well as the quality-assurance and quality-control measures that are necessary to address these uncertainties.

BACKGROUND

Simulation models, empirical models, regression models, and other statistical models historically have been used as approaches for predicting the quantity, quality, and loads of constituents in highway and urban runoff. Simulation models are used with model parameters that have a direct physical definition in an attempt to provide a detailed description of the physical processes and mechanisms that affect water quality. These models therefore require a considerable degree of detail in the description of the physical system. In simulation models, parameter estimation is not as data dependent as in statistical water-quality assessment models. On the other hand, empirical and statistical water-quality-assessment models provide a more general picture of water quality or changes in water quality. This picture could be in the form of a map of a water-quality statistic, or tables, or simply an equation with error bands on the parameters and predictions. Statistical regional water-quality models may also be used to estimate nonpoint-source loadings as inputs for more detailed water-quality simulation models (Ichiki and others, 1996). All these modeling techniques share one common aspect—the predictive ability of almost any model will be poor without suitable site-specific data for calibration (Shelley and Gaboury, 1986).

Criteria for model selection depend on modeling logistics as much as on data quality objectives (DQOs) and other technical considerations (Shoemaker and others, 1997). Logistical criteria include

- the availability of hardware and software to implement the model of choice,

- the availability of trained modelers to manipulate the model, to develop sound input parameters with an understanding of how they are used by the model, and to critically evaluate model results,

- organizational commitment to establish and support a model, to document the model, and to oversee subsequent applications of the model so that methods and results can be reviewed and accepted as valid, current, and technically defensible,

- organizational expertise with the model, to apply the model and to review applications of the model to maintain credibility of results, and

- available financial resources to support modeling efforts.

The resources needed to support a modeling effort increase in direct proportion to the complexity of the model chosen for analysis. A political criterion also is involved in the selection and use of modeling methods. To implement a successful modeling effort, the various interest groups involved in a project must be willing to accept modeling results. Even the most successful and sophisticated modeling effort will fail if the results are not understood and accepted by decision makers (Shoemaker and others, 1997).

A detailed description of each available runoff-quality model is beyond the scope of this report. Necessary information, however, is readily available in other publications. Driscoll and others (1990) discuss the application of different methods for national analysis of highway-runoff quality and related environmental effects. Hall and Hamilton (1991) describe many aspects of highway-runoff transport modeling. Bedient and Huber (1992) describe many of the hydrologic simulation models in use today. Young and others (1996) provide examples of simulation, empirical, and regression models, and the FHWA statistical model as applied to individual highway sites. Shoemaker and others (1997) provide a comprehensive guide describing most of the available predictive tools—including most applicable simulation, empirical, regression, and statistical models—for water-quality assessment within the TMDL process. A brief discussion of simulation, empirical, regression, and other statistical models, however, will provide information useful for statistical interpretation of local, regional, and national highway and urban stormwater-quality data.

Simulation models—including SWMM (Huber and Dickinson, 1988), STORM (U.S. Army Corps of Engineers, 1977), HSPF (Bicknell and others, 1993), and the Federal Highway Administration (FHWA) urban highway storm drainage model (Dever and others, 1983)—require detailed site-specific information and data to calibrate the model for current conditions. By using simulation models, highway engineers can evaluate the design of each highway—including the road surface, catch basins, and drainage structures—with respect to flow and water quality (Hall and Hamilton, 1991). Simulation models are also

useful in estimating relations between variations in input parameters and resulting flows, constituent concentrations, and constituent loads in runoff. The validity of these estimates, however, depends upon a robust calibration of the model with site-specific data (Driscoll and others, 1990). For example, Zarrielo (1998) compared results of simulations made with nine uncalibrated runoff models to observed flows in two urban watersheds; although the modelers had very detailed information about site characteristics and precipitation, the resulting estimates of peak flow rates and total stormflow volumes differed from measured data by as much as 260 and 240 percent, respectively. Simulations are considered to be highly useful in the design phase or post construction-analysis phase of specific sections of a highway, but collection of the site-specific data and the high level of effort necessary to successfully implement these models are not practical for planning or management on a regional or national scale (Driscoll and others, 1990).

Empirical models include the U.S. Environmental Protection Agency (USEPA) Screening Procedures (Mills and others, 1985) and the Simple Method (Schueler, 1987). Empirical models involve the use of estimated concentrations (presumably derived from local studies and (or) the literature) multiplied by some fraction of local precipitation volumes (used to estimate runoff volumes) to estimate annual or storm loads of constituents of concern. Empirical models provide order-of-magnitude estimates of loads but do not indicate correlation among variables and do not provide estimates of the uncertainty in predictions without detailed site-specific data needed to estimate probability distributions of the precipitation/runoff volumes and concentrations used in the equation.

Highway-runoff regression models include those developed by the FHWA (Kobriger and others, 1981; Driscoll and others, 1990); State departments of transportation in Washington (Chui and others, 1982), California (Kerri and others, 1985), and Texas (Irish and others 1998); and the Ontario Ministry of Transportation (Thomson and others, 1996; 1997a). These models use information such as precipitation characteristics, highway-design features, traffic volumes, and interrelations between measured constituents to predict concentrations and (or) loads of highway-runoff constituents. The USGS regression method (Tasker and Driver, 1988; Driver and Tasker,

1990) uses the percentage of impervious area, rainfall statistics, and in some cases the mean minimum January temperature to estimate concentrations and (or) loads of urban-runoff constituents. This model also may be used for estimating highway-runoff quality in urban areas (Young and others, 1996).

Other statistical techniques have also been used as approaches for prediction of modeling urban- and highway-runoff quantity, quality, and loads. These techniques use measures of the central tendency of available data (such as the mean or median) and measures of the variability of data (the variance or coefficient of variation) to predict model outputs and the uncertainty thereof. The model formulated by Driscoll and others (1990), uses storm event statistics and the probability distribution of site event-mean concentrations (EMCs) to estimate runoff volumes, concentrations, and loads in runoff. Driscoll and others (1990) then use these loads and the probability distribution of streamflow volume at a given site to estimate potential dilution in receiving waters. Statistical models that use readily available rainfall statistics and water-quality data to produce a frequency distribution of concentrations, loads, and potential for receiving-water effects are useful because assessments of risk and return periods can be calculated (Driscoll and others, 1990).

The existing FHWA statistical pollutant loading and impact model approach described by Driscoll and others (1990) is, given appropriate input data, generally valid for intended purposes. This model uses site characteristics and other factors to calculate estimates of runoff volumes, loads, and receiving-water concentrations as a probability distribution. This model, however, is not designed for scientific or technical interpretation of study-site data in terms of the potential relations between constituents and (or) study-site characteristics. Therefore, other statistical models are necessary for scientific or technical interpretation of local, regional, and national highway- and urban-stormwater data.

Statistical techniques are commonly best suited to highway-runoff modeling needs at any scale (local, regional, or national). Population statistics and standard techniques for the analysis of error in predictions can be used to assess risk of decision error. Estimation of the error in predictions from empirical models or uncalibrated simulation models is, in reality, impossible, because inputs are complex and the effects on the

error of outputs can be multiplicative. Use of simulation models requires a high degree of institutional expertise and experience, as well as a substantial modeling effort, for each site of interest. Many simulation models are suitable for specific highway applications, but the detailed input requirements and extensive modeling efforts required may preclude use for planning-level estimates. By comparison, use of existing statistical models minimizes the logistical burdens required for model application (Shoemaker and others, 1997). If the statistical model used is based on accepted statistical methods, is formulated from a valid, current, and technically defensible data set, and is documented and communicated in an accessible format, then these steps will ensure that model results will be accepted. The existence of a quantitative data set further lends credence to the acceptability of model results. The complexity of simulation models and the large range of reasonable input parameters inherent in the model calibration process can lead to differences in professional judgment, which can negatively affect acceptance of simulation-model results. Therefore, simple statistical models may be preferable to simulation models for many highway-runoff modeling needs.

Statistical analysis is also important to the design and implementation of highway- and urban-runoff data-collection programs that will yield results suitable for inclusion in local, regional, and national-synthesis efforts. Statistical analysis of available information is necessary to help determine the design of the sampling effort (random as opposed to systematic), the frequency of sampling, the number of and location of sites, and the quality of the resultant data (Averett and Schroder, 1994). For example, Thomson and others (1996; 1997b) used an extensive highway-runoff data set from Minnesota to determine that EMC samples from at least 15 to 20 storms are required to provide reasonable estimates of mean total suspended solids, total dissolved solids, total organic carbon, and zinc concentrations from each study site. Statistical methods can also be used to design and (or) optimize stormwater data-collection networks (Tasker and Raines, 1995). Statistical-analysis techniques are also necessary for the design, implementation, and interpretation of quality-assurance and quality-control (QA/QC) programs necessary to demonstrate that data collected are valid (useful for the intended purposes), technically defensible, and complete (Jones, 1999).

BASIC STATISTICAL CONSIDERATIONS

Proper classification of variables and an understanding of the characteristics of water-resources data are crucial for interpreting the results of individual studies and for combining these results in a regional or national synthesis of stormwater-quality data. Classification of the different types of variables being used to analyze data may determine which statistical methods are appropriate for data analysis. An understanding of the unique characteristics of water-resources data is necessary to evaluate the applicability of various statistical techniques, to interpret the results of these techniques, and to use tools and techniques that account for the nature of water-resources data sets. Understanding the methods and measures used to determine and analyze the population structure is also important. When necessary (as is common for water-resource and particularly for stormwater data sets), one must choose appropriate methods of data transformation to enable use of statistical techniques without violating the statistical assumptions underlying the methods chosen for analysis. Therefore, classification of variables of interest, an understanding of the statistical characteristics of water-resources data, a familiarity with the population structure and basic methods of analysis, and (when necessary) selection and proper use of population-transformation techniques are necessary to form valid, current, and technically defensible stormwater-runoff models.

Classification of Variables

Classification of variables is useful in several ways because the proper method of data analysis often depends on variable type. A variable can be classified by inherent mathematical structure (discrete or continuous), by statistical objectives of the study (response or predictor), and by level of measurement (nominal, ordinal, interval, or ratio).

A variable is discrete if there is a gap between two successively observable values in which an observed value is not possible (for example, an integer scale). A variable is continuous if there is always another observable value between any two observed values (for example, a real number scale). Examples

of a discrete variable are the number of floods above a threshold, the number of exceedences of a water-quality standard, or the identification of a group. It is not possible to have 3.2 floods above a threshold, 11.7 exceedences of a standard, or to have a value between group A and group B. Examples of continuous variables are chloride concentration or the number of pounds of road salt applied per lane-mile of roadway per season. Discrete variables, however, can sometimes be treated as continuous if they cover a wide range with small gaps between observations. For example, the number of days per year with measurable rainfall could vary between 0 and 365. Furthermore, continuous variables may be grouped into categories and treated as discrete variables. For example, pH readings may be grouped into the categories of low, medium, and high.

A variable may also be classified according to the statistical objectives of the study without regard to the variable's mathematical structure. A variable that is described in terms of other variables in a regression model is called the response variable (or the dependent variable, or the predicted variable). Variables used to describe the response variable in a model are called predictors (or explanatory variables, carriers, or independent variables). For example, if one wishes to predict average nitrate concentration at a streamflow site on the basis of land-use characteristics of the upstream watershed, then average nitrate is the response variable, and land-use variables, such as percentage of urban or industrial land, would be predictors.

The level of measurement may also be used to classify variables. Nominal variables are variables whose observed values are labels for different unordered categories. For example, the "type" variable for a basin may be 1 for urban or 0 for nonurban. Ordinal variables are variables whose observed values are ordered with no implication of distance between different points on the scale. For example, Driscoll and others (1990) classified highway sites as either "urban"—average daily traffic (ADT) counts that are greater than or equal to 30,000 vehicles per day (VPD)—and "nonurban"—ADTs that are less than 30,000 VPD. Another example of an ordinal variable would be the use of the previously mentioned pH categories of low, medium, and high. Interval variables have equal differences between successive points on the measurement scale, but the zero point is arbitrary. In this case, one can compare differences in observed

values. An example of an interval variable would be the increase in population density since the last census. Ratio variables have equal differences between successive points and a fixed zero. They allow one to compare differences in observed values and relative magnitude. For example, nitrate concentration measured at several monitoring sites in a region could be treated as a ratio-scale variable.

Characteristics of Water-Resources Data

The applicability of any given statistical procedure depends directly on assumptions about the characteristics of the data being explored. Results from statistical analysis may be meaningless or, even worse, misleading if data do not conform to the design assumptions of the statistical method used. Therefore, one must carefully consider which statistical methods are appropriate in terms of the data being evaluated. The unique features of water-resources data affect the suitability of interpretations made by the application of many classical statistical techniques. Pertinent characteristics of water-resources data (Helsel and Hirsch, 1992; Helsel, 1993; Hirsch and others, 1993), include the following:

- Nonrepeatability—Measurements are usually observational, not experimental. More specifically, statistical regularity cannot be demonstrated for water-resources data by repeating a controlled experiment because exactly recreating the many natural and anthropogenic influences that affect each measurement is impossible.
- A lower limit of zero—Negative values are not possible for many water-resources characteristics. For example, negative values for measured precipitation, flow, or constituent concentrations do not have meaning.
- Censored data sets—Limits in methods for sample collection and analysis cause data to be reported as either above or (more typically) below one or more reporting limits, which produce a censored population of data.
- Meaningful outliers—Valid measurements that are considerably higher or lower than most of the measured population are common among hydrologic data sets.

- Positive skewness—Data sets that are not symmetrical around mean or median values are typical for water-resources data sets because the combined effects of a lower bound of zero, censoring, and meaningful outliers tend to produce data sets in which the right tail of the distribution is extended and the left tail is truncated.
- Nonstandard distributions—Many statistical techniques are based on the assumption that the data (or in the case of regression models, the residuals) are normally distributed (the "bell-shaped" curve). The nonstandard distributions characteristic of water-resources data sets necessitate alternative tools for analysis.
- Autocorrelation—Natural and anthropogenic effects tend to cause conditions in which consecutive measurements tend to be strongly correlated.
- Interdependence—Changes in one characteristic of interest (such as rainfall intensity) cause changes in other characteristics (such as measured flows and concentrations).
- Temporal variation—Measured water-resources characteristics vary cyclically at timescales of hours, days, weeks, seasons, years, and even decades because of both natural and anthropogenic influences.

Therefore, data sets must be collected in a manner that will represent the underlying population distribution, and the statistical methods used to characterize the data must be appropriate for populations with these characteristics. Ancillary information and metadata (information about a given data set, including explanatory information and data-quality information) pertinent to the statistical characteristics of the sampled population need be documented and communicated in an accessible format with monitoring-study data to support interpretation methods.

Ancillary data are important to quantify possibly confounding variables that may preclude meaningful interpretation of data because statistical regularity cannot be demonstrated through controlled experiments. For example, Driscoll and others (1990) noted the effect of local land use on regression analysis to predict the median EMC of zinc from measures of average daily traffic on a site-by-site basis. When all sites (with average daily traffic greater than 30,000 vehicles per day) were included, regression analysis indicated a small negative slope with increasing traffic

and an R^2 of about 0.04, indicating that EMCs for zinc would decrease with increasing traffic but that this relation was very weak. When one site (which was heavily influenced by a local zinc-smelting operation) was omitted, however, regression analysis indicated a strong positive slope and relatively strong relation between increasing traffic volume and increasing EMCs for zinc (R^2 was about 0.7). In this case, the local land use is the ancillary information necessary for meaningful interpretation of data based on the assumption that traffic characteristics would affect (if not control) zinc concentrations in highway runoff.

The effects of censored data can be especially problematic for interpretation of water-quality data. Laboratory detection limits change with time, can be dramatically different from laboratory to laboratory, and may even be different from method to method within a laboratory. For example, Garbarino and Struzeski (1998) indicate that detection limits for total recoverable copper in whole-water samples are 0.4, 5., and 0.3 micrograms per liter for the graphite furnace-atomic absorption spectrophotometry (GF-AAS), inductively coupled plasma-optical emission spectrometry (ICP-OES), and inductively coupled plasma-mass spectrometry (ICP-MS) methods, respectively. A number of methods may be used to compensate for the effects of one or more detection limits while applying statistical analysis tools to a data set (Gilliom and Helsel, 1986; Helsel and Gilliom, 1986; Helsel and Cohn, 1988; Helsel, 1990; Helsel and Hirsch, 1992). These methods of statistical analysis require knowledge of the detection limits truncating each data set. Detection-limit information, however, may not be available. For example, in compiling data to develop a pollutant-loading model, Driscoll and others (1990), noted that "it was virtually impossible to unequivocally determine the actual detection limit associated with each pollutant concentration." Driscoll and others (1990) could not determine the detection limits because this information was not explicitly documented in available ancillary data and because the data were produced by a number of different analytical laboratories over the years between the mid 1970's and the mid 1980's. Thompson and others (1996) also noted the difficulty in identifying detection-limit data in an extensive highway-runoff data set (416 storms from four sites monitored during 1976–83). Detection-limit artifacts also affect

statistical properties of individual data sets. When a data set contains values reported as less than one or more detection limits an overestimation of central-tendency measures and an underestimation of dispersion measures will be caused by truncation of the lower tail of the true population (Driscoll and others, 1990). Also, because the relative uncertainty in the accuracy and precision of individual values tend to increase as reported concentrations approach the detection limit, the percent error expected for measurements near detection limits is much higher than for values well within the measurement range of the method of analysis (Granato and others, 1998).

The presence of meaningful high-end outliers (actual but extreme values) contributes to the positive skew and is a factor producing nonstandard distributions. High-end outliers represent times when, for example, regulatory criteria may be exceeded and the health of local ecosystems may be affected. These outliers, however, produce a host of potential problems for interpretation of data sets assembled for use in a regional or national synthesis.

Certain assumptions and conditions need to be considered in terms of the decision to include or exclude individual data points or even entire data sets, among which are the following:

- It is often assumed that outliers are not meaningful. This assumption may be misleading unless documented QA/QC information indicates problems with sample collection, processing, and (or) analysis that would justify elimination of suspect data.
- An outlier may be meaningful, but it may represent the effects of a process that should not be considered in a synthesis designed to characterize normal highway-runoff quality. For example, a chemical spill that occurs before or during a runoff-quality study may be representative of runoff quality and subsequent environmental effects at sites affected by spills, but the data may not be characteristic of "normal sites." In fact, the rate of substantial spills is about 0.0019 incident per lane kilometer per year (as estimated from the median of highway hazardous-materials incidents and public road mileage compiled by U.S. Environmental Protection Agency, 1999). Therefore, data from a spill site would incorrectly bias a national data set unless a sufficient number of sites were monitored to properly represent the probability of a spill at any given site. In another example, Driscoll and

others (1990) eliminated data sets from sites affected by fallout from the eruption of Mt. Saint Helens in Washington State. Although these "meaningful outliers" were representative of the effects of volcanic eruption on the quality of highway runoff, this effect was not deemed suitable for estimation of typical highway-runoff quality in the United States.

- An extreme outlier may be meaningful and may represent the effects of a process that should be considered in a synthesis designed to characterize normal highway-runoff quality, but this effect may obscure effects of other process-related variables. The example of the effect of the highway site under influence of a local zinc smelting operation indicates the necessity for detailed documentation to describe surrounding land use, but this effect precludes development of meaningful relations between metal concentrations and average daily traffic for more normal highway conditions across the United States (Driscoll and others, 1990).

