BNL 51068

#240
4/28/80

ph. 1102

MASTER

NEW TECHNIQUES FOR ANALYZING RELATIONSHIPS BETWEEN ENERGY AND WATER QUALITY

E. KAPLAN AND H.C. THODE, JR.

BIOMEDICAL AND ENVIRONMENTAL ASSESSMENT DIVISION
NATIONAL CENTER FOR THE ANALYSIS OF ENERGY SYSTEMS

BROOKHAVEN NATIONAL LABORATORY
ASSOCIATED UNIVERSITIES, INC.

UNDER CONTRACT NO. EY-76-C-02-0016 WITH THE

UNITED STATES DEPARTMENT OF ENERGY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

BNL 51068

UC-11

(Environmental Control Technology
and Earth Sciences - TID-4500)

NEW TECHNIQUES FOR ANALYZING RELATIONSHIPS BETWEEN ENERGY AND WATER QUALITY

E. KAPLAN AND H.C. THODE, JR.

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

BIOMEDICAL AND ENVIRONMENTAL ASSESSMENT DIVISION
NATIONAL CENTER FOR THE ANALYSIS OF ENERGY SYSTEMS

BROOKHAVEN NATIONAL LABORATORY
UPTON, NEW YORK 11973

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Printed in the United States of America
Available from
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161
Price: Printed Copy \$5.25; Microfiche \$8.00

A04

A01

ABSTRACT

Water quality data for 65 variables were obtained for the period 1955-1977 and aggregated on a county basis. Measurements were taken primarily in New England and the Middle Atlantic States. When a subset of 138 counties with complete data was used, it was found that county aggregation statistical procedures resulted in data still able to describe the chemical characteristics of natural waters. Energy and socioeconomic data were merged with water quality data for these 138 counties. The path analytic methodology used by geneticists was adapted for use with these combined data to investigate for potential interactions between energy-related activities and water quality. A path diagram was proposed to provide insight into the possible causal nature of these interrelations. Direct and indirect pathways from energy production and use were traced to three factors describing functional attributes of water: conductivity, hardness, and dissolved metallic ions. This analysis explained 25 to 40% of the variance in three water quality factors and indicated the applicability of this technique to regional assessments of water quality impacts due to many human activities.

ACKNOWLEDGMENTS

The authors express their indebtedness to Dr. Stephen Gage, Assistant Administrator for Research and Development, U.S. Environmental Protection Agency, for providing access to the EPA/STORET data base. Mr. Robert Weiner, a summer research assistant, performed most of the data extraction from STORET as well as several preliminary data analyses. Dr. John Eric Edinger provided invaluable insight during several consultations, especially in terms of chemical characteristics of natural waters.

Special appreciation is given to Professor Stephen Finch, State University of New York at Stony Brook, who provided important consultation in state-of-the-art statistical techniques, and to Leonard D. Hamilton, M.D., Head, Biomedical and Environmental Assessment Division, BNL, who maintained continuing moral support for our work.

The work on which this paper is based was carried out under the auspices of the U.S. Department of Energy, Division of Regional Assessments, Office of Technology Impacts, under contract No. EY-76-C-02-0016.

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

CONTENTS

ABSTRACT

ACKNOWLEDGMENT

1	INTRODUCTION.....	1
2	APPROACH AND ANALYSES.....	3
	2.1 Methodology for Data Collection and Aggregation.....	3
	2.2 Exploratory Data Analyses.....	6
	2.3 Analysis of Aggregation Procedure and Comparison with Professional Literature.....	9
3	RELATIONSHIP OF ENERGY SUPPLY AND DEMAND TO WATER QUALITY: A FIRST EXAMPLE.....	33
4	SUMMARY AND CONCLUSIONS.....	45
5	REFERENCES.....	47
6	APPENDICES.....	49
	A. Lower Detection Limits of Trace Elements.....	49
	B. Exploratory Data Analysis.....	50
	C. Complete Path Diagrams.....	63

TABLES

1.	Water Quality Variables in County Data Base.....	4
2.	States Included in Data Base.....	5
3.	Transformed and Untransformed Variables.....	8
4.	Comparison of Aggregated Data With Reported Values in Natural Waters.....	13
5.	Shared Variance Among Variables.....	15
6.	Results of Canonical Correlation Analysis on Ionic Species.....	15
7.	Comparison of Specific Conductivity Regression Equations With Reported Literature.....	30
8.	Loading of Variables on Unrotated (A) and Rotated (B) Water Quality Factors, and Factor Score Coefficients for Rotated Factors (C).....	35

FIGURES

1. Bicarbonate Frequency Plots	7
2. Medians vs. Number of Observations	10
3. 138 County Subset With 19 Non-Missing Water Parameters	12
4. Relation of Cations to Anions for Aggregated Data	16
5. Relation of Dissolved Solids to Sum of Anions and Cations for Aggregated Data	18
6. Relation of Bicarbonate to Total Alkalinity for Aggregated Data	20
7. Relation of Calcium to Specific Conductance for Aggregated and Reported Data	21
8. Relation of Magnesium to Specific Conductance for Aggregated Data ...	22
9. Relation of Sodium to Specific Conductance for Aggregated Data	23
10. Relation of Potassium to Specific Conductance for Aggregated Data ...	24
11. Relation of Bicarbonate to Specific Conductance for Aggregated Data..	25
12. Relation of Sulfate to Specific Conductance for Aggregated and Reported Data	26
13. Relation of Chloride to Specific Conductance for Aggregated Data	27
14. Relation of Total Dissolved Solids to Specific Conductance for Aggregated Data	28
15. Relation of Specific Conductance to Dissolved Solids for Aggregated Data	29
16. Relation of Conductance to Sum of Anions and Cations for Aggregated Data	31
17. Conceptualized Path Diagram	32
18. Distribution of Factor 1 for Northeastern U.S.	37
19. Distribution of Factor 2 for Northeastern U.S.	38
20. Distribution of Factor 3 for Northeastern U.S.	39
21. A Path Diagram Using Water Quality Factors, Socioeconomic Data, and Energy Data	40
22. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #1	41
23. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #2	43
24. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #3	44

1 INTRODUCTION

Since passage in 1972 of landmark water quality legislation,^{*} numerous studies have been made of emissions by various types of industrial activities and of the types, costs, and energy requirements of available pollution control. A number of federally sponsored reports have also discussed projected impacts on water quality from postulated future energy scenarios. Lacking, however, has been an underlying framework which, based upon existing data, describes the interrelationships between water quality, energy production, energy use, and other human activities.

This report investigates the distributions of various water quality parameters in the surface waters of counties in the Middle Atlantic and Northeastern States. Water quality data are aggregated on a county basis and are compared with demographic, economic, and energy information to explain effects of man on water quality.

Traditional water quality analyses have emphasized use of conventional statistics such as means and standard deviations. This has been particularly true of regional assessments covering wide geographical areas. Many such analyses used these mean values in various statistical calculations such as χ^2 tests of significance and multivariate regressions. Implicit in such procedures is the underlying assumption of normality in the distribution of the random variates. For water quality variables these distributions are usually skewed, oftentimes so much so that no approximation to normality may reasonably be made. Where nonparametric methods are used, the impacts of comparing different statistics each drawn from different populations can be minimized. Even in such analyses, however, more information can be extracted with greater confidence if the underlying distributions are known, if they are similar, and if, as one hopes, they are of the same general type.

Given the hundreds of thousands of measurements of water quality variables taken over time at different places within any county, it was desired to find surrogate statistics that could be substituted for one or more of these parameters. These surrogates would be quality-of-life indicators and would be analyzed for correlations among themselves, for their distributions across

---

^{*}Public Law 92-500, Federal Water Pollution Control Act Amendments of 1972.

large geographical regions, and for their potential interdependencies with energy, demographic, industrial, economic, and health indicators.

Statistical procedures used in hydrologic and water quality analyses have proliferated in recent years, primarily because of the increasing availability of data and ease of analysis. Snyder²⁰ illustrated the extension of regression and multivariate analyses to hydrologic studies, emphasizing the latter when significant correlations are found between variables treated as independent. The lengthier and somewhat more intricate calculations required for multivariate analyses (e.g., factor and canonical correlation analyses) are now handled routinely by using readily available statistical computer packages (SPSS,¹⁷ SAS,² BMDP⁵).

The large number of dissolved ionic and undissociated species present in natural waters presents a situation unusually well suited to more sophisticated statistical analyses. Dawdy and Feth⁴ used principal component analyses to investigate interrelationships between several ionic constituents in the groundwaters of California. Principal components were also utilized by Steele²¹ to examine chemical quality data of surface waters in the Pecos and Arkansas Rivers. Both analyses focused on relatively confined geographic areas.

