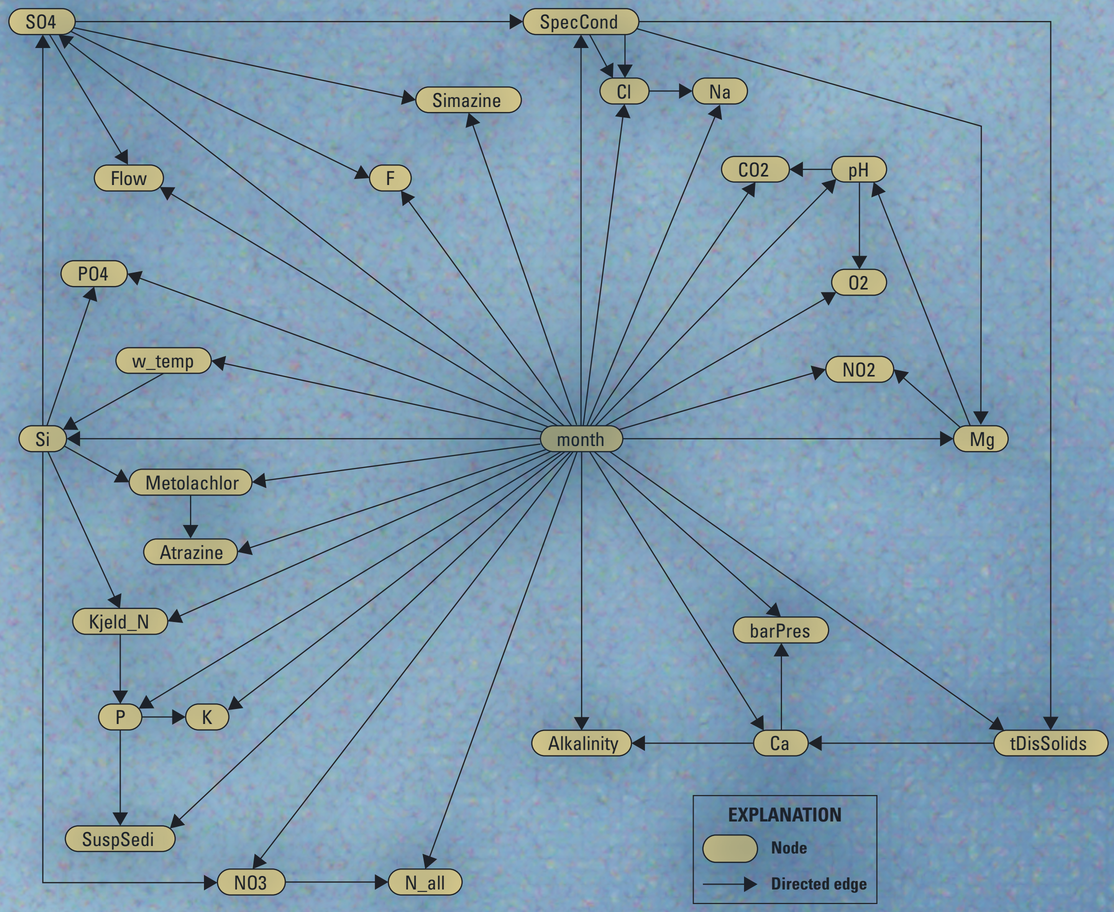
USGS
science for a changing world

National Water-Quality Program

An Exploratory Bayesian Network for Estimating the Magnitudes and Uncertainties of Selected Water-Quality Parameters at Streamgage 03374100 White River at Hazleton, Indiana, from Partially Observed Data



Scientific Investigations Report 2018–5053

U.S. Department of the Interior
U.S. Geological Survey

Cover. Directed acyclic graph of the Bayesian network developed to describe associations among selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1991–2016.

An Exploratory Bayesian Network for Estimating the Magnitudes and Uncertainties of Selected Water-Quality Parameters at Streamgage 03374100 White River at Hazleton, Indiana, from Partially Observed Data

By David J. Holtschlag

National Water-Quality Program

Scientific Investigations Report 2018–5053

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
RYAN K. ZINKE, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2018

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Suggested citation:
Holtschlag, D.J., 2018, An exploratory Bayesian network for estimating the magnitudes and uncertainties of selected water-quality parameters at streamgage 03374100 White River at Hazleton, Indiana, from partially observed data: U.S. Geological Survey Scientific Investigations Report 2018–5053, 30 p., https://doi.org/10.3133/sir20185053.

ISSN 2328-0328 (online)

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
	Length	
inch (in.)	25.4	millimeters
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
	Area	
square miles (mi^2)	2.590	square kilometers (km^2)
	Volume	
cubic foot (ft^3)	0.02832	cubic meter (m^3)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

$$°F = (1.8 \times °C) + 32.$$

Temperature in degrees Fahrenheit (°F) may be converted to degrees Celsius (°C) as follows:

$$°C = (°F – 32) / 1.8.$$

Supplemental Information

Concentrations of chemical constituents in water are given in either milligrams per liter (mg/L) or micrograms per liter (μg/L).

Abbreviations

Alkalinity	Alkalinity, water, filtered, inflection-point titration method (incremental titration method), field, milligrams per liter as calcium carbonate (parameter 39086)
Atrazine	Atrazine, water, filtered, recoverable, micrograms per liter (parameter 39632)
ASCII	American Standard Code for Information Interchange
barPres	Barometric (atmospheric) pressure, millimeters of mercury (parameter 00025)
BN	Bayesian network
Ca	Calcium, water, filtered, milligrams per liter (parameter 00915)
Cl	Chloride, water, filtered, milligrams per liter (parameter 00940)
ContMeasSet	Continuously Measureable Set refers to the conditioning variable set that includes month of measurement (month), barometric pressure (barPres), streamflow (Flow), dissolved oxygen (O2), pH (pH), specific conductance (SpecCond), and water temperature (w_temp).
CO2	Carbon dioxide, water, unfiltered, milligrams per liter (parameter 00405)
CPT	conditional probability table
DAG	directed acyclic graph
d-connected	Directionally connected or conditionally d-connected given measurements
d-separated	Directionally separated or conditionally d-separated given measurements
E	When associated with a water-quality parameter, an E indicates that the reported values were estimated.
EM	Expectation Maximization
F	Fluoride, water, filtered, milligrams per liter (parameter 00950)
Flow	Streamflow, in cubic feet per second (parameter 00060)
K	Potassium, water, filtered, milligrams per liter (parameter 00935)
Kjeld_N	Kjeldahl nitrogen, the sum of ammonia, ammonium, and organic nitrogen, water, unfiltered, milligrams per liter as nitrogen (parameter 00625)
M	When associated with a water-quality parameter, an M indicates that the presence of the parameter was verified, but not quantified.
Metolachlor	Metolachlor, water, filtered, recoverable, micrograms per liter (parameter 39415)
Mg	Magnesium, water, filtered, milligrams per liter (parameter 00925)
NA	Indicates a missing value (not available) in water-quality data.
Na	Sodium, water, filtered, milligrams per liter (parameter 00930)
NAWQA	National Water-Quality Assessment
NO2	Nitrite, water, filtered, milligrams per liter as nitrogen (parameter 00613)
NO3	Nitrate, water, filtered, milligrams per liter as nitrogen (parameter 00618)
NWIS	National Water Information System

NWQP	National Water Quality Program
N_all	Total nitrogen [nitrate + nitrite + ammonia + organic-N] (TN), water, unfiltered, milligrams per liter (parameter 00600)
O2	Dissolved oxygen (DO), water, unfiltered, milligrams per liter (parameter 00300)
P	Phosphorus, water, unfiltered, milligrams per liter as phosphorus (parameter 00665)
PCB	Polychlorinated Biphenyl
PDF	probability density function
pH	pH, water, unfiltered, field, standard units (parameter 00400)
PO4	Orthophosphate, water, filtered, milligrams per liter as phosphorus (parameter 00671)
PMF	probability mass function
Si	Silica, water, filtered, milligrams per liter as SiO_2 (parameter 00955)
Simazine	Simazine, water, filtered, recoverable, micrograms per liter (parameter 04035)
SO4	Sulfate, water, filtered, milligrams per liter (parameter 00945)
SpecCond	Specific conductance (SC), water, unfiltered, microsiemens per centimeter at 25 degrees Celsius (parameter 00095)
SuspSedi	Suspended sediment concentration (SSC), milligrams per liter (parameter 80154)
TAN	Tree Augment Naïve
tDisSolids	Dissolved solids dried at 180 degrees Celsius, water, filtered (TDS), milligrams per liter (parameter 70300)
USGS	U.S. Geological Survey
w_temp	Temperature, water, degrees Celsius (parameter 00010)
<	A qualifier for a water-quality parameter indicating that the concentration is less than the reported value
>	A qualifier for a water-quality parameter indicating that the concentration is greater than the reported value

An Exploratory Bayesian Network for Estimating the Magnitudes and Uncertainties of Selected Water-Quality Parameters at Streamgage 03374100 White River at Hazleton, Indiana, from Partially Observed Data

By David J. Holtschlag

Abstract

An exploratory discrete Bayesian network (BN) was developed to assess the potential of this type of model for estimating the magnitudes and uncertainties of an arbitrary subset of unmeasured water-quality parameters given the measured complement of parameters historically measured at a U.S. Geological Survey streamgage. Water-quality data for 27 water-quality parameters from 596 discrete measurements at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, were used to develop this BN. Data for each of the water-quality parameters were discretized into five intervals based on the quintiles of the measured values. The 596 discrete measurements were randomly partitioned into a training set with 80 percent of the data and a testing set with 20 percent of the data to identify, estimate, and assess the training and testing accuracy of the Bayesian network.

A BN with 28 nodes was formed from the 27 water-quality parameters and the month of sample collection. Based on data in the training set, a network with 53 directed edges and *month* as the target node was identified by minimizing the negative log-likelihood function for all nodes treated, in turn, as the target variable. The edge structure determines the number and magnitude of elements in conditional probability tables associated with all nodes.

The effectiveness of the BN was assessed on the basis of correct classification rates to one of the five discrete intervals, which were computed separately for the training and testing datasets and for two conditioning variable sets. The selected sets of conditioning variables represent two of many possible sets of measured parameters on which to base estimates of unmeasured parameters. The first set includes only the month of sample collection (*month*), and an expanded set includes *month* and six other continuously measurable parameters, referred to as the *ContMeasSet*, all of which were obtained from the discrete data.

Results indicated that the training dataset had average correct classification rates of 41.7- and 61.2-percent rates conditioned on the *month* and *ContMeasSet* sets, respectively. The testing dataset had somewhat lower average correct classification rates of 40.8 and 56.5 percent for the two conditioning variable sets. When conditioned on *month* only, the average correct classification rate for the testing dataset was only slightly lower than the average correct classification rate in the training dataset, indicating little model overfitting. When using the *ContMeasSet*, however, the average decrease in accuracy between training and testing sets was 4.9 percent. The training and testing datasets and both sets of conditioning variables, however, indicate that the BN would substantially outperform a random assignment model, which would be expected to have a 20-percent correct classification rate. In addition, the edge structure of the BN depicts how information can flow through the network, which may help prioritize parameters for measurement to facilitate estimation of unmeasured parameters. Finally, extension of a static BN, like the one developed in this report, to a dynamic BN may provide a basis for using high-frequency or continuous water-quality data to extend information in time between discrete water-quality samples, and this integration could mitigate some of the limitations of high-frequency and discrete water-quality sampling methods.

Introduction

The U.S. Geological Survey (USGS) National Water Quality Program (NWQP) collects, analyzes, and interprets water-quality data on the Nation's streams and aquifers to assess their status and trends and to characterize how natural processes and human interventions affect the Nation's water resources. Local, regional, and national studies are enabled by consistent methods of water-quality data collection and analysis. These data are made available for public access and future uses by archiving in the USGS National Water Information System (NWIS; U.S. Geological Survey, 2018a).

Stream water-quality data are routinely collected at fixed-location monitoring stations distributed across the Nation. At each station, water-quality information is commonly obtained at differing intervals as discrete samples, which can be analyzed for a large number of parameters in a laboratory. Only data from discrete samples are discussed in this report. It is technically possible, however, to measure a subset of the parameters in discrete samples by in situ methods at high frequencies, typically every 15 to 60 minutes. These continuously measurable (high frequency) parameters are of special interest because they have the potential to update water-quality information in Bayesian networks (BNs) between discrete measurements.

Numerous water-quality parameters can be measured in the laboratory. More than 21,500 five-digit parameter codes (U.S. Geological Survey, 2011a) are assigned to water-quality parameters in NWIS. Parameter codes are organized within 17 parameter groups characterizing physical properties, nutrient concentrations, organic and inorganic constituents, informational, and other data. The precise definition of each parameter code, including parameter name and description, the Chemical Abstracts Service Registry Number of the parameter where available, the U.S. Environmental Protection Agency Substance Registry Services name where available, and the parameter units of measurement, can be obtained by use of the NWIS website (U.S. Geological Survey, 2018b).

As a part of the NWQP, the National Water-Quality Assessment (NAWQA) Project operates long-term fixed-site monitoring stations. Repetitive water samples from monitored streams are commonly obtained at irregular intervals. These frequencies range from weekly to quarterly samples and may change over extended periods of time. Individual discrete water samples may be formed by compositing multiple point samples throughout a channel cross section. Measurements on these integrated water samples are more likely to represent the average concentration in a stream than in situ point measurements.

Limitations of discrete samples include (1) the expense of mobilizing and deploying a sampling team to a site with the appropriate measurement equipment, (2) the time delay between sampling and receiving the analytical results from a laboratory, (3) possible diel sampling bias due to the need to work in daylight hours for efficiency and safety, (4) possible day-of-week bias due to sampling primarily during week days for economy, and (5) limitation of sampling frequency to one or two samples per hour per team, given the flow or width integrated sampling approach commonly used.

A complementary program of high-frequency and discrete water-quality sampling provides a basis for mitigating some of the limitations of both sampling protocols. Stream water-quality sites are usually co-located at or near USGS streamgages, which provide a continuous record of streamflow. Streamgages are commonly equipped with monitoring and telemetry equipment to measure and transmit environmental data in real time. In addition to streamflow, streamgages monitor water-level information and, in some cases, multiple measurements per day of water temperature, specific conductance, pH, turbidity, and other continuously measurable physical characteristics or properties of water. Some streamgages are equipped to continuously measure surrogate water-quality parameters, such as specific conductance or turbidity, which can be used to estimate parameters of primary interest (Schenk and others, 2016).

Recent progress in instrumentation for in situ water-quality monitoring has expanded the set of continuously measurable parameters, including forms of nitrogen and phosphorus that were previously only measurable in a laboratory. Such high-frequency data can provide insight on dynamic processes occurring over sub diel time scales or document water-quality changes in response to rapidly changing hydrologic conditions.

High-frequency measurements, however, introduce a potential spatial sampling bias because the concentration or property measured at a fixed sampling point may not represent the average property of the stream at that time or may vary systematically from the mean as a function of stream water level. Also, high-frequency water-quality monitoring equipment is expensive to acquire and maintain. In particular, more site visits may be needed to mitigate effects of probe fouling or electronic drift of a continuously monitoring instrument than required for maintaining streamflow-gaging operations. Used together, however, discrete and high-frequency monitoring may improve sampling accuracy and efficiency and mitigate limitations of both sampling methods. Bayesian networks may provide a basis for integrating discrete and high-frequency data by increasing the understanding of the relations among water-quality parameters and identifying how information on one set of parameters may flow to other parameters.

