APPLICATION OF POISSON REGRESSION TO THE ANALYSIS OF BACTERIOLOGICAL DATA

A.H. El-Shaarawi[1], A. Maul[2] and J.-C. Block[2]

[1]National Water Research Institute

Canada Centre for Inland Waters

Burlington, Ontario, Canada, L7R 4A6

[2]Centre des Sciences de l'Environnement

Université de Metz

1 rue des Récollets-57000, METZ, FRANCE

Application of Poisson Regression
to the Analysis of Bacteriological
Data

A.H.El-Shaarawi, A.Maul, J.C.Block

NWRI # 86-19

SUMMARY

This paper presents and suggests the use of Poisson regression analysis for modelling the association between bacterial counts observed in a drinking water distribution system and a set of explanatory variables. When the dependent variable is a count that follows the Poisson distribution, the procedure developed in this work is much more appropriate than the conventional method of applying ordinary regression analysis after transforming the counts using the square root transformation, since such a transformation may not satisfy all the conditions needed for performing regression. A detailed description of the procedure used for calculating the maximum likelihood estimates of the unknown parameters of the model and their standard errors is given. The method is illustrated by studying the spatial and temporal variation of heterotrophic bacterial counts in the drinking water distribution system of the city of Morsang in France. The results indicated significant variations due to both the spatial and temporal characteristics of the distribution of bacteria in the network and also that there was a strong interaction between those characteristics. Further an attempt to model the temporal and spatial variation showed that a sizeable portion of the total variability could be explained by using temperature and turbidity as predictor variables.

SOMMAIRE

Dans cette étude, on suggère d'effectuer l'analyse de régression de Poisson pour modéliser les relations qui existent entre la numération bactérienne dans un système de distribution d'eau potable et un ensemble de variables explicatives. Si on emploie comme variable dépendante la numération bactérienne et que les valeurs de celle-ci correspondent à une distribution de Poisson, la procédure que nous présentons convient nettement mieux que la méthode classique, laquelle consiste à effectuer une analyse de régression ordinaire après avoir transformé les valeurs de la numération au moyen de transformations de racines carrées, car cette dernière ne permet pas nécessairement de satisfaire toutes les conditions requises pour l'analyse de régression. On décrit en détail la procédure employée pour calculer les probabilités maximales estimatives des paramètres inconnus du modèle et on indique les erreurs-types. On illustre la méthode en prenant pour exemple l'étude des variations dans le temps et dans l'espace de numérations de bactéries hétérotrophes dans le système de distribution d'eau potable de la ville de Morsang en France. Les résultats ont révélé des variations importantes dues aux caractéristiques spatiales et temporelles de la distribution des bactéries dans le système de distribution et mis en évidence l'intensité des interactions qui jouent entre ces caractéristiques. De plus, en tentant de modéliser les variations spatiales et temporelles de la distribution on a pu démontrer qu'une portion importante de la variabilité globale pouvait s'expliquer en prenant le température et la turbidité comme variables explicatives.

INTRODUCTION

The common approach for studying the dependence of bacterial counts on a set of physical and chemical predictor factors usually requires prior transformation of the crude counts. Regression analysis is then performed with the transformed values as the dependent variables and the other factors as the independent variables. The choice of a specific transformation is based on the variance-mean relationship, which in turn is a function of the probability distribution of the observed values. The square root transformation, for instance, is appropriate when the distribution of the untransformed dependent variable in Poisson which is commonly assumed in applied microbiology when the bacteriological data consists of plate counts. The justification of such transformations lies in the fact that the variance of the transformed variables will be approximately constant which is one of the requirements for the application of the standard regression methods. However, this is not enough, to justify the application of the usual regression technique since the other requirements for performing regression analysis may not be satisfied. Thus, several kinds of transformations of the variables are therefore available to achieve normality and homogeneity of the variances, which are both necessary, on strictly mathematical grounds, to perform regression analysis. For example, a transformation as that suggested by Anscombe [1953] is more successful in