The present study seeks to extend the use of regression and multivariate techniques to examine interrelations among the various chemical constituents of surface waters, among the various economic activities using these waters, and between both groups. Sections 2.1 and 2.2 detail methodologies of data collection, aggregation, and transformation used in a regional assessment of waters in the Northeastern United States. This is followed in Section 2.3 by detailed analyses demonstrating the ability of the aggregated and transformed data to explain observed interdependencies among chemical species on the basis of known relationships. This report is thus divided into two separate parts. Section 2 describes how the water quality data base was collected, analyzed, and verified to produce results consistent with known relationships. Section 3 presents a methodology useful in examining the ways in which water quality may be causally affected by various human activities.

2 . APPROACH AND ANALYSES

2.1 Methodology for Data Collection

Data for the water quality parameters shown in Table 1 were obtained by using the EPA/STORET²⁷ information system. Each parameter was sought primarily from federal and state measuring stations which monitored ambient conditions* during the period 1955-1977. This interval was chosen because it includes the time before improvements due to mandated effluent controls (e.g., PL92-500), because it spans the period covering energy and socioeconomic data available to the authors, and because it provides a large number of reliable observations. Various screening techniques were used to guard against the retrieval of erroneous data which would likely result from the process of entering raw data into STORET (e.g., upper and lower bounds within which data for a given parameter were deemed acceptable).

Data for each parameter were aggregated on a county basis, so that spatial and temporal information was masked. The aggregated data for each parameter in each county included the total number of observations, period of record, mean, variance, minimum and maximum values, and also 15th, 50th, and 85th percentiles. Table 2 identifies the states and gives the number of counties from which data were obtained.

Aggregating water quality measurements taken at various times and places in either a river basin or county is an unusual and risky procedure. STORET data are collected by numerous organizations each with its own needs, priorities, capabilities, and monitoring programs. A common situation is the monitoring of water quality at a site where some contaminant has been accidentally spilled into a waterway. In such an event the monitoring agency may sample one or more locations along a short length of the waterway for a small period of time (a few weeks or months). Where an agency may be interested in establishing trends, such as downstream from an industrial site, a monitoring program may include sampling at a single location once every

---

*As opposed to stations monitoring point and non-point source emissions.

Table 1		
Water Quality Variables in County Data Base		
Functional group	Variable names	
Oxygen related	Dissolved oxygen, B.O.D.	
Bacteriological	Coliforms: fecal and total	
Nutrients	Nitrogen (as N, NO ₂ , NO ₃)	
	Phosphorus (as P, PO ₄)	
Hardness related	Alkalinity	Potassium
	Bicarbonate	Sodium
	Calcium	Sulfate
	Chloride	Solids (TDS, TSS)
	Iron	Conductivity
	Magnesium	Turbidity
	Manganese	
Trace and toxic elements	Arsenic	Lead
	Barium	Mercury
	Beryllium	Nickel
	Cadmium	Selenium
	Chromium	Silver
	Copper	Zinc
Toxic nonmetals	Ammonia	Cyanide
	Chloride	Sulfide
Organics	PBBs; PCBs	Phthalate esters
	Phenols	Oil and grease
Radionuclides	Alpha	Deia
	²²⁶ Ra	⁹⁰ Sr
Pesticides	DDD; DDE	2,4-D; 2,4,5-T
	DDT	Heptachlor
	Aldrin	Malathion
	Dieldrin	Parathion
	Chlordane	Toxaphene
	Lindane	Silvex
General	pH	Fluoride

Table 2	
States Included in Data Base*	
State	Number of counties
Connecticut	8
Delaware	3
Maine	11
Maryland**	23
Massachusetts	13
New Hampshire	9
New Jersey	20
New York	58
Ohio	87
Pennsylvania	67
Rhode Island	5
Vermont	13
Total	317
*Note that no data could be obtained for the parameters of interest in 16 counties.	
**Includes independent cities.	

three or four months; rare is the case of such data existing in STORET with a period of record greater than a few years.*

*The entire body of data relating to physical, chemical, and biological characteristics of U.S. surface water is characterized by short (and sporadic) periods of record with changing locations and frequencies of observations. Wolman²⁸ discussed these problems in a paper which led to the creation of a recent federal program to establish an ongoing network of stations at which systematic and continuing measurements would be taken of important water quality variables. This system (NASQAN⁹) began operation in 1973, and some of its observations are contained in the data presented in this report.

Measurements used herein have been made by various agencies such as the U.S. Geological Survey, the U.S. Environmental Protection Agency, and other federal, state, and local agencies. Measurements of the same parameter by different methods will vary in accuracy. Even among agencies using the same methods, some are known to take greater care with the data, both in actual measurement and in entry into the STORET data base. During the past 20 years technologies for pollution detection have improved markedly (see Appendix A). This affects both the mean (ability to detect higher or lower thresholds of a substance) and the variance (improved accuracy) of the data. Pollutant concentrations may have changed markedly during this period. Locations of actual measurements also present problems, since the mean or median value of a variable in a given county should reflect in some manner the water quality in that county. This assumes measurements are taken homogeneously over the waters of the county. If most measurements in a county are taken immediately downstream (or upstream) from a point source of pollution, the concentration of the parameter under investigation could be much higher (or lower) than a more representative level in that county. The data retrieved from STORET were therefore limited to measurements of ambient water quality data gathered from surface waters rather than from immediately above or below point-source discharges. This could be done by using extraction keywords of the STORET system.

2.2 Exploratory Data Analyses

Distributions of mean values for most substances were characterized by high skewness and kurtosis together with long tails. Figure 1 illustrates this situation for the aggregated bicarbonate frequency distribution. Also shown are what the distribution would look like had it been drawn from a normally distributed population with the same mean and variance. The median in a county rather than the mean is suggested for use: if a county has few observations taken in a very polluted region, or made by a faulty instrument, or incorrectly entered into STORET, then these observations will affect the median for that county less than they will affect the mean. Several transformations were applied to the medians to obtain better behaved distributions (see Appendix B). Taking logarithms was found to produce desired results for some but not all parameters (Table 3). Correlation coefficients were computed between water quality parameters (for raw, logarithmic, and

Figure 1. Bicarbonate Frequency Plots

Table 3			
Transformed and Untransformed Variables			
Variable	Units of untransformed data	Raw median	Log(raw median + 1)
Turbidity	JTU		✓
Conductivity	µmho/cm		✓
Mn	µg/l		✓
Total suspended solids	mg/l		✓
Total dissolved solids	mg/l		✓
F	mg/l		✓
SO ₄	mg/l		✓
Ca	mg/l		✓
Mg	mg/l		✓
Na	mg/l		✓
Cl	mg/l		✓
NO ₃	mg/l		✓
PO ₄	mg/l		✓
HCO ₃	mg/l	✓	
Cu	µg/l	✓	
Fe	µg/l	✓	
K	mg/l	✓	
Total alkalinity	mg/l	✓	
pH	Std. units	✓	

square-root transformations) and non-water quality variables. (Energy and socioeconomic data were incorporated into the water quality data file. Section 3 describes how this information was used by a new methodology for integrated water-energy-socioeconomic analyses.) It was found that when the logarithmic transformation produced a large improvement in behavior of distribution it also increased the number of statistically significant correlations with non-water quality variables. Transformations were not used when they failed to produce better behaved distributions or when more significant correlations were found by using raw data.

Another finding (Figure 2) was that counties having few observations also tended to have high median values for water quality data. Monitoring of water quality during the period 1955-1977 was characterized chiefly by a rule of crisis. In contrast to measurements of stream flow taken at fixed locations and time intervals, much STORET water quality data resulted from measurements taken in response to specific problem such as contamination from accidental spills or from the point-source discharge of industrial wastes. Until recently⁹ there has been no large-scale program of water quality monitoring at fixed stations, at regularly scheduled intervals, by standard analytical methods. (It is this situation, in fact, which prompted the aggregation of data on a county basis regardless of place or time.) Thus a county with few observations for a given parameter could represent a location where the reported data represented an acute situation. Such was in fact the case for a number of toxic metals, for which the authors ascertained that the state agencies actually making the measurement did so in response to reported discharges. It was desired to assess a longer-term (i.e., chronic) situation, and it was found that the difficulty with outliers could be relieved simply by excluding counties with fewer than five observations for each of the water quality variables of interest* (see Appendix B).