Purpose and Scope

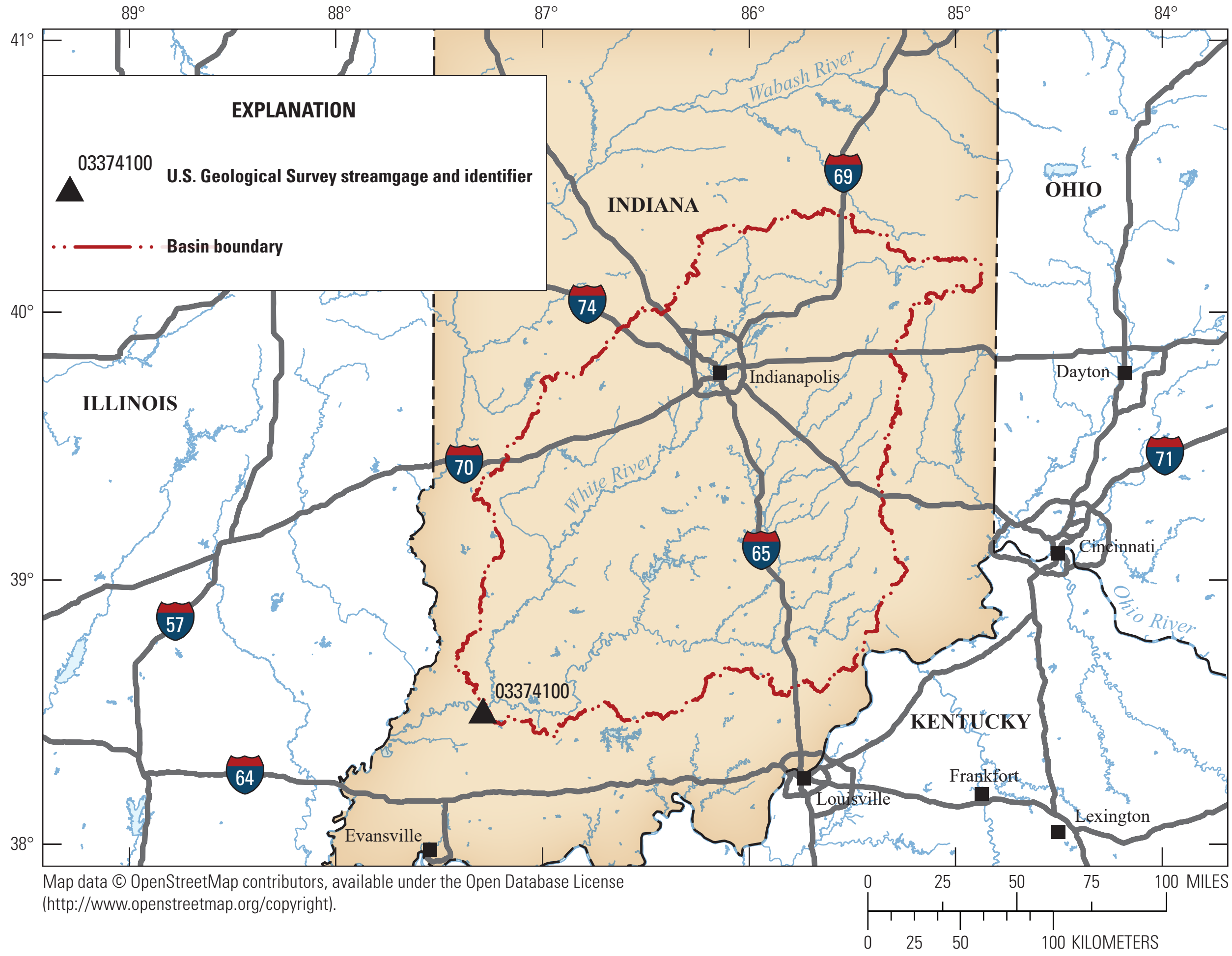
This report describes the development of a discrete BN that depicts the conditional dependency structure and information flow among 27 historically measured water-quality parameters and month at a selected stream site. The utility of the BN is assessed for estimating the magnitudes and uncertainties of an unmeasured subset of these water-quality parameters given only data on a measured complement. Also, analysis of the directional separation (*d*-separated) and *d*-connected of a BN describes how information can flow through a BN conditioned on the directionality of the edges and the measured parameters. This analysis may help identify parameters that are critical to the flow of information and may help prioritize parameters for measurement to facilitate estimation of unmeasured parameters.

The reliance on discrete measurements for development of the BN restricts model applications to static conditions. Although a static BN can be extended to a dynamic BN with time-delay coefficients, sufficient high-frequency data needed to estimate these coefficients are not yet widely available. Although high-frequency data may be used to update corresponding parameters and resultant information propagated through the network, the estimates reflect static relations.

Initialization, training, and verification data for the BN are available separately (Hopple, 2018).

Study Area

U.S. Geological Survey streamgage 03374100 White River at Hazleton, Ind., was selected for this exploratory assessment (fig. 1). White River drains 11,305 square miles (mi^2) of central and southwestern Indiana. The upper basin is situated in part of the Cornbelt and Northern Great Plains aggregate ecoregion; the lower basin drains part of the Southeast Temperate Forested Plains and Hills aggregate ecoregion (Risch and others, 2014). The streamgage monitors flow and water-quality constituents from the White River. From this site, White River flows westward 18.8 miles (mi) and empties into the Wabash River at the Indiana-Illinois border. When flows associated with discrete water-quality measurements were not measured at streamgage 03374100, which was about 16 percent of the time, they were estimated on the basis of flows at streamgage 03374000 White River at Petersburg, Ind., with a drainage area of 11,125 mi^2, which is located 26.8 mi upstream from 03374100. Daily flows at 03374000 generally have greater magnitudes and variabilities from February through May than from August through October (fig. 2).



Map data © OpenStreetMap contributors, available under the Open Database License (http://www.openstreetmap.org/copyright).

Figure 1. Drainage basin and location of water-quality monitoring streamgage 03374100 White River at Hazleton, Indiana, in central-southwestern Indiana.

Methods of Bayesian Network Analysis

The following paragraphs discuss the identification of the directed edge structure connecting nodes in a BN, estimation of conditional probabilities at each node given an edge structure, and the resulting implications for information flow through the network.

Structure of Bayesian Networks

A BN is a probabilistic graphical model that has an information-flow structure described by a directed acyclic graph (DAG) (Koller and Friedman, 2009). In this report, the model is depicted graphically by nodes representing water-quality parameters and by directed edges (links) indicating statistical associations among parameters. This directionality enables estimation of the magnitudes and uncertainties of

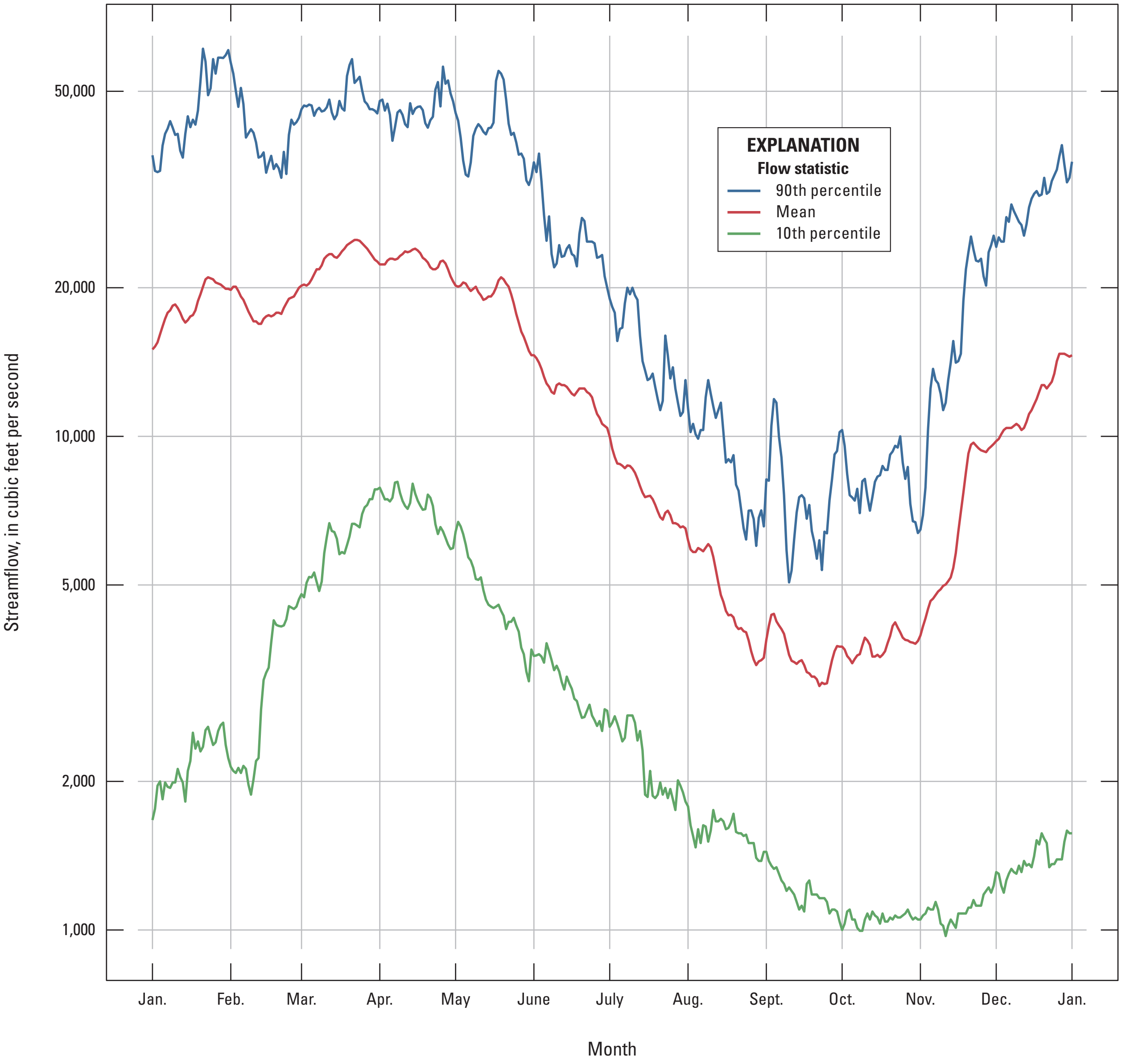


Figure 2. Daily flow characteristics at U.S. Geological Survey streamgage 03374000 White River at Petersburg, Indiana, from 1929 to 2014.

unmeasured parameters on the basis of measured parameters and conditional probabilities but prevents cycles in the graph, which make it impossible to return to an originating node by following the *d*-connected edges.

Nodes may be classified by their edge configuration. A node without an in-coming edge is a root node, which has no parent (predecessor) nodes. Thus, parameters associated with root nodes are not conditionally dependent on other parameters but are described by their marginal distributions. Root nodes commonly may have one or more child (descendent) nodes, which are connected by out-going edges. Child nodes have one or more parent nodes and may themselves have descendent nodes. Terminal nodes have one or more in-coming edges but no out-going edges. Nodes without edges are considered to be disconnected from the network.

Degree is a measure of the centrality of the node in the network and is determined by the total number of incident edges. Thus, water-quality parameters for nodes with high degree centrality are associated with more parameters than those of lower degree centrality. Knowing which parameters have a high degree of centrality may help prioritize parameters for measurement. Degree centrality can be refined into in-degree and out-degree measures by considering the directionality of the edges. The maximum degree of a graph is the degree of the highest of any node in the graph, and the minimum degree is the degree of the lowest of any node.

Discrete Probabilities and Discretization of Water-Quality Data

A BN can use either a continuous probability density function (PDF) or, as in this report, a discrete probability mass function (PMF) for probabilistic computations. The PMFs accumulate probability densities depicted in PDFs within a fixed set of intervals that span the range in water-quality parameters. Each discrete interval in a PMF has a non-negative probability, with probabilities summing to one over all intervals. The PMFs that are conditioned on the multiple intervals of one or more other parameters are referred to as conditional probability tables (CPTs).

Mathematical operations on PMFs are much faster than comparable integrations over PDFs, especially when these distributions are conditioned on one or more parameters. Discretization does result in some loss of precision, particularly when the discretization is coarse relative to the number of measurements. A discretization that is too fine for the variability and number of observations, however, can produce irregular (noisy) CPTs that could degrade the estimation process.

Finding the optimal number of intervals and procedures for discretizing continuous variables is one of the more challenging tasks in developing discrete BN and is still an active area of research. Alameddine and others (2011) experimented with 3 to 7 bins per variable and determined that changing the number of bins had a profound effect on the identified edge structure; for example, more discretized intervals resulted in having fewer directed edges in the network.

The distribution of parameter values in root nodes is not conditioned on other parameters, so it is simply a (marginal) PMF, which can be visualized as a histogram or a table with one row and k columns, corresponding to each interval of discretization for the parameter. In this report, the range of each parameter was initially subdivided at five quantiles using probabilities from 0 to 1 by quantile increments of 0.2. Thus, the range in parameter values within each interval would likely vary, but the average probability within each interval is nominally constant.

The distribution of parameter values in descendent nodes is described by a CPT, where the number of conditioning variables is determined by the number of edges originating from parent nodes. For a child node with one parent (incoming edge), the CPT would contain k_{parent} rows and k_{child} columns. For example, a water-quality parameter discretized into five intervals that was conditioned on month, which discretizes a year into 12 intervals, would have a conditional probability table containing 12 rows and 5 columns, where each row would sum to one. Similarly, a child node, in which the parameter range was discretized into 5 intervals and was conditioned on month (12 intervals) and a water-quality parameter discretized into 5 intervals, would form a probability table with 60 rows and 5 columns. Thus, the number of elements in a CPT grows rapidly with the number of conditioning variables.

Information Flow and Directional Connectedness

A DAG provides a basis for characterizing the flow of information through a network. Pearl (2009) presents three rules that describe conditions for information to propagate through a DAG, which are based on the argument that directional connectedness is a sufficient condition for describing information flow. Rule 1 states that two graph nodes are directionally connected (*d*-connected) if there is an unblocked path between them. In a probabilistic sense, information flows through a *d*-connected path. A path refers to a sequence of distinct edges through a set of adjacent nodes. An unblocked path indicates that a path between the two nodes can be traversed without encountering a pair of convergent edges colliding head-to-head at a node. For example, in figure 3*A*, the node sequences $\left[x \rightarrow r \rightarrow s \rightarrow t\right]$ and $\left[t \leftarrow u \leftarrow v \rightarrow y\right]$ are *d*-connected, but $\left[s \rightarrow t \leftarrow u\right]$ is *d*-separated by the head-to-head collision at t. Probabilistically, if t is unknown, then s and u are conditionally independent. Similarly, for node v unknown, information flows from u and y. All adjacent node tuples in a *d*-connected sequence are also *d*-connected.

Rule 2 states that a set of nodes $\left\{x, y\right\}$ are *d*-connected and conditioned on a set of measured nodes Z if there is a collider free path between x and y that does not traverse a member of set Z. In figure 3*B*, the measurement set Z with circled elements $\left\{r, v\right\}$ *d*-separates nodes x and s, and nodes u and y, whereas s and u are *d*-separated by the collider at t according to rule 1. Observing r, however, makes x and s conditionally independent (*d*-separated), whereas observing v makes u and y conditionally independent. From a water-quality network perspective, a measurement at r informs x and s despite the loss of information between x to s. Likewise, information measured at node v can propagate to y and u.

Finally, rule 3 states that if a collider is a member of a conditioning set Z or has a descendent in Z, then the collider no longer blocks any path that traces the collider. In figure 3*C*, the set Z contains circled nodes $\left\{r, p\right\}$. Because the collider at t has descendant p in Z, the path from r to y is unblocked. The path from x to s, however, is still blocked by r in Z, according to rule 2, although information at r can flow throughout the graph except to node p.