normalizing the distribution of the transformed variables. Nevertheless, a single transformation cannot both stabilize the variance and give approximate normality as effectively as these objectives can be achieved separately [Plackett, 1981]. Transformation of the data should therefore not be used incautiously. Further, whichever transformation is used, the assumption has to be made that the effects of the independent variables are additive on the transformed scale. The present paper sets forth an alternative approach for performing regression analysis which does not require any transformation of the dependent variables since it is based on the exact assumed probability distribution. This approach will be called in the paper Poisson regression. Note that Poisson regression models have recently been applied to epidemiologic follow-up studies while estimating the unknown parameters of the regression function by means of iteratively reweighted least squares (IRLS) [Frome, 1983]. Further, the IRLS was shown to be equivalent to the maximum likelihood procedure [Frome et al., 1973] which is developed and emphasized in the present work, using a nonlinear (i.e. the log-linear) regression model. The computation of the maximum likelihood estimates of the parameters when Poisson regression is used has to be done numerically using a computer. Newton-Raphson method is suggested here for performing the computation because it offers the advantage for also producing the variance-covariance matrix of the estimates. Hence different tests could be assessed and confidence intervals could be constructed for comparing and estimating the values of the parameters of the

regression function. Moreover, the likelihood ratio test [Rao, 1973] is presented which allows the assessment of the effects of introducing new factors in the model.

The Poisson regression technique is used to study the spatial and temporal variabilities of heterotrophic bacterial counts in a drinking water distribution system. The data were generated from the analysis of water samples which were collected at five different locations and at different times during the summer, fall and winter of 1981 from the Morsang drinking water distribution system. The object of the study was to determine the changes in the water bacteriological quality in the network both spatially and temporally which would undoubtedly help to understand the mechanisms of the incidence and the spreading of bacteria in a network. Further, an attempt is made to model the spatial and temporal variation using temperature and turbidity measurements as predictor variables.

STATISTICAL METHODS

Poisson Regression Model

Let $r_1, r_2, \ldots, r_n$ be n independent Poisson random variables and let λ_i be the expected value (the theoretical mean) of r_i. The aim of the Poisson regression is to assess the dependence of λ_i on a set $x_1, \ldots, x_p$ of p explanatory variables where n is sufficiently greater than p to ensure estimability of the parameters.

In order to apply this technique, the natural logarithm of λ_i is expressed as a linear function of the explanatory variables, i.e.,

$$\ell n\ \lambda_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + B_p x_{ip} \qquad (1)$$

where $x_{i1}, \ldots, x_{ip}$ are the values of $x_1, \ldots, x_p$ which are associated with the random variable r_i and $\beta_1, \ldots, \beta_p$ are the p unknown regression parameters. Note that when all the x_{i1}'s (i=1,...,n), for example, are set equal to one, the right side of equation (1) will contain a constant related to the general level of the process. When p = 1, equation (1) gives the usual linear regression model, and with p > 1 the above model is similar to multiple regression. Further, when the values of $x_1, \ldots, x_p$ are binary (i.e., zero or one) then the model can be considered to correspond to that of different types of experimental design such as one way and two way analysis of variance. Finally, when some of the x's are binary while the others can assume any value, then model (1) corresponds to the analysis of covariance. It should be noted that above expression for λ_i does not tranform or change the probability distribution for r_i and it is more general than that which expresses λ_i as a linear function of $x_{i1}, x_{i2}, \ldots, x_{ip}$ since for small range of λ_i model (1) will give approximately linear expression for λ_i as a function of the explanatory variables.