Elimination of outliers is considered a dangerous and unwarranted practice for the interpretation of water-quality data, unless one has substantial objective evidence demonstrating that the outliers are not representative of the population under study (Helsel and Hirsch, 1992). When statistical tests are used to detect outliers, these tests do not indicate that outliers represent errors; they do indicate that the population of measured vales is not necessarily a normal distribution. Excessive numbers of extreme values may cause significance levels of tests requiring the normality assumption to be in error. Therefore, use of a test requiring the normality assumption will produce inaccurate results when outliers affect population structure. Outliers may have high leverage and thus a strong potential for influencing the slope of a regression line (Helsel and Hirsch, 1992). If an outlier is discovered to have a strong influence on the slope of a regression line (the slope and (or) the correlation coefficient changes significantly when the point is omitted), then one must determine whether the outlier represents extreme values for a single process or if a secondary process is characterized by the outlier. Measurement and documentation of explanatory variables such as precipitation and flow (Church and others, 1999); real-time measures of water-quality characteristics such as specific conductance, pH, temperature, and turbidity (Spangberg and Niemczynowicz, 1992; Whitfield and

Wade, 1992; Granato and Smith, 1999); use of ratios between constituents of interest (Granato, 1996); and results from a comprehensive QA/QC program (Jones, 1999) can be used to identify and explain outliers in terms of the potential effect of real physicochemical processes as opposed to the effects of sampling artifacts.

The natural and anthropogenic processes controlling runoff quality and the methods for sampling, processing, and analysis often cause problems with autocorrelation (also referred to as serial correlation or correlation—the dependence of residuals in a time sequence because data reflect the effects of preceding conditions). One of the assumptions inherent in many regression techniques is that the residuals are independent (Helsel and Hirsch, 1992). Autocorrelation can be a problem within stormwater data sets because many of the variables used for analysis are pairs of data in a time series. For example, precipitation and flow, flow volume and concentration, and relations between measured constituents (total suspended solids and lead, for example) may be pairs of data in a time series. Time-series effects may also occur between subsequent storms. For example, Irish and others (1998) indicate that the duration, the volume of runoff per unit area, and the intensity of runoff per unit area of the preceding storm are significant causal variables in a regression model developed for highway-runoff loads of suspended solids and metals in Texas. Autocorrelation can be important because it affects the optimization of regression coefficients, affects estimates of population variance (invalidating results of hypothesis tests), and produces confidence and prediction intervals that are too narrow for the real population being sampled. To address autocorrelation problems, one may group data into time periods and use summary statistics in an analysis, use methods that are robust with respect to autocorrelation (Helsel and Hirsch, 1992), incorporate explanatory variables into predictive models that will account for potential effects of autocorrelation (Irish and others, 1998), or subsample from large data sets to eliminate autocorrelation.

Temporal variation may also increase variability in data and affect the comparability of data between sites. Seasonality is an obvious factor that may affect a population of stormwater samples at any given site. For example Driscoll and others (1990) segregated "snow washoff events" from other events for analysis and found that, for many constituents, snow washoff events had substantially higher median site EMCs, much

wider confidence intervals, and a relatively few number of events than for the rest of the available data. More subtle temporal variations may affect relations between predictors and response variables and (or) contribute to the variability in measured water quality in a population of stormwater-quality samples. For example, Whitfield and Wade (1992) used results from automatic water-quality monitoring stations to detect daily cycles in receiving-water quality, as well as the effects of storms. In addition, the magnitude of seasonal variation would be expected to be a function of local climate and may therefore partly obscure relations between predictors (such as ADT) and response variables (such as constituent concentrations) if a statistical analysis includes sites with different patterns of temporal variation. For example, Driver and Tasker (1990) used the mean minimum January temperature as a predictor variable to partly account for differences in the magnitude of seasonal variation among sites in an analysis of National Urban Runoff Program (NURP) data.

Population Structure and Analysis

Sampling theory—the concept that one can monitor a given number of events and with this information estimate the properties of the underlying distribution—is based upon the concept of a probability distribution. The structure of a given population (the probability distribution) will determine which methods will be appropriate for statistical analysis and interpretation of water-resources data. If, for example, measured data fit a normal distribution, one measure of location (central tendency) and one measure of spread (variability) can be used to define the entire population. However, the applicability, robustness, and relative power of different measures of location, measures of spread, and measures of skewness depend upon the structure of data and the objectives of the analytical process.

The structure of data is often described by means of population-frequency distributions. If variables can be ascribed to a particular frequency distribution, then the known structure of the distribution has many potential uses (Athayde and others, 1983; Driscoll and others, 1990; Helsel and Hirsch, 1992), including the following:

- concisely reporting data in terms of population measures rather than total range (which may be misleading for data with outliers and multiple detection limits),
- examining the characteristics of the data (for example, the mean and standard deviation can be used to define the location and shape of a normal distribution),
- establishing the probability of any given value in the distribution (for example the probability of exceeding a water-quality standard),
- comparing results from different sites on a common basis,
- providing a framework for examining the transferability of data quantitatively, and
- testing hypotheses (for example, establishing whether concentrations of metals at sites with an ADT of less than 30,000 vehicles per day are statistically different from concentrations of metals at sites with an ADT greater than 30,000 vehicles per day).

The number of potential distributions is infinite (Ott, 1993). For example, McLaughlin (1999) provides equations for the probability distribution and cumulative distribution functions for more than 50 distributions in terms of the location, shape, and scale of each distribution. Figure 1 indicates generalized shapes of the probability-distribution and cumulative-distribution functions for the exponential, gamma, and normal distributions. Commonly, water-resources data can be characterized by relatively few distributions, and many of the available distributions are specialized variations of more general distributions (Helsel and Hirsch, 1992; Ott, 1993). For example, the lognormal frequency distribution is simply a normal distribution for data that have been transformed to logarithmic space so that the resulting distribution approximates that of the theoretical normal distribution.

Statistical-analysis methods can generally be classified as either parametric (methods in which a specified data distribution is necessary to support design assumptions) or nonparametric (methods that do not depend on a specified data distribution to establish their meaning). Nonparametric tests, because they are not as dependent on an assumed population distribution, may be more robust for data analysis. The power of parametric techniques is generally higher
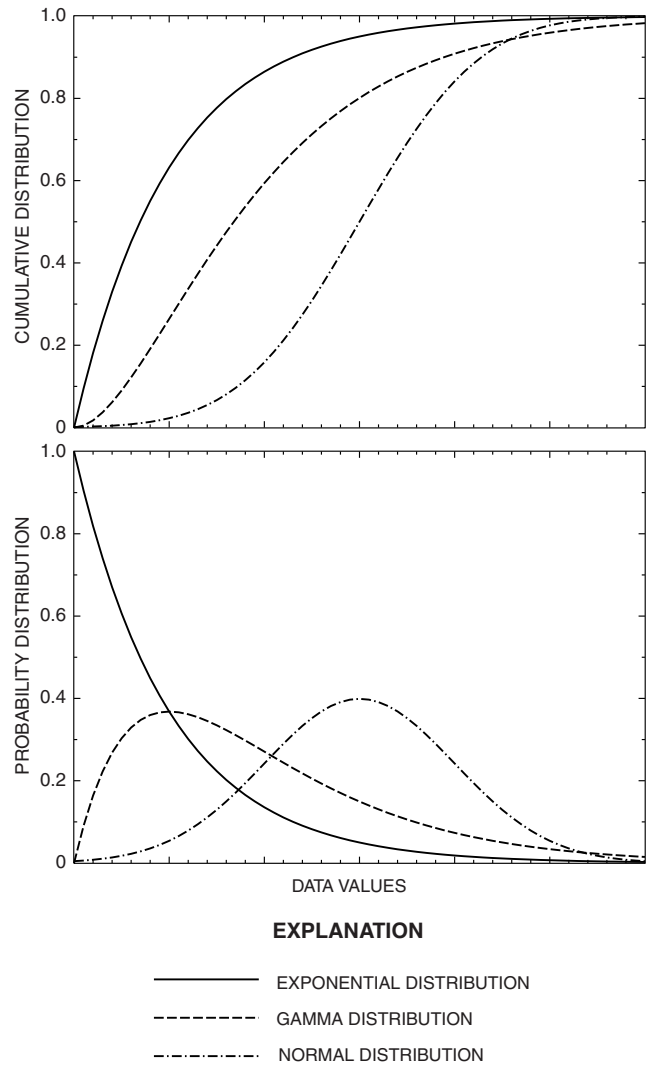
Figure 1. Example of the generalized shape of the probability and cumulative distributions of unitless, exponential, gamma, and normal populations.

when the distributional assumptions are correct. In many cases, however, the relative advantage of parametric techniques decreases with increasing population size (Hirsch and others, 1993).

Statistical measures of population location, spread, and skewness are summarized and defined in terms of the statistical basis and design assumptions that define each method in table 1. Helsel and Hirsch (1992) and Hirsch and others (1993) use these techniques to describe mathematical information for statistical analysis of water-resources data.

Table 1. Basic statistical techniques for parametric and nonparametric data analysis

[N, nonparametric; P, parametric]

Technique	Measure	Basis	Definition	Comments
Arithmetic mean (Average)	Location	P	The sum of all data divided by the sample size.	Can be affected by the presence of and (or) changes in the magnitude of one or more outlying observations. Representativeness depends upon the assumption that the data are normal (or at least unimodal and symmetric).
Median	Location	N	The middle value when data are ordered from lowest to highest.	When there is an even number of data points the median is the average of the two central observations. The median is also refered to as the 50th percentile.
Mode	Location	N	The data value that occurs with the highest frequency.	A data set may have more than one modal value.
Geometric mean	Location	P	The mean of the logarithms of data that is transformed back into original units.	Representativeness depends on the assumption that the data are normal (or at least unimodal and symmetric) in log space.
Trimmed (or weighted) mean	Location	P	The mean of censored data divided by the sample size after censoring.	Trimmed means (or weighted) means are computed once values judged as outliers have been eliminated (or weighted with a value of zero). It is typical to trim a given percentage from the bottom and top of the data in an attempt to apply systematic methods.
Range	Spread	N	The difference between the largest and smallest measurements in a set.	Although nonparametric, the range is affected by the presence of and (or) changes in the magnitude of one or more outlying observations.
Variance	Spread	P	The sum of the squared deviation of all measurements divided by one less than the total number of data points.	The variance can be unduly affected by the value of one or more outlying observations.
Standard deviation	Spread	P	The positive square root of the variance.	The standard deviation can be unduly affected by the value of one or more outlying observations.
Coefficient of variation (COV)	Spread	P	The ratio of the standard deviation over the mean.	A measure of spread normalized to the magnitude of the mean.
Interquartile range (IQR)	Spread	N	The difference between the 75th and 25th percentile values (by number of measurements) when data are ordered from lowest to highest.	Typically used as a measure of central spread because the 25th, 50th (median), and 75th percentiles split the data into four equal-sized quarters (by number of measurements). Other percentile ranges may be used as well.
Median absolute deviation (MAD)	Spread	N	The median of the absolute values of the difference between each data point and the data-set median.	The MAD, because it is the median of the population of absolute differences, is resistant to the effects of outliers.
Coefficient of skewness	Skewness	P	The third central moment divided by the variance cubed.	A positive value indicates that the population is right-skewed, and a negative value indicates left skew.
Quartile skew coefficient	Skewness	N	The difference between the range of each quartile (25th to 50th and the 50th to 75th) divided by the IQR	A positive value indicates that the population is right-skewed, and a negative value indicates left skew. Other percentile ranges may be used as well.

An understanding of the assumptions inherent in statistical measures, the methods used to calculated these measures, and their sensitivity to changes in location, spread, and skewness are important for interpretation of results of more complex statistical analysis. For example, in the calculation of a simple regression, the means and variances of each population of interest determine the slope and intercept of the resultant line equation (Hirsch and others, 1993). Furthermore, the simple measures of population characteristics covary with differing population structure. If a data distribution fits the normal distribution, then the mean, median, mode, and trimmed means should be equivalent if not equal (Helsel and Hirsch, 1992; Ott, 1993). If, however, a population is lognormal, then the geometric mean and median should be equivalent if not equal (Helsel and Hirsch, 1992). As unimodal populations become increasingly skewed to the left or right, the mean and trimmed mean will fall increasingly to the left or right of the median, respectively, and the mode will fall increasingly to the right or left of the median, respectively (Ott, 1993). Indices of spread also will have unique relations when data are normally distributed. For example, if data are normally distributed, then dividing the range by 4 should produce a value approximately equal to the standard deviation because about 95 percent of values should lie in the range of plus or minus 2 times the standard deviation (Ott, 1993).

For water-resources data, graphical analysis is an essential first step in the interpretation process. Use of graphical analysis can provide a visual summary of the data and can help reveal the most appropriate population structure and methods for analysis (Helsel and Hirsch, 1992). Graphical tools such as the scatterplot, the histogram, the boxplot, and the probability plot can be used to characterize data and find potential problems. For example, Driscoll and others (1990) present the event-mean suspended-solids concentrations and runoff volumes for the I-794 data set from Milwaukee, Wisconsin. A scatterplot (fig. 2), a boxplot (fig. 3), two histograms (fig. 4), and a probability plot (fig. 5) indicate that event mean suspended-solids concentrations are not normally distributed in linear space and that runoff volume does not control event mean suspended-solids concentrations at this site. Driscoll and others (1990) used probability plots to establish that the lognormal distribution was a sufficiently close approximation for highway-runoff-quality data to be included in
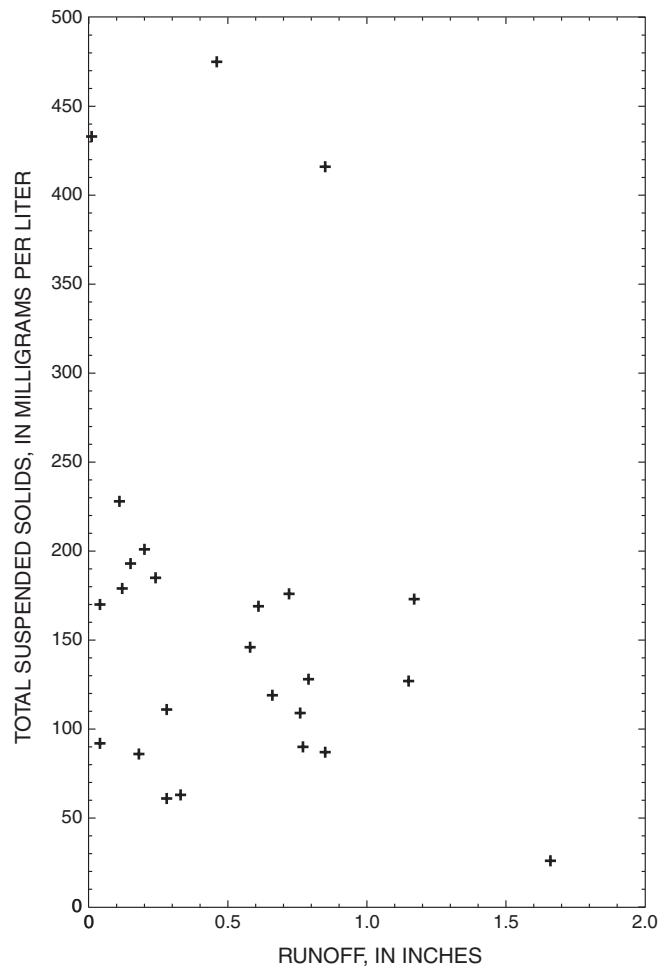

Figure 2. Example of a scatterplot of total runoff volume and event mean sediment concentrations from a highway runoff monitoring study along highway I-794 in Milwaukee, Wisconsin (data from Driscoll and others, 1990).

their national highway-runoff-quality model. In this case, use of graphical analysis helped identify that a substantial number of below-detection-limit values were incorporated (without special notation) into some of the historical data sets and indicated that the below-detection-limit data did not fit the lognormal probability distribution and would affect population statistics (Driscoll and others, 1990). Helsel and Hirsch (1992) provide detailed descriptions about the use of graphs in exploratory data analysis (EDA). The National Institute of Science and Technology provides text and software for mathematical and graphical techniques to approach EDA (Croarkin and Tobias, 2000). A number of specification tests (methods to check that assumptions of the
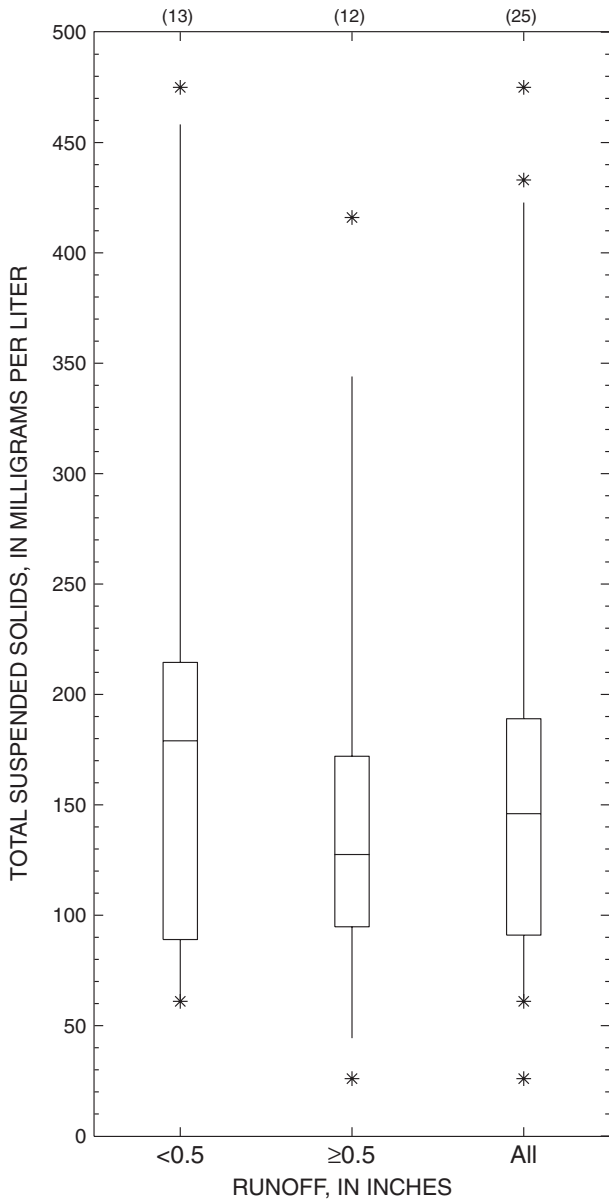
Figure 3. Example of a boxplot of total runoff volume and event mean sediment concentrations from a highway runoff monitoring study along highway I-794 in Milwaukee, Wisconsin (data from Driscoll and others, 1990).
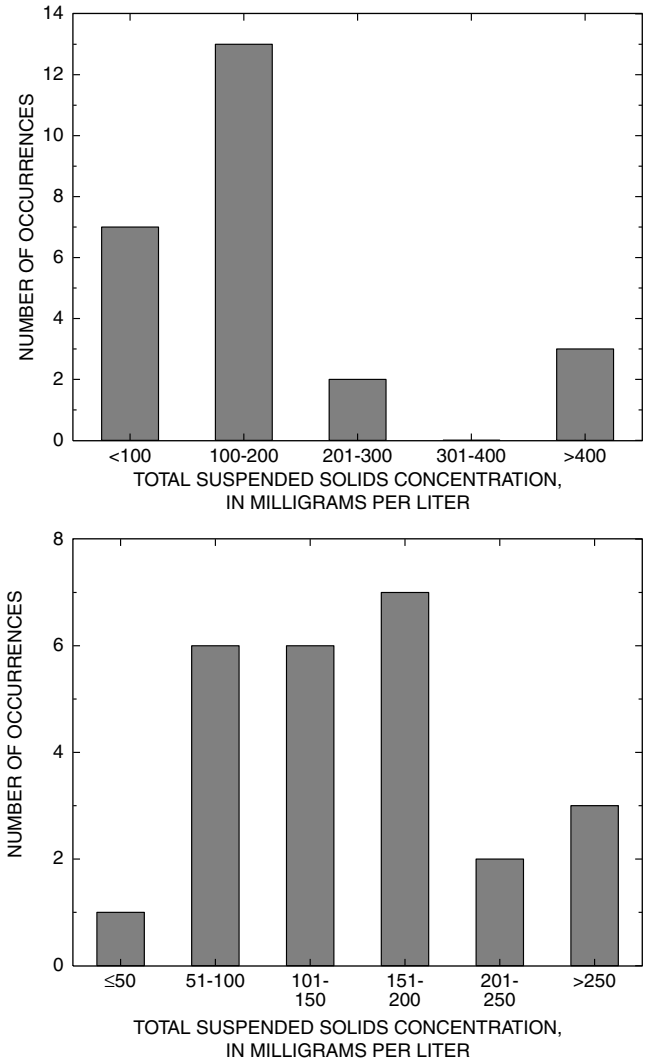
Figure 4. Example of two histograms (in groups of 50 and 100 milligrams per liter) of event mean sediment concentrations from a highway-runoff monitoring study along highway I-794 in Milwaukee, Wisconsin (data from Driscoll and others, 1990).

estimation method are valid) have been developed in recent years and may also be useful for examining runoff data before application of statistical-analysis techniques (Godfrey, 1988).