Software for Bayesian Networks

The Netica application, version 5.24 for Windows, by Norsys Software Corp. (Norsys Software Corporation, 2016), was used to develop the BN in this report. Netica provides a graphical user interface that facilitates the development and applications. Algorithms for identification of the edge structure and for estimation of CPT from data are provided. Netica also is able to use data from fully and partially quantified sets of observations, where elements of partially quantified observations may include missing, censored, or estimated (interval) values.

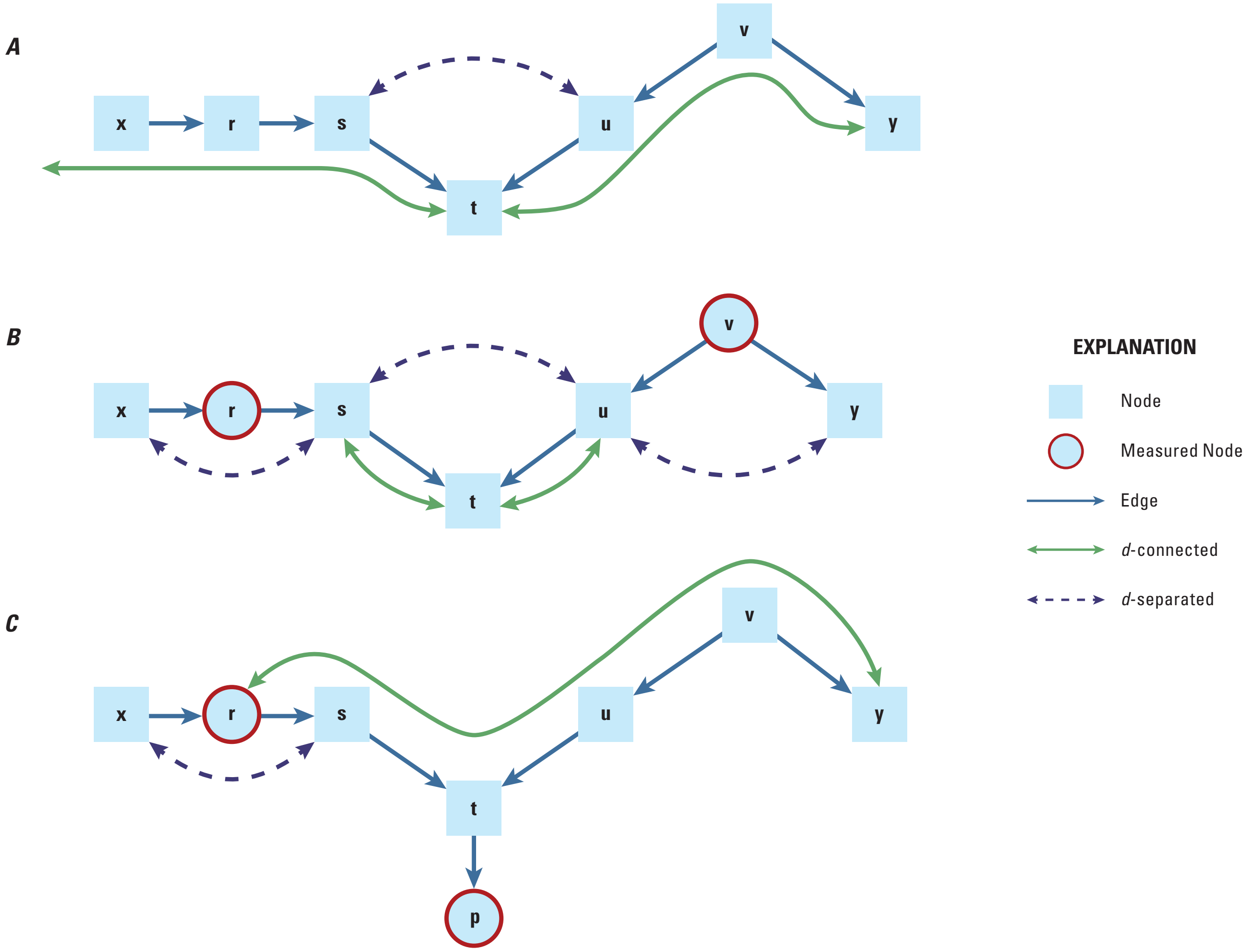


Figure 3. Information flow in a Bayesian network. *A*, A head-to-head collision at node *t* creating *d*-separation between the subgraphs to the right and left of node *t*. *B*, Measurements at nodes *r* and *v* creating *d*-separation between nodes *x* and *s*, and nodes *u* and *y*, respectively. *C*, A measurement at node *p*, which is a descendant of *t*, unblocking information flow at the collider.

Nonproprietary alternatives to Netica include the statistical programming environment R (R Core Team, 2016) packages abn (Additive Bayesian Network) developed by Fraser (2016), and bnlearn developed by Scutari (2010). The alternatives, however, are not known to provide a basis for computing with missing, censored, or estimated observations, which are ubiquitous in water-quality data.

Implementing a Bayesian Network for Water-Quality Data

Implementing a BN involves (1) selecting a set of water-quality parameters with a sufficient number of measurements as nodes in the network, (2) identifying the statistical dependencies (edges) among nodes in a network with a set of directed edges, (3) computing the CPTs for the network, and (4) evaluating the network performance.

Selection of Water-Quality Parameters

Existing water-quality data at streamgage 03374100 White River at Hazleton, Ind., were evaluated to identify parameters that would be likely to contain sufficient information to support the development of a BN. Data selection occurred in two stages. The first stage preselected a subset of all measured parameters on the basis of the number of discrete measurements by use of data summaries from the water-quality data inventory. The second stage retrieved time series of measured parameters for this subset to identify measurement characteristics, such as the number of fully quantified, censored, estimated, and missing values, evaluation of trends, and elimination of potentially redundant parameters.

Water-Quality Inventory Characteristics

The sampling frequency characteristics of water-quality data at streamgage 03374100 were initially characterized on the basis of water-quality inventory data retrieved from NWIS using the function whatNWISdata in the R package

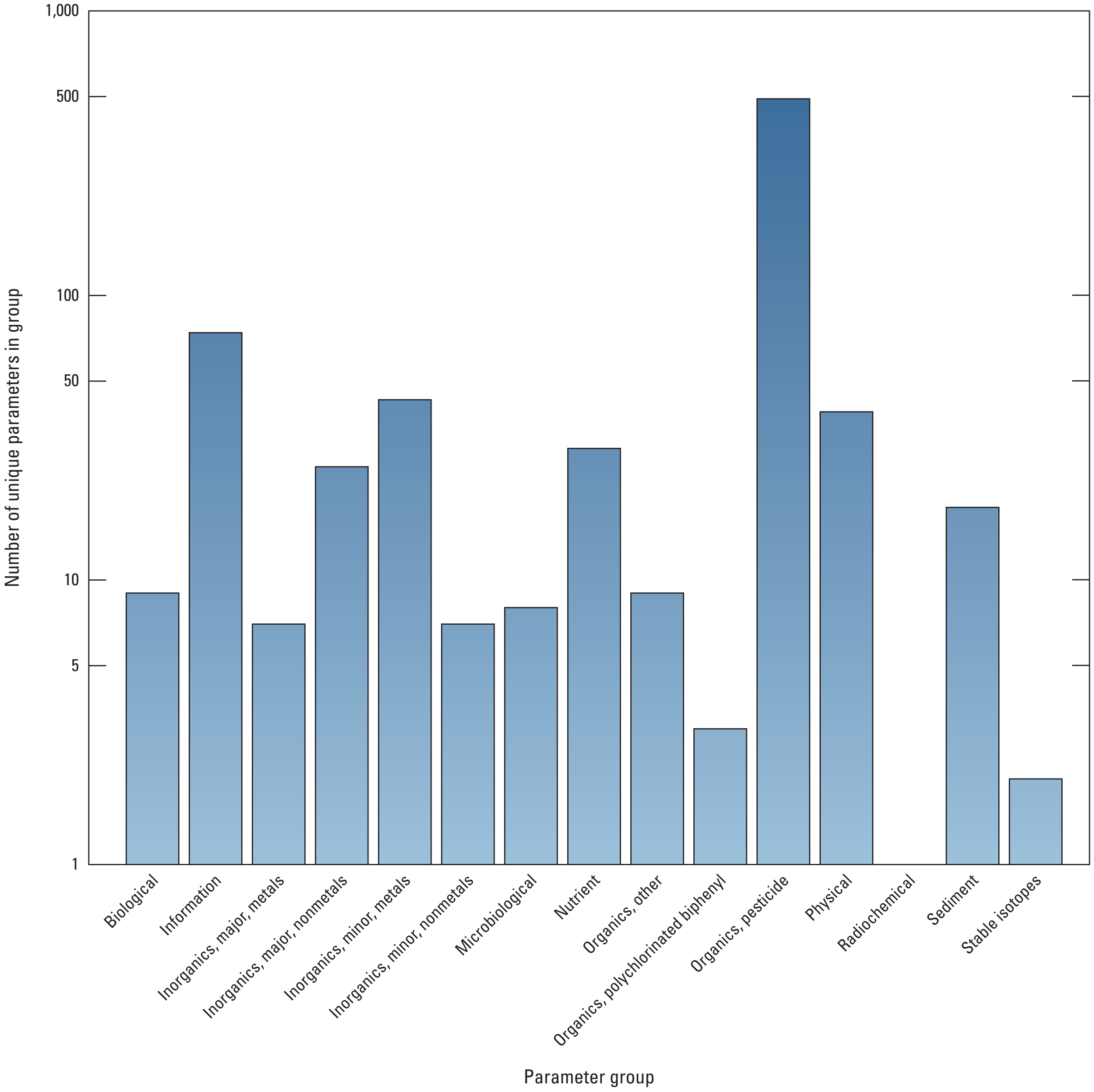


Figure 4. Frequency of unique water-quality parameters within parameter groups at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1973–2016.

dataRetrieval (Hirsch and De Cicco, 2015). This function was applied using parameters *siteNumbers* = '03374100' and *service* = 'qw,' where "qw" refers to data from discrete water-quality samples (U.S. Geological Survey, 2016).

The inventory data contain 29 fields (columns) that summarized measured data for each parameter. Inventory data include descriptors for the medium sampled (*medium_grp_cd*), the name of the parameter group (*parameter_group_nm*) associated with each parameter, the number of times each parameter was measured (*count_nu*) within an interval described by beginning and ending sampling dates (*begin_date* and *end_date*, respectively).

A data retrieval for streamgage 03374100 on January 20, 2017, returned a dataset containing 1,133 observations with 1,083 unique parameter codes. Sampling medium groups (*medium_grp_cd*) included biological ('bio'), sediment ('sed'), and with water ('wat'). Omitting all but the water samples resulted in a dataset with 764 unique parameter codes (*parm_cd*) distributed among 15 parameter groups (fig. 4). Results indicate that organic pesticides contained the largest number of unique parameters (491) measured among all parameter groups. Only 1 unique parameter was included in the radiochemical group, 2 were included in the stable isotopes group, and 3 unique parameters were measured in the organic polychlorinated biphenyl (PCB) group.

At streamgage 03374100, the 764 unique water-quality related parameters were measured one or more times on water samples from 1973 to 2016, potentially providing 91,417 parameter measurements. The distribution of these parameter measurements among 15 parameter groups are

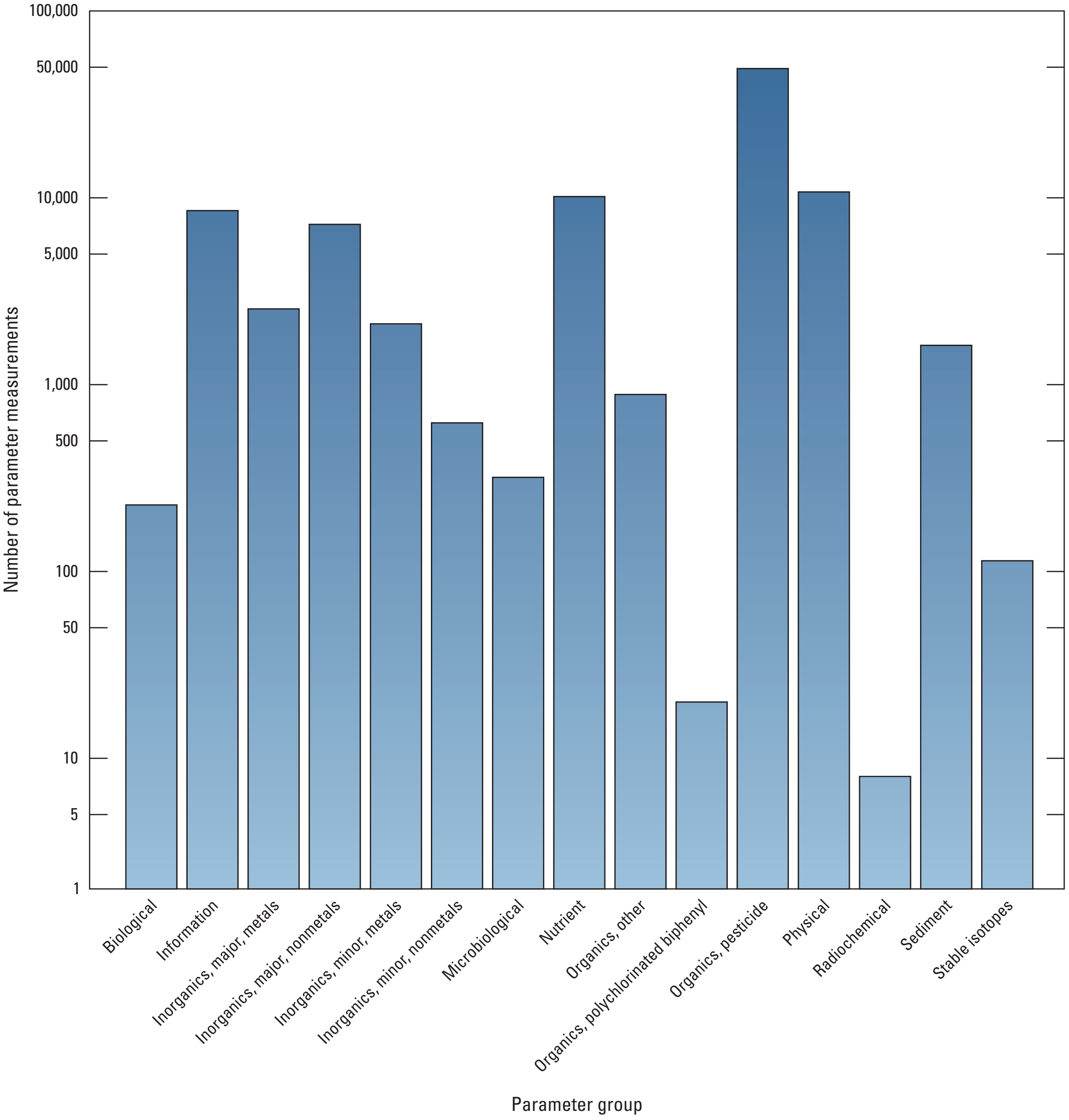


Figure 5. Counts of parameter measurements among parameter groups at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1973–2016.

shown on figure 5. The most frequently measured groups include organic pesticides, with 46,890 measurements; physical parameters, with 10,640 measurements; and nutrients with 10,088 measurements. In contrast, the radiochemical group included 8 measurements, and organic PCB included 20 measurements. The informational (termed "information") parameter group consists of metadata, such as agency analyzing the sample, project number, number of sampling points, site visit purpose, and details that may support analysis of individual results rather than quantify specific water-quality parameters.

The number of repetitive measurements for each parameter varied widely among groups (fig. 6). For example, the average number of repeated measurements of specific pesticides within the organic pesticide group is 95.5, computed by dividing the total number of pesticide measurements, 46,890, by the number of unique pesticides measured, 491. The range of repetitive pesticide measurements varied among individual pesticides from 1 (several compounds) to 554 (atrazine). Some of the more frequently measured parameters occurred within the nutrient, major inorganics non-metals, and physical parameter groups. Metadata associated with parameters from the information group were not used in the development of a BN.