Estimation of the parameters of the model

The likelihood function L for $\beta_1, \ldots, \beta_p$ is obtained by writing $P(r_i)$ for the probability distribution of r_i, multiplying these probabilities for $r_1, \ldots, r_n$ and substituting expression (1) for $\ell n\ \lambda_i$, i.e.,

$$L = \prod_{i=1}^{n} P(r_i) = \prod_{i=1}^{n} e^{-\lambda_i} \frac{\lambda_i^{r_i}}{r_i!}$$

$$L = \frac{e^{-\sum_{i=1}^{n} e^{\beta_1 x_{i1} + \ldots \beta_p x_{ip}}} \; e^{\beta_1 q_1 + \ldots + \beta_p q_p}}{r_1! \ldots r_n!} \qquad (2)$$

where $q_j = x_{1j}r_1 + x_{2j}r_2 + \ldots + x_{nj}r_n$ $(j = 1,2, \ldots, p)$.

The estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$ of $\beta_1, \ldots, \beta_p$ are obtained by maximizing L for the variations of $\beta_1, \ldots, \beta_p$. This can be done by solving the set of p equations

$$\frac{\partial \ell}{\partial \beta_j} = -\sum_{i=1}^{n} x_{ij}\, e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}} + q_j = 0 \quad (j=1,2, \ldots, p) \qquad (3)$$

where $\ell = \ell nL$. Equations (3) are non-linear and their solution can be obtained by iteration. One convenient method for doing this is Newton-Raphson method which requires the evaluation of the information

matrix I. This matrix is of order p x p with the element in the rth (r = 1, 2, ..., p) row and sth (s = 1, 2, ..., p) column is given by

$$i_{rs} = \sum_{i=1}^{n} x_{ir} x_{is} e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}$$

It is more convenient to present the foregoing results using matrix notation. Let X be the matrix of order (n x p) with x_{ij} as the element in the ith row and jth column and let $\underline{Q}$ be the column vector of length p and elements $q_1, \ldots, q_p$. Further denote to the vector of $\beta_1, \ldots, \beta_p$ by $\underline{\beta}$ and let $D(\underline{\beta})$ be a diagonal matrix of order (n x n) with its ith diagonal element $e^{\beta_1 x_{i1} + \ldots + \beta_p x_{ip}}$. Then it is easy to show that the system of equations (3) is

$$X' D(\underline{\beta}) \underline{1} = \underline{Q} \tag{4}$$

where $\underline{1}$ is a column vector of length n with all its elements equal to one and X' is the transpose of the matrix X. . The information matrix is

$$I = X' D(\underline{\beta}) X \tag{5}$$

Note that the matrix $D(\underline{\beta})$ is written in this way to emphasize its dependence on $\underline{\beta}$. Suppose that an initial estimate $\underline{\beta}_0$ for $\underline{\beta}$ is available then the iterative process is given by

$$\hat{\underline{\beta}}_k = \hat{\underline{\beta}}_{(k-1)} + I^{-1}(\underline{\beta}_{(k-1)}) \; (\underline{Q} - X'D(\hat{\underline{\beta}}_{(k-1)}) \; \underline{1}) \qquad (6)$$

where $\hat{\underline{\beta}}_k$ (k = 1, 2, ...) is solution at the stage k in the iteration process. The iteration continues until the difference between $\hat{\underline{\beta}}_k$ and $\hat{\underline{\beta}}_{(k-1)}$is very small, i.e., $(\hat{\underline{\beta}}_k - \hat{\underline{\beta}}_{(k-1)})' \; (\hat{\underline{\beta}}_k - \hat{\underline{\beta}}_{(k-1)})$ is less than a prespecified small value. At this stage the value of $\hat{\underline{\beta}}$ is taken as $\hat{\underline{\beta}}_k$.

