Comparative Analysis of Modeling Techniques for Coliform Organisms in Streams

JEROME L. MAHLOCH
Mississippi State University, State College, Mississippi 39762

Received for publication 19 June 1973

The use of models for predicting changes in water quality parameters is currently considered an integral part of river basin management. The application of modeling techniques to coliform organisms is in its infancy due to the complexities involved and the lack of definitive information on coliform populations in natural environments. The purpose of this study was to make a comparative analysis of the available models for coliform organisms in order to improve on the state of the art of this subject. The available coliform models may be classified into deterministic or statistical types. In this study, six different models, three of each type, were selected for analysis and were applied to coliform data available on the Leaf River. Results of comparing the models indicated that a deterministic model was best suited for total coliform and a statistical model was best suited for fecal coliform. Ultimate selection of a model for coliform organisms is dependent not only on the accuracy of the model but on ease of implementation. Current technology would probably dictate the use of a deterministic model because of the lack of a complete data base on which to base statistical models.

The presence of coliform bacteria in a natural waterway is generally regarded as evidence of pollution by warm-blooded animals. Testing for coliform bacteria usually encompasses both total coliform and fecal coliform measurements, recognizing the influence of natural pollution on the former. The use of the fecal coliform as an indicator organism is currently the most popular method; however, its use as an indicator organism is in a constant state of flux and debate (1, 6). Mechalas et al. (12) recently reviewed the value of total and fecal coliform organisms as indicators of recreational water quality. Since both total and fecal coliform measurements are still the most prevalent bacterial indicator organisms, this study will be confined to these organisms.

Because of the recent interest in basin wide planning for water quality management, the development of mathematical models for describing the various water quality parameters has been encouraged. The use of these models for describing certain water quality parameters such as biochemical oxygen demand (BOD) and dissolved oxygen has been successful to a certain degree. In contrast, attempts to model coliform bacteria in natural waterways have been limited, doubtlessly due to the complexity inherent in describing biological populations in a natural environment. Dixon et al. (8) made an exploratory effort to model coliform organisms in a stream, but the results of the study were inconclusive. Currently most of the work relating to coliform population in streams has been limited to attempts to describe survival and growth rates in the natural environment (10, 11). Brasfield (4) has related bacterial populations in a stream to environmental factors and developed multiple regression models for prediction of populations.

The two types of models used for describing coliform populations are deterministic and statistical. Deterministic models present an exact mathematical relationship between coliform population and a given set of variables, whereas statistical models present a probabilistic relationship between coliform population and a given set of variables. The purpose of this study was to summarize the state of the art for modeling coliform organisms in natural waterways and to make a comparative analysis between the two types of aforementioned models. The term "modeling" will be used in the context of predicting coliform populations using a prescribed model, and fitting the model to observed data.

MATERIALS AND METHODS

The growth of microorganisms in a stream is a function of the following effects: (i) birth and death,
(ii) Emigration and immigration, (iii) diffusion and advective transport, and (iv) sources and sinks. The birth and death processes are generally dependent on the environmental conditions and predation. Emigration and immigration are due to the motility of the bacteria. Diffusion and advective transport are characteristic of the movement of the fluid in which the bacteria reside. Sources and sinks are additions or withdrawals to the stream and may or may not affect the concentration of organisms. Since the statistical model is concerned only with probabilistic dependencies between the coliform concentration (criteria) and the assigned variables (predictors), the aforementioned factors have no direct bearing on this type of modeling. In the case of deterministic models, the following assumptions were made as to the above effects on coliform populations. (i) Emigration and immigration were considered insignificant. (ii) Diffusion was considered insignificant because only fresh water systems were studied. (iii) The only sources considered were point sources such as domestic and industrial effluents or tributary inflow. (iv) Sinks had no effect on the coliform concentration. (v) The net effect of birth and death processes was assumed to be described by the deterministic model selected. One possible source of error in this assumption is the neglect of overland runoff, particularly as it affects total coliform populations. The reason for discarding this factor was the inavailability of data as to the contribution from overland runoff in the study area.

**Study area and data.** The stream used for this study was the Leaf River which is located in the Pascagoula River Basin in the southeastern section of Mississippi. Figure 1 presents a schematic diagram of the Leaf River along with the location of all sources used in the model, as well as the location of all sampling points. Data for use in the models were obtained from a Federal Water Quality Administration (FWQA) survey and the United States Geological Survey (USGS) in Mississippi. Fecal and total coliform data were available from each survey and were used independently. Coliform concentrations were expressed as log number per 100 ml.