Within a discussion of population structure and statistical analysis, it should be noted the event mean concentration, or EMC, is not a statistical mean based on any implied probability distribution; rather the EMC is an operational definition used to characterize individual storm-event water quality. The EMC is defined as the total load of a stormwater-quality constituent divided by the total flow that occurs during the storm of
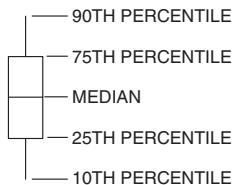
Figure 5. Example of a probability plot of event mean sediment concentrations from a highway runoff monitoring study along highway I-794 in Milwaukee, Wisconsin (data from Driscoll and others, 1990).
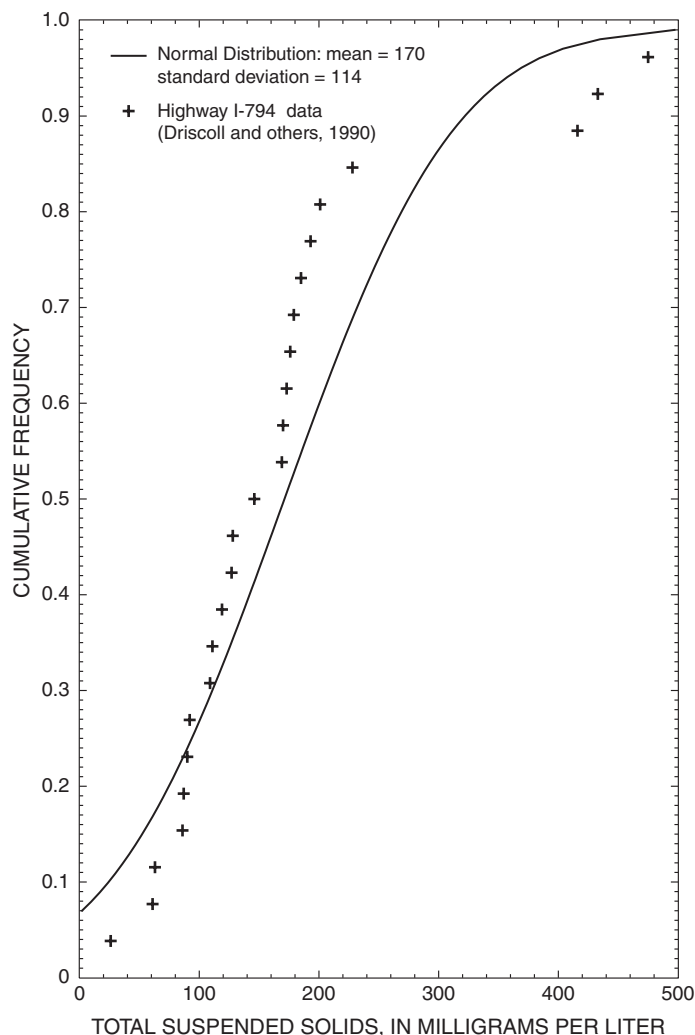
either case, the EMC is not a parametric average, but a time- and flow-integrated estimate of stormwater quality.

Transformations

The general versatility, power, and mathematical elegance of parametric procedures designed for data that fit a normal distribution are usually advantageous in the application of statistical analysis. Mathematical transformations (used to redistribute population characteristics to approximate a normal distribution) are often employed to facilitate the use of normal (parametric) statistical-analysis techniques. Transformations are usually made for one or more of the following reasons: to simplify the model; to stabilize the variance; to normalize the data; or, for regression analysis, to define a transformed model with error distributions that fit the assumptions of the model. For regression models, transformation of the response variable is often desirable if the response variable is nonnegative and the range of observed values is one or more powers of 10. Log transformations are often useful for minimizing the standard error of the estimate (Driver and Tasker, 1990). If all the values are far from zero and the range of values is relatively small, however, transformation will have little effect. Nonparametric methods are generally invariant to measurement scale, so transformations do not alter data with respect to these methods (Helsel and Hirsch, 1992).

Transformations typically improve the symmetry of data by means of a mathematical function designed to alter the distance between observations on a line plot. The goal is to expand or contract the distance between the median and extreme values in the population of interest. Power functions (in which the transformed variable is raised to an exponent) and logarithms (natural and base 10) are usually used to transform data (Helsel and Hirsch, 1992). To reduce negative (left) skewness, powers of greater than 1 are used. To address positive (right) skewness powers less than 1 (the square or cubed roots) or logarithms are used.

interest (Huber, and others, 1979; Driscoll and others, 1990; Huber, 1993). In theory, an EMC would be obtained by collecting and homogenizing all runoff from a given event and then sending a representative subsample for analysis. In practice, the EMC representing each storm is determined from analysis of one flow-weighted composite sample or from a number of discrete samples that are flow-weighted mathematically (Driscoll and others, 1990; Huber, 1993). In

Logarithmic transformations are especially useful for normalizing values that vary by orders of magnitude, and they dampen the effect of very large outliers on statistical estimators. In addition, the standard error of the transformed population can be expressed as a percent error when transformed back into original space. Logarithmic transformations, however, increase the relative weight of small data values and can magnify uncertainties in the true (or assigned) value of concentrations near detection limits (Stedinger and others, 1993).

Historically, populations of urban- and highway-runoff data can be characterized by the log-normal distribution (Athayde and others, 1983; Driscoll and others, 1990). EMCs of highway-runoff constituents indeed vary by more than one and as much as four orders of magnitude at and between sites. For example, Smith and Lord (1990) report that total suspended solids concentrations range from 4 to 1,156 mg/L (milligrams per liter), total organic carbon concentrations range from 5 to 290 mg/L, zinc concentrations range from about 0.01 to 3.4 mg/L, and chloride concentrations range from 5 to 13,300 mg/L. Driscoll and others (1990) used the natural logarithmic (ln) transform to normalize data used to develop their national highway-runoff-quality model. Athayde and others (1983) also indicated that urban-runoff data collected for the National Urban Runoff Program (NURP) generally could be characterized by use of the lognormal distribution. Thompson and others (1996) examined the effects of logarithmic and exponential transformation on the R^2 of various regression models designed to predict constituents from surrogate parameters; they found that these transformations demonstrate the best explanation (highest R^2) for the water-quality predictor variables but that untransformed models worked best for the constituents predicted by traffic volume. Driscoll and others (1990) graphically compared the probability distribution defined by the mean and standard deviation of the transformed population with individual points in the data set and tested the fit by means of the probability plot correlation coefficient (PPCC) test (Vogel, 1986). In that study, the lognormal distribution was found to be a satisfactory model for most runoff constituents at most runoff sites (Driscoll and others, 1990).

Athayde and others (1983) examined the mathematical properties of the lognormal distribution in terms of the ratio of the mean and percentile values in relation to the median as a function of typical coefficients of variation (COV; the ratio of the standard deviation to the mean). For a population with a COV of 1, the mean would be expected to be about 1.4 times the median, 90 percent of the data would fall in a range from about 0.25 to 3.9 times the median, and the interquartile range (about 50 percent of the data) would fall in a range from about 0.6 to 1.7 times the median (fig. 6). Historically, COVs of highway- and urban-runoff-quality constituent populations have ranged from about 1/4 to about 3, with typical values that are about 1 (Athayde and others 1983; Driscoll and others, 1990).

Mathematical properties of transformations may introduce error into the interpretation process if these properties are not properly addressed. When using logarithmic and exponential transformations, one must realize that statistical values such as the mean and standard deviation need to be calculated within the transformed data and then translated to original units (Helsel and Hirsch, 1992). Specifically, parametric statistics such as the mean and standard deviation cannot be transformed and subsequently used to estimate population characteristics in linear space. For example, the lognormal suspended-solids EMC population from the I-794 Milwaukee Wisconsin site described by Driscoll and others (1990) has a geometric mean of about 140 and a 95-percent confidence interval from about 40 to about 510. The standard deviation and COV (transformed from logarithmic space), however, are only about 1.9 and 1.3, respectively. It is apparent, therefore, that confidence-interval values must be calculated in log space and then transformed. Nonparametric statistics such as the median and quartiles, however, may be taken directly from the original or transformed data set because these statistics are associated with individual values in the data set.
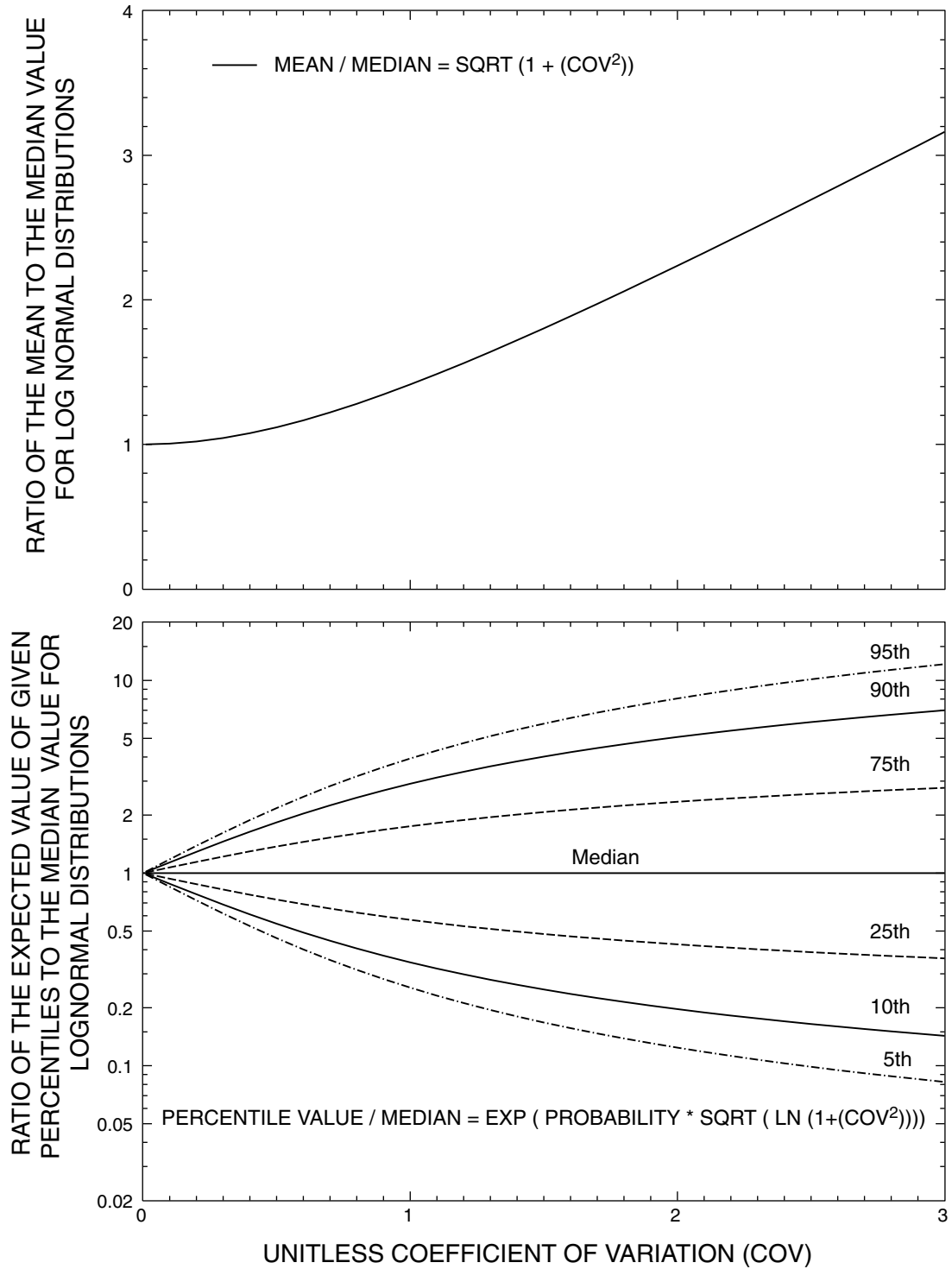
Figure 6. Properties of the lognormal (natural logarithm) distribution as a function of the population coefficient of variation (COV) modified from Athayde and others (1983).

Mathematical artifacts introduced by use of log transformation in regression analysis may also affect the predictive ability of resultant models. Transforming an unadjusted log-regression equation back into linear space (the original units, for example, milligrams per liter of suspended sediment) provides a median (the geometric mean) estimate, not a mean estimate of concentrations and (or) loads (Helsel and Hirsch, 1992). Ferguson (1986) summarized the problem in the context of estimating river loads from flow and concentration data and proposed a method for a bias correction factor (BCF). The BCF is a multiplicative term included in regression models that are formulated in logarithmic space and then transformed into original units; it is designed to prevent underestimation of concentrations or loads as an artifact of transformation. Further research indicates that several methods may be used to estimate a BCF and that a nonparametric method developed by Duan (1983) generally provides reasonable estimates for a BCF that is not affected by the structure of the data (Driver and Tasker, 1990; Helsel and Hirsch, 1992). Driver and Tasker (1990) used the Duan (1983) method to estimate BCF for storm-runoff loads, volumes, and selected constituent concentrations from NURP data and calculated BCFs that ranged from about 1.1 to 2.8 for their runoff models.

In the construction of regression models, the Box-Cox transformation (Box and Cox, 1964) can be used to stabilize variance and correct for nonnormality of a strictly positive response variable. John and Draper (1980) define a transformation that can be used when the response variable is not strictly positive. Aranda-Ordaz (1981) and Guerrero and Johnson (1982) define transformations that can be used when the response is binary (0 or 1) or a proportion between 0 and 1. For multiple regression models, a graphical device that can be helpful in deciding on a transformation for a predictor is a partial residual plot (Larsen and McCleary, 1972) in which the partial residuals (computed by taking the difference between the observed value and the components of the predicted value that exclude the variable of interest) are plotted against the predictor of interest. If the plot looks linear, no transformation is needed. If the plot shows some curvature, a transformation may be helpful. The practice of transformation optimization (finding the perfect root or power for transformation), however, is not encouraged because it is never known how well a sample represents the underlying population, and a generalized transformation that works reasonably well for all data of interest is better than multiple, slightly different transformations for each data set (Helsel and Hirsch, 1992). Alternatively, one may use other methods such as White's Specification Test, which is done by regressing squared residuals on the predictor variables and cross-products of the predictors. In this test, a significant regression implies that the specification is wrong or that heteroscedasticity of the residuals is related to the predictors (White, 1980).

REGRESSION ANALYSIS

Regression analysis is an accepted method for interpretation of water resources data and for prediction of current or future conditions at sites with characteristics that fit the input data model. The FHWA, State departments of transportation, and watershed managers need models to interpret data; predict runoff volumes, concentrations, and loads; and predict potential effects of runoff on receiving waters at sites for which data do not exist. In the following discussions, it is assumed that the purpose of the regional regression is for interpretation of data and for prediction of future responses, including possible extrapolation (that is, prediction outside the range of the sample data). If the regression model is not a reasonable representation of reality, then extrapolation could be erroneous. The purpose of the regression analysis will influence how predictors are selected for the regression model. Therefore, in selecting predictors for a possible regression model, one should choose variables that have a physical basis for explaining variations in the response, and the final regression coefficients should have a logical algebraic sign.

Historically, regression analysis was used for interpretation and prediction in several studies that represent the primary efforts for regional and (or) national

characterization of highway runoff in the United States (table 2). To date, five studies (Kobriger and others, 1981; Chui and others, 1982; Kerri and others, 1985; Driscoll and others, 1990; Thompson and others, 1996) represent regional analysis of highway-runoff flow and quality based on data collected during the 1970's and early 1980's. Additionally, Young and others (1996) identified the urban-runoff regression equations developed by Driver and Tasker (1990)—identified as "the USGS method"—as applicable for estimating highway-runoff quantity and quality (table 2). The USGS method was developed from NURP data collected throughout the United States during the 1970's and early 1980s, but it does not include highway runoff as a specific land use. More recently, Irish and others (1996; 1998) interpreted data collected in the 1990's using regression analysis, but they were clear that these equations and related interpretations may be valid only in the local Austin, Texas, area. Typically, for all the models featured in table 2, runoff data from a substantial number of storms at a number of sites were compared to predict runoff coefficients, stormflow volume, constituent EMCs, storm loads, and (or) annual loads from a number of explanatory variables. These response variables are predicted from runoff constituents, hydrologic variables, highway-design features, land-use characteristics, and climate (table 2). Tasker and Driver (1988) used both ordinary and generalized least-squares regression methods. The highway studies featured in table 2, however, were limited to ordinary least-squares regression analysis.

The Analytical Process

Regression analysis is designed to provide an estimate of the average response of a system as it relates to variation in one or more known variables. Regression equations are often derived by use of computer programs that calculate the regression parameters and provide an estimate of the correlation coefficient (R^2— the proportion of the variability in the dependent variable explained by the predictor variable) without an investigation of the validity of the selected model. When this exercise yields a correlation coefficient that is close to 1, then it is often assumed that a good regression model has been selected. Many factors may produce high but invalid correlation coefficients. Regression analysis, therefore, should include visual analysis of scatterplots, examination of the regression equation, evaluation of the method design assumptions, and regression diagnostics (Helsel and Hirsch, 1992).

Examination of scatterplots is necessary to examine the relations between predictor and response variables and to assess response variability (Helsel and Hirsch, 1992). Scatterplots of interest include graphs of the response as a function of each predictor variable and graphs of the residuals as a function of predictor variables. For most regression methods, one must ensure that a linear relation exists between predictor and response variables. If relations are not linear, other, more linear predictors may be chosen; or as previously discussed, transformation methods may be used to increase the linearity of relations between variables. For some regression methods, one must determine whether the variability in the response variables is also a function of the magnitude of the predictor. Transformations are also used to reduce or eliminate problems of nonconstant variance. Partial residual plots (described in Appendix 1*A*) and the White Specification Test (White, 1980) are often useful in determining the appropriate transformation on the basis of the structure of residuals.