Inferences about the parameters $\underline{\beta}$

Large sample confidence interval can be constructed by noting that $I^{-1}(\hat{\underline{\beta}})$ is the estimate of the variance-covariance matrix and hence an estimate of the the variance of $\hat{\beta}_j$ is the jth diagonal element $I^{-1}(j,j)$ of I^{-1}. The 95 percent confidence interval for $\hat{\beta}_j$ is then $\hat{\beta}_j \pm 1.96 \sqrt{I^{-1}(j,j)}$. In addition, to test if a subset of the β's, say $\beta_{\ell+1}, \beta_{\ell+2}, \ldots \beta_p$ (with $\ell < p$) can be regarded as zero's (i.e., not important parameters) in the regression equation then, the likelihood ratio test could be used which is given by

$$-2\ell n \; \Lambda = 2 \; \{\hat{\beta}_1 q_1 + \ldots + \hat{\beta}_p q_p - \tilde{\beta}_1 q_1 - \ldots - \tilde{\beta}_\ell q_\ell$$

$$- \sum_{i=1}^{n} e^{\hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}} + \sum_{i=1}^{n} e^{\tilde{\beta}_1 x_{i1} + \ldots \tilde{\beta}_\ell x_{i\ell}} \}, \qquad (7)$$

where $\tilde{\beta}_1, \ldots, \tilde{\beta}_\ell$ are the maximum likelihood estimates of $\beta_1, \ldots, \beta_\ell$ when $\beta_{\ell+1}, \ldots, \beta_p$ are set equal to zeros in (1). The asymptotic distribution of $-2\ell n\Lambda$ is χ^2 with $p-\ell$ degrees of freedom.

The case when the values of the random variable r can be classified according to the levels, for example, of two factors and the interest is to test if the two factors operate independently of each other is important in many applications. Model (1) is specialized here to deal with this case. Let r_{ij} be the observed value of the random variable r at level i ($i = 1,2, \ldots, \ell$) of the first factor and level j ($j = 1,2, \ldots, m$) of the second factor. Further, the likelihood ratio test is used to test the null hypothesis, H_o, that $\ell n\ \lambda_{ij}$ can be expressed as an additive linear combination of the ith level of the first factor and jth level of the second. Under H_o model (1) can be written as

$$\ell n\ \lambda_{ij} = \mu + \alpha_i + \beta_j, \quad (i = 1,2,\ldots,\ell;\ j = 1,2,\ldots,m) \qquad (8)$$

where μ is the general level of the process, α_i the effect due to the ith level of the first factor and β_j is the jth level of the second factor. Note that the parameters μ, $(\alpha_1, \ldots, \alpha_\ell)$ and $(\beta_1, \ldots, \beta_m)$ are not independent and to obtain a unique estimate for these parameters it is assumed that $\sum\alpha_i = \sum\beta_j = 0$.

Under the alternative hypothesis, model (8) is replaced by

$$\ell n\ \lambda_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (9)$$

where γ_{ij} represents the interaction between the ith level of the first factor with the jth level of the second factor. So the likelihood ratio test as given by (7) for testing the adequacy of model (8) is equivalent to testing that $\gamma_{ij} \equiv 0$ for all i and j.

APPLICATIONS

The method is illustrated by a numerical example based on total bacterial counts observed in water samples from five fixed points in the Morsang drinking water distribution system [Sle, 1983]. The general structure of the network and the position of the sampling stations are shown in Figure 1. Samples were collected from the exhaust-pipe of the treatment plant, above and below the Linas reservoir and from two other stations located in the Bondoufle network and also on a long dead-end at a place called "Ferme des Folies". Retention time of the water in the pipes before reaching the stations below the plant has been estimated to be 5 hours, 17 hours, 1 day and 1 month, respectively. Bacteriological data were obtained during special study which was conducted from May to December 1981. These data consist of total counts expressed as the number of colony-forming-units per mL after 72 hours of incubation at 20°C using the standard pour plate procedure. The aim of the study was to determine how the bacteriological water quality changes both spatially and temporally. So, the hypothesis of interest were to test: i) if the bacterial density in the water varies at any fixed time from one

sampling point to another, ii) if it varies at a fixed location from one time to the other and iii) for the presence of interactions between time and space.