All statistical analyses employed in this study followed the methods of Cooley and Lohnes (7). Hereafter, for ease of discussion, the models under consideration will be identified by Roman numerals as introduced below.

**Deterministic models.** Three types of deterministic models were used in this study. The first model (I) used is presented in equation 1 (13):

$$N_t = N_0 e^{-\lambda t}$$  \hspace{1cm} (1)

where $N_t$ = coliform bacteria concentration at time $t$; $N_0$ = initial coliform concentration; $\lambda$ = net death rate (1/day); $t$ = time of travel in the stream (days). The time of travel was considered to start at either the head of the river or at a point of source addition. Advective transport or hydraulic dilution was assumed to occur only at source addition and was therefore not considered as an integral part of the model since time was reinitialized at each source addition. The rate coefficient ($\lambda$) was determined by an iterative fit of the model equation to the observed data.

The second model (II), equation 2, incorporated a growth curve into the functional relationship (9). Equation 2 is:

$$N_t = N_0 \left( \frac{N}{N_0} e^{-\lambda t} \right)$$  \hspace{1cm} (2)

where $N_t$ = coliform bacteria concentration at time $t$; $N_0$ = initial coliform concentration; $t$ = total residence time (days); $\alpha, \beta$ = constants. Total residence time is composed of time of travel in the stream plus an additional lag time factor which allowed positioning along the growth curve. The lag time was determined by an iterative fit of the model equation to the observed data.

The third, deterministic model (III) used was derived from the literature (9) and is presented in equation 3:

$$N_t = N_0 (1 + nkt)^{-\lambda}$$  \hspace{1cm} (3)

where $N_t$ = coliform bacteria concentration at time $t$; $N_0$ = initial coliform concentration; $t$ = time of travel; $k$ = die-away coefficient; $n$ = coefficient of nonuniformity. The values of $k$ and $n$ were varied in this study to obtain the best results using model III.
The hydrologic parameters for determining advective transport of the coliform organisms were ascertained from stream data of the river basin. Tributary inflow as sources of coliforms were available for both sources of data (USGS and FWQA) and were added at the appropriate distances along the stream modeled. Other sources such as municipal and industrial waste effluents were adjusted for removal of coliform organisms due to any treatment and were added by the model.

**Statistical models.** Three types of statistical models were evaluated in this study. The first model (IV) was a multiple regression model developed from literature data (2). The model assumed a dependency between coliform population and time of travel and temperature. The multiple regression equations thus developed for total and fecal coliform are presented in equations 4 and 5, respectively:

\[ N_t = -0.221 - 0.4577 t + 0.0095 T \]
\[ R^2 = 0.573 \]  \hspace{1cm} (4)

where \( N_t \) = coliform bacteria concentration at time \( t \); \( N_e \) = initial coliform concentration; \( t \) = time of travel (days); \( T \) = temperature (C).

\[ N_t = N_e - 0.276 - 0.4784 t + 0.0044 T \]
\[ R^2 = 0.520 \]  \hspace{1cm} (5)

The second statistical model (V) was developed from the data obtained from the United States Geological Survey for the entire Pascagoula River Basin. In this model a regression relationship was explored between the coliform population and environmental factors in the stream. Fifteen different environmental variables were used in the analysis, and the dimensionality of the problem was reduced by prior extraction of four orthogonal factors. The extracted factors were weighted on the following variables: (i) temperature, (ii) flow, (iii) dissolved oxygen and BOD, and (iv) nitrogen forms. This procedure allowed the generation of four sets of factor scores which were used in combination to elucidate the best set of original environmental variables to use in the multiple regression model. A listing of the original environmental variables used along with the factoring results is presented in Table 1.

The third type of statistical model (VI) to be explored was a regression model which employed canonical variates. The canonical variates were obtained by using canonical weights derived from the analysis to transform the original measurement variables. Table 2 presents a summation of the canonical analysis on the original data. The result obtained was two linear regression equations in terms of the transformed variables. The dependent variables from these equations can be transformed back to the original measurement space (i.e., coliform population) by using the inverse of the canonical weight matrix.