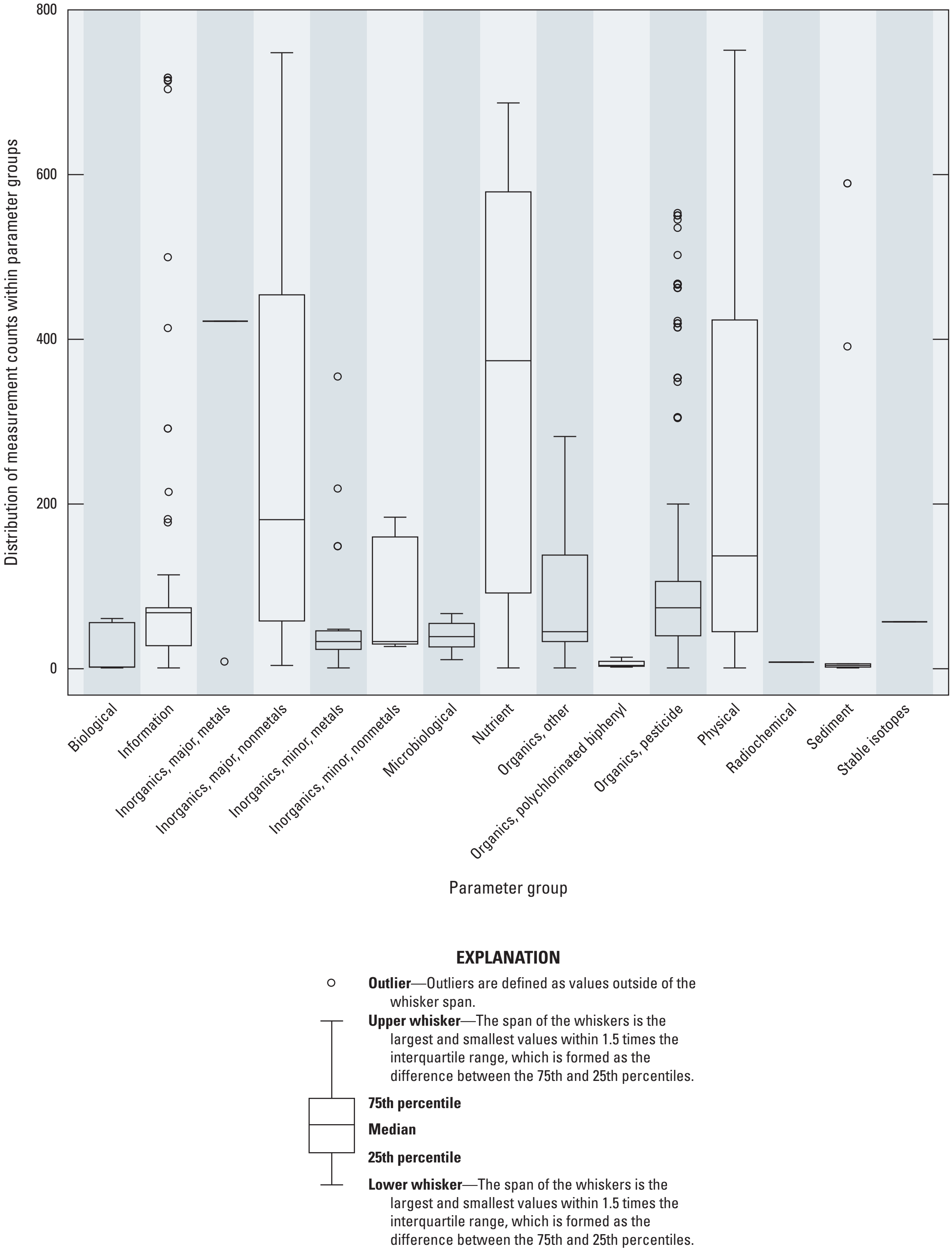


Figure 6. Distribution of repetitive parameter measurements within parameter groups at U.S. Geological Survey streamgage 03374100 White River near Hazleton, Indiana, water years 1973–2016.

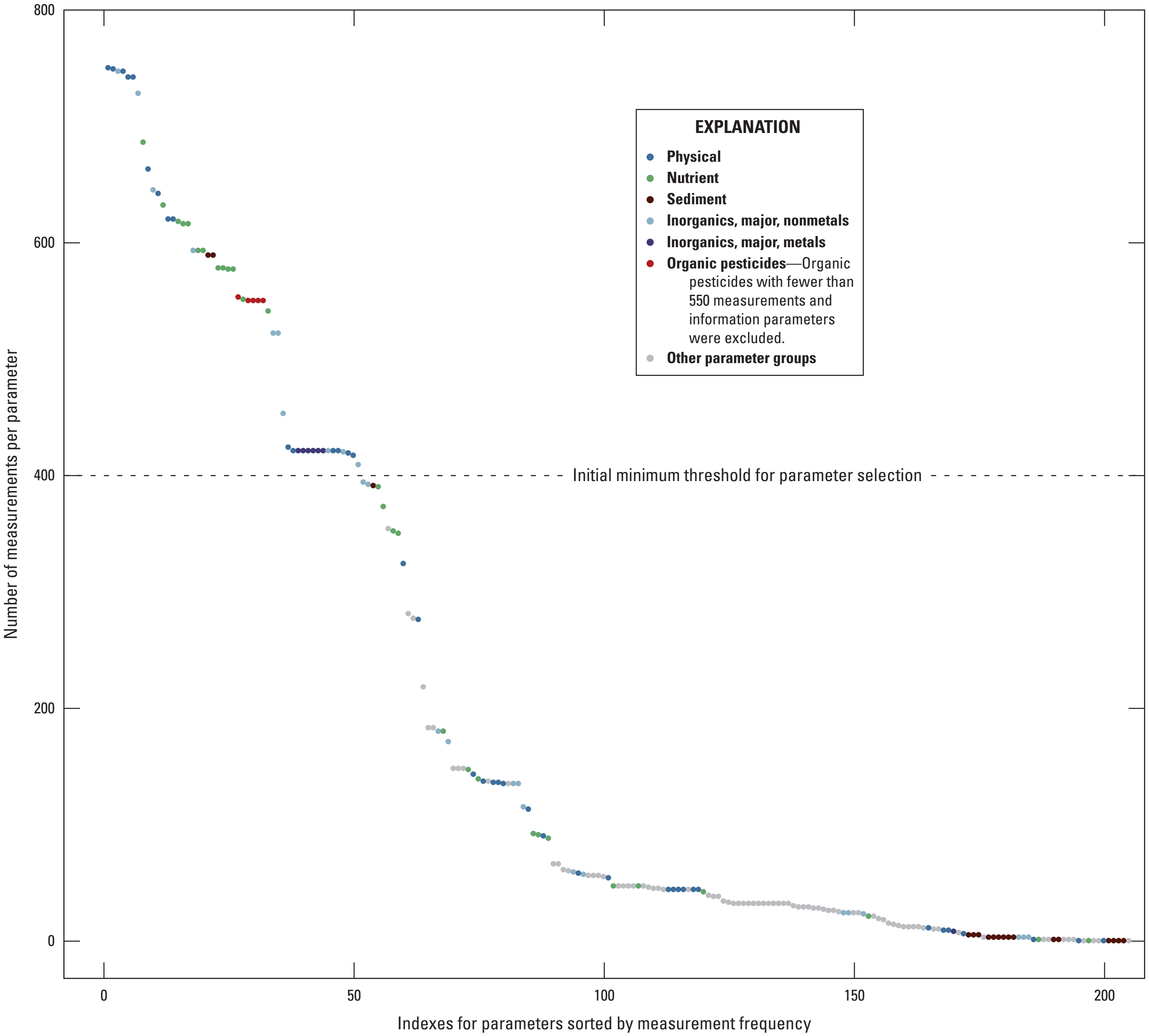


Figure 7. Sorted frequencies of individual parameter measurements color-coded by parameter group for samples collected at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1973–2016.

The ability to characterize statistical interrelations among water-quality parameters in a BN is dependent on the availability of a sufficient number of measurements in which sets of parameters have been measured on the same water sample. Although this information is not directly available from the inventory, a minimum frequency of measurements based on a measurement count greater than 400 field was used as a minimum threshold for initial selection of parameters. A total of 51 parameters satisfied this threshold (fig. 7).

Additional details on the 51 parameters that were initially selected on the basis of the inventory data are shown in table 1.1. These 51 parameter codes were considered for possible redundancies. For example, instantaneous discharge, in cubic feet per second (parameter 00061) was considered to be redundant with instantaneous discharge, in cubic meters per second (parameter 30209). Similarly, instantaneous discharge was considered to be redundant with gage height, in feet (parameter 00065), and with gage height above datum, in meters (parameter 30207), because of the curvilinear relation between gage height and discharge. Thus, for the purpose of this report, instantaneous discharge, in cubic meters per second (parameter 30209), and gage height, in feet or meters (parameters 00065 or 30207) were considered redundant with instantaneous discharge, in cubic feet per second (parameter 00061), and omitted from the BN. Redundancies outlined in table 1.1 identified 17 parameters that were omitted from the BN.

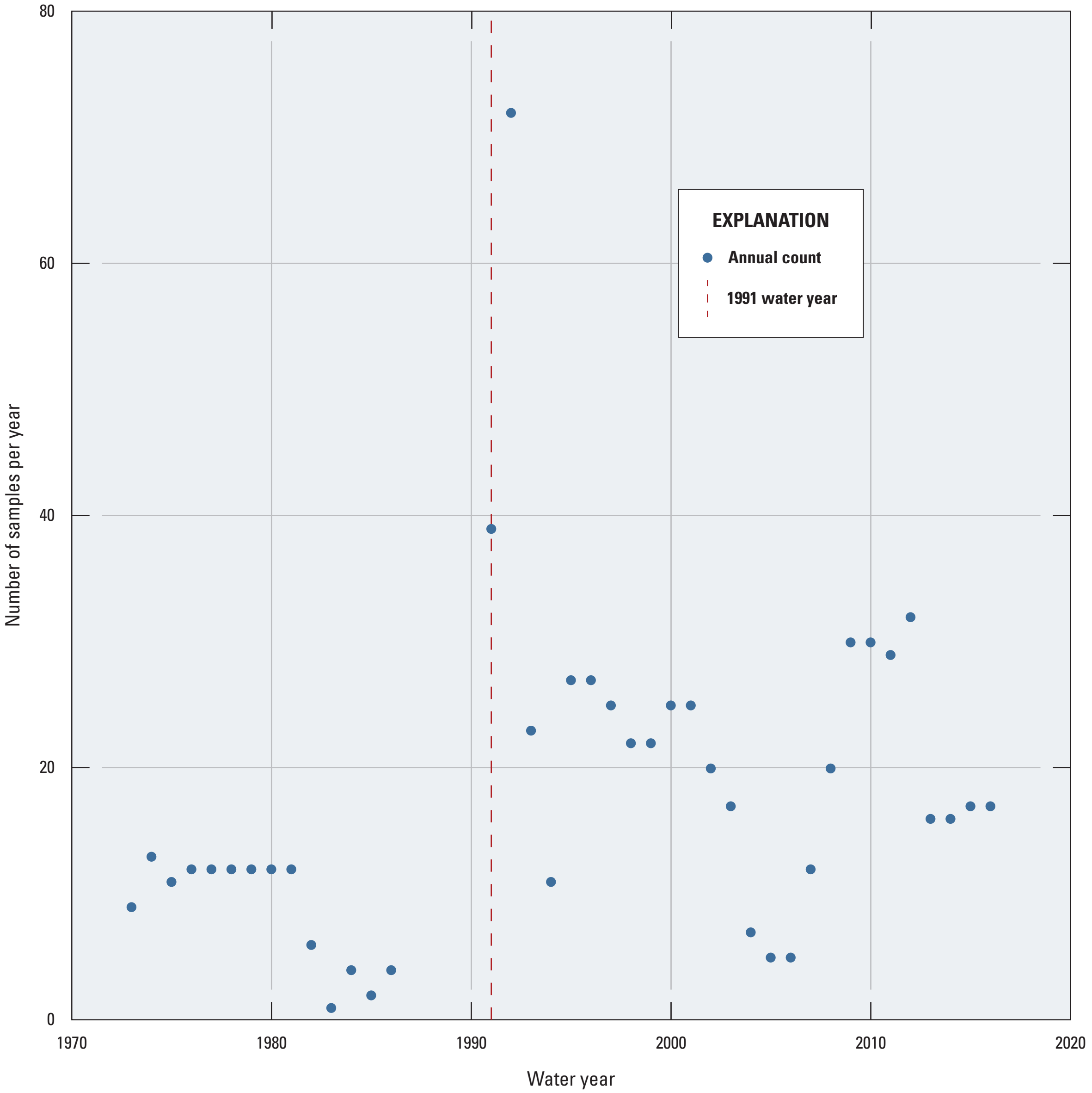


Figure 8. Frequency of discrete water-quality samples at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1973–2016.

[“discrete water-quality samples” refer to a sample_type_code equal to 9 (https://help.waterdata.usgs.gov/code/sample_type_query?fmt=html), with a medium_code of ‘WS’ (https://help.waterdata.usgs.gov/code/medium_cd_query?fmt=html)]

Water-Quality Time Series Characteristics

The readNWISqw function from the dataRetrieval package was used to retrieve time series of discrete water-quality measurements from NWIS for streamgage 03374100 using 34 of the 51 parameters that were not considered redundant. The argument ‘*startDate*’ was specified as ‘1973-01-01’ and ‘*endDate*’ was specified as ‘2016-09-30,’ with ‘expanded’ and ‘reshape’ set to true. The retrieval returned a table of time series data with 839 rows and 496 columns.

Among the time series information data returned was the medium code (‘*medium_cd*’) U.S. Geological Survey (2011b), which described the medium sampled. Based on this code, 832 surface-water samples, 3 bottom-material samples, 3 plant tissue samples, and 1 animal tissue were obtained. Only the surface-water samples were retained for analysis. In addition, sample type codes (“*samp_type_cd*”), U.S. Geological Survey (2011c), included 713 regular water samples (code 9), 116 replicate samples (code 7), and 3 composite samples (code H). Only the regular water samples were retained.

The varying frequencies of water-quality sampling from 1973 to 2016 are shown in figure 8 for the 713 regular surface-water samples analyzed in this report. Data from 1973 to 1981 were obtained at a median rate of 12 times per year but diminished to about 4 times per year from 1982 to 1986. No samples were obtained from 1987 to 1990. Although frequencies varied from 5 to 72 samples per water year, the median frequency increased to 22 samples per year from 1991 to 2016. A water year refers to the 12-month period beginning on October 1 and ending on September 30; the year is designated on the basis of the calendar year in which the water year ends.

Quantified, Censored, Estimated, and Missing Data

For each retrieved water-quality parameter, the table of time series data included a column quantifying the concentration or level for each parameter, identified as ‘p[#####],’ and a corresponding column of remarks identified as ‘r[#####],’ where [#####] indicates the five-digit parameter code (U.S. Geological Survey, 2011a). If the results column contained a numerical value and the remark column contained a missing value indicator (*NA*), then the result column was interpreted as a fully quantified numerical value. If the result column contained a numerical value and the remark column contained a less than value indicator (‘<’), the result column was interpreted as upper censoring limit for the parameter, and a value of zero was assumed for the lower censoring limit. Alternatively, if the remark column contained a greater than value indicator (‘>’), the result column was interpreted as the lower censoring limit for the parameter. An estimated value indicator (‘*E*’) in the remarks column was interpreted as a result with uncertainty, which might be expressed as an interval containing the reported value. The code ‘*M*,’ which indicates that the presence of a parameter was verified, but not quantified, was coded once for oxygen (parameter code 00300) on October 15, 2002, and for carbon dioxide (00405) in samples on December 12, 2007, and January 23, 2008. In this report, the three ‘*M*’ codes were treated as missing values. If both the result and remark fields contained missing value indicators, the parameter was considered a missing value.