The analysis started by fitting model (8) to the data. The estimated time effects $\hat{\beta}_j$ (j = 1,2, ..., 14) are plotted in Figure 2, while the location effects $\hat{\alpha}_i$ (i = 1,2, ..., 5) are plotted in Figure 3. The values of $\hat{\beta}_j$ appear to have maximums in the first half of July and in the beginning of September and it reaches its minimum in the middle of November. It is of interest to note that the values of $\hat{\beta}_j$ are nearly constant for July, August and September. Thereafter the value of $\hat{\beta}$ drops drastically to reach its minimum. To test for the significance of this pattern, the likelihood ratio test (equation (7)) is applied to test the hypothesis $H_1: \beta_1 = \beta_2 = \ldots = \beta_{14}$. The value of $-2 \ln \Lambda$ is 49428.78, which is very highly significant ($P<.01$). Hence H_1 is rejected (i.e., the temporal variabilities expressed in terms of the β's are accepted).

The values of $\hat{\alpha}_i$ (i = 1,2, ..., 5) indicate an increase with the distance from the treatment plant (Figure 3) and the maximum difference between two successive α's occurring between location one and location two which is followed by the difference between location four and five. The differences between locations two, three and four are of least importance. The likelihood ratio test gave $-2 \ln \Lambda = 48112.95$ which is very highly significant ($P < .01$). Hence the differences

between locations is accepted as real. Further, the interaction between the temporal and spatial variabilities was tested by comparing model (9) with model (8) (i.e., testing that all $\gamma_{ij} = 0$ for $i = 1,2, \ldots, 5$ and $j = 1,2, \ldots, 14$). The likelihood ratio test gave $-2 \ell n \Lambda = 10424.02$ which is highly significant ($P < .01$) when compared with χ^2 on 52 degrees of freedom. This suggests that there is a strong interaction between the spatial and temporal variabilities of bacterial counts.

An attempt is made here to correlate the spatial and temporal variabilities with temperature and turbidity where these measurements are available. First, it was decided to see if the inclusion of temperature will explain the temporal differences; hence in model (8) the parameter β_j has been replaced by θx_{ij} where x_{ij} is the temperature of the water in the ith location and during the jth time and θ is an unknown regression parameter. Under this assumption the value of θ was estimated as $\hat{\theta} = .3667$ with standard error .0029 which shows that the effect of temperature is significant and is positively correlated with bacterial density. To determine if the inclusion of temperature could explain the temporal variabilities, the likelihood ratio test was applied and showed that temperature could not completely explain the variation in the β_j's ($-2 \ell n \Lambda = 20152.47$). Since a linear term in temperature was not adequate, it was decided to consider a quadratic term, hence in model (8), β_j has been replaced by $\theta_1 x_{ij} + \theta_2 x^2_{ij}$. The values of θ_1 and θ_2 were

estimated as $\hat{\theta}_1 = .7582$ and $\hat{\theta}_2 = -.0857$, respectively. The standard error of $\hat{\theta}_1$ is .0112 and that of $\hat{\theta}_2$ is .0018 which indicates that the inclusion of θ_2 is significant ($-2\ \ell n\Lambda = 15920.94$). Note that the value of $-2\ \ell n\Lambda$ is reduced quite substantially as the result of adding the quadratic term. The effect of turbidity was next introduced into the model by adding an interaction term between turbidity and temperature. Thus, another term in the form $\gamma\ x_{ij}\ T_{ij}$ has been considered in the regression equation, where T_{ij} is the turbidity in the ith location at the jth time. The likelihood ratio statistic, $-2\ \ell n\Lambda$, amounted to 5267.77 which is highly significant. This indicates that, although the interaction between temperature and turbidity is very marked, there is still unexplained temporal variability. Finally, the regression was repeated by adding another linear term for the turbidity. The inclusion of this new factor resulted in the decrease in $-2\ell n\Lambda$ from the previous value substantially to 341.97, which is still significant. From this analysis it was concluded that temperature and turbidity cannot explain all the spatial and temporal variation in the bacterial density, although they explained a very large portion of it. Figure 4 displays the bacterial counts observed and the corresponding estimated value when the turbidity and temperature are both included in the model. The figure shows that the model appears to fit the experimental observations reasonably well except for the last three sampling dates.