**Evaluation of models.** The evaluation of the models was accomplished by using the absolute value of the difference between observed and calculated values. These values were summed and divided by the number of points used to yield a mean residue for each model considered, which allowed a comparison of the models to be made. The use of a mean residue also allowed evaluation of the first two deterministic models (I and II) at each sampling point along the steam by minimization of this residue.

**RESULTS**

A comparison of models to predict coliform population based on only the ability to accurately predict results is inane, since other considerations such as ease of computation, availability of data, and soundness of method are also involved. In this particular study, a deterministic model based on first-order kinetics gave the best results for total coliforms and gave poor results for predicting fecal coliforms. In contrast, models V and VI, which were statistical in nature, gave good results when applied either to fecal or total coliform populations. It may be argued that relating fecal coliform population to environmental conditions in the stream will give a better fit of the observed data over kinetic models since fecal coliforms will be more sensitive to a changing environment than total coliforms. This would indicate that two models are needed for modeling both total and fecal coliforms, yet this would probably be impracticable in terms of effort expended. Both the statistical and deterministic models employed in this study had advantages and disadvantages of merit in considering their adoption for use in any river basin study. The results of applying the models are presented in terms of mean residues for the data sets employed in Tables 3 and 4.

**Model I.** This model yielded the best results for the total coliform population. Since the rate of coefficients were determined by an iterative procedure, application of this model for predictive purposes requires verification. This is particularly true since the rates for the two sets of total coliform data employed differed; however, the rates for the fecal coliform data were equal.
suggesting that this particular constant would yield reproducible results. The rate constants (1/day) determined for the FWQA data and USGS data were 0.0 and 0.40, respectively, for total coliform; they were 0.40 and 0.40, respectively, for fecal coliform. An iterative approach for determination of rate constants has the advantage of allowing individual consideration of points along the stream to point out anomalies or inconsistencies in the data. In this particular study, examination of individual points along the stream for the best rate coefficient indicated that regrowth was occurring at the upper portion of the stream and normal die-away was occurring at the lower end of the stream. In addition, the rate coefficients varied greatly with distance along the stream. A plot of the rate constants for a specific location is presented in Fig. 2. The availability of this type of data allows the investigator to determine points along the streams which will allow coliform regrowth, to show possible errors in the observed data, and to determine problem areas in the stream with respect to coliform quality. Rate coefficients determined by other investigators include 0.64 (1/day) by Canale et al. (5) and 0.22 (1/day) by Bhagat et al. (3) at 20°C. These coefficients lie within the range of the rate constants determined in this paper.

**Model II.** The results from applying the regrowth equation to the coliform data gave poorer results for the USGS data set and comparable results for the FWQA data set. The values for lag times to be imposed on each system were equal between data sets and were

### Table 1. Results of factoring USGS water quality data

| Test                  | Factor pattern | Communality |
|-----------------------|----------------|-------------|
|                       | Factor I | Factor II | Factor III | Factor IV |           |
| Temperature           | 1.00     | 0.00      | 0.00       | 0.00      | 1.00      |
| pH                    | 0.29     | 0.07      | 0.10       | 0.16      | 0.15      |
| Flow                  | 0.37     | 0.93      | 0.00       | 0.00      | 1.00      |
| Dissolved oxygen      | -0.29    | 0.16      | 0.53       | -0.36     | 0.52      |
| Chloride              | -0.05    | 0.08      | -0.20      | 0.08      | 0.05      |
| Dissolved solids      | -0.06    | -0.03     | 0.29       | 0.24      | 0.14      |
| BOD                   | 0.06     | -0.03     | 0.62       | 0.36      | 0.52      |
| Specific conductance  | -0.10    | 0.06      | 0.12       | 0.19      | 0.06      |
| Color                 | -0.12    | -0.04     | -0.13      | 0.31      | 0.13      |
| Ammonia-N             | 0.28     | 0.10      | -0.13      | -0.31     | 0.21      |
| Nitrite-N             | 0.21     | -0.23     | 0.11       | 0.59      | 0.46      |
| Organic-N             | -0.08    | 0.23      | 0.22       | 0.50      | 0.38      |
| Nitrate-N             | 0.13     | -0.06     | -0.32      | 0.47      | 0.34      |
| Phosphate             | -0.26    | 0.04      | -0.02      | 0.44      | 0.26      |
| Sodium                | 0.00     | -0.02     | -0.22      | 0.25      | 0.11      |