Often, it is also prudent to examine the effect of seasonality on highway- and urban-runoff data by use of scatterplots of each variable of interest and the regression residuals during the monitoring period. If seasonality exists, then explaining and quantifying this factor may increase the linearity in response to other predictors by removing seasonality from the response variable. Methods for quantifying seasonality are described in Appendix 1*B*.

Examination of the regression equation ensures that the model is logical, useful, and quantitative (Helsel and Hirsch, 1992). In a logical regression model, the coefficients will have a sign and magnitude that can be explained by a reasonable scientific hypothesis. When explanatory variables in a multiple regression equation covary, this multicollinearity will cause some predictors to have illogical values and (or) sign (Helsel and Hirsch, 1992). Methods for analysis of multicollinearity are described in Appendix 1*C*.

Regression statistics also provide information about the suitability of the model. Examination of the regression equation includes interpretation of the correlation coefficient to determine whether the resulting equation explains much of the variance in the data.

Table 2. Documented metadata for selected reports that document highway-runoff regression analysis

[**A**, area; **ADP**, antecedent dry period duration; **ADT**, average daily traffic; **Al**, aluminum; **AL**, annual loads; **AP**, annual precipitation; **ATC**, antecedent dry period traffic count; **As**, arsenic; **B**, boron; **BODx**, biochemical oxygen demand; **Cd**, cadmium; **Cl**, chloride; **COD**, chemical oxygen demand; **Cr**, chromium; **Cu**, copper; **D**, storm duration; **DQ**, discussed qualitatively; **DP**, dissolved phosphorus; **DS**, dissolved solids; **EMC**, event mean concentration; **Fe**, iron; **FR**, filterable residue (similar to TSS); **GLS**, generalized least squares regression; **Hg**, mercury; **HT**, highway type (indicating the degree of urbanization and (or) impervious area); **I**, storm intensity (flow divided by duration); **IA**, impervious area; **OG**, oil and grease; **OLS**, ordinary least squares regression; **LR**, TSS loading rate by climate; **LU**, land use; **MJT**, mean January temperature; **N**, total nitrogen; **Na**, sodium; **ND**, not documented; **NFR**, nonfilterable residue (dissolved solids); **NOx**, nitrate and (or) nitrite; **Ni**, nickel; **P**, precipitation volume; **PA**, pollution accumulation rate; **Pb**, lead; **PD**, population density; **PDUR**, preceding storm duration; **PI**, preceding storm intensity; **PQ**, preceding stormflow; **Q**, stormflow; **QD**, stormflow duration; **RC**, runoff coefficient; **SL**, storm loads; **SO4**, sulfate; **SS**, suspendid solids; **TDS**, total dissolved solids; **TKN**, total kjeldahl nitrogen; **TOC**, total organic carbon; **TP**, total phosphate; **TR**, total residue (similar to total solids); **TS**, total solids; **TSS**, total suspended solids; **TVS**, total volatile solids; **VDS**, vehicles during storm; **VSS**, volatile suspended solids; **Zn**, zinc; *, indicates that different equations were developed for different categories or a unified equation was developed with this variable as a nominal variable]

Reference	Year of data collection	Location/ environmental setting	Seasonality	Number of samples	Number of sites	Model	Output
Chui and others, 1982	1979–81	Washingon: eastern (semi-arid) and western (wet)	Winter sanding events considered	500	9	OLS	SL, AL
Driscoll and others, 1990	1976–84	National: 8 sites in Washington, 3 sites each in California and Wisconsin, 2 sites each in Florida, Minnesota, Pennsylvania, and 1 site each in Arkansas, Colorado, North Carolina, and Tennessee	Separated storms into snowmelt and nonsnowmelt events	Used site median EMCs from about 900 storms	24	OLS	Q, RC, EMC
Driver and Tasker, 1990	1977–83	30 urban areas nationwide in 3 regions designated by precipitation statistics	Developed equations for seasonal and annual loads	2,813	173	OLS GLS	Q,EMC, SL,AL
Irish and others, 1998	1993–95	Austin, Texas (semi-arid)	Date of storm and temperature were not significant	58	1	OLS	SL
Kerri and others, 1985	1975–81	California (arid to semi-arid)	ND	ND	3	OLS	SL
Kobriger and others, 1981	1976–77	National: 3 sites in Wisconsin (humid), and 1 site each in Pennsylvania (humid), Tennessee (humid), and Colorado (arid)	Deicing mentioned but not quantified	159	6	OLS	Q, QD, SL
Thompson and others, 1996	1976–83	Minnesota	Classified storms as rainfall, snowmelt, or mixed	416	4	OLS	Q, EMC

Table 2. Documented metadata for selected reports that document highway-runoff regression analysis—*Continued*

| Reference | Response variables | | Predictor variables | | | | | | Uncertainty of estimates |
	Runoff constituents	Hydrologic variables	Runoff constituents	Hydrologic variables	Highway characteristics	Land use	Climate	Other	
Chui and others, 1982	TSS			RC	VDS		LR*		Mean: ~200% Range: 0.3–1,650%
	COD, Cu, NOx, Pb, TKN, TP, TOC, VSS, Zn		TSS		ADT				ND
Driscoll and others, 1990		Q, RC		P				QD	Yes
		RC			Percent impervious				Yes
	BODx, Cd, Cl, COD, Cr, Cu, Fe, Hg, NOx, OG, Pb, TOC, TP, TKN, TS, TSS, VSS, Zn				ADT, HT*				
Driver and Tasker, 1990		Q		P		IA	R*, AP	A	Yes
	Cd, Cu, COD, DP, DS, N, Pb, SS, TKN, TP, Zn			P, D, I		A, IA, LU	R*, MJT		Yes
Irish and others, 1998	BODx, COD, Cu, Fe, NOx, OG, Pb, TP, TSS, VSS, Zn			ADP, D, I, PDUR, PI, PQ, Q	VDS, ATC				Yes
Kerri and others, 1985	COD, FR, Pb, TKN, Zn				VDS				Yes
	COD, NFR, Zn				TR				Yes
Kobriger and others, 1981		Q		ADP, P	HT*				DQ
		QD		D, ADP*	HT*				DQ
	PA				ADT				DQ
	TS			I	HT*				DQ
	BODx, Cd, Cl, Cr, Cu, COD, Fe, Hg, Pb, TKN, TOC, TP, TSS, TVS		TS		HT*				DQ
Thompson and others, 1996		Q		P					
	Al, As, BODx, Cd, Cl, COD, Cr, Cu, Fe, Hg, N, Na, Ni, NOx, Pb, SO4, TKN, TP, Zn		TDS, TOC, TSS, TVS		ADT				Examine potential leverage of outliers
	TDS, TOC, TSS, TVS			ADP, D, I, P, Q	ADT				ND

Table 2. Documented metadata for selected reports that document highway-runoff regression analysis—*Continued*

Reference	Comments
Chui and others, 1982	The regression equations provided adequate estimates of the central tendency of all storm loads, but were not accurate for predicting individual storm loads.
Driscoll and others, 1990	Regression analysis was used to examine factors that influence highway runoff characteristics. It was determined that there were not enough sites with consistent explanatory variables to quantitatively explore the effects of climate, atmospheric deposition, configuration, pavement (composition, quantity, or condition), design geometrics, right-of-way characteristics, drainage features, vehicle characteristics, maintenance practices, regulations, or surrounding land use characteristics.
Driver and Tasker, 1990	Flows and loads during individual storms, seasons and years were predicted for a number of constituents as a planning tool rather than for interpretation of cause and effect relations. Explanatory variables that would be readily available for planning purposes on a regional scale were used. Highways, as an individual land use, were not included in the analysis.
Irish and others, 1998	Regression analysis indicated that over 90 percent of the variation observed for most constituent loads may have been explained by in-storm-, antecedent dry period-, and preceding storm-variables. Date and time of storm, temperature, wind speed, and wind direction were not statistically significant predictors. Traffic mix, surrounding land use, curb height, guard-rail height, and maintenance activities could not be evaluated.
Kerri and others, 1985	ADP and ATC showed no statistical significance. B, Cd, NOx, OP, OG, TP, and TR cannot be estimated from explanatory variables
Kobriger and others, 1981	Sites were classified into 3 groups for regression: Type I—urban elevated bridge deck; Type II—curbed highways; and Type III—rural highways with flush shoulders. Equations were developed to estimate runoff volume, runoff duration, constituent accumulation during an ADP, pollution washoff, and constituent loadings (as a function of TS).
Thompson and others, 1996	Surrogate parameters were established to estimate most constituents. Also, models to predict surrogate parameters were developed. Additional data sets that were not used in the formulation of the model, were used to test and verify applicability of the models.

There is no general rule of thumb to determine a minimum acceptable correlation coefficient, but the risk of posing the wrong model must be considered in relation to the ability of the model to explain variance (Helsel and Hirsch, 1992). The *t*-ratio (the value of the estimate divided by the standard error of the estimate) also is a useful statistic. The *t*-ratio will usually indicate the significance of each coefficient so that the analyst can ensure that an apparent relation did not arise by chance when there is no real linear relationship. For OLS regression, *t*-ratio statistics that have an absolute value greater than 2 are generally considered to indicate a statistically significant non-zero relation between individual predictors and response variables (Helsel and Hirsch, 1992).

Regression diagnostics include methods designed to determine whether the equation posed as a model is dominated by a few outliers in the data set. An analyst can use regression diagnostic methods to find influential observations and study their effects. Examination of a scatterplot of the residuals will often identify the effect of outliers when there is only one predictor variable, but influential outliers are more difficult to recognize in plots from multiple regression analysis. Typical problems that affect the validity of regression models are curvature, outliers, and high-leverage points. Outliers, observations (or a subset of observations) that appear to be inconsistent with the remainder of that set of data, are fairly common in hydrologic data (Hirsch and others, 1993). Outliers should be checked for possible gross errors in measurement or mistakes in recording the observations, but rejecting them out of hand is not a prudent practice. Regression diagnostics are discussed further in Appendix 1*D*.

Linear Regression Methods

As previously discussed, water-resources data—and in particular, populations of urban- and highway-stormwater data—have statistical properties that require special treatment. Predictive modeling and interpretation of cause and effect relations of runoff quality and quantity may benefit from use of techniques applied in other water-resources studies. Some of the techniques and practical alternative methods for dealing with the realities of hydrologic data are discussed in the following sections. Applicable methods, classification of appropriate response and predictor variables, and general purpose and assumptions of each method are listed in table 3 with a reference to appendix that describe some mathematical details for each of the methods discussed. This table, which distinguishes among several statistical methods discussed herein on the basis of the type of data involved, can serve as a rough guide to help the researcher choose a method for study.

A method often used for predicting water quantity and quality is ordinary-least-squares (OLS) regression (Haith, 1976; Lystrom and others, 1978; Peters, 1984; Driver and Tasker, 1990; Irish and others, 1996; Jordan and others, 1997). The OLS model, however, requires several restrictive assumptions about the parameters and errors in the model, which are often not valid for hydrologic data. To fully implement OLS regression, one must demonstrate that the response variable (or the transformed values) is linearly related to predictors (or the transformed values), the data used to fit the model are representative of the population of interest, the variance of the residuals is constant, the residuals are independent, and the residuals are normally distributed (Helsel and Hirsch, 1992). Normally distributed errors are required, even with large samples, for the determination of prediction intervals - although certain empirical methods allow one to generate robust prediction intervals from OLS-determined residuals. Appendix 2*A* more fully describes the OLS regression model.

In some cases, transformations either may not adequately transform the error structure to fit the required distributional assumptions or may be undesirable because of the possible transformation artifacts (Helsel and Hirsch, 1992). Nonparametric regression provides a distribution-free alternative to OLS regression that does not require errors to be normally distributed. Nonparametric regression can refer to models with a prespecified functional form but with an unspecified error distribution or models with neither a prespecified functional form nor error distribution. Appendix 2*B* contains more details.

Stormwater data often include outliers, which may or may not be meaningful. Although it is often imprudent to remove outliers, they often exert a leverage that will limit the accuracy and precision of regression results for the bulk of data included in the regression model. Robust regression procedures deal with outliers by reducing their influence without necessarily rejecting them entirely from the analysis. In one sense, regression diagnostics and robust regression

Table 3. General guide to regression methods

Method	Response classification	Predictor classification	General purpose and assumptions	Appendix
Ordinary least-squares regression	Continuous	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Errors are independent and identically distributed with no outliers. Normality of errors is required for hypothesis testing.	2A
Nonparametric regression	Continuous	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Error distribution unspecified. Functional form may or may not be specified. Useful when errors are not approximately normally distributed.	2B
Robust regression	Continuous	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Useful for detecting outliers and highly influential observations. Fits main portion of data, giving outliers little or no weight.	2C
Generalized least-squares regression	Continuous	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Errors can be correlated and variances of errors may be different. Useful when observations of response variable are not independent or not measured with equal accuracy.	2D
Tobit regression	Part continuous, part nominal	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Useful when response variable is censored below a detection limit.	2E
Logistic regression	Nominal	Usually continuous but nominal can be used in addition	Predicts probability of response being in one category or another.	2F
Contingency tables	Nominal	Nominal or ordinal groups	Describes the relation between nominal response and nominal or ordinal predictors.	2G
Ridge regression	Continuous	Usually continuous but nominal can be used in addition	Describes the relation between response and predictors. Useful when predictors exhibit high multicollinearity. Regression coefficients are biased.	2H
SPARROW	Continuous	Continuous	Nonlinear regression method to predict water quality for a stream reach based on spatially referenced predictors. The predictors are a function of the point and nonpoint sources and their location relative to the stream reach.	3A
Artificial neural networks	Continuous or nominal	Continuous or nominal	Flexible nonlinear nonparametric model for prediction. Any underlying model or functional relations may be impossible to extract. Data-in / predictions out black box.	3B

have the same goal of detecting outliers, but they approach the problem from different ends. Diagnostics use the classical fit of data to detect outliers and influential observations, whereas robust regression fits most of the data and detects outliers by their large residuals from the robust model. Robust regression techniques are discussed further in Appendix 2*C*. Some of the nonparametric methods discussed in Appendix 2*B* (Kendall-Theil and LOWESS smooth) also can be resistant to outliers (Helsel and Hirsch, 1992).

In analysis of stormwater quality, observations may not be independent in time and space. For example, Irish and others (1998) noted the effect of the volume, intensity, and duration of storm rainfall characteristics on the water quality measured during the next storm event. In using OLS regression, one assumes that observed values of the response variable are independent, resulting in independent residuals. In cases where this assumption is not approximately true, estimated generalized least-squares regression (GLS) can be used if the dependence of the residuals can be

estimated from the data. For example, Tasker and Driver (1988) and Tasker and Raines (1995) show that when the observed response in a regional regression to estimate mean annual loads is obtained from at-site rainfall-load models, responses may have nonconstant errors and correlated errors. Appendix 2*D* describes one estimated GLS regression approach.

As previously mentioned, concentrations of water-quality constituents below one or more detection limits in water-resources data sets are not uncommon. These values are considered as censored values for statistical analysis. Concentrations reported as less than a detection limit censor the data at the limit of detection and relegate all values below the limit to the nominal scale. When only a few observations are in the censored range, either fabricating values at or below the censoring threshold or ignoring the values in the censored range has been done but generally, these practices are not acceptable. When the response variable is moderately censored (below 20 percent censoring), Hirsch and others,(1993) recommend the Kendall-Theil robust regression method (Appendix 2*C*) or Tobit regression (Appendix 2*E*). When the level of censoring exceeds about 20 percent, logistic regression (Appendix 2*F*) or contingency tables (Appendix 2*G*) are recommended (Hirsch and others,1993).

As discussed, multicollinearity will affect the results of regression analysis, causing some predictors to have a regression coefficient with large standard errors as an artifact of regression-equation optimization. Large standard error may result in coefficient estimates having the wrong sign or unreasonable values, and it generally results in coefficients being insignificant. When using OLS or GLS regression, one must often eliminate some potentially valuable and logical predictors from a regression model. A method for dealing with multicollinearity without predictor elimination is ridge regression (Hoerl and Kennard, 1970). A successful ridge regression analysis (Appendix 2*H*) produces slightly biased regressor coefficients with smaller variance. Thus, one may trade absence of bias for a stable set of regression coefficients in the presence of multicollinearity.

Nonlinear Regression Methods

In the previous section, the structure of the regression models was assumed to be linear in the predictor coefficients. In some areas of hydrology, the assumption of linearity may represent a distortion of the physical process being modeled. In such cases, a more theoretically based nonlinear model may be appropriate. In other cases, the analyst may wish to relax the linearity constraint for a more flexible model without knowledge of the form of the model.

Many regional regression studies relating water quality to basin attributes treat contamination sources as homogeneously distributed throughout the watershed in defining the predictors (basin attributes) (Lystrom, and others1978; Peters, 1984; Driver and Tasker, 1990). This treatment limits the usefulness of the models because it fails to account for spatial differences between sources and the water-quality monitoring points within a watershed. For example, consider the predictor "urban land use (in percent)" for two watersheds both with 10 percent urban land use. The urban land in one is in the lower part of the basin, immediately upstream from the water-quality monitoring point. In the other watershed, the urban land is in the upper part of the basin, far from the monitoring point. The simple predictor of percent urban land use fails to account for loss of contaminant mass during instream and overland transport, a factor that may be substantially different between the two watersheds.

Smith and others (1997) deal with this problem by developing water-quality predictors for point and nonpoint sources as functions of both river reach and land-surface attributes, as well as by considering rates of material transport. Predictor formulas describe the transport of contaminant mass from source to the end of a reach. This innovative technique, called SPARROW (SPAtially Referenced Regressions On Watershed attributes), has the potential for greatly improving the usefulness of regional regression models for water quality because it is able to include specific sources of contaminants and their location relative to the stream reach. It includes substantial refinements of

a prototype method described in Smith and others (1993). Appendix 3*A* provides a more complete description of the method.

Artificial neural networks (ANN) include a class of flexible nonlinear models that can be used, like regression models, to predict responses from a set of predictors. Lingras and Adamo (1996) use ANN to estimate average and peak traffic volumes on the basis of road classes and short-duration counts. Zhang and Stanley (1997) and Thirumalaiah and Deo (1998) use ANN to forecast water quality and river stage on the basis of previous time steps, respectively. Artificial neural networks attempt to simulate the manner in which humans think (Hertz, Krogh, and Palmer, 1991). An ANN is composed of simple processing units, called neurons, arranged in layers. Each unit receives input from other units and converts the input to a single output, which it sends to other units. The conversion takes place in two stages: first, a net input is computed as a weighted sum of inputs, then an activity function transforms the net input into an output. The flexibility of ANN comes from one's being able to specify multiple layers of neurons with nonlinear activity functions and alternative methods for computing the net input. Observed values of predictors (inputs) and responses (targets) are used to "train" the ANN by iteratively adjusting the weights used by the neurons to produce output so that the sum of squared differences between output and target data is small.

ANN can over train (fit the observed data well, but not predict well for new data). For this reason, one should always set aside a portion of the observed data as a test data set that is not involved in any way with training the ANN, so that the ANN can be tested for predictive ability on new data. Just as in linear regression analysis, omission of important variables or inclusion of unimportant variables can be a problem. Artificial neural networks are data-in/predictions-out black boxes. Any underlying model or functional relation may be impossible to extract from the network. Appendix 3*B* gives more details about ANNs.