DISCUSSION AND CONCLUSION

In this paper, the method of maximum likelihood is used to obtain estimates of the parameters in a log-linear regression model when the experimental observations are independent and assumed to follow the Poisson distribution. The application of the usual regression methods assumes: normality, constancy of variance and additivity of the effects of the explanatory variables, and therefore often requires transformation of the experimental data. Consequently, since no single transformation might be capable of fulfilling all these requirements, Poisson regression, which requires no previous transformation of the data, offers an appropriate way to avoid all these difficulties. Hence, this statistical method is much more appropriate than conventional regression analysis for assessing the dependence of Poisson distributed data on a set of categorical and/or continuous variables. The improvement resulting from the inclusion of new factors in a given regression model can be evaluated by means of the likelihood ratio test; furthermore, examination and comparison of the likelihood ratios corresponding to different regression equations may progressively achieve an adequate model for fitting the data. In this regard, it must be emphasized that the sequential order of introducing new independent variables into the model is not unique and may not be prespecified. Thus, an interesting approach to attain an appropriate model containing a set of variables with great predictor potentials can be achieved by performing a stepwise regression

analysis by means of the likelihood ratio used as the discriminating function [Draper and Smith, 1966].

One important area of application of Poisson regression models is the analysis of bacteriological quantitative data. The statistical procedures developed in section 2 have great potential for improving estimation of parameters and fitting models to such data sets. The results of the analysis of the Morsang's distribution system data, which is presented in this work as a numerical example for illustrating the method, indicated the presence of strong spatial and temporal variation in the bacterial density with a marked interaction between those two components. Factors which may have contributed to influence the temporal variability are turbidity and water temperature. Both of these variables were highly positively correlated to bacterial incidence in the network. These observations are in accord with those of other studies, showing a structured pattern of the spatial and temporal heterogeneity of the bacteria in distribution systems. In particular, high densities of heterotrophic bacteria are: i) more likely to occur in peripheral locations far from the treatment plant rather than close to it; and ii) concomitant with warmest water temperatures [Maul et al., 1985a]. The understanding of the variability in bacterial density through appropriate regression models can be useful for i) studying the relation between bacterial incidence and some physical and chemical characteristics of the body of water considered; and ii) determining the spatial and temporal dispersion pattern of bacteria in light of those characteristics.

This information, thereafter, can be used for identifying problem areas or periods and will thus facilitate taking remedial action. In addition, it will help to improve the design of future monitoring sampling programs [Maul et al., 1985b].

Although Poisson regression requires more computation than ordinary regression analysis, it is worth bringing to the attention of microbiologists for the existence of computers makes this represent no real disadvantage. In fact, the analysis of the data presented here took only a few seconds using APL computer system.

ACKNOWLEDGEMENTS

The bacteriological data for this work were provided through the courtesy of Michèle Rizet, Société Lyonnaise des Eaux, France.

REFERENCES

Anscombe, F.J., Discussion of Hotelling's paper, J.R. Statist. Soc., B15, 229-230, 1953.

Draper, N.R. and H. Smith, Applied regression analysis, 407 pp., John Wiley, New York, 1966.

Frome, E.L., The analysis of rates using Poisson regression models, Biometrics, 39, 665-674, 1983.

Frome, E.L., M.H. Kutner and J.J. Beauchamp, Regression analysis of Poisson-distributed data, J. Am. Stat. Assoc., 68, 935-940, 1973.

Maul, A., A.H. El-Shaarawi and J.C. Block, Heterotrophic bacteria in water distribution systems. I. Spatial and temporal variation, Sci. Total Environ., 44, 201-214, 1985a.