The type and amount of data qualifications and missing values are shown by parameter code in figure 9. Water-quality parameters of water temperature (*w_temp*, parameter 00010), specific conductance measured in the field (*SpecCond*, parameter 00095), and pH (*pH*, parameter 00400), respectively, had the most fully quantified measurements, while specific conductance measured in the laboratory (parameter 90095), bicarbonate (parameter 00453), the pesticide Alachlor (parameter 46432), organic nitrogen (parameter 00605), ammonia and ammonium as ammonium, NH4, (parameter 71846), ammonia and ammonium as nitrogen (parameter 00608), and the pesticide Metribuzin (parameter 82630) had fewer than 330 fully quantified measurements. The 27 water-quality parameters having more than 330 fully quantified values after October 1, 1990, were used in subsequent analysis.

Statistical Characteristics of Water-Quality Parameters

Risch and others (2014) analyzed water-quality trends in selected nutrients, metals, and ions in Indiana streams. In their statistical analysis of data from 2000 to 2010, no statistically significant trends were detected for streamgage 03374100 White River at Hazleton, Ind. in any of the constituents tested, including annual median nitrate or organic nitrogen, phosphorus, suspended solids, copper, iron, lead, chloride, zinc, hardness, sulfate, or dissolved solids concentrations.

For selected parameters used in the development of a BN in this report, measurements of barometric pressure, nitrite and nitrate nitrogen, orthophosphate, simazine, metolachlor, atrazine, and alkalinity are generally sparse or non-existent before 1991, whereas other parameters were measured more consistently during the period of record. Some parameters, such as fluoride and orthophosphate, show decreased concentration granularity (attributed to increased precision of measurements) for samples collected beginning October 1, 1990.

In particular, water temperature, specific conductance, pH, and dissolved oxygen are fairly uniformly distributed over the vertical axis, considering annual variations, whereas discharge, simazine, metolachlor, and atrazine appear more frequently at lower concentrations, indicating a positive skewness coefficient. Censored and estimated data are common in nitrite and nitrate nitrogen concentrations and orthophosphate. One measurement of dissolved oxygen and two measurements of carbon dioxide were indicated by the qualifier ‘*M*’ as present but not quantified. These data were treated as missing values. One measurement of barometric pressure of 10.20 inches of mercury, measured on December 12, 1985, was deleted as a low outlier. The outlier classification was based on an inner 99 percent quantile range (from 0.5 to 99.5 percentiles) of barometric pressures from 726 to 783 millimeters of mercury for 639 pressure measurements at the streamgage. It is possible that a leading ‘7’ in the pressure value was mistaken for a ‘2.’

Identifying a Bayesian Network for Selected Water-Quality Parameters

In this report, identifying a BN for the selected water-quality parameters refers to a two-step process that includes identifying an appropriate set of directed edges connecting nodes (the edge structure), and estimating the corresponding CPT, by use of water-quality data. The minimum number of edges in a network is zero, implying that all parameters are independent, and the maximum number of directed edges, $\#E_d^{\max}$, is $\#E_d^{\max} = \left(\left(\#N \cdot \left(\#N - 1\right)\right)/2\right)$ for *#N* parameters. For a network with 28 nodes, this maximum is 378 edges, implying that all parameters are conditionally dependent. The contents of CPTs are determined by the edge structure.

The following paragraphs discuss the format of case files of water-quality data used for identification, the parameter-specific process for identifying the edge structure of the DAG, the estimation of corresponding CPTs, and the selection of a target parameter for the BN.

Data for Development of the Bayesian Network

To provide greater consistency in the temporal and statistical distribution of measured parameters used to develop the

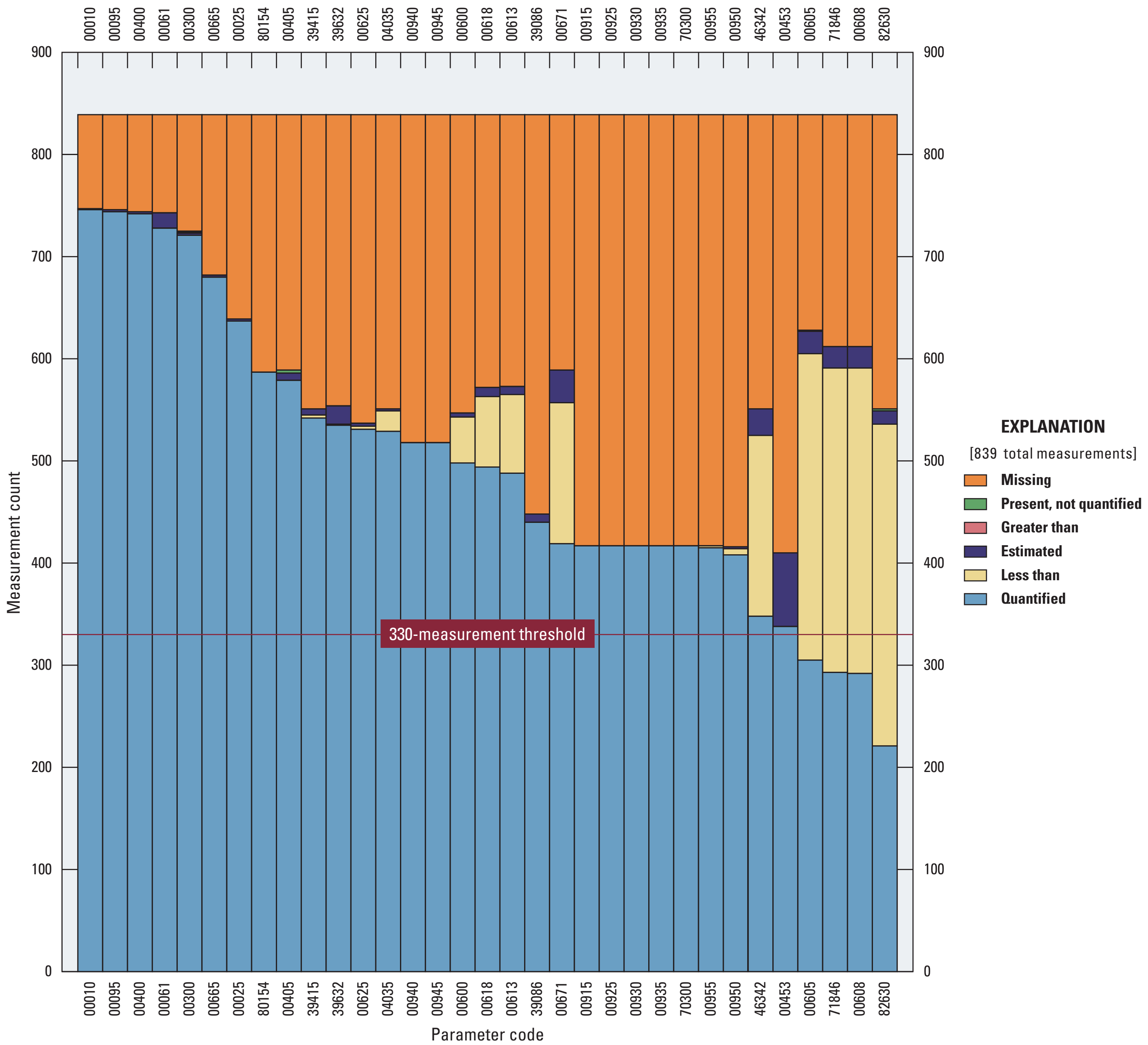


Figure 9. Distribution of qualifying codes and missing values among selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, 1990–2016 (full parameter names and parameter codes are given in table 1.1).

BN, only data measured on or after October 1, 1990 (beginning of water year 1991) were used in this report. Due to the timing of this study, no data after September 30, 2016 (end of water year 2016), were included. The resulting dataset contained 596 observations on 27 parameters, not including the month indicator. Within this matrix, about 64.8 percent were fully quantified, 32.4 percent were missing values, and 2.7 percent were censored or estimated values.

To provide data for developing and testing the BN, the available 596 observations of water-quality parameters were randomly partitioned into a training dataset of 479 observations (80 percent) and a testing dataset of 117 observations (20 percent). This partitioning provides independent sets of data to estimate the parameters of BN and to assess the accuracy of estimation. Without this portioning, it can be difficult to control or assess the potential overfitting of the model to the idiosyncrasies of a particular dataset. A USGS data release prepared by Hopple (2018) can be used to access the data used in the development of the BN.

Case Files with Water-Quality Parameter Data

Netica can use measured data formatted as case files to identify the network structure and to compute CPTs. Netica's case files are formatted with the American Standard Code for Information Interchange (ASCII) and are tab-delimited with a header row identifying columns of data. The first column is the observation number (IDnum) followed by a set of identifiers corresponding to the nodes in the BN. In this report, the parameter short name (table 1.1), including *month*, *w_temp*, *PO4*, and so forth, are the node identifiers. Each row after the header represents the water-quality parameters measured for an individual discrete sample. Numerical values represent fully quantified results. Missing values are indicated by an asterisk. Censored values are preceded by a less-than (<) or greater-than (>) symbol preceding the numerical censoring limit. Estimated values are represented by a pair of numerical values between braces '[*low, high*].' Interval estimates are based on empirical quantiles computed using fully quantified values for the parameter. The *low* estimate is the inverse empirical quantile at the max (0, empirical quantile of the estimate – 0.1). Similarly, the *high* estimate is the inverse empirical quantile at the min (1, empirical quantile of the estimate + 0.1). The selection of a plus or minus 0.1 quantile about the quantile of the estimated value was selected to illustrate model flexibility. Interval estimates based on the expected percent error would also have been viable.

Identifying the Edge Structure of a Bayesian Network

Within the Netica (version 5.24) application, the process for initiating a network project begins with selecting the menu item 'File,' then the subitem 'New,' and then the subitem 'Network.' A blank window labeled 'Untitled-1' was displayed. To initialize the network with the selected parameters and set the levels of discretization for each parameter, the menu item 'Cases' was first selected, then the subitem 'Learn,' and then the subitem 'Add Case File Nodes…' A window labeled 'Case file to obtain nodes from:' is shown, where the list of selectable files in the current directory displayed. In this report, the file baye_network_initialize.cas was selected to initialize the network. This case file specifies a range of parameter values that spans the range for all parameters in both the training (baye_network_training.cas) and testing (baye_network_testing.cas) datasets. These case files are accessible in the USGS data release by Hopple (2018).

After selecting baye_network_initialize.cas file, an unlabeled pop-up window appears displaying the text 'How many states would you like continuous node *month* to have (0 for no discretization)?' The default value of '5' was replaced with 12, and the 'OK' button was selected. A pop-up window labeled 'Netica' contained the message: 'You have requested 12 states for node month, but only 12 unique values were observed. Would you like this to be a DISCRETE numeric node, instead of continuous?' The button labeled 'YES' was selected, and an unlabeled pop-up displayed the text 'How many states would you like continuous node *w_temp* to have (0 for no discretization)?' The default value of 12 was replaced with 5, so that 5 intervals at the quintiles of distribution of water temperature were created. A value of 5 was selected as the discretization level for all parameters other than month.

Netica provides the Tree Augmented Naïve (TAN) Bayes algorithm (Friedman and others, 1997; Chow and Liu, 1968) to identify the edge structure. To apply the TAN algorithm, a specific target node must be selected, where the target node corresponds to the parameter of primary interest. In the BN developed in this report, no parameter was considered to be of primary interest, where the objective was to provide an estimate of the magnitude and uncertainty of an arbitrary set of parameters given data on the measured complement; therefore, each parameter in the BN was, in turn, used to identify 28 alternative edge structures.

As an example of identifying the network structure by use of training data, the *month* node was arbitrarily selected (by left clicking on the node corresponding to month), where by default the node is displayed as a table icon with 12 rows in column 1, labeled 1 through 12 and corresponding to months from January to December. The second column of the node icon contains the value 8.33, which corresponds to the expected frequency of data for each month. Then, 'Case' menu item is re-selected, followed by the subitem 'Learn,' and then the subitem 'Learn TAN structure.' After this subitem was selected, a pop-up window labeled 'Case file to learn structure from:' appeared with the list of files in the directory, from which the file baye_network_training.cas was selected. If a pop-up window labeled 'Netica error #3550' is displayed with the message 'You must select a single node to classify, or there must be a node-set named 'target' containing only it.', click the button labeled 'OK,' select the target node (*month*), and re-read the training case file as described above. The view of the BN was updated with a set of 53 directed edges displayed among the 28 nodes.

Estimating Conditional Probability Tables

Once a set of edge structures was identified by taking each parameter, in turn, as the target node, the corresponding conditional probability tables were computed. In Netica, the menu item 'Cases' was selected, followed by the subitem 'Learn,' and then by the item 'Learn by EM.' Within Netica, a pop-up window labeled 'Case file to learn CPTs from:' is displayed along with a list of selectable files in the current directory. In this report, the file 'baye_network_training.cas' was always selected. Then, Netica displays an unlabeled pop-up window that displays the text 'Enter degree (normal is 1):' with a text box for data entry. In this report, the default value of 1 was used for all parameters. Once the 'OK' button is pressed on this pop-up window, the iterative expectation maximization (EM) algorithm computes CPTs for all nodes and a negative log-likelihood for the corresponding model. Progress of the EM algorithm in minimizing the negative log-likelihood for each iteration can be viewed in Netica by selecting the menu item 'Report' and the submenu 'Network.'

The table icons representing nodes are updated with horizontal bars showing the percentage of data for each of the nominal quantile limits displayed in the first column. Left clicking on any table icon will set the observed quantile for that parameter and update the distribution of conditional probabilities in all child nodes. Quantiles in multiple nodes can be set by left clicking, which updates CPTs in all child nodes.

In the absence of an *a priori* preferred target node, the edge structure of the BN was selected on the basis of the target node with the minimum value of the converged negative log-likelihood value (fig. 10). Based on these results, the most likely (smallest negative log-likelihood) BN for the 28 selected parameters had the *month* parameter selected as the target node. Because the *month* parameter is associated with seasonal changes in temperature, light, precipitation, and cultural practices, such as farming, it is likely associated with numerous water-quality processes.

The selected BN is depicted in figure 11. The network with 53 directed edges, which is about 14 percent of the number of edges in a complete graph, has all nodes connected by an edge. The network of water-quality parameters is connected in that all nodes are connected to the network by one or more edges.

The *month* node has an out-degree of 27 for the 27 edges directed from the *month* node, making it a parent to all other nodes in the network. There are no parent nodes to *month*, making it the only zero in-degree (root) node in the network. The parameter silica (*Si*) has the next highest out-degree of three. Specific conductance (*SpecCond*), sulfate (*SO4*), pH, magnesium (*Mg*), calcium (*Ca*), and phosphorus (*P*) all have the next highest out-degree of two. Of the remaining nodes, 6 have an out-degree of one, and 14 have an out-degree of zero, which are referred to as terminal nodes. With the exception of water temperature (*w_temp*), which has one parent node (*month*), all other water-quality parameters have an in-degree of two. Given the 53 directed edges in the BN, estimation of a total of 7,872 elements in the 28 CPT was required.

Although not explicitly indicated in the Netica documentation, it may be the case that the identification algorithm limits the maximum in-degree for all nodes to two, at least for the number of observations available in this training dataset. Such a restriction would limit the number of CPT elements, which increase rapidly with the in-degree of the node. This apparent limit for in-degree of 2 can be exceeded by adding edges manually. Given that this report focused on developing an exploratory BN to assess the general feasibility of the modeling approach, no further refinement of the network was attempted.