UNCERTAINTY, QUALITY ASSURANCE, AND QUALITY CONTROL

Uncertainty is an important part of any decision-making process. Success of a water-quality interpretive model depends on uncertain future meteorological, demographic, political, and technical conditions, all of which may affect future costs and benefits. In order to deal with problems of external uncertainty, the analyst first needs to know the severity of the statistical uncertainty inherent in the methods used to predict water quality. Statistical models need to be based on information that is meaningful, representative, complete, precise, accurate, and comparable to be deemed valid, up to date, and technically supportable. If sensitivity analyses reveal too much uncertainty in the predictions, new data and new methods may be needed, or safety factors based on prediction-interval estimates may be used. These criteria will also determine whether interpretations based on the model will be admissible as legal evidence.

Statisticians must deal with expected and unexpected uncertainties that can potentially affect interpretations made from a given data set. Expected uncertainty arises because many of the factors that affect the process are unknown or cannot be known with certainty. Statisticians can often address expected uncertainties, but these efforts are usually based on the assumption that data are collected and recorded correctly, thereby minimizing bias. Statistical measures of uncertainty are good for determining the level of noise (variability) in a data set, but they cannot detect bias without a population of "true" values with which to test the hypothesis. Unexpected uncertainty can arise from faulty computer programs, faulty application of inappropriate statistical techniques, and faulty data sets. Classical statistical measures cannot detect problems caused by unexpected uncertainty because statistical analysis is done under the assumption that the calculations are done correctly, that the correct model has been selected for the data, and that the data are representative of the environmental system under study. Thorough documentation, quality assurance, and quality

control are necessary to ensure that bias is not introduced by the errors in the computer programs used for analysis, in the modeling effort, and in the data used for analysis. Prediction errors arise from natural heterogeneity, measurement errors, and structural differences between the real world and the methods used for predictions; therefore, quality-assurance and quality-control programs must be designed to quantify these sources of uncertainty. Quality assurance and quality control are necessary throughout any study-from design through data entry and interpretation (Jones, 1999). Increasingly legal and financial liability is driving modelers to implement rigorous quality-assurance and quality-control procedures at all stages of a modeling project (Van der Heijde, 1990).

Benchmarking of Analytical Tools

Results generated by the complex computer programs currently used for statistical analysis often are assumed to be correct because it is expected that the software companies have thoroughly tested their computer code under a number of different conditions (Landwehr and Tasker, 1999). If statistics that look reasonable but are, in fact, grossly incorrect are computed, this dangerous error is liable to remain undetected until applied to a real-world problem where the model noticeably fails. To prevent this situation, one can benchmark the statistical software to assess its reliability.

Benchmarking consists of applying a suite of statistical analyses to various standard data sets for which the values of the statistics are known with great precision and assessing whether the resulting values are in conformance. A discussion and review of how to do such assessments can be found in Sawizki (1994a) and McCullough (1998), who use the Statistical Reference Datasets (StRDs) recently published by the National Institute of Standards and Technology (1998). Wilkinson (1985) also proposed a collection of simple tests designed to uncover common flaws in statistical programs, including an example in which the variables are collinear and the difference in magnitude between variables was extreme but the magnitude of the observations for each variable was not unlike that found among common statistical problems.

Benchmarking studies to provide quality assurance and quality control to verify the operation of statistical software packages sound esoteric; but when such studies have been done, real problems have been discovered. For example, Sawitzki (1994b) reported on a joint effort by members of two working groups ("Computational Statistics of the International Biometrical Society" and "Statistical Analysis Systems" of the "GMDS," Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie) to apply the Wilkinson tests to nine data-analysis systems, including some running on multiple platforms. The group demonstrated performance difficulties with each system, even between platform implementations of the same package. More recently Landwehr and Tasker (1999) completed a benchmark study of commercial statistical-software packages commonly used by the USGS and found that several analysis packages had more difficulty in providing computationally precise and (or) correct values than did others, and some were cumbersome to use in obtaining specific statistics. Therefore, quality assurance and quality control for interpretive efforts that make use of existing software should include documentation of the commercial software (including version and platform), documentation of benchmarking efforts, as well as scientific and technical reviews of the software selection and benchmarking efforts.

In general, the analyst should be aware of the specific question or questions to be answered by the analysis, as well as how a specific package is computing the answer it provides (Landwehr and Tasker, 1999). Quality assurance and quality control (using tools such as the ANASTY data set) is especially important when an analyst will write computer code to implement a statistical process, even if the code is written within the interface of an existing commercial software product that has been previously benchmarked. Quality assurance and quality control for code development includes verification of the structure and coding, model validation, record keeping, software documentation, and scientific and technical reviews.

Uncertainty in Modeling Efforts

Results generated by statistical models may appear to be more certain, more precise, and more authoritative than they really are because design assumptions and results (even realistic ones) can be stated with illusive precision and seeming accuracy. Models—like scientific theories—cannot truly be validated; they can only be invalidated when an exception

or error is revealed by new data or circumstances. For example, Driscoll and others (1990) note that regression models have been criticized as poor predictors when applied beyond the original data set used to create the model. The analyst, therefore, has the responsibility to document efforts to

- examine the representativeness of data used to construct statistical models,
- assess uncertainties in models, and
- evaluate the potential predictive ability for sites not included in the construction of the model.

The application and documentation of these steps in the modeling procedure are essential for the transfer of a model to potential model users. Quality assurance and quality control in the processes used for data reduction, evaluation, and interpretation is as important as quality assurance and quality control for field and laboratory data-collection efforts (Brown and others, 1991). Quality assurance and quality control for modeling efforts include the procedural and operational framework used by an organization to ensure the technical and scientific adequacy of the tasks, and documentation thereof, to ensure that the results are fully reproducible and defensible. Modeling efforts are being used in the regulatory and legal domain, and the needs for model documentation describing the inherent uncertainty in predictions are increasing (Haan and others, 1990). Jones (1999) describes the quality assurance and technical review necessary to prepare data in a computerized database used to develop water-quality models.

A number of statistical tools are available to assess the uncertainties inherent in models during model development. The fit of a regression model is usually measured by means of the correlation coefficient, but a better measure of the model's predictive ability at unmonitored stations is the variance of prediction (Gilroy and others, 1990). This statistic is computed by estimating the mean or median variance of prediction for individual stations while using every station in the regression model. This computation is made on the basis of the assumption that the available stations are representative of the entire population of potential sites to be modeled. Irish and others (1996) note, however, that the standard error of forecasts (a measure of the spread of data points not used to formulate a regression model indicating the predictive ability of the coefficients in the model for data not included in the monitored population) is always larger than the standard error of the regression for modeled data. When the errors in a regression model are approximately normally distributed, standard errors of prediction and prediction intervals, which serve as a measure of the uncertainty in the predictions, can be computed on the basis of normal theory. Appendix 4*A* provides some details. When using methods that do not involve the assumption that errors are normal or do not specify an error distribution, the analyst may be able to use the bootstrap method (Efron and Tibshirani, 1986). The bootstrap is a simple, straightforward method for computing biases, standard deviations, and confidence intervals for almost any nonparametric problem (Efron, 1982). More details are given in Appendix 4*B*. Documenting a sensitivity analysis—determining how input parameters control model output—is necessary to indicate how uncertainty in input values will affect uncertainty in computed results. For example, Driver and Tasker (1990) provided uncertainty estimates for regional regression equations, which had coefficients of variation ranging from 0.2 to 0.65, standard errors of the estimate ranging from 79 to 145 percent, and average prediction errors ranging from -67 to 203 percent.

When developing a regional hydrologic regression for a fairly large geographic area, such as a state or several states, it is sometimes advantageous to subdivide the region into several homogeneous subregions in which the basic regression assumptions are more likely to be true than for the whole region. These regions may or may not have geographic boundaries. Geographic regions may be based on some general topographical or geological feature of the region or may be based on ecoregions (Omernik, 1995). Regions can also be defined on the basis of values of the predictors. Multivariate techniques of cluster and discriminant analysis have been used to define regions based on basin attributes (Tasker, 1982). Another possible method for dividing a large area into regions is referred to as the "region of influence method" (Tasker and Slade, 1994; Tasker and others, 1996) a method in which a unique regression equation is estimated for each site where a prediction is to be made. The regression equation is based only on data observed for sites with basin characteristics similar to the site where a prediction is to be made. Appendix 5 gives more details. Many of the characteristics that may be useful for regionalizing highway-runoff-quality data (including maps of climatic characteristics, receiving-water characteristics, and ecoregions) are provided as

Geographic Information System (GIS) coverages for the conterminous United States by Smieszek and Granato (2000). The process of regionalization needs to be documented, as well as the regional characteristics, so that practitioners using the model may properly assign any given site within an appropriate region.

Proper model application as part of a planning, design, or assessment effort also requires substantial documentation. Information that should be documented includes the statistical characteristics and design assumptions supporting application of the model. Model-performance parameters as defined in the previous paragraph should be documented in terms of an uncertainty estimate in the predictions made by the model when the model is applied. For example, Young and others (1996) apply the USGS method (Driver and Tasker, 1990) to estimate an annual load of suspended solids in a hypothetical case study to illustrate use of the method. Young and others (1996) use watershed area (one significant figure), land use by percent (one significant figure), and the impervious area (two significant figures) to compute a storm load of 397.4 kg (four significant figures). This storm load is then multiplied by the average number of storms (two significant figures) to compute an annual load of 55,636 kg (five significant figures). Young and others (1996) provide the exact results of these hypothetical computations to illustrate the method unambiguously. In reality, however, model results are expected to have an uncertainty that is compounded by use of input values with only one significant figure. Computed storm-loads for this hypothetical site are about 400 kg per storm with a 90-percent confidence interval from about 40 to about 2,000 kg per storm. Computed annual-loads for this hypothetical site are about 60,000 kg per year with a 90-percent confidence interval that ranges from about 6,000 kg to about 300,000 kg per year. Reporting the final rounded numbers and estimates of the uncertainty of the calculations provides the information necessary to evaluate the potential results from decisions based on these loads.

The analyst should also document an assessment of model suitability for site-specific conditions when models are applied. Regional regression models are designed to provide estimates for the average site with characteristics identified by predictor variables. One must determine that values of predictor variables for the site to be modeled fall within the range of data used to construct the models, and that no other distinguishing site characteristics would differentiate the site from the modeled population.

Uncertainty in Input Data

The quality of interpretations depends directly upon the quality and representativeness of available data. Statistical models are empirical models generally requiring large amounts of water-quality, land-use, and highway-related data for parameter estimation. Although a large amount of such data exist in this country, the data are in disparate databases and the comparability of this data is in doubt. De Vries and Klavers (1994) demonstrate that the reliability of modeling estimates is determined primarily by the quality of the monitoring strategy and that computation methods can be much less important than the data incorporated into a given model. Models, at best, are only as good as the uncertainty in the input data (Montgomery and Sanders, 1985). Furthermore, there is no guarantee that water-quality data, no matter how carefully collected, will be transferable to other areas and other circumstances (Sonnen, 1983). Harrop (1983) observed that the high uncertainty in highway-runoff-quality models was caused by "too much analysis being applied to too little data." Driscoll and others (1990) also noted that even with 31 sites and hundreds of monitored storms, many of the investigated factors—which theoretically should affect the quality of runoff—could not be quantitatively defined. Specifically, individual relations could not be defined because each site had a number of explanatory variables (including climate, traffic, highway-design features, and surrounding land-use characteristics) that were not held constant from site to site. In other words, it is difficult to develop meaningful models to quantitatively predict water quality from physical or chemical differences between sites unless the "noise" introduced by the sampling effort is much smaller than the "signal" produced by differences between sites. The NURP program recognized that interpretation of data would be questionable unless field programs at different sites provided consistent and sound data. Therefore, quality assurance and quality control elements were adopted as integral parts of each site/project workplan, including elements to address

potential problems with the program, field monitoring and sample collection, laboratory analysis, data management, and data analysis (Athayde and others, 1983).

Standard least-squares methods for regression analysis depend on the assumption that the predictors are known without error. Nevertheless, measurement errors are inherent in most predictors used in regression analysis of environmental data. Measurement errors in the predictors may or may not be a problem in regional regression analyses. The effect of measurement errors can be ignored with little consequence when the variance of the measurement error is very small in relation to the variance in the predictor itself. Weisburg (1980), Seber (1977), and Davies and Hutton (1975) provide more details and methods for determining whether measurement errors are small enough to ignore.

When significant measurement errors are in the data set of predictors, use of the regional least-squares regression for prediction requires the predictions to be made by means of the same methods for measuring the predictors as used in the determination of the regression coefficients. For example, consider a regional regression in which the response, S, is annual sulfate load and predictors P and T are mean annual precipitation and average annual daily traffic flow, respectively. In addition, P is estimated from a contour map of the region based on 1930–60 data, and T is estimated from a regression on a sample short period count (Erhunmwunsee, 1991). As long as predictions from the regional model are made by use of the same methods for determining P and T, the measurement errors in P and T present little problem. However, if predictions from the regional model are made using P estimated from a nearby rainfall record or from a different contour map or if T is measured by means of some method different from the short-period-count regression, then the predictions from the regional least-squares model will not be appropriate. It follows that a problem exists when the observed predictors used to estimate the regression coefficients are measured using significantly different methods with different measurement errors. For example, consider a regional regression study covering several states in which each state uses a different method to estimate average annual daily traffic flow, T. In these cases, it is necessary to adjust the regression model for errors in the predictors.

The problem with using standard least-squares methods when measurement errors are in the predictors is that the observed predictors will correlate with the regression errors, resulting in biased estimates of the regression coefficients. An alternative to standard least squares that deals with the measurement errors in predictors is the method of instrumental variables (Johnston, 1972). Instrumental variables are variables that correlate with the predictor that contains measurement error and but do not correlate with the regression errors. Johnston (1972) and Fuller (1987) describe several methods for instrumental-variable estimation; SAS/ETS procedure MODEL (SAS Institute, 1988) can be used for the computations.

Therefore, to assemble a regional or national data set, one must ensure that the methods used to define both the predictor (explanatory) and response variables are the same, or that the different methods will produce results that are neither substantially or statistically different. Furthermore, it has been demonstrated that differences in monitoring objectives of past studies will affect the suitability of available data for inclusion in a regional or national synthesis because the monitoring objectives of those studies will affect the representativeness of the selected sites when compared to the total population of existing sites (Norris and others, 1990). This is because local studies are often designed for addressing local problems rather than for national characterization (Norris and others, 1990).

Availability of reliable runoff-quality data in an electronic format is necessary to facilitate future use of and interpretation of data collected. Driscoll and others (1990) and Thompson and others (1996) indicate the substantial difficulties involved in the collection, examination, quality assurance, quality control, and (when necessary) data entry of historical runoff data in their efforts toward local, regional, and national interpretation of highway-runoff data. In comparison, Driver and Tasker (1990) were able to assemble a much larger National Urban Runoff Program (NURP) data from the USGS and the USEPA with less effort because these programs were supported by quality-assurance and quality-control measures, and the data were stored in readily available national water-quality databases. The Transportation Research Board (1997) determined that development of consistent electronic data structures is a primary research

need to standardize results of hydraulics, hydrology, and water-quality research efforts for future use. Data-collection activities, therefore, need be conducted within a framework of an established quality-assurance and quality-control program to demonstrate that data collected meet data-quality objectives and will be meaningful for interpretation of the characteristics of runoff quantity and quality (Jones, 1999). Many of the criteria for basic information, acceptable uncertainty, and quality-assurance and quality-control documentation described by Granato and others (1998) are essential for reliable interpretation of local, regional, and national runoff-quality data sets.

SUMMARY

Engineers, planners, economists, regulators, and other decision makers concerned with stormwater runoff need viable methods for the interpretation of local, regional, and national highway-runoff and urban-stormwater data. Stormwater-quality models have, historically, been used to characterize stormwater flow and quality, predict pollutant runoff loads, assess impacts on receiving waters, and determine the effectiveness of various best management practices to mitigate possible impairment of designated beneficial uses of receiving waters. Valid, current and technically defensible stormwater-runoff models are needed to interpret data collected by field studies; support existing highway and urban runoff planning processes, meet National Pollutant Discharge Elimination System requirements, and provide methods for calculation of Total Maximum Daily Loads, in a systematic and economical manner.

Historically, conceptual models, simulation models, empirical models, and statistical models of varying levels of detail, complexity, and uncertainty have been used to meet various data-quality objectives in the decision-making processes necessary for the planning, design, construction, and maintenance of highways and for other land-use applications. Water-quality simulation models attempt a detailed description of the physical processes and mechanisms by means of model parameters with a direct physical definition, and require as input a considerable degree of detail in the description of the physical system. In simulation models, parameter estimation is not as data dependent as in statistical regional water-quality-assessment models. On the other hand, empirical and statistical regional water-quality-assessment models provide a more general picture of water quality or changes in water quality over a region. Statistical regional water-quality models may also be used to estimate nonpoint-source loadings as inputs for more detailed water-quality simulation models. All these modeling techniques share one common aspect—the predictive ability of almost any model will be poor without suitable site-specific data for calibration.

An understanding of the classification of variables, the unique characteristics of water-resources data, and the concept of population structure and analysis is necessary to properly interpret the results of individual studies and to combine these results to form meaningful interpretations as part of a regional or national synthesis of stormwater quality data. Classification of variables being used to analyze data may determine which statistical methods are appropriate for data analysis. An understanding of the fundamental characteristics of water-resources data is necessary to evaluate the applicability of various statistical techniques, to interpret the results of these techniques, and to use tools and techniques which account for the unique nature of water-resources data sets. Understanding the methods and measures used to determine the population structure and analyze population characteristics also is necessary to form valid, current, and technically defensible stormwater runoff models.

Regression analysis is an accepted method for interpretation of water-resources data and for prediction of current or future conditions at sites that fit the input data model. The Federal Highway Administration, state departments of transportation, have successfully implemented regression models to interpret data; identify quantitative relations between constituents; predict runoff volumes, concentrations, loads; and predict potential effects of runoff on receiving waters at sites for which data do not exist. Regression analysis is designed to provide an estimate of the average response of a system as it relates to variation in one or more known variables. To date, highway- and urban-runoff studies have generally been limited to ordinary least squares (OLS) and generalized least squares (GLS) regression techniques. There are, however, a number of linear and nonlinear regression methods that may be appropriate for interpretation of local, regional, and national highway-runoff and urban-stormwater data when the classification of variables and the structure of the data violate the design assumptions of the OLS and (or) GLS methods.

Uncertainty is an important part of any decision-making process. Success of a water-quality interpretive model depends on uncertain future meteorological, demographic, political, and technical conditions, all of which may affect future costs and benefits. In order to deal with uncertainty problems, the analyst needs to know the severity of the statistical uncertainty of the methods used to predict water quality. Statistical models need to be based on information that is meaningful, representative, complete, precise, accurate, and comparable to be deemed valid, up to date, and technically supportable. If sensitivity analyses reveal too much uncertainty in the predictions, new data and new methods may be needed, or safety factors based on prediction interval estimates may be used. To ensure that decision makers can assess uncertainty in the analytical tools, the modeling methods, and the underlying data set, the analyst must document and communicate each of these components in an accessible format within project publications. These criteria will also determine whether interpretations based on the models developed will be admissible in a regulatory framework and (or) as legal evidence when necessary.

REFERENCES

Acreman, M.C., and Wiltshire, S.E., 1987, Identification of regions for regional flood frequency analysis: EOS, Transactions, American Geophysical Union, v. 68, no. 44, p. 512.

Aitcheson, J., and Brown, J.A.C., 1957, The lognormal distribution, with special reference to its uses in economics: Cambridge, U.K., Cambridge University Press, 176 p.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W., 1972, Robust estimates of location--Survey and advances: Princeton, N.J., Princeton University Press, 373 p.