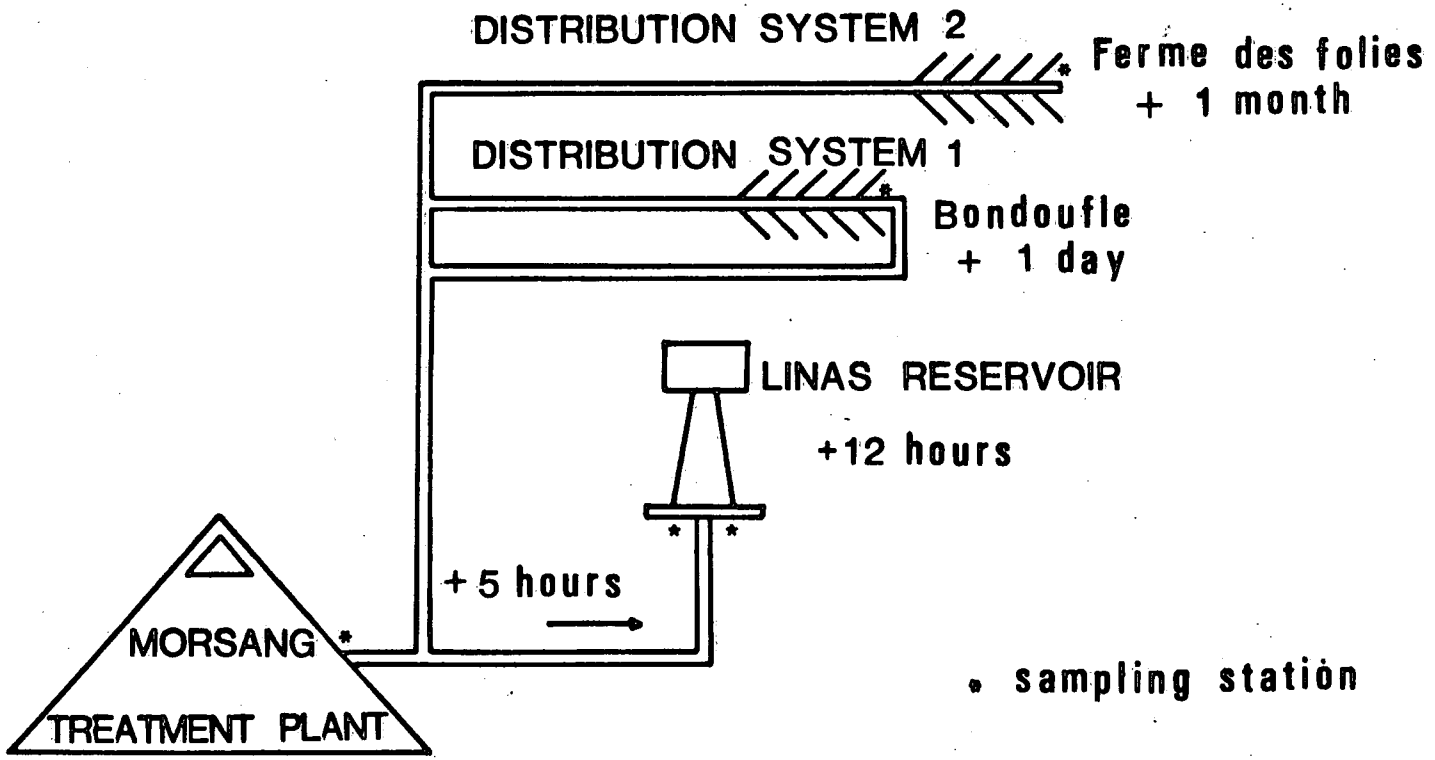
Maul, A., A.H. El-Shaarawi and J.C. Block, Heterotrophic bacteria in water distribution systems. II. Sampling design for monitoring, Sci. Total Environ., 44, 215-224, 1985b.