2.3 Analysis of Aggregation Procedure and Comparison With Published Work

2.3.1. Selection of Counties

The largest number of counties (138) satisfying the restriction of a minimum number of observations and complete data** resulted from using a reduced set of 19 water variables: turbidity, conductivity, pH, total suspended and dissolved solids (TSS and TDS), copper, iron, manganese, fluoride, sulfate, calcium, magnesium, sodium, potassium, chloride, nitrate, phosphorus, bicarbon-

---

* Future work will concentrate exclusively on those counties with very large median values. Again note that ongoing analyses are meant to assess chronic impacts on water quality.

** Providence, RI, satisfied these criteria but was not included in the analysis because of suspiciously high values of conductivity which require further investigation.

Figure 2. Medians vs. Number of Observations

ate, and alkalinity. This represented 42% of the total (333) counties in New England, Ohio, Pennsylvania, New York, New Jersey, Delaware, and Maryland. The question arose as to how typical these counties were regarding energy and socioeconomic characteristics of the region. A discriminant analysis was performed using all 317 counties in the data base: one group was composed of 138 counties comprising those areas having a complete set of water quality data, and all remaining counties were in a second group. This analysis showed that neither group could be distinguished on the basis of non-water quality data. Thus, socioeconomic and energy analyses using the 138-county subset can be generalized to the entire region. Figure 3 shows the counties that were used in all subsequent analyses.

2.3.2. Comparison of Means and Medians

Table 4 compares mean and order statistics for the data in the 138-county subset with ranges of values found in natural waters. As expected, means are higher than medians, and the standard deviation, estimated from the 85th and 15th percentiles is far less than the standard deviation estimated by moments. It is also seen that acute events have apparently affected medians less than means, with values in the 85th percentile well within reported ranges of variables in natural waters. It is clearly less desirable to use means and standard deviations than percentiles.

2.3.3. Comparison of Aggregated Data With Known Relationships

A measure of the validity of the county aggregation procedure is the extent to which resulting data describe known relationships between water quality variables. This section discusses anion-cation balances and relationships between ions and measurements of total dissolved solids and total alkalinity. Also shown are comparisons of aggregated data with known relationships describing ionic concentrations in terms of conductivity.

Fourteen of the nineteen water quality parameters represent ionic species: eight cations (Ca^{++} , Mg^{++} , Na^+ , K^+ , Cu^{++} , Fe^{++} or $+++$, Mn^{++} , H^+) and six anions (F^- , $\text{SO}_4^{=}$, Cl^- , NO_3^- , $\text{PO}_4^{=}$, HCO_3^-). To ascertain whether the aggregation methodology preserves the equality between the sum of the cations and the sum of the anions, it is necessary to convert concentrations in milligrams per liter (mg/l) to milliequivalents per liter (meq/l). The latter recognizes the different atomic weights and charges of the various ions, where the meq sum of cations should equal the meq sum of anions. Figure 4 indicates

Figure 3. 138 County Subset With 19 Non-Missing Water Parameters

Table 4							
Comparison of Aggregated Data With Reported Values in Natural Waters							
Variable (mg/l)	Mean	S.D.	Percentiles			Naturally occurring range	Ref.
			15%	50% (median)	85%		
pH (std units)	7.1	0.73	6.6	7.30	7.8	4.0-9.0 6.0-8.5	1 13
TSS	39	46	6.00	14	26	0-1000	23
TDS	278	295	65.00	164	406	<3000	24
Conductivity (μ mho/cm)	486	612	106.00	242	600	0-1000	1
Calcium	39	29	8.30	27	66	0-200 <600	1,23 24
Magnesium	12	12	1.80	6.7	21	0-300 0-50	1,24 23
Sodium	21	28	3.60	8.2	19	0-500 <1000	23 24
Potassium	2.7	2	1.00	2.0	3.4	<10	1,24
Phosphorus	0.49	0.56	0.04	0.12	0.48		
Copper (μ g/l)	35.14	71	0.00	10	24	<30 <600	1
Iron (μ g/l)	1107	1325	70.00	220	750	0-1000 <500	23 24
Manganese (μ g/l)	248	385	20.00	60	230	<200	23,24
Fluoride	0.26	0.18	0.10	0.20	0.30	0-2	24
Sulfate	70	70	13.00	34	95	0-200 <2000	23 1,24
Chloride	105	438	5.70	13	37	0-100 <10	23 24
Nitrate	5.2	5.8	0.05	2.1	8.6	0-5 <10	23 1,24
Bicarbonate	91	77	14.00	55	184	<500	24

that this equality has been very closely approximated, the regression line between the two classes of species given as:

$$\text{Sum (cations)} = 0.94 \text{ Sum (anions)} + 0.12$$

with an $R^2 = 0.85$ at the 0.001 significance level. The sum of cations (in meq/l) ranges from 0.21 to 16; with a standard error of 0.98.

Factor analysis^{3,17} was applied to both cation and anion groups to obtain additional insight into potential interactions between chemical species.

Beginning with a set of η variables, the technique yields a set of m factors, each a linear combination of the original η variables. The set $m \leq \eta$, and each factor is extracted such that it is orthogonal (i.e., uncorrelated) to the remaining $m - 1$ factors. This technique thus provides information on how variables tend to form groups or patterns based upon communality, that is, the extent to which different variables have a portion of their variance explained by the same unknown determinants. This is shown in Table 5. Among cations it was found that calcium, magnesium, potassium, and sodium grouped into one factor, with manganese comprising a second factor essentially by itself, the two factors together explaining 62% of the variance in the cation group.*

Interestingly, calcium, magnesium, and (to a lesser extent) potassium had most of their variance accounted for by other species in the cation group; copper was not significantly accounted for by either factor. Among anions there emerged a single factor with an eigenvalue** > 1.0 which by itself explained 54% of the variance among anions. Fluoride, nitrate, and sulfate ions had the highest amount of shared variance. These associations are typical for waters in equilibrium with various soil and rock types¹⁸ and are expected in terms of the carbonate and sulfur cycles which predominate in water chemistry. More detailed analyses will be required to determine whether the latter account for much of the unexplained variance in the cation/anion data. Such analyses will be the subject of a forthcoming paper in which county soil, rock, and land use data will be incorporated into this study.

---

* Only these factors had eigenvalues > 1 .

** The eigenvalues are calculated from the correlation matrix of independent variables and are associated with each factor to provide an indication of the degree of variance accounted for by the factor. See Harman.²⁹

Table 5
Shared Variance Among Variables

1. Shared Variance (estimated communalities)

Cations		Anions	
Mn(L)*	0.47	F(L)	0.63
Ca(L)	.86	SO ₄ (L)	.44
Mg(L)	.87	Cl(L)	.23
Na(L)	.50	NO ₃ (L)	.52
Cu	.05	PO ₄ (L)	.29
Fe	.41	HCO ₃	.35
K	.64		
H ⁺	.14		

*Signifies logarithmic transformations.

2. Factors With Eigenvalues >1.0 (and % variance explained)

Cations: Factor 1 = 0.43 Mg + 0.26 Ca + 0.25 K + 0.10 Na (40%)

Factor 2 = 1.03 Mn (22%)

Anions: Factor = 0.50 F + 0.22 NO₃ + 0.13 HCO₃ + 0.12 SO₄ (54%)

Canonical correlation analysis was employed to examine interactions both within each anion/cation group and between groups. In essence, the technique seeks linear combinations within each group which account for the greatest common variance within each group and also maximize between-group correlations. For the 138-county example three canonical variables emerge most strongly; see Table 6.

Table 6 Results of Canonical Correlation Analysis on Ionic Species			
Canonical variable	Cations (meq/l)	Anions (meq/l)	Correlation (at 0.001 significance)
1	0.49Ca + 0.40Mg	0.57HCO ₃ + 0.53SO ₄	0.97
2	0.65Mg - 0.64Na - 0.61K	1.05HCO ₃	0.85
3	0.82Mn + 0.66Ca	1.13SO ₄ - 0.76F	0.73

Figure 4. Relation of Cations to Anions for Aggregated Data

These sets indicate the influence of soils and rocks on water chemistry. The grouping of Ca, Mg, and SO_4 would reflect the ubiquitous nature of gypsum,¹³ whereas fluorite (CaF_2), apatite, and other resistate sediments would contribute to the presence of fluoride. The various feldspars would account for the presence of sodium and potassium ions. The bicarbonate ion is sharply influenced by many dissolved species, especially metal ions which may undergo oxidation and in turn lower the bicarbonate concentration. The tendency of magnesium compounds to hydrate could lower available H^+ concentrations and thereby raise bicarbonate levels. These are but a few of the reactions involved in the complex mechanism of the carbonate-bicarbonate equilibrium, but they illustrate in a qualitative manner the plausibility of the resultant canonical variates.