Aranda-Ordaz, F.J., 1981, On two families of transformations to additivity for binary response data: Biometrika, v. 68, p. 357–363.

Athayde, D.N., Shelly, P.E., Driscoll, E.D., Gaboury, D., and Boyd, G., 1983, Results of the Nationwide Urban Runoff Program—volume 1—final report: U.S. Environmental Protection Agency, WH-554, 186 p.

Averett, R.C., and Schroder, L.J., 1994, A guide to the design of surface-water-quality studies: U.S. Geological Survey Open-File Report 93-105, 39 p.

Barbe, D.E., and Francis, J.C., 1995, An analysis of seasonal fecal coliform levels in the Tchefuncte River: Water Resources Bulletin, v. 31, no. 1, p. 141-146.

Bedient, P.B., and Huber, W.C., 1992, Hydrology and floodplain analysis (2d ed.): Reading, Mass., Addison-Wesley, 692 p.

Belsley, D.A., 1991, Conditioning diagnostics--Collinearity and weak data in regression: New York, John Wiley and Sons, 396 p.

Belsley, D.A., Kuh, E., and Welsch, R.E., 1980, Regression diagnostics—Identifying influential data and sources of collinearity: New York, John Wiley and Sons, 292 p.

Bicknell, B.R., Imhoff, J.C., Kittle, J.L., Jr., Donigian, A.S., and Johanson, R.C., 1993, Hydrological simulation program—FORTRAN, Users manual for Release 10: U.S. Environmental Protection Agency Report EPA/600/R-93/174, 667 p.

Birkes, D., and Dodge, Y., 1993, Alternative methods of regression: New York, John Wiley and Sons, 228 p.

Box, G.E.P., and Cox, D.R., 1964, An analysis of transformations: Journal of Royal Statistical Society, B, v. 26, p. 211–246.

Breault, R.F., and Granato, G.E., 2000, A synopsis of technical issues of concern for monitoring trace elements in highway and urban runoff: U.S. Geological Survey Open-File Report 00-422, 67 p.

Breiman, L., Friedman, J., Olshen, R., and Stone, C.J., 1984, Classification and regression trees: Belmont, Calif., Wadsworth, 358 p.

Brown, R.D., Marinenko, G., and Egan, D.E., 1991, Quality assurance as viewed by a data user, *in* Friedman, D., ed., Waste testing and quality assurance: Philadelphia, Pa., American Society for Testing and Materials, ASTM STP 1075, v. 3, p. 391–404.

Burn, D.H., 1990a, An appraisal of the "region of influence" approach to flood frequency analysis: Hydrological Sciences Journal, v. 35, p. 149–165.

Burn, D.H., 1990b, Evaluation of regional flood frequency analysis with a region of influence approach: Water Resources Research, v. 26, no. 10, p. 2257–2265.

Chui, T.W.D., Mar, B.W., and Horner, R.R., 1982, Pollutant loading model for highway runoff: Journal of the Environmental Engineering Division, American Society of Civil Engineers, v. 108, no. EE6, p. 1193–1210.

Church, P.E., Granato, G.E., and Owens, D.W., 1999, Basic requirements for collecting, documenting, and reporting precipitation and stormwater-flow measurements: U.S. Geological Survey Open File Report 99-255, 30 p.

Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots: Journal of the American Statistical Association, v. 74, p. 829–836.

Cohn, T.A., Caulder, D.L., Gilroy, E.J., Zynjuk, L.D., and Summers, R.M., 1992, The validity of a simple statistical model for estimating fluvial constituent loads--An empirical study involving nutrient loads entering Chesapeake Bay: Water Resources Research, v. 28, no. 9, p. 2353–2363.

Conover, W.J., 1980, Practical nonparametric statistics (2d ed.): New York, John Wiley and Sons, 493 p.

Cook, R.D., and Weisburg, S., 1982, Residuals and influence in regression: New York, Chapman and Hall, 230 p.

Cook, R.D., and Weisburg, S., 1994, An introduction to regression graphics: New York, John Wiley and Sons, 253 p.

Cox, D.R., and Snell, E.J., 1989, Analysis of binary data (2nd ed.): London, Chapman and Hall, 236 p.

Croarkin, Carroll, and Tobias, Paul, eds., 2000, Engineering statistics handbook: National Institute of Science and Technology, Statistical Engineering Division, accessed July 3, 2000, at URL http://www.itl.nist.gov/div898/handbook/

Davies, R.B., and Hutton, B., 1975, The effects of errors in the independent variables in linear regression: Biometrika, v. 62, no. 2, p. 383–391.

Dever, R.J., Roesner, L.A., and Aldrich, J.A., 1983, Urban highway drainage model--Surface Runoff Program user's manual and documentation: Federal Highway Administration Final Report FHWA-RD-85-001, 173 p.

De Vries, A., and Klavers, H.C., 1994, Riverine fluxes of pollutants monitoring strategy first, calculation methods second: European Water Pollution Control, v. 4, no. 2, p. 12–17.

Driscoll, E.D., Shelly, P.E., and Strecker, E.W., 1990, Pollution loadings and impacts from highway stormwater runoff--Volume III, analytical investigation and research report: Federal Highway Administration, FHWA-RD-88-008, 150 p.

Driver, N.E., and Tasker, G.D., 1990, Techniques for estimation of storm-runoff loads, volumes, and selected constituent concentrations in urban watersheds in the United States: U.S. Geological Survey Water-Supply Paper 2363, 44 p.

Duan, Naihua, 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610.

Efron, B., 1982, The jackknife, the bootstrap and other resampling plans: Philadelphia, Society for Industrial and Applied Mathematics, 92 p.

Efron, B., and Tibshirani, R., 1986, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy: Statistical Science, v. 1, no. 1, p. 54–77.

Erhunmwunsee, P.O., 1991, Estimating average annual daily traffic flow from short period counts: Institute of Transportation Engineers Journal, November, p. 23–30.

Ferguson, R.I., 1986, River loads underestimated by rating curves: Water Resources Research, v. 22, no. 1, p. 74–76.

Friedman, J., and Stuetzle, W., 1981, Projection pursuit regression: Journal of the American Statistical Association, v. 76, p. 817–823.

Fuller, W.A., 1987, Measurement error models: New York, John Wiley and Sons, 440 p.

Garbarino, J.R., and Struzeski, T.M., 1998, Methods of analysis by the USGS National Water Quality Laboratory—Determination of elements in whole-water digests using inductively coupled plasma-optical emission spectrometry and inductively coupled plasma-mass spectrometry: U.S. Geological Survey Open-File Report 98-165, 101 p.

Gillom, R.J., and Helsel, D.R., 1986, Estimation of distributional parameters for censored trace level water quality data—1. Estimation techniques: Water Resources Research, v. 22, no. 2, p. 135–146.

Gilroy, E.J., Hirsch, R.M., and Cohn, T.A., 1990, Mean square error of regression-based constituent transport estimates: Water Resources Research, v. 26, no. 9, p. 2069–2077.

Godfrey, L.G., 1988, Misspecification tests in econometrics the language multiplier principle and other approaches: New York, Cambridge University Press, Econometric Society Monographs, no. 16, 252 p.

Granato, G.E., 1996, Deicing chemicals as a source of constituents of highway runoff: Transportation Research Board, Transportation Research Record 1533, p. 50–58.

Granato, G.E., Bank, F.G., and Cazenas, P.A., 1998, Data quality objectives and criteria for basic information, acceptable uncertainty, and quality-assurance and quality-control documentation: U.S. Geological Survey Open-File Report 98-394, 17 p.

Granato, G.E., and Smith, K.P., 1999, Estimating concentrations of road-salt constituents in highway runoff from measurements of specific conductance: U.S. Geological Survey Water-Resources Investigations Report 99-4077, 22 p.

Guerrero, V.M., and Johnson, R.A., 1982, Use of the Box-Cox transformation with binary response models: Biometrika, v. 69, p. 309–314.

Haan, C.T., Solie, J.B., and Wilson, B.N., 1990, To tell the truth—Hydrologic models in court, *in* Janes, E.B., and Hotchkiss, W.R., eds., Symposium on Transferring Models to Users, Denver, Colo. Nov. 4–8, 1990, Bethesda, Md. Proceedings:American Water Resources Association, p. 337–348.

Haith, D.A., 1976, Land use and water quality in New York rivers: Journal of the Environmental Engineering Division, American Society of Civil Engineers, v. 102, no. EE1, p. 1–15.

Hall, M.J., and Hamilton, R.S., 1991, Highway runoff transport modeling, *in* Hamilton, R.S., and Harrison, R.M., eds., Highway pollution, New York, Elsevier Science, Studies in Environmental Science 22, p. 131–164.

Härdle, W., 1990, Applied nonparametric regression: Cambridge, U.K., Cambridge University Press, 333 p.

Harrop, O., 1983, Stormwater pollution from highway surfaces—A review: Middlesex Polytechnic Research and Consultancy, Urban Stormwater Pollution Research Report 6, 109 p.

Helsel, D.R., 1990, Less than obvious—Statistical treatment of data below the detection limit: Environmental Science and Technology, v. 24, no. 12, p. 1766–1774.

Helsel, D.R., 1993, Statistical analysis of water-quality data, *in* Paulson, R.W., Chase, E.B., Williams, J.S., and Moody, D.W., compilers, National Water Summary 1990–91—Hydrologic Events and Stream Water Quality: U.S. Geological Survey Water-Supply Paper 2400, p. 93–100.

Helsel, D.R., and Cohn, T.A., 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Resources Research, v. 24, no. 12, p. 1997–2004.

Helsel, D.R., and Gilliom, R.J., 1986, Estimation of distributional parameters for censored trace level water quality data--verification and applications: Water Resources Research, v. 22, no. 2, p. 147–155.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: Amsterdam, Elsevier, 522 p.

Hertz, J., Krogh, A., and Palmer, R.G., 1991, Introduction to the theory of neural computation: Reading, Mass., Addison-Wesley Publishing, 327 p.

Hirsch, R.M., Helsel, D.R., Cohn, T.A., and Gilroy, E., 1993, Statistical analysis of hydrologic data, *in* Maidment, D.R., ed), Handbook of hydrology: New York, McGraw-Hill, p. 17.1–17.55.

Hoerl, S.D., and Kennard, R.W., 1970, Ridge regression—Biased estimation for nonorthogonal problems: Technometrics, v. 12, p. 55–67.

Hoos, A.B., and Sisolak, J.K., 1993, Procedures for adjusting regional regression models of urban-runoff quality using local data: U.S. Geological Survey Open-File Report 93-39, 39 p.

Huber, P.J., 1973, Robust regression—Asymtotics, conjectures, and Monte Carlo: Annals of Statistics, v. 1, p. 799–821.

Huber, W.C., 1993, Contaminant transport in surface water, *in* Maidment, D.R., ed., Handbook of hydrology: New York, McGraw-Hill, p. 14.1–14.50.

Huber, W.C., and Dickinson, R.E., 1988, Storm water management model version 4, part A—Users manual: U.S. Environmental Protection Agency Report EPA/600/3-88/001a, 569 p.

Huber, W.C., Heaney, J.P., Smolenyak, K.J., and Aggidis, D.A., 1979, Urban rainfall-runoff-quality database—Update with statistical analysis: U.S. Environmental Protection Agency Final Report EPA 600/8-79-004, 283 p.

Ichiki, A., Yamada, K., and Ohnishi, T., 1996, Prediction of runoff pollutant load considering characteristics of river basin: Water Science and Technology, v. 33, no. 4-5, p. 117–126.

Irish, L.B., Jr., Barrett, M.E., Malina, J.F., Jr., and Charbeneau, R.J., 1998, Use of regression models for analyzing highway storm-water loads: Journal of Environmental Engineering, v. 124, no. 10, p. 987–993.

Irish, L.B., Lesso, W.G., Barrett, M.E., Malina, J.F., Jr., Charbeneau, R.J., and Ward, G.H., 1996, An evaluation of the factors affecting the quality of highway runoff in the Austin, Texas, area: Federal Highway Administration, Texas State Department of Transportation Interim Report FHWA/TX-96/1943-5, 246 p.

John, J.A., and Draper, N.R., 1980, An alternative family of transformation: Applied Statistics, v. 29, p. 190–197.

Johnston, J., 1972, Econometric methods (2d ed.): New York, McGraw-Hill, 437 p.

Jones, B.E., 1999, Principles and practices for quality assurance and quality control: U.S. Geological Survey Open File Report 98-636, 24 p.

Jordan, T.E., Correll, D.L., and Weller, D.E., 1997, Effect of agriculture on discharges of nutrients from coastal plain watersheds of Chesapeake Bay: Journal of Environmental Quality, v. 26, p. 836–848.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H, Lee, T., 1985, The theory and practice of econometrics (2d ed.): New York, John Wiley and Sons, 1019 p.

Kerri, K.D., Racin, J.A., and Howell, R.B., 1985, Forecasting pollutant loads from highway runoff: Washington, D.C., Transportation Research Board, Transportation Research Record 1017, p. 39–46.

Kleinbaum, D.G., and Kupper, L.L., 1978, Applied regression analysis and other multivariable methods: North Scituate, Mass., Duxbury Press, 556 p.

Kobriger, N.P., Meinholz, T.L., Gupta, M.K., and Agnew, R.W., 1981, Constituents of highway runoff--Volume III, predictive procedure for determining pollution characteristics in highway runoff: Federal Highway Administration Final Report FHWA/RD-81/044, 205 p.

Landwehr, J.M., and Tasker, G.D., 1999, Notes on numerical reliability of several statistical analysis programs: U.S. Geological Survey Open-File Report 99-95, 19 p.

Larsen, W.A., and McCleary, S.J., 1972, The use of partial residual plots in regression: Technometrics, v. 14, p. 781–790.

Lingras, P., and Adamo, M., 1996, Average and peak traffic volumes—Neural nets, regression, factor approaches: Journal of Computing in Civil Engineering, v. 10, no. 4, p. 300–306.

Liu, S., Yen, S.T., and Kolpin, D.W., 1996, Atrazine concentrations in near-surface aquifers—A censored regression approach: Journal of Environmental Quality, v. 25, p. 992–999.

Lystrom, D.J., Rinella, F.A., Rickert, D.A., and Zimmermann, L., 1978, Regional analysis of the effects of land use on stream-water quality, methodology and application in the Susquehanna River Basin, Pennsylvania and New York: U.S. Geological Survey Water-Resources Investigations 78-12, 60 p.

McCullough, B., 1998, Assessing the Reliability of Statistical Software—Part I: American Statistician, v. 52, no. 4, p. 358–366.

Mclaughlin, M.P., 1999, Appendix A—A compendium of common probability distributions, in Regress+ online site accessed on July 3, 2000, at URL http://www.geocities.com/~mikemclaughlin/math stat/Dists/Compendium.html.

Mills, W.B., Porcella, B.B., Ungs, M.J., Gherini, S.A., Summers, K.V., Lingsung, M., Rupp, G.L., Bowie, L.G., and Haith, D.A., 1985, Water quality assessment--A screening procedure for toxic and conventional pollutants in surface and ground waters Part 1: U.S. Environmental Protection Agency Report EPA/600/6-85/002a, 638 p.

Montgomery, D.C., and Peck, E.A., 1982, Introduction of linear regression analysis: New York, John Wiley and Sons, 504 p.

Montgomery, R.H., and Sanders, T.G., 1985, Uncertainty in water quality data, in El-Shaarawi, A.H., and Kwiatkowski, R.E., eds., Statistical Aspects of Water Quality Monitoring, Proceedings of the workshop held at the Canada Center for Inland Waters, Oct. 7–10, 1985, New York, Elsevier, p. 17–29.

Myers, R.H., 1986, Classical and modern regression with applications: Boston, Mass., Duxbury Press, 359 p.

National Institute of Standards and Technology, 1998, Statistical reference datasets, accessed July 4, 2000, at URL http://www.nist.gov/itl/div898/strd/.

Norris, J.M., Hren, J., Myers, D., Chaney, T.H., and Childress, C.J.O., 1990, Water-quality data-collection activities in Colorado and Ohio, Phase III—Evaluation of existing data for use in assessing regional water-quality conditions and trends: U.S. Geological Survey Water-Supply Paper 2295-C, 46 p.

Omernik, J.M., 1995, Ecoregions—a spatial framework for environmental management, in Davis, W.S. and Thomas, P.S., eds., Biological assessment and criteria: Boca Raton, Fla., Lewis Publishers, Tools for Water Resources Planning and Decision Making, p. 49–62.

Ott, R.L., 1993, An introduction to statistical methods and data analysis (4th ed.): Belmont, Calif., Duxbury Press, 1183 p.

Peters, N.E., 1984, Evaluation of environmental factors affecting yields of major dissolved ions of streams in the United States: U.S. Geological Survey Water-Supply Paper 2228, 39 p.

SAS Institute, 1988, SAS/ETS User's Guide, Version 6: Cary, N.C., SAS Institute, 559 p.

Sarle, W.S., 1994, Neural Networks and Statistical Models: in Proceedings of the Nineteenth Annual SAS Users Group International Conference, April 3–4, 1994, Cary, NC, SAS Institute, p. 1538–1550. accessed on April 26, 2000, at URL ftp://ftp.sas.com/pub/neural/neural1.ps.

Sawitzki, G., 1994a, Testing numerical reliability of data analysis systems: Computational Statistics and Data Analysis, v. 18, p. 269–286.

Sawitzki, G., 1994b, Report on the reliability of data analysis systems: Computational Statistics and Data Analysis, v. 18, p. 289–301.

Schueler, T.R., 1987, Controlling urban runoff—A practical manual for planning and designing urban BMPs: Metropolitan Washington Council of Governments Publication 87703, 275 p.

Seber, G.A.F., 1977, Linear regression analysis: New York, John Wiley and Sons, 465 p.

Shelley, P.E., and Gaboury, D.R., 1986, Estimation of pollution from highway runoff-Initial results, in Urbonas, B., and Roesner, L.A., eds., Urban Runoff Quality--Impact and Quality Enhancement Technology: Henniker, N.H., Proceedings of an Engineering Foundation Conference, New England College, June 23–27, p. 459–473.

Shoemaker, L., Lahlou, M., Bryer, M., Kumar, D., and Kratt, K., 1997, Compendium of tools for watershed assessment and TMDL development: U.S. Environmental Protection Agency, Office of Water, EPA 841-B-97-006, 221 p.

Smieszek, T.W., and Granato, G.E., 2000, Geographic information for analysis of highway runoff-quality data on a national or regional scale in the conterminous United States: U.S. Geological Survey Open-File Report 00-432, 15 p. with CD-ROM insert. [GIS information available at URL http://water.usgs.gov/GIS/]

Smith, D.L., and Lord, B.M., 1990, Highway water quality control—Summary of 15 years of research: Washington, D.C., Transportation Research Board, Transportation Research Record 1279, p. 69–74.

Smith, R.A., Alexander, R.B., Tasker, G.D., Price, C.V., Robinson, K.W., and White, D.A., 1993, Statistical modeling of water quality in regional watersheds: Alexandria, Va., Proceedings of Watershed '93, A National Conference on Watershed Management, March 21–24, 1993, p. 751–754.

Smith, R.A., Schwarz, G.E., and Alexander, R.B., 1997, Regional interpretation of water-quality monitoring data: Water Resources Research, v. 33, no. 12, p. 2781–2798.

Sonnen, M.B., 1983, Guidelines for the monitoring of urban runoff quality: U.S. Environmental Protection Agency EPA-600/2-83-124, 128 p.