### Table 2. Canonical analysis on the measurement variables

| Measurement variable* | Canonical variate I | Canonical variate II |
|-----------------------|---------------------|---------------------|
| Fecal coliform        | Temp (°C)           |                     |
| Total coliform        | Flow (cfs)          |                     |
|                       | Dissolved oxygen (mg/liter) |             |
|                       | Ammonia nitrogen (mg/liter) |            |
|                       | Nitrite (mg/liter)   |                     |
|                       | Organic nitrogen (mg/liter) |           |
|                       | Nitrate (mg/liter)   |                     |
|                       | BOD (mg/liter)       |                     |

* $R_1^2 = 0.76; R_2^2 = 0.48; \text{total redundancy} = 0.52.$

### Table 3. Modeling results, total coliform

| Model type | Mean residue (log no.) | USGS Data | FWQA Data |
|------------|------------------------|-----------|-----------|
| I          | 0.35                   | 1.06      | 0.96      |
| II         | 1.54                   | 1.56      | 1.56      |
| III        | 1.69                   | 2.02      | 2.02      |
| IV         | 1.60                   | 1.44      | 1.44      |
| V          | 0.48                   | 0.61      | 0.61      |

### Table 4. Modeling results, fecal coliform

| Model type | Mean residue (log no.) | USGS data | FWQA Data |
|------------|------------------------|-----------|-----------|
| I          | 1.22                   | 1.25      | 1.25      |
| II         | 2.00                   | 1.68      | 1.68      |
| III        | 1.95                   | 1.97      | 1.97      |
| IV         | 1.73                   | 1.88      | 1.88      |
| V          | 0.59                   |           |           |
| VI         | 0.52                   |           |           |
in the declining portion of the growth curve. The lag times (day) determined for the FWQA data and the USGS data were 0.75 and 0.75, respectively, for total coliform; they were 1.00 and 1.00, respectively, for fecal coliform. The imposed lag time was greater for the fecal coliform data than for the total coliform data, indicating a greater rate of die-away for the fecal coliform. An investigation of individuals points along the stream with respect to this model did demonstrate regrowth in certain areas, but the overall usefulness of this type of model over a simple kinetic model was not demonstrated. Furthermore, the growth equation model assumed in this study was empirical in nature and may not represent the growth of coliform organisms in a stream. The effort required for refinement of this model to give an improved deterministic model is probably not warranted since it would be limited in application.

Model III. The application of model III to the observed data did not yield good results. This model required prior evaluation of constants which were an integral part of the equation (namely, equation 3). These coefficients were varied within the suggested limits (9) but with little improvement in the performance of the model. This fact would suggest that the use of this model for describing coliform population is not universally applicable.

The use of deterministic models for predicting coliform population in streams generally involves some prior determination of a constant or modeling parameter and then verification of the model against a replicate set of data. During this study, verification of models I and II was possible since two sets of data were available. All cases, except for model I for total coliform, yielded good verification of modeling constants which were determined. The problem of verification of these modeling constants is one which must be checked constantly to insure the constants have not changed. The factors affecting one type of deterministic modeling constant, namely kinetic rate constants, have been discussed by Canale et al. (5). Despite the fact that verification was made for these deterministic models, the accuracy of these models, as determined by mean residue from observed data, was comparable or worse than the accuracy of the statistical models.

Model IV. Model IV, which incorporated rate coefficients and temperature relationships derived from data presented by other investigators, did not yield good results when applied to the study area. This fact would appear to demonstrate that the model constants for a particular stream may not be universally applied or that the model itself is not correct.

Model V. The multiple regression procedure used for deriving model V yielded the following equations:

\[ Y = 2.488 - 0.065 \text{(NH}_3\text{)} + 12.32 \text{(NO}_2\text{)}^{-} + 0.503 \text{(org-N)} + 1.108 \text{(NO}_3\text{)}^{-} \]  
where \( Y = \log_{10} \text{total coliform; } R^2 = 0.37. \)

\[ Y = 1.394 + 0.211 \text{(NH}_3\text{)} + 5.18 \text{(NO}_2\text{)}^{-} + 0.558 \text{(org-N)} + 1.25 \text{(NO}_3\text{)}^{-} \]  
where \( Y = \log_{10} \text{fecal coliform; } R^2 = 0.33. \)

The multiple regression models derived from the data available on the river basin yielded fairly good results. The statistical soundness of these multiple regression models was not too good, but the accuracy based on mean residue from the observed data indicated that the coliform populations were mainly dependent on the concentration of the various forms of nitrogen present in the stream. This fact indicated the strong dependency of coliform survival on the environmental condition of a stream; a fact which is not readily apparent by using deterministic models. Inclusion of additional independent variables in multiple regression models improves the statistical validity of the model. The number of data points requisite for consideration of large numbers of independent variables is usually large, and generally these types of data are not available. Additionally, there are inherent problems in applying multiple regression models for predictive purposes (7).