A sample's total dissolved solids (TDS) concentration is often approximated by using the sum of the concentrations of the major ionic and undissociated constituents.¹⁹ Figure 5 illustrates this with the aggregated data for our subset of counties, for which the regression line of TDS on the sum of the previously mentioned ionic species (without H^+) results in

$$\text{TDS} = 0.82\text{SUM}(13 \text{ ions}) + 31.1$$

with an $R^2 = 0.82$ and a significance of 0.001.

Regressing measurements of TDS on concentrations of both cation and anion groups as well as measurements of pH leads to a linear function explaining 85% of the variance in TDS and comprised primarily of Mn, Cl, HCO_3 , and SO_4 ions (Hem¹³ has indicated the presence of the complex ion MnHCO_3^+ in waters containing bicarbonate):

$$\text{TDS} = 2.2\text{SO}_4 + 0.43\text{Cl} + 0.51\text{HCO}_3 - 0.086\text{Mn} + \Sigma(10 \text{ lesser variables}).$$

Total alkalinity* in most natural waters is contributed to primarily by dissolved carbonate and bicarbonate ions. The scattergram in Figure 6 illustrates this. Also expected is the contribution of calcium ions, followed by

---

*The capacity of a solution to neutralize acid.

Figure 5. Relation of Dissolved Solids to Sum of Anions and Cations for Aggregated Data

those of other ions such as calcium and orthophosphates (Hem¹³). A stepwise regression of total alkalinity (TALK) on cation and anion groups leads to the following linear relationships with all variables included at the 0.001 significance level:

TALK = 0.76HCO ₃	R ² = 0.95
TALK = 0.69HCO ₃ + 0.37Ca	R ² = 0.96
TALK = 0.68HCO ₃ + 0.516Ca - 0.20Cl	R ² = 0.97
TALK = 0.64HCO ₃ + 0.636Ca - 0.20Cl - 0.01Mn	R ² = 0.97

Use of all cations and anions gives

$$\text{TALK} = 0.60\text{HCO}_3 + 0.43\text{Ca} - 0.22\text{Cl} + 10.0\text{PO}_4 + \Sigma(10 \text{ lesser variables})$$

with R² = 0.98, where each indicated ion is significant at least at the 0.005 level, with the overall model at the 0.001 level.

Relationships between ionic concentrations and specific conductance have been discussed at length (e.g., Steele^{21,22}). For solutions of single salts, the ratio is linear over very extended concentrations,¹³ a situation reflected in Figures 7 to 14 (each significant at the 0.001 level) and Table 7. Since natural waters contain a number of dissolved (ionic and undissociated) species, conductance is often compared with either measured or calculated total dissolved solids. With the latter expressed in milligrams per liter, a regression line of conductance on total dissolved solids can be expected to have a slope within the range 1.33 to 1.82.¹³ Regressing conductivity on measured values of TDS (Figure 15) yields a slope of 1.38, and on computed values (Figure 16) it yields a slope of 1.31.

This aggregation methodology sought to characterize a county's waters in a qualitative form that would enable further analyses of water quality data together with other county information (socioeconomics, energy, and health). Clearly, a major underpinning of such analyses is confidence that aggregation does not impair the description of qualitative characteristics of water.

Aggregation of water quality data taken at various locations and over a 20-year span in counties of the Northeast and Middle Atlantic States has pro-

Figure 5. Relation of Bicarbonate to Total Alkalinity for Aggregated Data

Figure 7. Relation of Calcium to Specific Conductance for Aggregated and Reported Data

Figure 8. Relation of Magnesium to Specific Conductance for Aggregated Data

Figure 9. Relation of Sodium to Specific Conductance for Aggregated Data

Figure 10. Relation of Potassium to Specific Conductance for Aggregated Data

Figure 11. Relation of Bicarbonate to Specific Conductance for Aggregated Data

Figure 12. Relation of Sulfate to Specific Conductance for Aggregated and Reported Data

Figure 13. Relation of Chloride to Specific Conductance for Aggregated Data

Figure 14. Relation of Total Dissolved Solids to Specific Conductance for Aggregated Data

Figure 15. Relation of Specific Conductance to Dissolved Solids for Aggregated Data

Table 7
Comparison of Specific Conductivity Regression Equations With Reported Literature

Constituent	Regression equation**	Correlation coefficient	Std. error	Steele*		
				Regression equation	Correlation coefficient	Std. error
Ca ⁺⁺	Ca = 0.10S.C. + 2.20	0.90	11.3	Ca = 0.05S.C. + 31.4	0.84	4.91
				Ca = 0.11S.C. - 2.36	.97	2.74
Mg ⁺⁺	Mg = 0.03S.C. - 1.21	.89	4.09	Mg = 0.02S.C. - 2.19	.95	1.12
				Mg = 0.03S.C. - 0.10	.99	0.38
Na ⁺	Na = 0.05S.C. - 4.72	.68	13.8	Na = 0.13S.C. - 37.0	.97	5.22
				Na = 0.06S.C. + 0.73	.92	2.71
K ⁺	K = 0.004S.C. + 0.75	.76	.89	K = 0.005S.C. + 0.43	.82	0.43
				K = 0.002S.C. + 1.67	.50	0.34
HCO ₃ ⁻	HCO ₃ = 0.26S.C. + 1.98	.74	56.6	HCO ₃ = 0.31S.C. + 34.8	.93	18.7
				HCO ₃ = 0.34S.C. + 1.73	.98	6.93
SO ₄ ⁼	SO ₄ = 0.16S.C. + 0.07	.78	30.2	SO ₄ = 0.007S.C. + 66.7	.20	5.39
				SO ₄ = 0.13S.C. + 1.83	.97	3.48
Cl ⁻	Cl = 0.09S.C. - 6.39	.68	22.7	Cl = 0.16S.C. - 51.3	.95	7.96
				Cl = 0.06S.C. - 0.21	.84	3.89
TDS	TDS = 0.62S.C. + 14.8	.92	60.2	TDS = 0.53S.C. + 30.3	.99	13.9
				TDS = 0.57S.C. + 2.09	.99	6.89

*Steele²¹ reported regressions based on two flow intervals. This information is represented here by two sets of values for each constituent.

**S.C. = Specific conductance (μmho/cm).

Figure 16. Relation of Conductance to Sum of Anions and Cations for Aggregated Data

ENERGY DEMAND

ENERGY SUPPLY

Figure 17. Conceptualized Path Diagram

duced results consistent with previous findings and basic definitions with respect to the qualitative characteristics of natural waters. The use of soil, rock, and land use data helps clarify the relationships of the aggregated data.^{10,12,13,18,23}

3 RELATIONSHIP OF ENERGY SUPPLY AND DEMAND TO WATER QUALITY:

A FIRST EXAMPLE

A causal diagram is used to illustrate one manner in which various human activities affect water quality. Figure 17 illustrates this in the form of a path diagram.^{6,7,14} The approach assumes linear, additive relationships between a set of variables that may include both measured and hypothetical variables (e.g., factors). It is emphasized that given a set of variables, one may draw a number of possible path diagrams (i.e., models).^{*} The concept is to postulate a model that is plausible and, in terms of preselected criteria, is better than alternative models which are also consistent with the same set of data.

One-way arrows lead from each determining variable to the variables depending on it. Associated with each arrow is a path coefficient. These are determined by performing a multiple regression of the standardized dependent variable on the standardized determining variables; the resulting regression coefficients are the path coefficients.

Where it is not possible to build a completely self-contained system of dependent and independent variables, that is, where complete determination does not hold, residual variables are introduced which are assumed to be uncorrelated with other determining variables and which account for the remaining variance in the model. Thus, it is presupposed that residual variables include all variables that may affect the given variable but which are not included in the data upon which the model is based.

Figure 17 thus postulates that it is the demand for energy, based upon population, which drives the overall model, though both supply and demand sectors may also affect water quality.

---

^{*}"The choice of a particular causal ordering cannot be based on the correlations among the variables because any correlation matrix will be consistent with any causal ordering one may suggest." (Duncan,⁷ p. 25.)