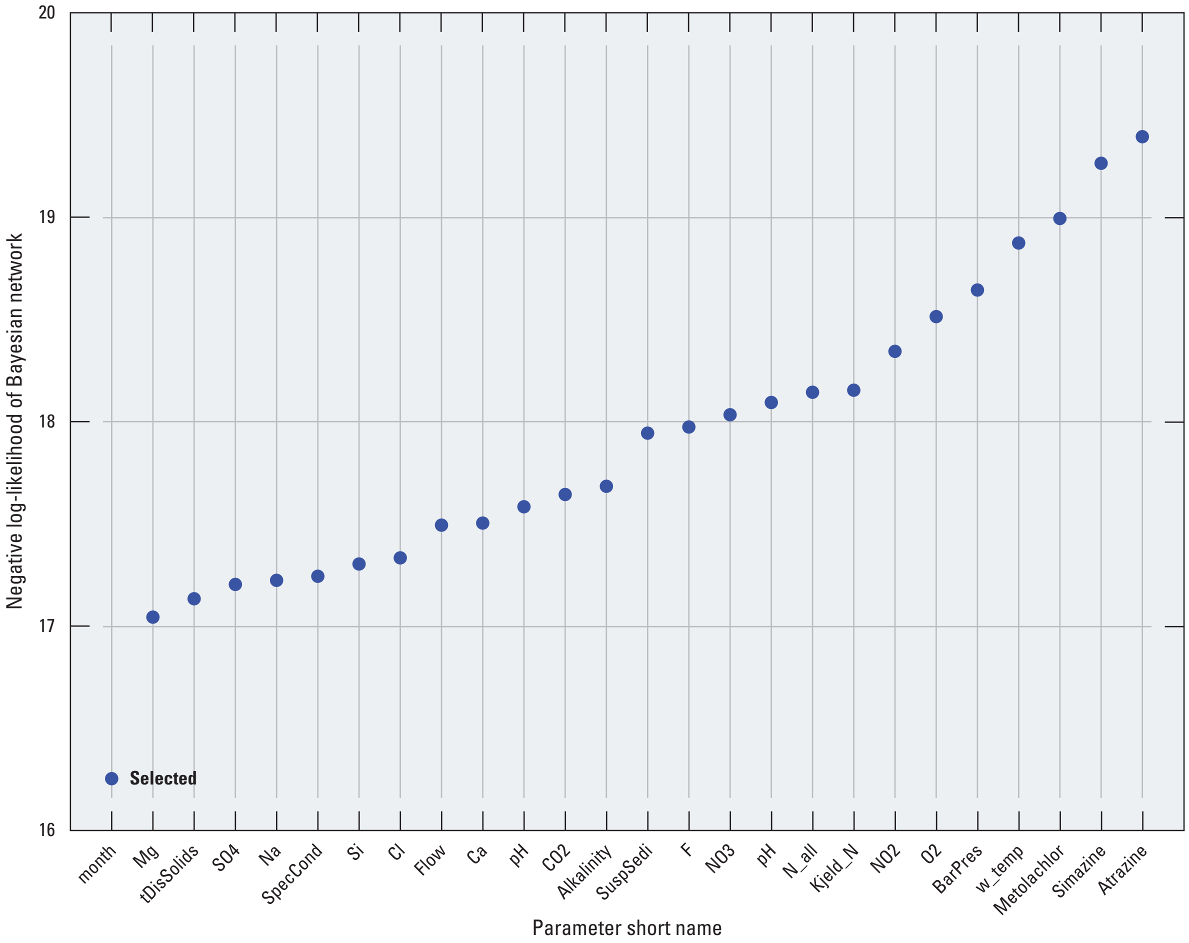


Figure 10. Negative log-likelihoods of the Bayesian network as a function of the selected target node (full parameter names and parameter codes are given in table 1.1).

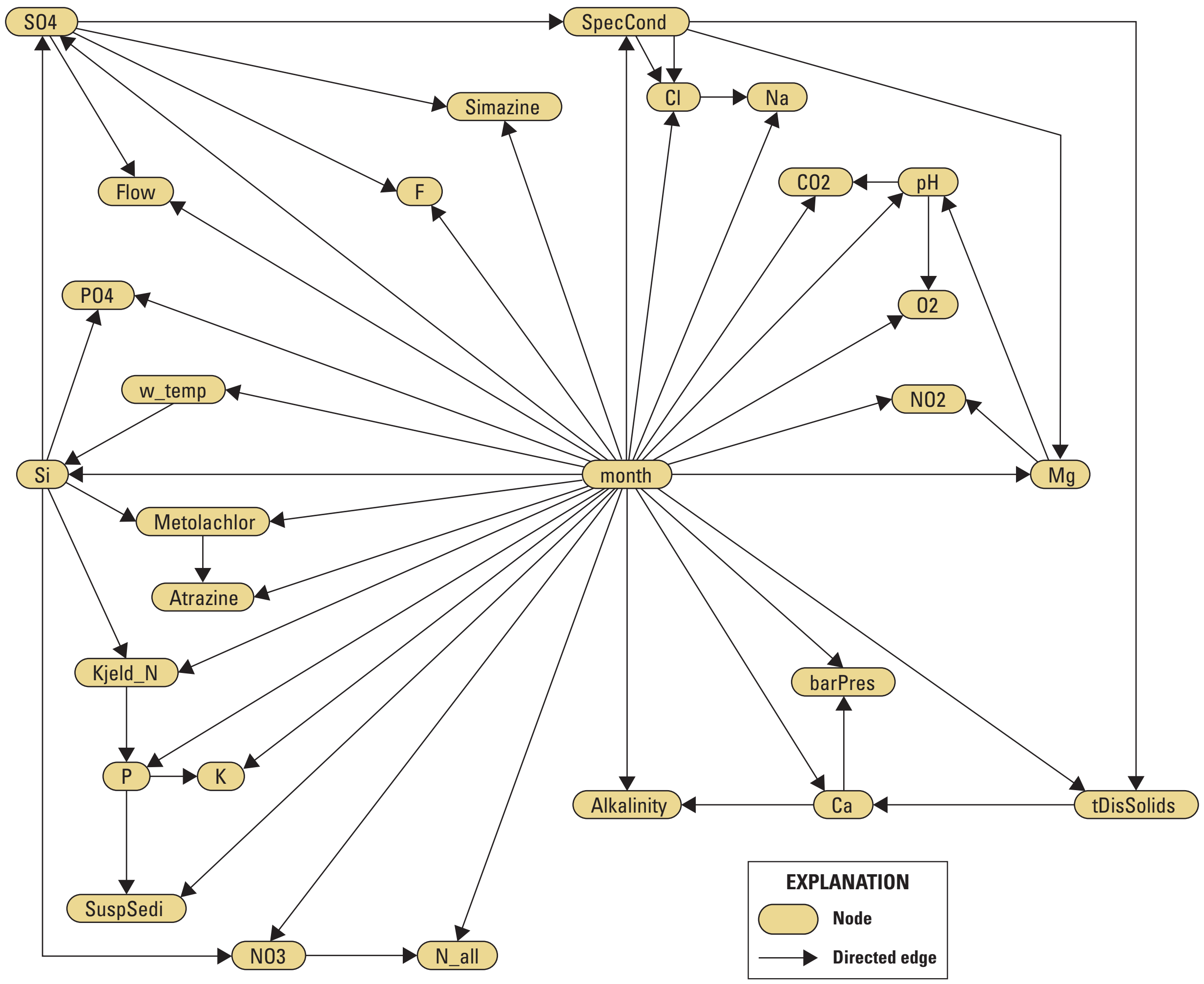


Figure 11. Directed acyclic graph of the Bayesian network developed to describe associations among selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, water years 1991–2016 (full parameter names and parameter codes are given in table 1.1).

The arrangement of edges and levels of discretization for individual parameters results in 28 CPTs with a total of 7,872 elements. Elements of the CPTs are computed during the EM step of model development. An example of a CPT for this network for nitrate nitrogen (*NO3*) concentration decile ranges conditioned on month number and silica (*Si*) concentration decile ranges is shown in table 1.1.

Computing Magnitudes and Uncertainties of Selected Parameters

This section provides the equations and a numerical example for computing the magnitudes and uncertainties of water-quality parameters on the basis of the edge structure and conditional probability tables estimated in the development of the BN. The magnitude will be estimated on the basis of the expected value, and the uncertainty will be expressed as the standard deviation.

Conditioning water-quality parameters for any parameter of interest can be identified by inspection of the BN. For example, if nitrate nitrogen (*NO3*) was selected as the parameter of interest, then on the basis of the BN (fig. 11), it is apparent that month (*month*) and silica (*Si*), are the two conditioning parameters. To characterize the conditioning parameters, let the month of measurement be August, *month* = 8, and consider that a silica concentration of 2.9 milligrams per liter (mg/L) was measured. Based on the discretization of silica concentrations, the measured value falls within silica interval 2, represented as $[2.9]_2$,which spans concentrations from 2.4 to 4.8 mg/L (table 1.3). For convenience, the corresponding row from table 1.3 is reproduced below as table 1, with probabilities converted to decimal from percent.

Table 1. Conditional probabilities for nitrate nitrogen concentrations for the month of August and a silica level ranging from 2.4 to 4.8 milligrams per liter.

[xmid in the table header refers to the midpoint of the silica concentration range within the discretized interval (quintile), in milligrams per liter (mg/L); the month number of 8 refers to August; dimensionless probabilities are reported in the body of the table; ~, approximately]

		Probability, conditioned on month equals 8 and silica range in level 2				
Month (number)	**Silica range (mg/L)**	**0.0 to 0.4 (xmid = 0.2)**	**0.4 to 1.2 (xmid = 0.8)**	**1.2 to 1.9 (xmid = 1.55)**	**1.9 to 2.4 (xmid = 2.15)**	**2.4 to 6.8 (xmid = 4.6)**
8	2.4 to 4.8	0.7048	0.1478	0.1474	$\sim 10^{-6}$	$\sim 10^{-6}$

The expected value, $E[\bullet]$, as an indicator of magnitude, is computed using the following equation:

$$E\left[NO3 \mid month = 8,\ Si = [2.9]_2\right] = \sum_{k=1}^{5} xmid_k \cdot p_k = 0.488 \text{ mg/L} = \\ 0.2 \cdot 0.7048 + 0.8 \cdot 0.1478 + 1.55 \cdot 0.1474 + 2.15 \cdot 10^{-6} + 4.60 \cdot 10^{-6} \tag{1}$$

where

$xmid_k$ is the midpoint of the k^{th} nitrate interval, and

p_k is the k^{th} conditional probability from table 1.

The variance, $Var[\bullet]$, as a measure of uncertainty for *NO3* conditioned on the month and the silica interval, has two components. The first component characterizes the uncertainty of *NO3* being observed in any of the five intervals, and the second component characterizes the uniform variance of *NO3* with an interval specified by endpoints *xmin* and *xmax*. Both components are computed individually and weighted by the common probability for each interval. This variance can be written using the following equation:

$$Var\left[NO3 \mid month = 8,\ Si = [2.9]_2\right] = \\ \sum_{k=1}^{5} p_k \cdot \left(\left(xmid_k - E\left[NO3 \mid month = 8, Si = [2.9]_2\right] \right)^2 + \frac{1}{12}\left(xmax_k - xmin_k\right)^2 \right) = 0.512^2 = \\ 0.7048 \cdot \left((0.2 - 0.488)^2 + \frac{1}{12}(0.4 - 0.0)^2 \right) + \cdots + 10^{-6}\left((1.55 - 0.488)^2 + \frac{1}{12}(1.9 - 1.2)^2 \right), \tag{2}$$

where

$xmin_k$ and $xmax_k$ are the minimum and maximum nitrate concentrations for the k^{th} nitrate interval.

Thus, in August, for a measured silica concentration of 2.9 mg/L, which is in silica interval 2, the expected nitrate nitrogen concentration is 0.488 mg/L with a standard deviation of 0.512 mg/L.

Classification Rates for the Bayesian Network

In Netica, a correct classification indicates that a measurement is predicted to occur, based on its expected value, in the interval in which it was measured. Given the five equiprobable intervals assigned to each water-quality parameter, a correct classification rate of 20 percent would be expected by chance. Correct classification rates greater than 20 percent indicate some predictive ability.

Correct classification rates in a BN would be expected to vary based on the conditioning information provided for prediction. In general, this conditioning set could be selected from a large number of alternative non-empty subsets of parameters historically measured at a monitoring site. In this report, two conditioning parameter sets are discussed. The first parameter set consists of the month of sample collection (*month*); the second set consists of the *month* and six selected continuously measurable parameters (*ContMeasSet*). These selected continuously measurable parameters, and their percent numerical values in the training dataset, were barometric pressure (*barPres*, 85.0), streamflow (*Flow*, 83.5), dissolved oxygen (*O2*, 84.6), pH (*pH*, 85.2), specific conductance (*SpecCond*, 85.2), and water temperature (*w_temp*, 85.0).

The correct classification rates may be expected to vary between training and testing datasets. In particular, overfitting of the training data may be evident by substantially lower correct classification rates in testing data as compared to training data. It is possible due to random sampling, however, that error rates in some of the testing dataset could be lower than the training dataset.

Correct classification rates for all parameters, conditioning variable sets, and datasets are depicted in figure 12. When conditioned on *month* only, water temperature (*w_temp*) had the maximum correct classification rates of 62.3 and 56.4 percent in the training and testing datasets, respectively. When conditioned on the *ContMeasSet* set of parameters, total dissolved solids (*tDisSolids*) had the maximum correct classification rates of 86.4 and 84.2 percent in the training and testing datasets, respectively. When conditioned on *month* only, silica (*Si*) had the minimum correct classification rate of 31.9 percent in the training dataset, and carbon dioxide (*CO2*) had minimum correct classification rate of 35.2 percent in the testing dataset. When conditioned on the *ContMeasSet* set of parameters, phosphorus (*P*) had the minimum correct classification rates of 38.8 and 35.2 percent in the training and testing datasets, respectively.

The average correct classification rate in the training dataset for all parameters conditioned on *month* was 41.7 percent. With the *ContMeasSet* set, the average correct classification rate increased to 61.2 percent for the training dataset. In comparison, the average correct classification rate in the testing dataset for all parameters conditioned on *month* was 40.8 percent. With the *ContMeasSet* set, the average correct classification rate in the testing dataset was 56.5 percent. It may be worth noting that all parameters had higher correct classification rates when conditioned on the *ContMeasSet* set than the single condition parameter of *month* in both training and testing datasets (fig. 12), except for total nitrogen, *N_all*, data in the testing dataset. Here, the average correct classification rate was 4.62 percent lower when conditioned on the *ContMeasSet* set than when conditioned on month alone.