Model VI. The last statistical model tried, model VI, employed canonical transforms of variables to maximize the correlation between the coliform population and environmental conditions of the stream. Computationally, the solution of the canonical analysis and derivation of the subsequent model is simple and provides a simultaneous solution between both total and fecal coliform population and the independent variables. This model provided the best solution for fecal coliform population. These models were based on data from the entire river basin but yielded good results when applied to a restricted portion of the basin (e.g., the Leaf River).

It is apparent from the above discussion that the use of statistical versus the use of deterministic models for determining coliform populations is dependent on the type of coliform organism to be modeled. A deterministic model yielded the best result for total coliform, presumably because of a lack of a strong dependency between environmental conditions in the
stream and total coliform population. Conversely, since a dependency did exist for fecal coliform, a statistical model incorporating this relationship proved to be the best model. The deterministic model may be modified to include the effect of environmental parameters, but sufficient data are not yet available to warrant such a modification. A further question as to the general applicability of the models developed also requires clarification. The results of this work demonstrate that models or constants previously developed do not apply to other streams; consequently, the choice and evaluation of a model should be accompanied by extensive verification.

The use of coliform models is usually limited to predicting coliform populations in streams. Since the deterministic model is simple to implement for predictive purposes, it is usually preferred over the statistical models which require input of several independent variables which compound the error in using statistical models for predictive purposes. The advent of additional amounts of water quality data on river basins may well merit reexamination of statistical models for predictive purposes. Since these statistical models have an increased sensitivity to environmental conditions in streams, they would make definite contributions to water quality studies using coliform models.

ACKNOWLEDGMENT

A portion of this study was supported through a research contract from the Pat Harrison Waterway District.

LITERATURE CITED

1. Adams, J. C. 1972. Unusual organism which gives a positive elevated temperature test for fecal coliforms. Appl. Microbiol. 23:172-173.

2. Ballentine, R. K., and F. W. Kittrell. 1969. Observation of fecal coliforms in several recent stream pollution studies. Federal Water Pollution Control Administration, Cincinnati, Ohio.

3. Bhagat, S. K., W. H. Funk, and D. C. Johnstone. 1972. Correlated studies of Vancouver lakewater quality prediction study. Environmental Protection Agency. Report No. R2-72-111. Washington, D.C.

4. Brasfield, H. 1972. Environmental factors correlated with size of bacterial populations in a polluted stream. Appl. Microbiol. 24:349-352.

5. Canale, R. P., R. L. Patterson, J. J. Gannon, and W. F. Powers. 1973. Water quality models for total coliform. J. Water Pollut. Cont. Fed. 45:325-336.

6. Cherry, W. B., J. B. Hanks, B. M. Thomason, A. M. Murtin, J. W. Biddle, and J. M. Croom. 1972. Salmonellae as an index of pollution of surface waters. Appl. Microbiol. 24:334-340.

7. Cooley, W. W., and P. R. Lohnes. 1971. Multivariate data analysis. John Wiley & Sons, Inc., New York.

8. Dixon, N. F., D. W. Hendricks, A. C. Huber, and J. M. Bagley. 1970. Developing a hydro-quality simulation model. Utah Water Research Laboratory, Logan, Utah.

9. Frankel, R. J. 1965. Economic evaluation of water quality, an engineering-economic model for water quality management. SERL Report No. 65-3. University of California, Berkeley.

10. Hendricks, C. W. 1972. Enteric bacterial growth rates in river water. Appl. Microbiol. 24:168-174.

11. McFeters, G. A., and D. G. Stuart. 1972. Survival of coliform bacteria in natural waters: field and laboratory studies with membrane-filter chambers. Appl. Microbiol. 24:805-811.

12. Mechalis, B. J., K. K. Hekimian, C. A. Schinazi, and R. H. Dubey. 1972. Water quality criteria data book, vol. 4. Environmental Protection Agency, Washington, D.C.

13. Pielou, E. C. 1969. An introduction to mathematical ecology. Wiley-Interscience, New York.