Spangberg, A., and Niemczynowicz, J., 1992, High resolution measurements of pollution wash-off from an asphalt surface: Nordic Hydrology, v. 23, no. 4, p. 245–256.

SPSS, Inc., 1997, Neural Connection 2.0: Chicago, Illinois, SPSS, Inc. also available at URL accessed on April 26, 2000, at http://www.spss.com/neuro/nc2info.htm

Stedinger, J.R., and Tasker, G.D., l985, Regional hydrologic analysis 1—Ordinary, weighted, and generalized least squares compared: Water Resources Research, v. 2l, no. 9, p. l42l–l432.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, Efi, 1993, Frequency analysis of extreme events, in Maidment, D.R., ed., Handbook of hydrology: New York, McGraw-Hill, p. 18.1–18.66.

Tasker, G.D., 1982, Comparing methods of hydrologic regionalization: Water Resources Bulletin, v. 18, no. 6, p. 965–970.

Tasker, G.D., and Burns, A.W., 1974, Mathematical generalization of stream temperature in central New England: Water Resources Bulletin, v. 10, no. 6, p. 1133–1142.

Tasker, G.D., and Driver, N.E., 1988, Nationwide regression models for predicting urban runoff water quality at unmonitored sites: Water Resources Bulletin, v. 24, no. 5, p. 1091–1101.

Tasker, G.D., Hodge, S.A., and Barks, C.S., 1996, Region of influence regression for estimating the 50-year flood at ungaged sites: Journal of the American Water Resources Association, v. 32, no. 1, p. 163–170.

Tasker, G.D., and Raines, T.H., 1995, An analysis of the storm-water data collection network for the Dallas–Fort Worth Metroplex, in Loethen, M.L., ed., Water Management in Urban Areas: Herndon, Va., American Water Resources Association, p. 1–10.

Tasker, G.D., and Slade, R.M., Jr., 1994, An interactive regional regression approach to estimating flood quantiles, in Fontane, D.G., and Tuvel, H.N., eds., Water Policy and Management--Solving the Problems: Proceedings of the 21st annual conference of the Water Resources Planning and Management Division, American Society of Civil Engineers, p. 782–785.

Tasker, G.D., and Stedinger, J.R., l989, An operational GLS model for hydrologic regression: Journal of Hydrology, v. 111, p. 36l–375.

Teso, R.R., Poe, M.P., Younglove, T., and McCool, P., 1996, Use of logistic regression and GIS modeling to predict groundwater vulnerability to pesticides: Journal of Environmental Quality, v. 25, p. 425–432.

Thirumalaiah, K., and Deo, M.C., 1998, River stage forecasting using artificial neural networks: Journal of Hydrologic Engineering, v. 3, no. 1, p. 26–32.

Thomson, N.R., McBean, E.A., and Mostrenko, I.B., 1996, Prediction and characterization of highway stormwater runoff quality: Ontario, Ministry of Transportation, Research and Development Branch, 98 p.

Thomson, N.R., McBean, E.A., Snodgrass, W., and Monstrenko, I.B., 1997a, Highway stormwater runoff quality—Development of surrogate parameter relationships: Water, Air, and Soil Pollution, v. 94, p. 307–347.

Thomson, N.R., McBean, E.A., Snodgrass, W., and Mostrenko, I.B., 1997b, Sample size needs for characterizing pollutant concentrations in highway runoff: Journal of Environmental Engineering, v. 123, no. 10, p. 1061–1065.

Transportation Research Board, 1997, Environmental research needs in transportation: Washington, D.C., Transportation Research Board, National Research Council, Circular no. 469, 98 p.

U.S. Army Corps of Engineers, 1977, Storage treatment, overflow, runoff model (STORM) user's manual: Davis, Calif., U.S. Army Corps of Engineers, Hydrologic Engineering Center, Generalized Computer Program 723-S8-L7520, [variously paged].

U.S. Environmental Protection Agency, 1999, Indicators of the environmental impacts of transportation: U.S. Environmental Protection Agency EPA/230-R-99-001, 189 p.

Van Der Heijde, P.K.M., 1990, Quality assurance in the application of groundwater models, *in* Janes, E.B., and Hotchkiss, W.R., eds., Symposium on Transferring Models to Users, Denver, Colo. Nov. 4–8, 1990, Bethesda, Md. Proceedings:American Water Resources Association, p. 97–109.

Vogel, R.M., 1986, The probability plot correlation coefficient test for the normal, lognormal, and Gumbel distribution hypotheses: Water Resources Research, v. 22, no. 4, p. 587–590.

Weisburg, S., 1980, Applied linear regression: New York, John Wiley and Sons, 283 p.

White, H.L., Jr., 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity: Econometrica, v. 48, p. 817–838.

Whitfield, P.H., and Wade, N.L., 1992, Monitoring transient water quality events electronically: Water Resources Bulletin, v. 28, no. 4, p. 703–711.

Wilkinson, L., 1985, Statistics quiz: Evanston, Ill., Systat, Inc., [Available on July 3, 2000 at URL http://www.tspintl.com/products/tsp/benchmarks/wilk.txt].

Young, G.K., Stein, S., Cole, P., Kammer, T., Graziano, F., and Bank, F., 1996, Evaluation and management of highway runoff water quality: Federal Highway Administration Final Report FHWA-PD-96-032, 480 p.

Zarriello, P.J., 1998, Comparison of nine uncalibrated runoff models to observed flows in two small urban watersheds, *in* Proceedings of the First Federal Interagency Hydrologic Modeling Conference, Las Vegas, Nev., April 19–23, 1998, v. 2, p. 7.163–7.170.

Zhang, Q., and Stanley, S.J., 1997, Forecasting raw-water quality parameters for the North Saskatchewan River by neural network modeling: Water Research, v. 31, no. 9, p. 2340–2350.

APPENDIX 1. Regression Tools

A. Partial Residual Plots

Transformation of predictors in a multiple regression model can be used to achieve linearity and simplify the model. A graphical device that can be helpful in deciding on a transformation for a predictor is a partial residual plot (Larsen and McCleary, 1972). A partial residual plot for a predictor, x_t, in a multiple linear regression is a plot of the partial residual, computed by subtracting the effects of all variables except x_t, against x_t. For example, suppose we have the three-variable model

$$y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + e_i \quad (i=1, n), \tag{1}$$

and we wish a partial residual plot for variable x_3. The partial residual, $(y_i - b_0 - b_1 x_{1,i} - b_2 x_{2,i})$, is plotted against x_3. If the plot appears to be linear, then no transformation is needed. If the plot shows some curvature, then a transformation may be helpful. This plot is sometimes called a component-plus-residual plot because the partial residual, $(y_i - b_0 - b_1 x_{1,i} - b_2 x_{2,i})$, is equal to $b_3 x_{3i} + e_i$, which is easier to compute.

B. Seasonality

Data may exhibit seasonal patterns. One method for dealing with seasonal patterns is to develop different regressions for the different seasons. Barbe and Francis (1995) use this method for coliform concentrations in a river in Louisiana. Driscoll and others (1990) classified snowmelt storms separately from rainfall-runoff events, but they did not develop seasonality as a quantitative variable in their highway-runoff-quality models. Effects of seasonality sometimes can be reduced by dealing with deviations from seasonal means or by use of periodic functions. When a simple periodic function such as a cosine function is used to describe the cyclic variation, the model is

$$y_i = \beta_o + \beta_1 \cos\left(\frac{2\pi t_i}{\tau}\right) + \beta_2 \sin\left(\frac{2\pi t_i}{\tau}\right) + \varepsilon_i \,, \tag{2}$$

in which t_i is a time unit and τ is the cycle length in the same time units. Cycle lengths may be known (such as annual cycles, diurnal cycles, or tidal cycles) or may be estimated from the data. Tasker and Burns (1974) use a periodic function with estimated cycle lengths to model stream temperatures in New England. Other variables also may be included in the model. For example, Cohn and others (1992) use a periodic function to remove seasonality from a model to estimate nutrient loading in Chesapeake Bay.

C. Collinearity

Correlation among the predictors in a regression results in near redundancies among the predictors, and inferences based on the model can be misleading or erroneous. Multicollinearity is the problem of linear dependencies between predictors. Multicollinearity in a regression can cause prediction problems when predictions are extrapolated beyond the sample space of the predictors. When the sample X-space is 3 or more dimensions, it is especially difficult to recognize a collinearity problem.

One diagnostic for detecting possible collinearity problems is the variance inflation factor (*VIF*). It is computed as

$$VIF_j = \frac{1}{1 - R_j^2} \,, \tag{3}$$

where R_j is the coefficient of determination from the regression of X_j on the other explanatory variables. A guideline for serious multicollinearity problems is $VIF > 10$. Belsley and others (1980) and Belsley (1991) describe other useful methods for detecting multicollinearity. Treatment of the multicollinearity problem depends somewhat on the source of the multicollinearity. Sources may be related to the sampling design, constraints on the model, and model specification.

It is sometimes possible to deal with multicollinearity created by the sampling design by collecting new data to fill in sparse areas on the predictor space. For example, consider a regression of sulfate load with two predictors, drainage area (A) and percent of basin urbanized (U). If in designing the monitoring network, sites with large A tended to have small U and sites with large U tended to have small A, a negative correlation in A and U would result. This could be fixed by adding new sites in the network with small A and small U and sites with large A and large U.

Multicollinearity problems created by constraints on the model or model specification can sometimes be dealt with by redefining the predictors. For example, consider two regressors, drainage area (A) and stream length (L). Suppose further that in this region all basins with a large A also have a large L and basins with small A have small L, so that A and L are constrained by the population to have a positive correlation. It may be possible to create a new dimensionless shape variable, $Sh=L^2/A$, which is not correlated with A or L, which can be substituted for either A or L.

It is also possible to deal with multicollinearity by means of predictor elimination. For example, if predictors x_1, x_2, and x_3 exhibit strong multicollinearity, eliminating one of the predictors may reduce the problem greatly. Predictor elimination, however, may not be an attractive alternative if the analyst wishes to extract information regarding the roles of individual predictors.

D. Regression Diagnostics

Regression diagnostics are used to identify possible outliers. An analyst can use regression diagnostic methods to find influential observations and study their effects. Regression diagnostics aid in the systematic location of data points that are unusual or are highly influential in estimating regression parameters and standard errors. Diagnostics help to avoid misinterpretation of the regression model. Cook and Weisberg (1982; 1984) and Belsley and others (1980) give comprehensive treatments of regression diagnostics. Partial regression leverage plots are an important part of regression diagnostics. Two commonly used diagnostics, leverage and Cook's D are briefly discussed below.

Observations that are far from the center of the X-variable space are considered high leverage points because of their great potential to influence the regression results. The leverage of a point is defined as

$$h_{ii} = x_i(X'X)^{-1}x'_i. \tag{4}$$

The limits on h_{ii} are $(1/n)<h_{ii}<1$ and

$$\sum_{i=1}^{n} h_{ii} = p'.$$

Therefore the average $h_{ii}=\dfrac{p'}{n}$, and a suggested value to identify a point with high leverage, is $h_{ii}=2\dfrac{p'}{n}$. Leverage plays an important role in the calculation of influence statistics and standardized residuals.

Cook's D is a measure of the shift in the vector of predicted values of y when the ith observation is not used. It shows the influence of the observation on the regression estimates. Computationally, it is obtained as follows

$$D_i = \frac{r^2_i}{p'}\left(\frac{h_{ii}}{1-h_{ii}}\right). \tag{5}$$

A suggested cutoff value to flag influential data points is $D_i>4/n$.

APPENDIX 2. Linear Regression Methods

A. Linear Ordinary Least-Squares Regression

The linear model for relating a response variable, Y, to p predictors is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \varepsilon_i. \tag{6}$$

Subscript i denotes an observation at site i. There are p predictors and $p' = (p+1)$ parameters to be estimated. Let n denote the number of observations or sites.

Denote the following matrices:
   Y, a (n x 1) column vector of observed response,
   X, a (n x p') matrix of a column of ones followed by p columns of predictors,
   β, is a (p' x 1) vector of parameters to be estimated, and
   ε, is a (n x 1) column vector of random errors.
The linear model can be written in matrix notation as

$$Y = X\beta + \varepsilon, \tag{7}$$

in which

$$
Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \quad
X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \quad
\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} \quad
\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}
$$

The usual assumptions about random vector of errors, ε, is that all the elements, ε_i, have a common variance, σ^{2}, and are statistically independent. These assumptions can be written in shorthand as

$$\varepsilon \sim N(\mathbf{0}, I\sigma^2). \tag{8}$$

If the model is correct (another assumption), then

$$Y \sim N(X\beta, I\sigma^2). \tag{9}$$

The regression coefficients, β, are best estimated as

$$b = (X'X)^{-1}(X'Y). \tag{10}$$

The predicted mean of the response variable at site k with basin characteristics $x_k = (1, x_{k,1}, x_{k,2}, \ldots, x_{k,p})$ is

$$\hat{y}_k = x_k b. \tag{11}$$

The variance of the prediction is $Var(\hat{y}_k) = \sigma^2[1 + x_k(X'X)^{-1}x'_k]$.

B. Nonparametric Regression

The Kendall-Theil method (Helsel and Hirsch, 1992) is a nonparametric method for fitting a prespecified linear equation of the form:

$$Y = B_0 + B_1 X. \tag{12}$$

For each pair of points (x_i, y_i) and (x_j, y_j), such that $i < j$, a slope, $S_{ij} = \dfrac{(y_i - y_j)}{(x_i - x_j)}$, is computed and B_1 is set equal to the median of the computed slopes. Let $Y_{.5}$ and $X_{.5}$ denote the median of X and Y, respectively. The intercept, $B_0 = Y_{.5} - B_1 X_{.5}$. Conover (1980) provides a method for estimating a confidence interval for B_1. Instead of using the median of the pairwise slopes, median (S_{ij}), one can use a weighted median of the slopes with weights proportional to the distance between the pairs of points. Birkes and Dodge (1993, p. 114) show that the nonparametric weighted median estimator is the value of B_1 that minimizes the sum

$$\sum_{i=1}^{n} \left(rank(y_i - B_1 x_i) - \frac{(n-1)}{2} \right)(y_i - B_1 x_i) \quad . \tag{13}$$

Equation 12 may be generalized to more than one predictor for a multiple regression problem. Nonparametric estimates of $B_1, B_2, \ldots, B_p$ are found by minimizing the sum

$$\sum_{i=1}^{n} \left(rank(y_i - (B_1 x_{1i} + B_2 x_{2i} + \ldots)) - \frac{(n-1)}{2} \right)(y_i - (B_1 x_{1i} + B_2 x_{2i} + \ldots)) \quad . \tag{14}$$

An estimate of B_0 is obtained as the median of $y_i - (B_1 x_{1i} + B_2 x_{2i} + \ldots)$. Birkes and Dodge (1993, p. 123) provide iterative procedures to find the values of $B_1, B_2, \ldots, B_p$.

Regression smooths are nonparametric local averaging methods that require no prespecified model functional form. Three major smoothing methods for problems with one predictor are kernel smoothing, k-nearest neighbor (k-NN) smoothing, and splines (Härdle, 1990). Kernel smoothing uses local observations within a bandwidth to compute a weighted average defined by the kernel. The k nearest neighbors are used to estimate the local weighted average in k-NN smoothing. Splines are piecewise polynomials of order k that are smoothly joined. A cubic spline ($k=3$) is usually good enough for most problems. Average smooths tend to follow outlying points and are not particularly robust against outliers. However, one may choose median smooths (Helsel and Hirsch, 1992, p. 286) or LOWESS (Cleveland, 1979; Helsel and Hirsch, 1992, p. 288) when outliers are perceived to be a problem. LOWESS is an iterative procedure that progressively downweights outliers in computing the local weights.

The regression smooths described above are for problems with one predictor. Extending the local averaging smooths to problems with multiple predictors raises the problem of sparse data in local neighborhoods. Breiman and others (1984) proposed regression trees as a type of nonparametric, multiple-predictor smooth. Regression trees define a piecewise constant regression surface based upon neighborhoods defined by hyper-rectangles with sides parallel to coordinate axes. Friedman and Stuetzle (1981) describe projection pursuit regression, an extension of regression trees, that uses smoothing methods on linear combinations of predictors to form the regression surface. Research is needed to determine whether regression trees and projection pursuit regression would be useful in regional water-quality-assessment models.

C. Robust Regression

Robust regression methods are insensitive to the effects of outliers in the data; they are useful for detecting outliers by accentuating observations with large residuals from the robust model. Observations that are down-weighted in the robust regression model require close examination for the reasons for downweighting. The parameter-estimation problem in regression analysis may be thought of as finding the estimates of β to minimize the sum of some function, ρ,

$$\sum_{i=1}^{n} \rho(z) , \qquad (15)$$

where $z = \dfrac{y_i - x_i \beta}{s}$ and s is a scale factor.

Robust procedures dampen the effects of outliers and tend to leave large residuals for the outliers. In robust estimation, the scale factor, s, must be a robust estimator of scale and not the standard deviation of the residuals because the standard deviation is relatively sensitive to outliers. Montgomery and Peck (1982, p. 367) suggest the robust estimator

$$s = \frac{median|e_i - median(e_i)|}{0.6745} , \qquad (16)$$

where $e_i = y_i - x_i \beta$.

The least-squares estimator, corresponding to $\rho(z) = z^2$, is sensitive to outliers because it gives relatively greater weight to large residuals. On the other hand, the least-absolute-deviation estimator, corresponding to $\rho(z) = |z|$, is resistant to outliers but may give too much weight to small residuals. Huber (1973) proposed a compromise estimator that is robust yet relatively efficient if data are normal. It weights small residuals as a least-squares estimator and large residuals as an absolute-deviation estimator. The ρ function for Huber's estimator is

$$\rho(z) = \begin{cases} \dfrac{1}{2} z^2 & |z| \leq t \\[2mm] \left(|z| t - \dfrac{1}{2} t^2 \right) & |z| > t \end{cases} , \qquad (17)$$

in which t is a constant usually equal to 2 or less. Parameter estimates are made from iteratively reweighted least squares. For example, the weights, w, in a weighted least-squares regression minimizing the sum of the function in equation 18 are

$$w(z) = \begin{cases} 1.0 & |z| \leq t \\[2mm] \dfrac{t}{|z|} & |z| > t \end{cases} . \qquad (18)$$

Because the weights depend on β, iteration is required until little or no changes in the parameters are observed. The parameters for many other functions that have been suggested for the robust regression problem (Andrews and others, 1972) also can be found by iteratively reweighted least squares.

One problem with robust methods is an apparent lack of agreement among authorities on how best to construct confidence intervals for the parameter estimates. Therefore, one important aspect of statistical regional models is not clearly determined for robust regression models.