3.1 Water Quality Factors

Performing a path analysis for each water quality variable one at a time would produce many fragmentary results, especially insofar as the existence of interrelationships of various chemical species in the water column has previously been demonstrated. The 138-county subset with 19 variables was used to perform a factor analysis in an attempt to derive a limited number of different linear combinations of variables which would be surrogates for use in the path diagram. Three factors were found (Table 8A), each with an eigenvalue > 1.0, explaining 53%, 12%, and 7%, respectively, of the variance in the water quality data. The factors are linear in terms of standardized variables, that is,

$$X_{\text{standardized}} = \frac{X - \bar{X}_{\text{mean}}}{\sigma_x}$$

Table 8A may be interpreted as showing how each water constituent determines (i.e., loads on) each factor. For example, the single most important determinant of bicarbonate is factor No. 1:

$$\frac{\text{HCO}_3' - \overline{\text{HCO}_3}}{\sigma_{\text{HCO}_3}} = 0.78F_1 - 0.46F_2 + 0.23F_3 + dv$$

where F_1 , F_2 , and F_3 are the common orthogonal factors, v is a unique (and unknown) factor, and d is a regression coefficient (as are the coefficients of the common factors). Note that, in this particular instance, the variance in HCO_3' accounted for by F_1 is 0.61 (i.e., 0.78^2). The variance in HCO_3' accounted for by all common factors (i.e., communality) is 0.87, whose complement, 0.13, represents the portion unaccounted for by the three common factors.

Table 8
Loading of Variables on Unrotated (A) and Rotated (B) Water Quality
Factors, and Factor Score Coefficients for Rotated Factors (C)

Variable	A			B			C		
	No.1	No.2	No.3	No.1	No.2	No.3	No.1	No.2	No.3
(% of variance)	(53%)	(12%)	(7%)						
Turbidity (L)*	0.053	0.01	0.34	0.16	0.55	0.26	-0.06	0.06	0.07
Conduct. (L)	.96	.08	.06	.67	.63	.31	.47	.25	.16
Mn (L)	.28	.72	.11	.22	-.06	.75	-.04	-.02	.18
TSS (L)	.70	.02	.24	.35	.60	.27	-.10	.13	.16
TDS (L)	.90	.12	.04	.64	.56	.32	.02	.04	.03
F (L)	.72	.02	-.24	.69	.32	.08	.02	-.02	.01
SO ₄ (L)	.84	.42	.06	.62	.40	.59	.08	-.23	.60
Ca (L)	.90	-.07	.16	.54	.72	.20	-.43	.32	-.14
Mg (L)	.94	.04	.13	.60	.68	.30	-.08	.12	.24
Na (L)	.84	.09	-.41	.90	.25	.10	.63	-.42	-.15
Cl (L)	.78	.02	-.30	.77	.32	.08	-.02	.002	-.14
NO ₃ (L)	.64	.09	-.11	.55	.30	.17	-.02	-.01	.03
PO ₄ (L)	.58	-.15	-.17	.52	.33	-.07	-.02	.006	-.007
HCO ₃	.78	-.46	.23	.35	.85	-.16	.08	-.30	.17
Cu	.19	-.01	.08	.08	.18	.07	.02	-.03	.05
Fe	.28	.49	.32	.04	.17	.63	-.09	.11	.17
K	.81	.12	-.40	.87	.22	.13	.42	-.16	-.19
Total alkalinity	.80	-.48	.21	.38	.86	-.17	-.20	.94	-.64
pH	.48	-.64	-.01	.28	.58	-.47	.09	.01	-.15

*Signifies logarithmic transformation.

A varimax rotation¹⁷ was performed to achieve an even simpler pattern among variables (Table 8B). Some variables may load equally heavily on more than one of the original factors, and rotation attempts to minimize this occurrence. Again, with bicarbonate as an example, it is seen that F₁ and F₃ are equally important though subordinate to rotated F₂, which by itself determines a significant proportion of the variance in this variable. After rotation,

three different patterns emerge as underlying the complete set of 19 variables:

1. Sodium, potassium, chloride, fluoride, sulfate, conductivity, and total dissolved solids.
2. Alkalinity, bicarbonate, calcium, magnesium, suspended solids, and, to a lesser extent, conductivity and pH.
3. Iron, manganese, and, to a lesser extent, sulfate and pH (the latter in a negative direction).

(Table 8C presents the actual coefficients in the linear relationships used to construct each rotated factor.)

Each pattern (i.e., factor) represents expected interrelationships among the various chemical species. "Ninety-five percent of the total ionic strength of aqueous solutions in (United States) streams is contributed by ... Ca^{++} , Mg^{++} , Na^+ as cations, HCO_3^- , $\text{SO}_4^{=}$ and Cl^- as anions. Other species, such as K^+ , F^{3-} , and NO_3^- commonly account for most of the remaining ionic strength...".²² Factor 1 presents this result most clearly. The second factor groups those ions traditionally associated with hardness, i.e., alkalinity, calcium, magnesium, sulfate, and the carbonate-bicarbonate buffer system (HCO_3^- , pH). The expected grouping of metal ions, Fe, Mn, with $\text{SO}_4^{=}$ and pH (especially with the inverse sign of the latter)²³ is found in factor 3. Figures 18 through 20 show the relative magnitudes of each of the factors as they are distributed throughout the study area.

(Other results consistent with chemical characteristics of natural waters may also be inferred from Tables 8B and C. For example, phosphorus in water frequently occurs primarily as phosphate and is a result of processes associated with the presence of Na^+ , K^+ , and $\text{SO}_4^{=}$: natural weathering and subsequent addition of minerals, soil erosion, use of soluble substances such as phosphate detergents and road salts, etc.)

These factors can now be used in path diagrams as indicators of water quality. However, since each factor represents so much of what might be attributed to natural processes, it is anticipated that any proposed model explaining the impact of man on water quality will necessarily account for only a very limited amount of the variance in each factor.

(VARIABLES ARE STANDARDIZED, FACTORS ARE DIMENSIONLESS)

Figure 18. Distribution of Factor 1 for Northeastern U.S.

3.2 Path Analysis

Figure 21 is a sample causal diagram based on activities occurring within each county. It is assumed that people may affect water quality directly or through various economic activities: commerce, farming, and industry. All these activities demand the use of energy, which in turn promotes the supply of energy, both of which may also affect water quality. Farming is accounted for in terms of the acreage allotted to this activity and the value added of

(VARIABLES ARE STANDARDIZED, FACTORS ARE DIMENSIONLESS)

Figure 19. Distribution of Factor 2 for Northeastern U.S.

farm products sold in each county. Commerce is approximated from data for a county's retail sales. Industrial activities are considered first in terms of the estimated employment in six energy-intensive standard industrial categories (foods; paper; chemicals; petroleum and coal; stone, clay and glass; and primary metals)³⁰ and then in terms of value added in manufacturing in each county. This information is derived primarily from United States census data.²⁶

(VARIABLES ARE STANDARDIZED, FACTORS ARE DIMENSIONLESS)

Figure 20. Distribution of Factor 3 for Northeastern U.S.

Each track leads to estimates of electric energy consumption, which promotes the production of electricity, which then leads to the production of raw fuels. This information was obtained primarily from the BNL energetics atlas.⁸

Availability of water quality data was the limiting factor in determining which counties were included in the path analysis. As described above, 138

Figure 21. A Path Diagram Using Water Quality Factors, Socioeconomic Data, and Energy Data

Figure 22. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #3

counties were found (Figure 3) having complete sets of data for the 19 water quality parameters used in the factor analysis.

Figures 22 to 24 are causal diagrams for the three water quality factors, with the amount of variance explained by the overall model for each variable shown in each box.* These diagrams lead to a number of both new and expected observations:

1. Counties with increased mineral shipments tend to have waters with higher iron, manganese, and sulfate concentrations and lower pH. This may be due primarily to the presence in the analysis of counties in Western Pennsylvania and Ohio which produce and ship large quantities of coal (see Figure 3).

2. Alkalinity, suspended solids, and hardness appear to be greater in counties with high levels of industrial electric consumption and with more acreage devoted to farming.

3. The direct production of electricity in a county does not significantly affect any water quality factor in the same county, whereas a rise in total production of energy tends slightly to decrease alkalinity-hardness. The latter may be due to inclusion in the analysis not only of numerous coal-producing counties but also of New England counties that produce clean hydroelectric power.

4. Counties with increased levels of conductivity and related dissolved ions also have higher levels of industrial electric consumption. However, compensatory mechanisms are also seen between the effects of commercial electric consumption and population (tending to raise factor 1) and those of retail sales and mineral shipments (tending to lower factor 1).

5. The models are all significant at the 0.001 level and account for 25%, 29%, and 40%, respectively, of the variance in water quality factors.