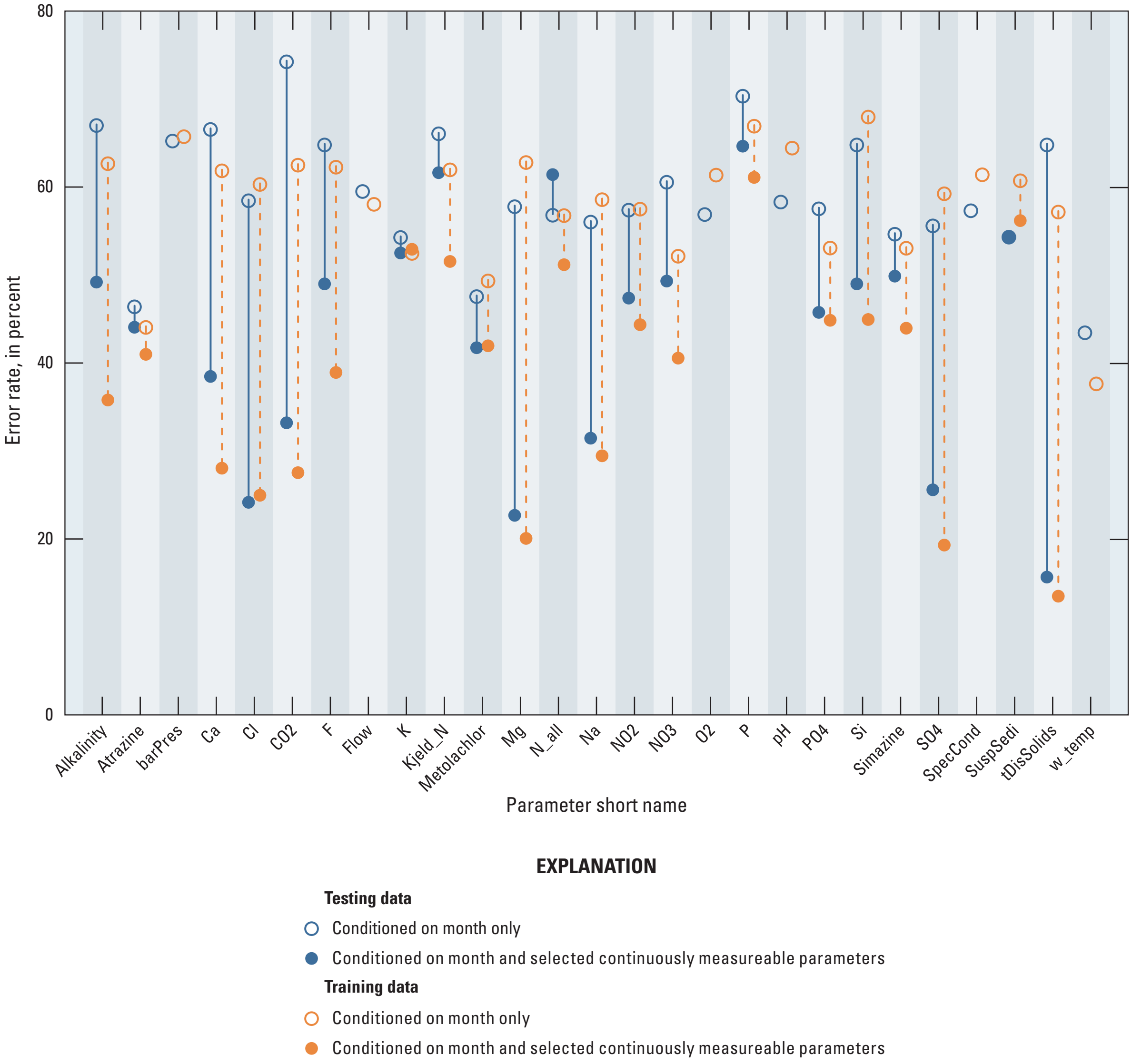


Figure 12. Bayesian network error rate for training and testing dataset conditioned on month and selected continuously measurable parameters (full parameter names and parameter codes are given in table 1.1).

Two primary conclusions are indicated from these results. First, even with an average rate of only 84.4 percent of conditioning variables in the *ContMeasSet* set quantified, the average correct classification rate in the training and testing datasets increased by 17.6 percent, from conditioning on *month* alone. Thus, data from continuously measurable parameters, such as the set of parameters that can increasingly be measured at high frequencies, increase the accuracy of a static BN, although a dynamic BN would be needed to fully utilize information from high-frequency data.

Second, the average correct classification rates decreased 0.9 and 4.9 percent in parameters between the training and testing datasets, respectively, for parameters conditioned on month, and the *ContMeasSet* set. Thus, when parameters were conditioned only on the single parameter *month*, there was no significant change (p-value = 0.4698) in accuracy between training and testing datasets based on the one sample Wilcoxon Test (Conover, 1980). When parameters were conditioned on the seven conditioning parameters in the *ContMeasSet* set, however, there was a significant (p-value < 0.0001) decrease in accuracy of 4.9 percent between the training and testing datasets. The 4.9-percent decrease in accuracy is at least partially attributed to the effects of model overfitting with the larger parameter set in *ContMeasSet*.

Application Potential

Discrete samples of water quality provide a basis for the development of BNs describing at-site associations among water-quality parameters. Such a model could be used to estimate the magnitude and uncertainty of unmeasured parameters given a measured complement from historically measured parameters. Thus, the utility of a set of measured parameters may be extended to estimation of a set of unmeasured parameters. In addition, the DAG associated with the BN can be used to better understand the flow of information through the network based on the rules of directional connectedness. Finally, the degree centrality of a parameter, as indicated by the number of edges incident upon a node, may help prioritize individual parameters that are mostly informative from a network perspective.

It is likely that the number of nodes of water-quality parameters in a network could be extended considerably beyond the 28 parameters (including *month*) used in this report. Such extensions may provide insight into higher dimensional associations that are not apparent from conventional techniques of multivariate analysis, which are commonly constrained by missing, censored, and interval data characteristic of water-quality data. Such extensions may benefit from resampling experiments to help assess the robustness of the edge structure and conditional probability tables defined from individual sample realizations.

In addition, static BNs could be developed for a common set of parameters at multiple stream sites, or over multiple time periods at the same site, to assess whether parameter associations are sensitive to measures or changes in measures of aquatic ecosystem health. Analysis of robust network structure might be used to determine what statistical associations are interrupted or initiated with changes in the ecosystem.

High-frequency (multiple daily or hourly) water-quality data are becoming increasingly available for an expanding set of water-quality parameters. These data could provide a basis for characterizing dynamic interrelations among parameters on a subdiel time scale. This extension would involve estimation of time-delay parameters in the network but may provide insight in process dynamics that could not be assessed on the basis of discrete sample information.

Despite the potential of BN for describing intrasite, or perhaps intersite relations among water-quality parameters, difficulties are recognized. Some of these restrictions may be associated with the present state of software development in Netica, or similar packages, whereas others may present more theoretical constraints.

Continuous variables are commonly discretized into a fixed set of class intervals to facilitate computations. Alameddine and others (2011) note, however, that the number of discretized intervals can affect the identification of the edge structure. In addition, the number of elements in CPTs, which would commonly need to be estimated from data, increase exponentially with the number of parent nodes. Further investigation into the possible relation between the number quantified measurements and the optimum number of class intervals needed to optimize model performance may be productive. Thus, the complexity of the network structure may need to be artificially constrained to compute viable estimates of discrete probabilities. Also, the directionality of edges identified from data may not be consistent with causal relations among parameters. Finally, not all software provides for the integration of known causal relations with edge structures inferred from data.

Summary and Conclusions

An exploratory discrete Bayesian network (BN) was developed to assess the potential for estimating the magnitudes and uncertainties of an arbitrary subset of unmeasured water-quality parameters given the measured complement. Data from 596 discrete water quality samples obtained from water year 1991 to 2016 at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana, were used in the analysis. Selected parameters included alkalinity (*Alkalinity*), atrazine (*Atrazine*), barometric pressure (*barPres*), calcium (*Ca*), chloride (*Cl*), carbon dioxide (*CO2*), streamflow (*Flow*), fluoride (*F*), Kjeldahl nitrogen (*Kjeld_N*), magnesium (*Mg*), metolachlor (*Metolachlor*), potassium (*K*), total nitrogen (*N_all*), sodium (*Na*), nitrite (*NO2*), nitrate (*NO3*), dissolved oxygen (*O2*), phosphorus (*P*), pH (*pH*), orthophosphate (*PO4*), silica (*Si*), simazine (*Simazine*), sulfate (*SO4*), specific conductance (*SpecCond*), suspended sediment (*SuspSedi*), total dissolved solids (*tDisSolids*), water temperature (*w_temp*), and month of sample collection (*month*). The parameters were selected because they were commonly measured in water-quality samples at the streamgage, tended to have fewer censored or missing observations than non-selected parameters, and did not contain obvious temporal trends.

All water samples used in this report were obtained with the interval from October 1, 1990, and September 30, 2016. Water-quality data from the 596 discrete samples were randomly partitioned into a training set, with 80 percent of the measurements, and a testing dataset, with the remaining 20 percent of the data. Data for 27 strictly water-quality parameters were discretized at their quintiles, providing five ordered levels each containing about 20 percent of the observations. The month of sampling (*month*) was discretized into 12 monthly levels. Water-quality data for the selected parameters included fully quantified numerical values; missing value indicators; censored data, represented by less than and greater than threshold values; and estimated data, represented by intervals.

The BN was estimated from the selected water-quality parameters following a two-step process using the training set data. The first step identifies a set of directed edges connecting nodes. The Netica software used in this report provides a Tree Augmented Naïve (TAN) Bayes algorithm to facilitate this step, which requires the specification of a target node (parameter). Because any set of nodes in a water-quality network might need to be estimated (as a target node), all nodes in the network were selected, in turn, to develop a set of edge structures for evaluation.

In the second step, the training dataset was applied to compute the negative log-likelihood for each edge structure defined from individual target nodes, and the corresponding elements in conditional probability tables for each node. The edge structure with the maximum likelihood (the minimum negative log-likelihood) was selected. This structure corresponded to the BN with *month* selected as the target node. This network included 53 edges connecting the 28 nodes, and required estimation of 7,872 elements in the collection of conditional probability tables given the discretization of parameters. The month node, the only node without one or more parent nodes, had an out degree of 27, which indicates that all other parameters were conditioned on month of sample collection. In contrast, the maximum in degree of any node was two.

The accuracy of the BN was evaluated separately on the basis of data in the training and testing datasets, and on two sets of conditioning variables. The conditioning variables included *month* only in one set, and *month* plus the selected continuously measureable parameters of barometric pressure (*barPres*), streamflow (*Flow*), dissolved oxygen (*O2*), pH (*pH*), specific conductance (*SpecCond*), and water temperature (*w_temp*), forming the *ContMeasSet* set. Accuracy was based on the correct classification rates.

Given the five equiprobable levels of discretization for each water-quality parameter, random assignment to an interval would have been expected to produce an average correct classification rate of 20 percent, given that month of sampling was always observed. For the training set, the average correct classification rates for parameters conditioned on *month* only was 41.7 percent, and 61.2 percent when conditioned on the *ContMeasSet* set. For the testing dataset, the average correct classification rate was 40.8 percent when conditioned on *month* only, and 56.5 percent when conditioned on the *ContMeasSet* set. Thus, conditioning on the *ContMeasSet* set rather than only *month* increased the correct classification rates in both training and testing data. Although some degradation between training and testing dataset accuracy occurred, the *ContMeasSet* variable set substantially improved estimation accuracy relative to *month* only. Conditioning variables from high-frequency data would likely similarly increase the accuracy of classifications.

This exploratory analysis demonstrated the potential utility of BNs for estimating the magnitudes and uncertainties of an unmeasured subset of historically measured water-quality parameters at U.S. Geological Survey National Water Quality Program stream sites, given the measured complement of parameters. The edge structure of the BN provides details on how information propagates through a water-quality network, and which parameters may be central to this flow of information. This structure may help identify parameters for measurement, either from discrete water-quality samples or from high-frequency monitoring.

Many important questions, however, remain unanswered and would benefit from further research. In particular, it would be helpful to know (1) how effectively information in a BN will scale from the tens of parameters used in this analysis to the hundreds of parameters commonly available at National Water Quality Program sites; (2) how robust network identification is to the particular sample of data used in development; (3) how sensitive the network structure may be to different software or levels of parameter discretization; (4) how process knowledge may be used to augment the statistical identification of network structure, particularly in the extension of static to dynamic BN; and (5) whether or not network structures may be used to help identify aquatic ecosystems that are stable and healthy.

This report was made possible by publicly available, quality-assured data obtained through the U.S. Geological Survey National Water Quality Program. The selected site is just one of many available for analysis across the United States. Data from this repository continues to support improved understanding of the static and dynamic relations among water-quality characteristics, which provide a basis for understanding and monitoring possible changes in aquatic ecosystem health as advancements in statistical methodology continue.

References Cited

Alameddine, I., Cha, Y.K., and Reckhow, K.H., 2011, An evaluation of automated structure learning with Bayesian networks—An application to estuarine chlorophyll dynamics: Environmental Modelling and Software, v. 26, no. 2, p. 163–172, accessed February 12, 2018, at https://doi.org/10.1016/j.envsoft.2010.08.007.

Chow, C.K., and Liu, C.N., 1968, Approximating discrete probability distribution with dependence trees: Institute of Electrical and Electronics Engineers, Transactions on Information Theory, v. 14, p. 462–467.

Conover, W.J., 1980, Practical nonparametric statistics (2d ed.): New York, John Wiley & Sons, 491 p.

Fraser, I.L., 2016, abn—Modelling multivariate data with additive Bayesian networks: R package version 1.0, accessed February 12, 2018, at https://CRAN.R-project.org/package=abn.

Friedman, N., Geiger, D., and Goldszmidt, M., 1997, Bayesian network classifiers: Machine Learning, v. 29, p. 131–163.

Hirsch, R.M., and De Cicco, L.A., 2015, User guide to Exploration and Graphics for RivEr Trends (EGRET) and dataRetrieval—R packages for hydrologic data (version 2.0, February 2015): U.S. Geological Survey Techniques and Methods, book 4, chap. A10, 93 p., accessed February 12, 2018, at http://dx.doi.org/10.3133/tm4A10.

Hopple, J.A., 2018, Datasets to reproduce the exploratory Bayesian network developed in USGS SIR 2018-5053 for estimating water-quality parameters at streamgage 03374100 White River at Hazleton, Indiana, 1973-2016: U.S. Geological Survey data release, at https://doi.org/10.5066/P9JJYKWD.

Koller, D., and Friedman, N., 2009, Probabilistic graphical models—Principles and techniques: Cambridge, Mass., Massachusetts Institute of Technology, 1217 p.

Norsys Software Corporation, 2016, Netica Application, version 5.24 for Microsoft Windows: Norsys Software Corporation [software program].

Pearl, J., 2009, Causality—Models, reasoning, and inference (2d ed.): Cambridge University Press, 484 p.

R Core Team, 2016, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, accessed February 12, 2018, at https://www.R-project.org/.

Risch, M.R., Bunch, A.R., Vecchia, A.V., Martin, J.D., and Baker, N.T., 2014, Water quality in Indiana—Trends in concentrations of selected nutrients, metals, and ions in streams, 2000–10: U.S. Geological Survey Scientific Investigations Report 2014–5205, 47 p., accessed February 12, 2018, at http://dx.doi.org/10.3133/sir20145205.

Schenk, L.N., Anderson, C.W., Diaz, P., and Stewart, M.A., 2016, Evaluating external nutrient and suspended-sediment loads to Upper Klamath Lake, Oregon, using surrogate regressions with real-time turbidity and acoustic backscatter data: U.S. Geological Survey Scientific Investigations Report 2016–5167, 46 p., accessed February 12, 2018, at https://doi.org/10.3133/sir20165167.

Scutari, M., 2010, Learning Bayesian Networks with the bnlearn R Package: Journal of Statistical Software, v. 35, no. 3, p. 1–22, accessed February 12, 2018, at http://www.jstatsoft.org/v35/i03/.

U.S. Geological Survey, 2011a, National Water Information System—Help system—Parameters: [last modified on November 9], accessed January 11, 2018, at https://help.waterdata.usgs.gov/codes-and-parameters/parameters.

U.S. Geological Survey, 2011b, National Water Information System—Help system—Medium codes, descriptions, and definitions: [last modified on November 9], accessed on January 11, 2018, at https://help.waterdata.usgs.gov/medium_cd.

U.S. Geological Survey, 2011c, National Water Information System—Help system—Sample type codes: [last modified on November 9], accessed January 11, 2018, at https://help.waterdata.usgs.gov/code/sample_type_query?fmt=%20rdb&inline=true.