D. Generalized Least-Squares Regression

In generalized least-squares (GLS) regression, β is estimated by

$$\tilde{\beta} = (X^T \Lambda^{-1} X)^{-1} X^T \Lambda^{-1} Y , \tag{19}$$

in which Λ is the covariance matrix of errors, $E(ee^T)$. The operational difficulty with this procedure is that Λ must be estimated from the data at hand. Stedinger and Tasker (1985) show that Λ can be estimated as

$$\Lambda = \gamma^2 I + \Sigma , \tag{20}$$

where γ^2 is an estimate of the variance of the error inherent in the model, Σ is an estimate of the sampling-error covariance matrix, and I is an (n by n) identity matrix. The model error variance, γ^2, and regression coefficients, b, are found by iteratively searching for the best non-negative solution to the equation

$$E\{(Y - X\beta)^T \Lambda^{-1}(Y - X\beta)\} = n - k - 1 \quad . \tag{21}$$

A leverage statistic in GLS analogous to leverage in OLS regression is the ith diagonal element of

$$H^* = X(X' \Lambda^{-1} X)^{-1} X' \Lambda^{-1} . \tag{22}$$

The sum of the diagonal elements of H^* is equal to the number of parameters in the model; and a high-leverage site would be one in which the associated diagonal element is greater than 2 times the number of parameters divided by the number of observations, as a rule of thumb. A GLS version of Cook's D is

$$D'_i = \frac{e_i^2 h'_{ii}}{p'(\gamma_i - h'_{ii})^2} , \tag{23}$$

where h'_{ii} are diagonal elements of

$$H' = X(X' \Lambda^{-1} X)^{-1} X' . \tag{24}$$

D'_i is large if it exceeds about $(4/n)$ (Tasker and Stedinger, 1989).

E. Tobit Regression

The tobit regression model for censored observations is

$$y_i = \left(\begin{array}{ll} (\boldsymbol{x}_i\beta + e_i) & \text{if } (y_i > \text{threshold}) \\ \text{nominal} & \text{otherwise} \end{array} \right. \cdot \qquad (25)$$

If one arbitrarily assumes a value for those observations of y below the threshold and uses the entire sample, or if one uses the subsample of observations when y_i is greater than the censoring threshold, the least-squares estimator of β is biased and inconsistent. However, maximum-likelihood estimators for the tobit regression model are available and are described in Judge and others (1985). Liu and others (1996) use tobit regression to predict atrazine concentrations in the Midwest. Hirsch and others (1993, p. 17.51) caution that application of tobit regression in hydrology is experimental.

F. Logistic Regression

In logistic regression, the response is a nominal variable with two possible values (0 and 1). For example, y_i is equal to zero if the value is below the detection limit and equal to 1 if the value is above the detection limit. In the model, the estimated response, $E(y|\boldsymbol{x})$, is a proportion between 0 and 1 and is given by

$$E(y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\beta)}{1 + \exp(\boldsymbol{x}\beta)} \cdot \qquad (26)$$

The unknown β's are estimated by maximizing the log likelihood function

$$L(\beta) = \sum_{i=1}^{n} \{y_i ln[\boldsymbol{x}\beta] + (1 - y_i)ln[1 - \boldsymbol{x}\beta]\} \text{ (Cox and Snell, 1989).} \qquad (27)$$

Teso and others (1996) use logistic regression to estimate probability of pesticide contamination. Major statistics packages, such as SAS, Minitab, and STATIT, include procedures for fitting the logistic regression function. Thus, the logistic regression procedure can be used to estimate the likelihood of a water-quality characteristic being above or below a censoring threshold at a site with basin characteristics equal to $\boldsymbol{x}$.

G. Contingency Tables

Contingency tables can be used when both response and predictors are nominal or censored. The number of observed values falling within a cell defined by the response and predictor groups divided by the total observations in a group provides an estimate of the probability of a value being in the cell. If, for example, the effect of traffic volume on the potential for detection of cadmium was of interest, then this effect could be examined by use of the following (hypothetical) contingency table:

Contingency-table example using hypothetical cadmium concentrations in urban and rural highway runoff

[ADT, Average Daily Traffic; Cd, cadmium concentrations in micrograms per liter; VPD, vehicles per day]

Cadmium concentrations	Rural highway (ADT ≤30,000 VPD)	Urban highway (ADT>30,000 VPD)	Total
Cd < detection limit	A=15	B=8	A+B=23
Cd ≥detection limit	C=5	D=20	C+D=25
Total:	A+C=20	B+D=28	

In this hypothetical example, the probability of cadmium being greater than or equal to the detection limit is $5/20 = 0.25$ for rural highways and $20/28 = 0.71$ for urban highways. Helsel and Hirsch (1992) and other texts provide details for statistical analysis using contingency tables.

H. Ridge Regression

The ridge estimator of regression coefficients, β, is

$$\hat{\beta}_R = (\mathbf{Z'Z} + \kappa\mathbf{I})^{-1}\mathbf{Z'}\mathbf{y}^o \quad , \tag{28}$$

in which constant κ is a biasing parameter to be determined, and Z and y^o are standardized versions of X and y, respectively. The choice of κ is the subject of several studies. Myers (1986) and Montgomery and Peck (1982) describe several methods for choosing κ.

Computations can be made by augmenting the standardized data and using ordinary least-squares methods as follows:

$$\mathbf{Z}_a = \begin{bmatrix} \mathbf{Z} \\ \sqrt{\kappa}\mathbf{I} \end{bmatrix} \qquad \mathbf{y}_a = \begin{bmatrix} \mathbf{y}^o \\ \mathbf{0} \end{bmatrix} , \tag{29}$$

where $\sqrt{\kappa}\mathbf{I}$ is a p by p diagonal matrix with diagonal elements equal to the square root of κ and $\mathbf{0}$ is a p by 1 vector of zeros (Montgomery and Peck, 1982). The estimate $\hat{\beta}_R$ is then computed as

$$\hat{\beta}_R = (\mathbf{Z}_a'\mathbf{Z}_a)^{-1}\mathbf{Z}_a'\mathbf{y}_a \quad . \tag{30}$$

The use of ridge regression requires thoughtful study of the data and careful analysis, but it can be an effective method for dealing with multicollinearity problems.

APPENDIX 3. Nonlinear Regression Methods

A. SPARROW

In the SPARROW (<u>SPA</u>tially <u>R</u>eferenced <u>R</u>egression <u>O</u>n <u>W</u>atershed attributes) model (Smith and others, 1997), the stream network in a region is divided into many stream reaches and the instream load, L_i, of a constituent in a stream reach indexed by i is equal to the sum of contributions to the load from all upstream sources, $S_{n,i}$, so that

$$L_i = \sum_{n=1}^{N} S_{n,i} , \tag{31}$$

where N is the number of sources. Let $J(i)$ represent the set of all stream reaches upstream from reach i and including reach i but downstream from all monitoring stations upstream from reach i. Let $K(i)$ represent the set of all monitoring sites directly upstream from reach i. The source terms are determined by

$$S_{n,i} = \beta_n \sum_{j \in J(i)} X_{n,i,j} + \sum_{k \in K(i)} X_{n,i,k} , \tag{32}$$

in which β_n is a coefficient for source n and $\sum_{j \in J(i)} X_{n,i,j}$ is the predictor for source n and reach i associated with the upstream reaches $J(i)$ or monitoring sites $K(i)$. The values for $X_{n,i,j}$ are given by

$$X_{n,i,j} = s_{n,j} \exp(-\alpha_1 Z_{1,j} - \alpha_2 Z_{2,j} \ldots - \delta_1 T_{1,i,j} - \delta_2 T_{2,i,j} \ldots) , \tag{33}$$

where $s_{n,j}$ is a measure of the contaminant mass from source n that is present in reach j or at monitoring site k, the α's are delivery coefficients associated with land-surface characteristics $Z_{i,j}$, and the δ's are decay coefficients associated with flowpath characteristics $T_{n,i,j}$. For point sources or monitoring sites, the α's are set equal to zero. For example, $s_{n,j}$ might be the mass of chlorides placed on the roads within the drainage of reach j or measured at monitoring site k; the $Z_{1,j}$ might be average soil permeability for the drainage area of reach j; and $T_{1,i,j}$ might be the stream length between reach i and reach j. The predictors are spatially referenced because the contribution from all reaches above a given reach is tied to the reach by the flowpath characteristic, $T_{i,j}$.

The parameters (β's, α's, and δ's) of this nonlinear regression technique are determined by means of SAS/ETS procedure MODEL (SAS Institute, 1988). Smith and others (1997) determine the standard errors of the parameters using bootstrap methods (Efron, 1982).

B. Artificial Neural Networks

An artificial neural network, or ANN (Hertz and others, 1991), is composed of simple processing units, called neurons, arranged in layers. Each unit receives input from other units and converts the input to a single output, which it sends to other units. The conversion takes place in two stages: first, a net input is computed as a weighted sum of inputs, then an activation function transforms the net input into an output. The flexibility of ANN comes from the analyst's being able to specify multiple layers of neurons with nonlinear activation functions and alternative methods for computing the net input. A multilayer perceptron (MLP) with three nodes in the input layer, three nodes in the hidden layer, and one output node is shown in figure A. The hidden layer is so named because it has no direct connection to the outside world.
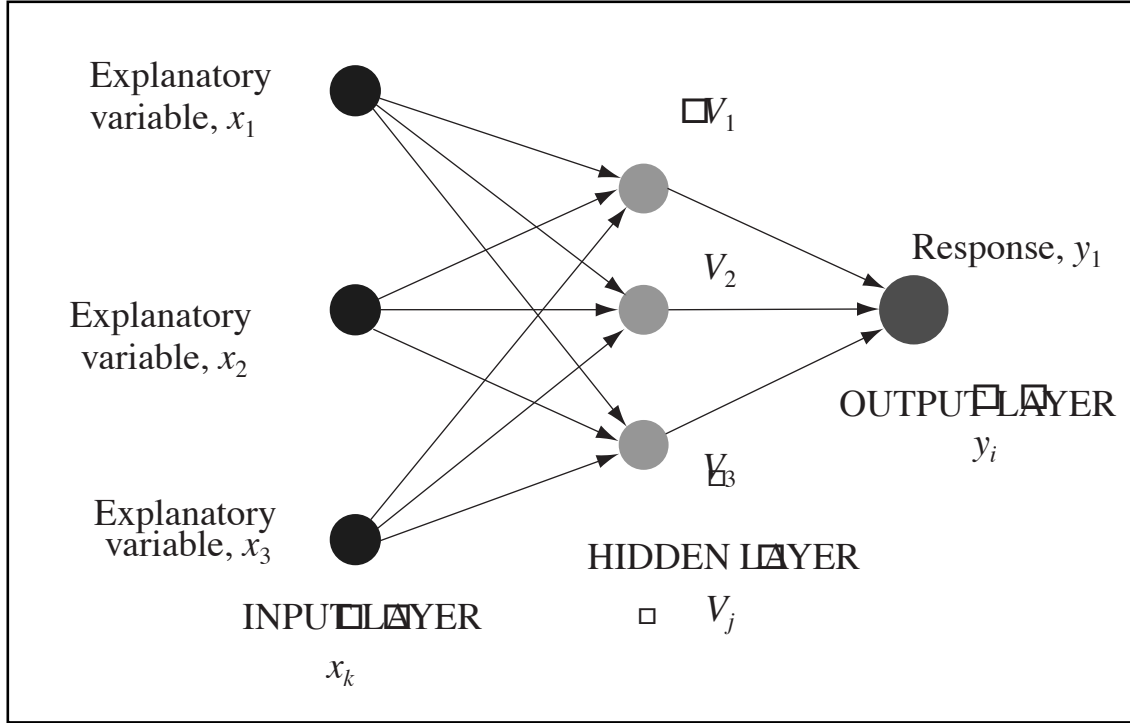
Figure A. Multilayer perceptron illustrating the function of an artificial neural network (ANN).

Each explanatory variable has one input node in the model. Denote the kth input variable as x_k and the number of hidden nodes as n_h. The net input to the jth hidden node is

$$a_j + \sum_{k=1}^{n_h} b_{jk} x_k , \qquad (34)$$

where a_j and b_{jk} are intercept and weights from input k to hidden node j. The activation function, $g(u)$, for the hidden nodes is usually a smooth nonlinear function with a single-valued first derivative, such as the sigmoid function, $g(u) = \frac{1}{1+e^{-u}}$, or the hyperbolic tangent function, $g(u) = \tanh(u)$. For example, the estimated response, y_1, for the MLP in figure A with three input nodes, three hidden nodes with a hyperbolic tangent activation function, and one output node is

$$y_1 = c_1 + \sum_{j=1}^{3} d_{1,j} \tanh\left(a_j + \sum_{k=1}^{3} b_{jk} x_k \right), \qquad (35)$$

where c_1 and $d_{1,j}$ are intercept and weights from hidden node j to output 1. The activation function for the output node in this example is the linear function, $g(u) = u$.

Observed values of predictors (inputs) and responses (targets) are used to train the ANN by iteratively adjusting the weights used by the neurons to produce output so that the sum of squared differences between output and target data is small. The method used is called back-propagation with a conjugate gradient training algorithm (Hertz and others 1991). Neural Connection 2.0 (SPSS, Inc., 1997) software was used for the calculations. Alternatively, one could estimate the intercepts and weights in equation 35 using nonlinear regression methods (Sarle, 1994). ANN can overtrain (fit the observed data well, but not predict well for new data). For this reason, a portion of the observed data is used as a validation data set that is not used in training the ANN. Artificial neural networks are data-in/predictions-out black boxes. Any underlying hydrologic model or hydrologically significant functional relation may be impossible to extract from the network.

APPENDIX 4. Uncertainty Analysis

A. Uncertainty—Normally Distributed Errors

The predicted response at unmonitored site k with basin characteristics $x_0 = (1, x_{0,1}, x_{0,2}, .., x_{0,p})$ is

$$\hat{y}_0 = x_0 b . \tag{36}$$

The standard error of the prediction in OLS regression is

$$S(\hat{y}_0) = \sigma \sqrt{1 + x_0 (X'X)^{-1} x'_0} . \tag{37}$$

In GLS regression the standard error of prediction is:

$$S(\hat{y}_0) = \sqrt{\hat{\gamma}^2 + x_0 X' \hat{\Lambda}^{-1} X^{-1} x'_0} . \tag{38}$$

A $100(1-\alpha)$ prediction interval would be

$$\hat{y}_0 - T \leq y_0 \leq \hat{y}_0 + T , \tag{39}$$

where

$$T = t_{\frac{\alpha}{2}, (n - p')} S(\hat{y}_0) , \tag{40}$$

and where $t_{a/2, n-p'}$ is the critical value from a t-distribution for n-p' degrees of freedom. The use of the t-statistic requires the errors to be approximately normally distributed. If a log transformation had been made so that $y_0 = \log_{10}(q_0)$, then the prediction interval would be

$$10^{\hat{y}_0 - T} \leq q_0 \leq 10^{\hat{y}_0 + T} . \tag{41}$$

When a log transformation has been made and the standard error in log units follows a normal distribution, the standard error may be expressed in percent of the predicted value in the original untransformed units. Denote σ as the standard error in log (base 10) units, S_{org} as the standard error in original units, and $E(q|x_k)$ as the predicted value of q, in original units, given x_k, and $x_k = (1, x_{k,1}, x_{k,2},..., x_{k,p})$ is a vector of basin characteristics at site k. The standard error in percent, $S_{percent}$, is given by

$$S_{percent} = 100 \frac{S_{org}}{E(q|x_k)} = 100 \sqrt{(e^{5.302\sigma^2} - 1)} \quad \text{(Aitcheson and Brown, 1957).} \tag{42}$$

Sometimes it is said in OLS that two-thirds of the points lie within one standard error of estimate of the regression function. This is true for the log unit standard error of estimate, σ, but it generally is not correct for $S_{percent}$, because the errors in log space are symmetrically distributed under the assumption of normality of the log errors, but the errors in original units are skewed. One can, however, calculate a +percent and -percent errors with the following formulas:

$$S_{plus} = 100(10^\sigma - 1) \quad \text{and} \tag{43}$$

$$S_{minus} = 100(10^{-\sigma} - 1) . \tag{44}$$

The three formulas (42, 43, and 44) above apply not only to the standard error of estimate for an OLS regression but also to the standard error of the model, $\hat{\gamma}$, in GLS regression, and standard error of a prediction in both OLS and GLS.

B. Uncertainty—Non-Normally Distributed Errors

Consider the general regression model

$$y_i = g(\boldsymbol{x}_i, \beta) + \varepsilon_i \qquad i = 1, 2, 3, ..., n \quad, \tag{45}$$

where $g(\boldsymbol{x}_i, \beta)$ is a function of known form, x_i is a vector of known predictors, and β is a vector of unknown coefficients. The ε_i are independent errors drawn from an unspecified distribution, F, centered at zero, that may not be normally distributed. Having observed y_i for $i=1, 2,..., n$, β is estimated by minimizing the sum of some function, ρ, of the errors

$$\hat{\beta} : \min \sum_{i=1}^{n} \rho(y_i - g(\boldsymbol{x}_i, \beta)) \quad. \tag{46}$$

Such a model may be too complicated for standard analysis, but a bootstrap method similar to that described below can be used.

1. Compute the observed residuals: $\hat{\varepsilon}_i = y_i - g(\boldsymbol{x}_i, \hat{\beta})$

2. Draw a bootstrap sample by randomly selecting, with replacement, from the observed residuals, a bootstrap sample of residuals, $\hat{\varepsilon}^{\circ}_1, \hat{\varepsilon}^{\circ}_2, ..., \hat{\varepsilon}^{\circ}_n$, compute $y_i^{\circ} = g(\boldsymbol{x}_i, \hat{\beta}) + \hat{\varepsilon}_i^{\circ}$ and calculate $\hat{\beta}^{\circ}$.

3. Repeat step 2 Z times to obtain bootstrap replications $(\hat{\beta}^{\circ})_1, (\hat{\beta}^{\circ})_2, ..., (\hat{\beta}^{\circ})_Z$.

Let $\hat{\beta}^{\circ}_{mean} = \sum_{b=1}^{B} \dfrac{(\hat{\beta}^{\circ})_b}{Z}$; then, an estimate of β's covariance matrix is

$$cov(\hat{\beta}) = \frac{1}{Z-1} \sum_{b=1}^{B} ((\hat{\beta}^{\circ})_b - \hat{\beta}^{\circ}_{mean})((\hat{\beta}^{\circ})_b - \hat{\beta}^{\circ}_{mean})' \quad. \tag{47}$$

A nonparametric $(1-\alpha)$ confidence interval for a prediction can be approximated by taking the $(1-\alpha)$ central portion of Z predictions based on the bootstrap replications from step 3.

APPENDIX 5. Region of Influence Method

In this method of developing site-specific predictions with a data set spanning a large geographic area, the regression equation for a site is computed using data from a unique region called the region of influence by Burn (1990a, 1990b) and suggested by Acreman and Wiltshire (1987). The unique subset of monitoring sites that make up the region of influence for each prediction site is made up of the Ns nearest neighbors. The method is an attractive alternative to the more traditional methods because it can be easily updated by simply updating the water-quality data in a database file from which the method draws its basic data; furthermore, extrapolation errors tend to be small because predictions by definition occur near the center of the space of the predictors. In this method, the nearness of two neighbors is not measured by the physical distance between the sites, but rather by a distance defined in terms of the watershed characteristics. This distance between any two sites, indexed by i and j, is determined by the Euclidean distance metric;

$$d_{ij} = \left(\sum_{k=1}^{p} \left(\frac{x_{ik} - x_{jk}}{sd(X_k)} \right)^2 \right)^{1/2} , \tag{48}$$

in which, d_{ij} is the distance between the watershed characteristics at sites i and j, p is the number of watershed characteristics needed to calculate d_{ij}, X_k represents the kth watershed characteristic, $sd(X_k)$ is the sample standard deviation for X_k, and x_{ik} is the value of X_k at the ith site. The d_{ij}'s between the prediction site i and all monitoring sites $j = 1, 2,..., n$ in a region is determined, and the Ns monitoring sites with smallest dij make up the region of influence for site i. For this method to work, the value of Ns should be large enough to have enough degrees of freedom in the regression to estimate two or three parameters. The method is computer intensive and requires some subjective judgement for selecting Ns and the attributes used in the distance metric.