---

*Note that path coefficients may take values greater than unity.

"Consequently, wherever there (are path coefficients of values greater than one) in a causal system, there must exist compensatory variables(s) in the system... In other words, the presence of path coefficient(s) greater than unity in a causal system is a sure signal for the existence of compensatory mechanism(s) (negative influence(s)) in that system." (Li,¹⁴ p. 170).

Figure 23. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #1

Figure 24. Causal Diagram for Energy, Socioeconomic, and Water Quality Factor #2

6. The high level of variance accounted for in the case of electric energy consumption is deceptive, in that the information was derived from census data.⁸

7. More affluent counties tend towards higher retail sales but lower mineral shipments and industrial electric consumption (perhaps a manifestation of the "bedroom community" concept).

8. The path diagram has disregarded other variables that are important in determining water quality (e.g., soil type, underlying geologic formation, land use). For energy production also, additional variables have been neglected. Further analyses defining the importance of these variables should provide interesting results.

9. Counties were uncoupled hydrologically. Further analyses should provide necessary resolution to understand how activities in one county may affect water quality in its neighbors. This concept can be expanded to include the manner in which a county may influence economic and energy developments in its neighbors.

It has been shown that, with sufficient attention paid to the distributional properties of the data, water quality can be characterized on a county basis, and can be used together with other environmental, socioeconomic, and energy information to study their interactions. Path analysis offers an opportunity to understand how water quality may be affected by the myriad human activities broadly classified in terms of socioeconomics and energy.

4 SUMMARY AND CONCLUSIONS

1. Water quality data for each of 19 parameters taken over a 22-year span (1955-1977) were aggregated on a county basis for use in regional analyses incorporating socioeconomic and energy data.

2. Despite spatial and temporal aggregation, many water quality data were unavailable for many parameters in many counties. Complete sets of data for 19 parameters were found for only 138 of 332 counties in the study area (New England, Middle Atlantic States, and Ohio).

3. Use of medians is preferred to use of means.

4. Logarithmic transformation of the medians was desirable for some parameters to obtain better behaved frequency distributions for subsequent analyses.

5. This aggregation methodology sought to characterize a county's waters qualitatively in a form that would enable further analyses in which water quality data are used together with other county information (e.g., socioeconomics, energy, and health). A major underpinning of such analyses is confidence that aggregation does not impair the description of qualitative characteristics of water. Aggregation and statistical transformations did not impair the usefulness of the data in correctly interpreting the expected qualitative chemical characteristics of natural waters.

6. Factor and path analytic techniques show interesting promise for use in describing many of man's impacts on water.

7. Additional information is required to further this approach, including water quality data in those counties with missing values for many parameters and also soil type, geologic, and land use data.

8. Given the availability of this additional information, it would be interesting to propose a model (path diagram) accounting for natural chemical quality on the basis of soil and rock type as well as vegetative cover. The natural levels could then be subtracted from the aggregated values used in the present study to produce a set of residuals, which would in turn be used in path diagrams assessing the ways man affects water quality.

5 REFERENCES

1. Standard Methods for the Examination of Water and Wastewater, 13th ed., American Public Health Association, Washington, DC, 1971.
2. Barr, A.J. et al., A User's Guide to SAS 76, SAS Institute, Inc., Raleigh, NC, 1976.
3. Cattell, R.B., The Scientific Use of Factor Analysis in Behavioral and Life Sciences, Plenum, New York, 1978.
4. Dawdy, D.R. and Feth, J.H., Applications of factor analysis in study of chemistry of groundwater quality, Mojave River Valley, California, Water Resources Res. 3:(2), 505-10 (1967).
5. Dixon, W.J. et al., BMDP-77, Biomedical Computer Program P-Series, U. of California Press, Berkeley, 1977.
6. Duncan, O.D., Path analysis: Sociological examples, Am. J. Sociol. 72:(1), 1-16 (1966).
7. Duncan, O.D., Introduction to Structural Equation Models, Academic, New York, 1975.
8. Drysdale, F.R. and Calef, C.E., The Energetics of the United States of America: An Atlas, BNL 50501-R, Oct. 1977.
9. Ficke, J.F. and Hawkinson, R.O., The National Stream Quality Accounting Network (NASQAN) - Some Questions and Answers, U.S. Geol. Survey Circ. 719, 1975.
10. Garrels, R.M. and Christ, C.L., Solutions, Minerals and Equilibria, Harper & Row, New York, 1965.
11. Gehm, W., Bregman, J.I., and Beeland, G.V., Editors, Handbook of Water Resources and Pollution Control, Van Nostrand Reinhold, New York, 1976.
12. Haupt, H.F. and Snyder, G.G., Modeling Natural Chemical Quality of High Mountain Streams Based on Geologic Parameters, USDA Forest Service, Review Draft, April 25, 1978.
13. Hem, J.D., Study and Interpretation of the Chemical Characteristics of Natural Water, 2nd Ed., U.S. Geol. Survey, Water Supply Paper 1473, 1970.
14. Li, C.C., Path Analysis - A Primer, Boxwood Press, Pacific Grove, CA, 1975.
15. Matalas, N.C. and Reihner, B.J., Some comments on the use of factor analysis, Water Resources Res. 3:(1), 213-23 (1967).

16. Mosteller, F. and Tukey, J.W., Data Analysis and Regression, Addison-Wesley, Reading, MA, 1977.
17. Nie, N.H. et al., Statistical Package for the Social Sciences, 2nd ed., McGraw-Hill, New York, 1975.
18. Ponce, S.L., Striffler, W.D., and Aubertin, G.O., A method to predict baseline water quality on wildland watersheds, presented at Fall Mtg, Am. Geophys. Union, Dec. 1978; Ponce, S.L., Personal communication.
19. Rainwater, F.N. and Thatcher, L.L., Methods for Collection and Analysis of Water Samples, U.S. Geol. Survey, Water Supply Paper 1454, 1960.
20. Snyder, W.M., Some possibilities for multivariate analysis in hydrologic studies, J. Geophys. Res. 67:(2), 721-29 (1962).
21. Steele, J.D., Digital-computer applications in chemical quality studies of surface water in a small watershed, in Symp. on Use of Analog and Digital Computers in Hydrology, Tucson, Dec. 1968.
22. Steele, T.D. and Matalas, N.C., Principal-component analysis of streamflow chemical quality data, in Proc. Symp. on Mathematical Models in Hydrology, Warsaw, July 1971, IAHS-ASH, Publ. No. 100, 1974.
23. Stumm, W. and Morgan, J.J., Aquatic Chemistry, Wiley-Interscience, New York, 1970.
24. Todd, D.K., Editor, The Water Encyclopedia, A Compendium of Useful Information on Water Resources, Water Information Center, Huntington, NY, 1970.
25. Tukey, J.W., Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.
26. County and City Data Book 1972 (A Statistical Abstract Supplement), U.S. Bureau of the Census, GPO, Washington, DC, 1973.
27. Handbook: Water Quality Control Information System, STORET, U.S. Environmental Protection Agency, Washington, DC, 1976.
28. Wolman, M.G., The Nation's rivers, Science 174, 905-18 (1971).
29. Harman, H.H., Modern Factor Analysis, U. of Chicago Press, 1967.
30. Sevia, W., Biomed. and Environ. Assess. Div., BNL, Personal communication.

APPENDIX A						
Lower Detection Limits ($\mu\text{g}/\ell$) *						
Substance	Atomic emission spectrometry			Atomic absorption spectrometry		
	$\text{N}_2\text{O-C}_2\text{H}_2$ flame	dc arc plasma jet	rf plasma torch	Flame	Non-flame	Atomic fluorescence
Al	10	8	1	20	0.001	100
Ag	20	2	0.1	1	1×10^{-4}	0.1
Ba	2.00	5	1	15	0.006	-
Be	4×10^4	0.5	-	2	3×10^{-5}	10
Ca	0.1	0.5	1	2	4×10^{-4}	1
Cd	2×10^3	0.5	0.01	0.2	8×10^{-5}	0.001
Co	50	1	4	2	0.002	5
Cr	5	1	1	2	0.002	40
Cu	10	2	-	4	6×10^{-4}	0.03
Fe	50	30	5	4	0.01	3
Mg	5	1	0.1	0.1	4×10^{-5}	0.2
Mn	5	0.5	1	0.8	2×10^{-4}	5
Ni	30	5	3	5	0.009	3
Pb	200	20	1	10	0.002	10
As	-	-	-	30	0.008	100
Hg	-	-	-	0.1	0.02	0.2
Zn	-	-	-	0.6	3×10^{-5}	5

* See Winefordner, J.D., editors, Trace Analysis Spectroscopic for Elements, Wiley, New York, 1976. Also see Quinby-Hunt, M.S., A survey of instrumentation used for monitoring metals in water, American Laboratory, Dec. 1978.