U.S. Geological Survey, 2011d, National Water Information System—Help system—Remark codes (remark_cd): accessed February 12, 2018, at https://help.waterdata.usgs.gov/remark_cd.

U.S. Geological Survey, 2016, USGS Site web service—Show one or more data types (outputDataTypeCd): accessed January 11, 2018, at https://waterservices.usgs.gov/rest/Site-Service.html#outputDataTypeCd, last modified on December 6.

U.S. Geological Survey, 2018a, National Water Information System—Web interface: accessed January 2018 at https://doi.org/10.5066/F7P55KJN.

U.S. Geological Survey, 2018b, Parameter code definition: [last modified February 12], accessed on January 10, 2018, at https://nwis.waterdata.usgs.gov/dc/nwis/pmcodes.

Appendix

Table 1.1. Redundancy and omission of selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana.

[--, no data]

Parameter group[1]	Parameter code	Reason for parameter omission in Bayesian network	Parameter short name	National Water Information System parameter name (parameter_nm)
Physical	00010	--	*w_temp*	Temperature, water, degrees Celsius
Physical	00020	[2]r(00010)	--	Temperature, air, degrees Celsius
Physical	00025	--	*barPres*	Barometric pressure, millimeters of mercury
Physical	00061	--	*Flow*	Discharge, instantaneous, cubic feet per second
Physical	00065	r(00061)	--	Gage height, feet
Physical	00095	--	*SpecCond*	Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees Celsius
Physical	00400	--	*pH*	pH, water, unfiltered, field, standard units
Physical	00900	r(39086)	--	Hardness, water, milligrams per liter as calcium carbonate
Physical	30207	r(00061)	--	Gage height, above datum, meters
Physical	30209	r(00061)	--	Discharge, instantaneous, cubic meters per second
Physical	70300	--	*tDisSolids*	Dissolved solids dried at 180 degrees Celsius, water, filtered, milligrams per liter
Physical	70301	r(70300)	--	Dissolved solids, water, filtered, sum of constituents, milligrams per liter
Physical	70302	r(70300)	--	Dissolved solids, water, tons per day
Physical	70303	r(70300)	--	Dissolved solids, water, filtered, tons per acre-foot
Physical	90095	r(00095)	--	Specific conductance, water, unfiltered, laboratory, microsiemens per centimeter at 25 degrees Celsius
Inorganics, Major, Metals	00915	--	*Ca*	Calcium, water, filtered, milligrams per liter
Inorganics, Major, Metals	00925	--	*Mg*	Magnesium, water, filtered, milligrams per liter
Inorganics, Major, Metals	00930	--	*Na*	Sodium, water, filtered, milligrams per liter
Inorganics, Major, Metals	00931	r(00930)	--	Sodium adsorption ratio (SAR), water, number
Inorganics, Major, Metals	00932	r(00930)	--	Sodium fraction of cations, water, percent in equivalents of major cations
Inorganics, Major, Metals	00935	--	*K*	Potassium, water, filtered, milligrams per liter
Inorganics, Major, Non-metals	00191	r(00400)	--	Hydrogen ion, water, unfiltered, calculated, milligrams per liter
Inorganics, Major, Non-metals	00300	--	*O2*	Dissolved oxygen, water, unfiltered, milligrams per liter

Table 1.1. Redundancy and omission of selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana.—Continued

[--, no data]

Parameter group[1]	Parameter code	Reason for parameter omission in Bayesian network	Parameter short name	National Water Information System parameter name (parameter_nm)
Inorganics, Major, Non-metals	00301	r(00300)	--	Dissolved oxygen, water, unfiltered, percent of saturation
Inorganics, Major, Non-metals	00405	--	*CO2*	Carbon dioxide, water, unfiltered, milligrams per liter
Inorganics, Major, Non-metals	00453	HPNQD[3]	--	Bicarbonate, water, filtered, inflection-point titration method (incremental titration method), field, milligrams per liter
Inorganics, Major, Non-metals	00940	--	*Cl*	Chloride, water, filtered, milligrams per liter
Inorganics, Major, Non-metals	00945	--	*SO4*	Sulfate, water, filtered, milligrams per liter
Inorganics, Major, Non-metals	00950	--	*F*	Fluoride, water, filtered, milligrams per liter
Inorganics, Major, Non-metals	00955	--	*Si*	Silica, water, filtered, milligrams per liter as SiO_2
Inorganics, Major, Non-metals	39086	--	*Alkalinity*	Alkalinity, water, filtered, inflection-point titration method (incremental titration method), field, milligrams per liter as calcium carbonate
Nutrient	00600	--	*N_all*	Total nitrogen [nitrate + nitrite + ammonia + organic-N], water, unfiltered, milligrams per liter
Nutrient	00605	HPNQD	--	Organic nitrogen, water, unfiltered, milligrams per liter as nitrogen
Nutrient	00608	HPNQD	--	Ammonia, water, filtered, milligrams per liter as nitrogen
Nutrient	00613	--	*NO2*	Nitrite, water, filtered, milligrams per liter as nitrogen
Nutrient	00618	--	*NO3*	Nitrate, water, filtered, milligrams per liter as nitrogen
Nutrient	00625	--	*Kjeld_N*	Ammonia plus organic nitrogen, water, unfiltered, milligrams per liter as nitrogen
Nutrient	00631	r(00613 + 00618)	--	Nitrate plus nitrite, water, filtered, milligrams per liter as nitrogen
Nutrient	00660	r(00671)	--	Orthophosphate, water, filtered, milligrams per liter as PO_4
Nutrient	00665	--	*P*	Phosphorus, water, unfiltered, milligrams per liter as phosphorus
Nutrient	00671	--	*PO4*	Orthophosphate, water, filtered, milligrams per liter as phosphorus
Nutrient	71846	HPNQD	--	Ammonia, water, filtered, milligrams per liter as NH_4
Nutrient	71851	r(00618)	--	Nitrate, water, filtered, milligrams per liter as nitrate
Nutrient	71856	r(00613)	--	Nitrite, water, filtered, milligrams per liter as nitrite
Organics, pesticide	04035	--	*Simazine*	Simazine, water, filtered, recoverable, micrograms per liter

Table 1.1. Redundancy and omission of selected water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana.—Continued

[--, no data]

Parameter group[1]	Parameter code	Reason for parameter omission in Bayesian network	Parameter short name	National Water Information System parameter name (parameter_nm)
Organics, pesticide	39415	--	*Metolachlor*	Metolachlor, water, filtered, recoverable, micrograms per liter
Organics, pesticide	39632	--	*Atrazine*	Atrazine, water, filtered, recoverable, micrograms per liter
Organics, pesticide	46342	HPNQD	--	Alachlor, water, filtered, recoverable, micrograms per liter
Organics, pesticide	82630	HPNQD	--	Metribuzin, water, filtered, recoverable, micrograms per liter
Sediment	80154	--	*SuspSedi*	Suspended sediment concentration, milligrams per liter
Sediment	80155	r(80154 × 00061)	--	Suspended sediment discharge, tons per day

[1] The parameter group ‘Informational’ was not included because informational parameters describe metadata about samples rather than chemical or physical properties of water samples.

[2] r(00010) implies that parameter 00020 is redundant with parameter 00010. Therefore, parameter 00020 was not included in the Bayesian network developed in this report. The notation was extended to multiple parameters that resulted in redundancy from a sum (+) or product (×) of two or more included parameters.

[3] HPNQD indicates a High Percentage of Non-Quantified Data is associated with the parameter, so it was not included in the Bayesian network developed in this report even though minimum data requirements were satisfied.

Table 1.2. Excerpt from training dataset[1] for a Bayesian network of water-quality parameters at U.S. Geological Survey streamgage 03374100 White River at Hazleton, Indiana.

[data in brackets indicate the expected range for estimated values; *, missing data; <, less than]

Parameter short name	Identification number				
	1	2	3	…	20
month	5	5	5	….	7
w_temp	*	*	*	….	27.7
SpecCond	*	*	*	….	600
pH	*	*	*	….	8.2
Flow	[11245, 19025]	[9481, 15232]	[9276, 14932]	….	1960
O2	*	*	*	….	7.3
P	*	*	*	….	0.18
barPres	*	*	*	….	756
CO2	*	*	*	….	*
SuspSedi	*	*	*	….	222
Cl	*	*	*	….	*
SO4	*	*	*	….	*
Metolachlor	0.19	0.15	0.22	….	0.17
Simazine	0.05	0.08	0.07	….	<0.050
Kjeld_N	*	*	*	….	1.8
Atrazine	0.27	0.31	0.53	….	0.61
N_all	*	*	*	….	<1.90
NO3	*	*	*	….	<0.040
NO2	*	*	*	….	0.01
Alkalinity	*	*	*	….	*
Ca	*	*	*	….	*
Mg	*	*	*	….	*
Na	*	*	*	….	*
K	*	*	*	….	*
tDisSolids	*	*	*	….	*
Si	*	*	*	….	*
F	*	*	*	….	*
PO4	*	*	*	….	<0.010

[1] Note: the training data have been transposed so that the columns would fit on the page.

Table 1.3. Conditional probability table for deciles of nitrate nitrogen concentration range conditioned on month number and silica range.

[NO_3, nitrate nitrogen; mg/L, milligram per liter; >, greater than]

Month number[1]	Silica range (mg/L)	Probability, in percent conditioned on month and silica NO_3 concentration range, in milligrams per liter				
		0 to 0.4	>0.4 to 1.2	>1.2 to 1.9	>1.9 to 2.4	>2.4 to 6.8
1	0 to 2.4	0.0010	0.0010	99.9960	0.0010	0.0010
1	2.4 to 4.8	0.0006	0.0006	0.0006	55.5255	44.4729
1	4.8 to 6.6	0.0003	33.5959	0.0003	66.4030	0.0003
1	6.6 to 7.3	0.0002	0.0002	66.5848	16.7074	16.7074
1	7.3 to 9.1	0.0001	0.0001	8.2181	8.1956	83.5862
2	0 to 2.4	0.0010	0.0010	0.0010	0.0010	99.9960
2	2.4 to 4.8	0.0010	0.0010	0.0010	0.0010	99.9960
2	4.8 to 6.6	0.0003	0.0003	33.3331	66.6659	0.0003
2	6.6 to 7.3	0.0005	0.0005	0.0005	0.0005	99.9980
2	7.3 to 9.1	0.0001	0.0001	0.0001	37.4999	62.4997
3	0 to 2.4	0.0002	0.0002	60.6619	39.3375	0.0002
3	2.4 to 4.8	0.0005	0.0005	99.9980	0.0005	0.0005
3	4.8 to 6.6	0.0001	0.0001	14.3378	15.2659	70.3961
3	6.6 to 7.3	0.0002	0.0002	0.0003	25.3489	74.6504
3	7.3 to 9.1	0.0002	0.0002	0.0002	67.1099	32.8896
4	0 to 2.4	0.0003	66.6659	0.0003	0.0003	33.3331
4	2.4 to 4.8	0.0003	26.3795	73.6197	0.0003	0.0003
4	4.8 to 6.6	0.0001	0.0001	28.5714	28.5713	42.8570
4	6.6 to 7.3	0.0002	0.0002	67.7893	32.2102	0.0002
4	7.3 to 9.1	0.0002	0.0002	20.0000	39.9998	39.9998
5	0 to 2.4	39.9999	30.0000	30.0000	0.0001	0.0001
5	2.4 to 4.8	0.0001	0.0001	46.4786	53.5211	0.0001
5	4.8 to 6.6	0.0001	0.0002	73.1765	17.3518	9.4714
5	6.6 to 7.3	0.0001	23.1152	35.1977	10.8666	30.8203
5	7.3 to 9.1	0.0003	0.0003	99.9989	0.0003	0.0003
6	0 to 2.4	42.6870	42.8760	14.4369	0.0001	0.0001
6	2.4 to 4.8	6.0194	11.8300	48.1360	28.0996	5.9151
6	4.8 to 6.6	0.0001	0.0002	41.1627	0.0004	58.8366
6	6.6 to 7.3	0.0001	8.9937	8.9938	11.2369	70.7755
6	7.3 to 9.1	0.0001	0.0001	0.0001	19.7902	80.2095
7	0 to 2.4	60.2634	39.7364	0.0001	0.0001	0.0001
7	2.4 to 4.8	8.7152	74.5614	8.7398	7.9836	0.0001
7	4.8 to 6.6	0.0001	24.8685	33.3732	27.8388	13.9195
7	6.6 to 7.3	0.0002	0.0002	53.9021	30.7317	15.3659
7	7.3 to 9.1	0.0001	14.2863	42.8550	42.8585	0.0001
8	0 to 2.4	89.8485	10.1514	0.0000	0.0000	0.0000
8	2.4 to 4.8	70.4803	14.7802	14.7392	0.0001	0.0001
8	4.8 to 6.6	25.0061	74.9931	0.0003	0.0003	0.0003
8	6.6 to 7.3	0.0002	99.9992	0.0002	0.0002	0.0002
8	7.3 to 9.1	0.0005	0.0005	49.9992	0.0005	49.9992
9	0 to 2.4	85.7727	14.2269	0.0001	0.0001	0.0001

Table 1.3. Conditional probability table for deciles of nitrate nitrogen concentration range conditioned on month number and silica range.—Continued

[NO_3, nitrate nitrogen; mg/L, milligram per liter; >, greater than]

Month number[1]	Silica range (mg/L)	Probability, in percent conditioned on month and silica NO_3 concentration range, in milligrams per liter				
		0 to 0.4	>0.4 to 1.2	>1.2 to 1.9	>1.9 to 2.4	>2.4 to 6.8
9	2.4 to 4.8	52.5806	47.4190	0.0001	0.0001	0.0001
9	4.8 to 6.6	0.0005	99.9980	0.0005	0.0005	0.0005
9	6.6 to 7.3	0.0010	99.9960	0.0010	0.0010	0.0010
9	7.3 to 9.1	0.0005	99.9980	0.0005	0.0005	0.0005
10	0 to 2.4	80.9044	19.0950	0.0002	0.0002	0.0002
10	2.4 to 4.8	0.0002	0.0002	99.9990	0.0002	0.0002
10	4.8 to 6.6	40.2395	20.8555	38.9046	0.0002	0.0002
10	6.6 to 7.3	0.0010	0.0010	99.9960	0.0010	0.0010
10	7.3 to 9.1	0.0002	24.9999	74.9993	0.0002	0.0002
11	0 to 2.4	16.6668	66.6663	0.0002	16.6666	0.0002
11	2.4 to 4.8	0.0003	0.0003	99.9987	0.0003	0.0003
11	4.8 to 6.6	0.0001	12.3398	87.6598	0.0001	0.0001
11	6.6 to 7.3	0.0002	0.0002	99.9993	0.0002	0.0002
11	7.3 to 9.1	0.0002	0.0006	49.9994	49.9995	0.0002
12	0 to 2.4	0.0005	0.0005	49.9992	0.0005	49.9992
12	2.4 to 4.8	0.0005	0.0005	99.9980	0.0005	0.0005
12	4.8 to 6.6	0.0010	0.0010	0.0010	99.9960	0.0010
12	6.6 to 7.3	0.0002	59.9996	20.0000	0.0002	20.0000
12	7.3 to 9.1	0.0002	0.0002	16.6667	33.3332	49.9997

[1] Month number represents consecutive months from January as 1 to December as 12.

For additional information, contact:
Director, Michigan Water Science Center
U.S. Geological Survey
6520 Mercantile Way, Suite 5
Lansing, MI 48911

or visit our website at https://mi.water.usgs.gov/

Publishing support provided by the Madison, Rolla, and West Trenton Publishing Service Centers

Holtschlag—**Estimating the Magnitudes and Uncertainties of Selected Water-Quality Parameters**—Scientific Investigations Report 2018–5053

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20185053