APPENDIX B

EXPLORATORY DATA ANALYSIS

Characteristics of Data Within Counties

For each county the data base contains summary statistics for each pollutant measured in the county: number of observations, maximum and minimum values, mean and standard deviation, 15th, 50th (i.e., median), and 85th percentiles. The number of observations in a given county varies widely among the parameters (i.e., water quality variables), ranging from less than ten to over a thousand. For many counties there were no STORET data for several parameters. The occurrence of many large values makes the mean an unsatisfactory estimator. The median is therefore indicated for use in the analyses because it is resistant to outliers and because many of the samples are small.

Table B1 reviews data for total dissolved solids, conductivity, and iron found in the waters of Connecticut. It is seen that in each county the distributions have long tails, where the maximum observations range over many standard deviations above the mean. Except for total dissolved solids in Litchfield County, means are always found to be greater than medians.

Meaningful comparisons of the various parameters necessitate a re-expression of the data. Both logarithmic and root transformations were applied,* with the results shown in Table B2. (Since there are zeroes in the

---

*Tukey²⁵ and Mosteller and Tukey¹⁶ illustrate a number of rules used to indicate the type of transformation required to symmetrize data. One computes the differences between the 15th and 50th percentiles and between the 50th and 85th percentiles for both raw and transformed data. These values are closer the greater the symmetry in the distributions. As a variance stabilizing technique, one computes the log median and log of the difference between 15th and 85th percentiles. To remove dependencies of the median on the difference, the ratio of the two is computed: with a value near one-half, a root transformation is indicated, whereas a value near unity indicates a preference for a logarithmic transformation.

Table B1
Aggregated Data for Connecticut

Variable	County	No. of obser- vations	Maximum	Minimum	Mean	S.D.	Percentile		
							15%	50%	85%
Conductivity ($\mu\text{mho/cm}$)	Fairfield	624	1,400	74	210	290	150	201	258
	Hartford	1333	850	2.5	154	261	73	119	242
	Litchfield	1976	645	7.2	140	217	76	110	230
	Middlesex	209	214	43	91	112	73	89	108
	New Haven	597	36,400	54	270	1753	110	187	318
	New London	457	14,700	9.3	136	819	75	99	131
	Tolland	467	290	14	77	105	55	74	93
	Windham	669	370	34	113	163	67	103	152
Iron ($\mu\text{g/l}$)	Fairfield	289	32,000	0	231	2115	35	70	170
	Hartford	661	2,700	0.02	212	457	60	150	330
	Litchfield	1,220	4,100	0	82	279	0.2	0.7	190
	Middlesex	84	1,500	30	205	393	90	170	300
	New Haven	338	3,000	0	218	512	0.6	150	390
	New London	211	1,200	2.6	243	383	100	230	360
	Tolland	349	680	0.02	67	200	0.2	0.62	200
	Windham	486	980	0.04	257	446	70	220	440
Total dissolved solids (mg/l)	Fairfield	276	552	17	137	194	99	125	177
	Hartford	575	615	25	106	182	52	74	183
	Litchfield	353	244	23	113	161	58	115	166
	Middlesex	65	80	36	58	67	49	56	67
	New Haven	243	372	14	142	198	77	138	198
	New London	175	9,500	18	122	834	45	65	82
	Tolland	90	101	27	53	67	41	51	70
	Windham	405	234	2.0	72	105	44	65	102

Table B2
Transformed Data for Connecticut

Variable	County	Logarithmic, $y = \log(x + 1)$				Square root $y = \sqrt{(x+1)}$			
		Percentile				Percentile			
		Max.	15%	50%	85%	Max.	15%	50%	85%
Conduc- tivity	Fairfield	3.2	2.2	2.3	2.4	37	12	14	16
	Hartford	2.9	1.9	2.1	2.4	29	8.6	11	16
	Litchfield	2.8	1.9	2.0	2.4	25	8.8	11	15
	Middlesex	2.3	1.9	2.0	2.0	15	8.6	9.5	10
	New Haven	4.6	2.0	2.3	2.5	191	10	14	18
	New London	4.2	1.9	2.0	2.1	121	8.8	10	11
	Tolland	2.5	1.8	1.9	2.0	17	7.5	8.7	10
	Windham	2.6	1.8	2.0	2.2	19	8.2	10	12
Iron	Fairfield	4.5	1.6	1.8	2.2	179	6.0	8.4	13
	Hartford	3.4	1.8	2.2	2.5	52	7.8	12	18
	Litchfield	3.6	0.08	0.23	2.3	64	1.1	1.3	14
	Middlesex	3.2	2.0	2.2	2.5	39	9.5	13	17
	New Haven	3.5	0.2	2.2	2.6	55	1.3	12	20
	New London	3.1	2.0	2.4	2.6	25	10	15	19
	Tolland	2.8	0.08	0.21	2.3	26	1.1	1.3	14
	Windham	3.0	1.8	2.3	2.6	31	8.4	15	21
Total dissolved solids	Fairfield	2.7	2.0	2.1	2.2	23	10	11	13
	Hartford	2.8	1.7	1.9	2.2	25	7.3	8.7	14
	Litchfield	2.4	1.8	2.0	2.2	16	7.7	11	13
	Middlesex	1.9	1.7	1.8	1.8	9	7.1	7.8	8.2
	New Haven	2.6	1.9	2.1	2.3	19	8.8	12	14
	New London	4.0	1.7	1.8	1.9	97	6.8	8.1	9.1
	Tolland	2.0	1.6	1.7	1.8	10	6.5	7.2	8.4
	Windham	2.4	1.6	1.8	2.0	15	6.7	8.1	10

data, logarithms and square roots are both "started," that is, x becomes $x + 1$ before transformation.)

For almost symmetric distributions the difference between the 85th and 50th percentiles should be a close approximation to the difference between the 50th and 15th percentiles. If the distributions were approximately normal, each of the differences would also be about one standard deviation (since the mean equals the median for the normal distribution). Except in the case of iron in three counties it is seen that either transformation produces more reasonably symmetric distributions than would obtain with raw values. Both transformations have maxima which stray far less than those with raw data; on this basis it is evident that the log transformation is preferred.

Cross County Comparisons

Table B3 illustrates the important properties of distributions of both means and medians for eight dissolved species found in the waters of the Northeast. Once again the non-normality of all distributions is seen, with each case having a skewness greater than that expected (at a 1% significance level) had the distribution been derived from a normal population. Large values (including many that exceed drinking water criteria) occur more frequently than expected under the normality assumption. With few exceptions it is seen that the distributions are also more peaked than they would be if normality were used as a working hypothesis (at a 1% significance level with the Geary test used for kurtosis). For these ions it is thus seen that the normality assumption for the distribution of either county aggregated means or medians is clearly untenable.

With nicely behaved data following bell-shaped distributions, either mean or median can be used as the measure for locating the centers of distributions.¹⁶ Table B3 also indicates that the distributions of the dissolved ions have long (i.e., stretched-out) tails.

"In the usual measurement situations, behavior in the tails is both the most vital and the least securely tied down of all. It is not well tied down because changes in tail size by large and important factors need involve only a few percent of all observations and are therefore both difficult to detect by sampling and easy for transient causes to produce. It is

Table B3								
Distribution of Means and Medians for County Aggregate Data ($\mu\text{g}/\text{L}$)								
Variable	No. of Counties	Water quality criteria*	Max. value	Mean	S.D.	95% confidence interval	Skew- ness	Kur- tosis
<u>Means</u>								
Arsenic	230	50	230	6.2	25	3-9	7	55
Barium	171	1000	1000	87	134	66-107	4	24
Cadmium	243	10	803	10	55	3-17	12	173
Copper	292	1000	615	37	71	29-46	5	27
Lead	268	50	854	27	84	17-37	6	48
Mercury	222	2	275	4.0	23	0-6	10	104
Nickel	230	100	425	31	50	24-38	5	30
Silver	173	50	98	5.0	12	3-7	4	23
<u>Medians</u>								
Arsenic	230	50	250	3.3	17	1.1-5.6	13	182
Barium	171	1000	1000	81	131	61-100	4	27
Cadmium	243	10	150	3.3	13	1.7-4.9	9	94
Copper	292	1000	200	18	28	15-21	16	257
Lead	268	50	700	12	46	7-18	4	15
Mercury	222	2	550	3.0	37	2-8	12	177
Nickel	230	100	430	26	45	20-32	4	30
Silver	173	50	30	3.9	9.6	2.4-5.3	2	3
* <u>Quality Criteria for Water</u> , U.S. Environmental Protection Agency, EPA-440/9-76-023, 1976.								

vital because the extreme tails of the distribution of individual values are likely to influence large portions of the distributions of many derived quantities."¹⁶

Behavior in the tails is especially important to aggregated water quality data as it indicates the locations that may be most seriously affected by

contaminants.* One would thus question the validity of a county mean or median which greatly exceeded criteria levels set by legislation where, in fact, large numbers of observations existed for the given parameter. Furthermore, the aggregation process utilized in this analysis characterizes chronic exposures to waterborne pollutants. As is often the case, an accidental spill of some substance will initiate a short-term monitoring program which would result in the appearance in STORET of a few extreme observations taken over a relatively short period.

The question therefore arises as to whether large values of either mean or median occur more frequently in counties with few observations. The scattergrams of Figures 1B through 4B indicate that this indeed appears to be true for many of the water quality parameters considered here. (This does not suggest any correlation, but merely that the number of observations found in a county appears to have a threshold below which there is a likelihood of experiencing many large, and possibly spurious, values.) A later discussion will focus directly on those counties characterized by large aggregated means or medians (with the simultaneous occurrence of either small or large numbers of observations). Because analyses were contemplated which would use transformations and combinations of the various water quality variables, it was therefore decided to trim each distribution in an attempt to make them more nicely behaved.**

Table B4 illustrates the manner in which trimming affects the properties of the ionic distributions. For each variable, counties were eliminated from consideration first if fewer than five observations were found and next if

---

* Future work will concentrate on counties with very large median values. Ongoing analyses are meant to assess chronic impacts on water quality.

** It should be emphasized that we are not preoccupied with normalizing per se: "The history of statistics and data analysis is a messy mixture of healthy skepticism and naive optimism about the exact shapes of the distributions of observations. Such optimism has often been inflated by the wonderful properties of a single family of distributions, the 'normal' distributions..."²⁵

Figure B-1. Medians vs. Number of Observations

Figure B-2. Medians vs. Number of Observations

Figure 3-3. Medians vs. Number of Observations

Figure B-4. Medians vs. Number of Observations

Table B4

Trimming the Distribution of Means and Medians Using Criterion of
Minimum Number of Observations in Each County

Variable	Threshold on min. No. of observations in each county	No. of counties meeting threshold	Mean ($\mu\text{g}/\ell$)	S.D.	Skew- ness	Kur- tosis
<u>Means</u>						
Arsenic	5	178	5.8	24	7.9	64
	10	141	6.8	27	7.1	50
Barium	5	109	90	91	1.1	0.5
	10	77	91	89	1.0	0.1
Cadmium	5	172	7.5	23	6.5	48
	10	133	8.0	25	6.4	45
Copper	5	246	40	76	4.5	24
	10	203	40	79	4.6	25
Lead	5	210	30	91	6.2	43
	10	166	28	90	6.7	51
Mercury	5	141	3.4	17	8.8	82
	10	110	4.2	19	7.7	63
Nickel	5	140	35	59	4.3	24
	10	102	36	62	4.6	25
Silver	5	104	8.1	15	3.1	13
	10	64	7.9	17	3.6	15
<u>Medians</u>						
Arsenic	5	178	1.9	2.8	2.9	11
	10	141	1.7	2.8	3.4	15
Barium	5	109	79	82	0.7	-1.3
	10	77	81	82	0.6	-1.4
Cadmium	5	172	2.0	3.8	2.8	9.1
	10	133	1.9	3.9	3.0	10
Copper	5	246	17	28	4.0	19
	10	203	17	29	4.1	19
Lead	5	210	9.0	20	6.1	43
	10	166	6.9	12	5.7	40
Mercury	5	141	0.5	1.1	2.3	4.5
	10	110	0.5	1.1	2.1	3.6
Nickel	5	140	24	51	4.9	32
	10	102	24	55	5.1	32
Silver	5	104	5.9	12	1.5	0.4
	10	64	4.4	10	2.0	2.1

fewer than ten observations were found. Trimming does reduce the maximum value found for each species. Table B4 indicates that all distributions of mean values are non-normal at the 1% significance level. Despite trimming the skewness and kurtosis of these distribution of means remain relatively unchanged. The analogous situation with distribution of medians is also shown in Table B4. Normality is still rejected at the 1% level on the basis of skewness, but note that in some cases (e.g., mercury and silver) the kurtosis does not depart greatly from what would be expected from a normal population. Furthermore, it is seen that both test parameters (skewness and kurtosis) are reduced in magnitude more for the distribution of medians than for means.

The sample mean is expected to be sensitive to a small number of values straggling far from the center of the distribution, and indeed this is the case for the aggregated water quality data retrieved from STORET. All counties with less than five measurements of a given parameter are thus excluded from all calculations involving that parameter. (This threshold also was found to prevent the occurrence of counties with so few observations that derived statistics are meaningless and also to eliminate only a relatively small number of counties from overall consideration.) Furthermore, on the basis of the foregoing calculations it was decided to utilize order statistics, specifically medians, in all subsequent analyses. (Note that order within a population is maintained when monotonic transformations are applied to distributions of medians. Such is not the case with means.)

As in the preceding section, the use of transformations is again suggested to produce a set of distributions that are better behaved and thus more useful for subsequent analyses (in which water quality variables will be compared with various demographic, economic, and energy indicators). Table B5 shows the results of both log and root transformations applied to the distribution of ionic medians for both trimmed and untrimmed data. It is seen that both transformations produce distributions with closer means and medians. Both skewness and kurtosis are again significant at both 1% and 5% levels for both transformations. It is clear, however, that the distributions of raw medians for some variables can be made better behaved via transformation. Table 3 (in the text) indicates the transformations applied to water quality data for use in subsequent analyses.

Table B5

Trimming the Distribution of Log (Median + 1.0) and
Square Root (Median + 1.0) Using Criterion of Minimum Number of
Observations in Each County

Variable	Threshold on min. No. of observations in each county	No. of counties meeting threshold	Mean ($\mu\text{g}/\text{l}$)	S.D.	Skew- ness	Kur- tosis
<u>Log</u>						
Arsenic	5	178	0.8	0.7	0.9	0.5
	10	141	0.7	0.7	0.9	0.8
Barium	5	109	3.4	1.8	-0.8	-0.3
	10	77	3.5	1.7	-0.9	-0.1
Cadmium	5	172	0.6	0.9	1.2	0.2
	10	133	0.6	0.8	1.3	0.6
Copper	5	246	2.1	1.3	-0.2	-0.6
	10	203	2.1	1.3	-0.1	-0.6
Lead	5	210	1.7	1.0	0.6	1.2
	10	166	1.6	0.9	0.3	1.1
Mercury	5	141	0.2	0.5	1.8	1.8
	10	110	0.3	0.5	1.7	1.3
Nickel	5	140	2.1	1.5	0.4	-0.6
	10	102	2.2	1.4	0.4	-0.4
Silver	5	104	0.7	1.4	1.5	0.2
	10	64	0.6	1.2	1.9	1.6
<u>Square Root</u>						
Arsenic	5	178	1.6	0.6	1.8	3.7
	10	141	1.5	0.6	2.0	5.1
Barium	5	109	7.5	4.9	0.3	-1.4
	10	77	7.7	4.8	0.3	-1.4
Cadmium	5	172	1.5	0.8	1.8	9.1
	10	133	1.5	0.8	2.0	2.9
Copper	5	246	3.5	2.3	1.8	3.6
	10	203	3.5	2.4	1.9	5.0
Lead	5	210	2.7	1.7	3.2	14
	10	166	2.5	1.3	2.8	12
Mercury	5	141	1.2	0.4	2.0	2.9
	10	110	1.2	0.4	1.9	2.2
Nickel	5	140	3.8	3.3	2.0	5.4
	10	102	3.8	3.3	2.4	7.4
Silver	5	104	1.9	1.8	1.5	0.3
	10	64	1.7	1.6	2.0	1.9

APPENDIX C

Figure C-1. Complete Path Diagram

APPENDIX C (cont'd)

Figure C-2. Complete Path Diagram

APPENDIX C (cont'd)

Figure C-3. Complete Path Diagram