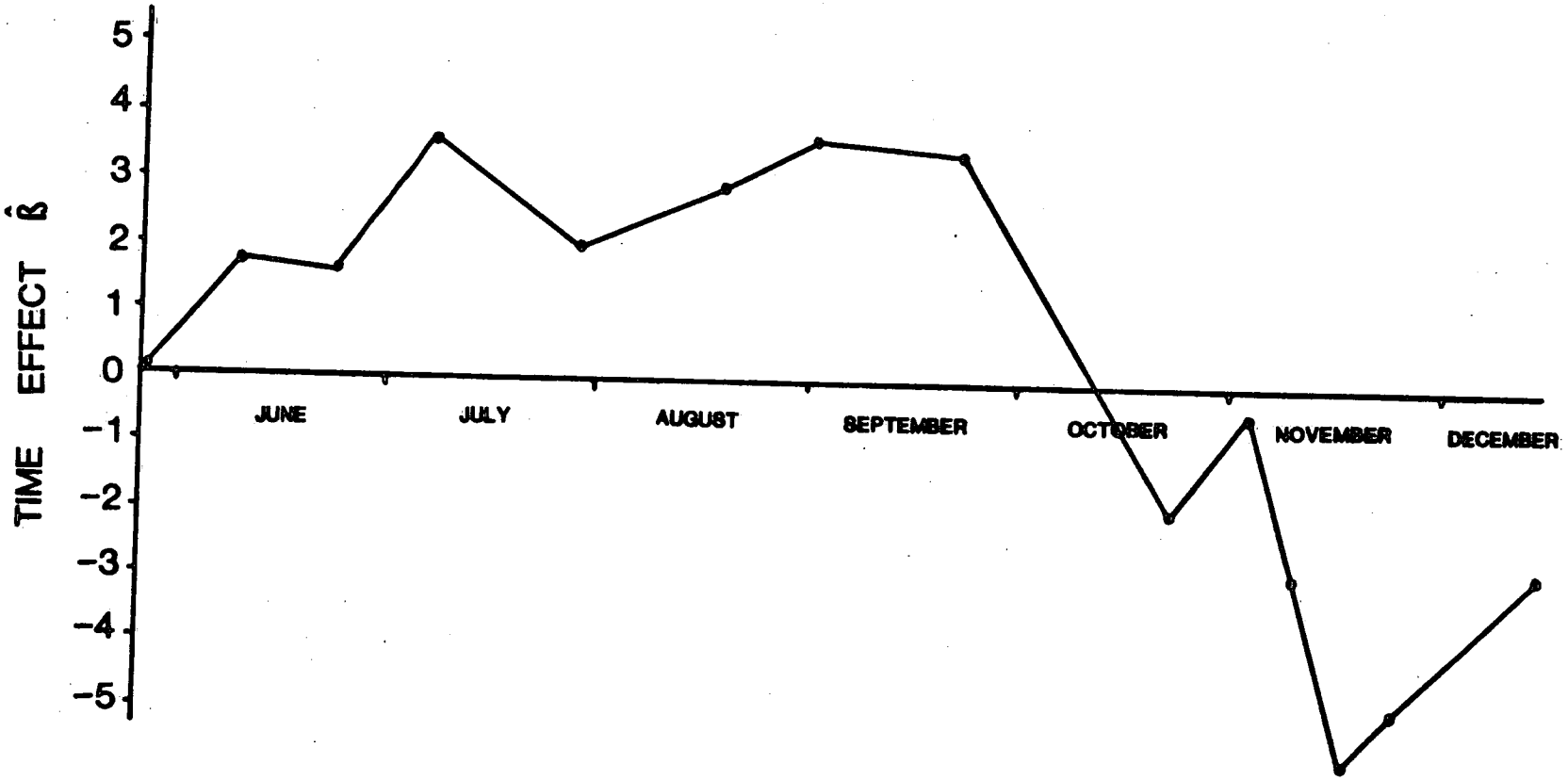
Plackett, R.L., The analysis of categorical data, 207 pp., 2nd edn., Macmillan Publishing Co., New York, 1981.

Rao, C.R., Linear statistical inference and its applications, 2nd edn., John Wiley, New York, 1973.

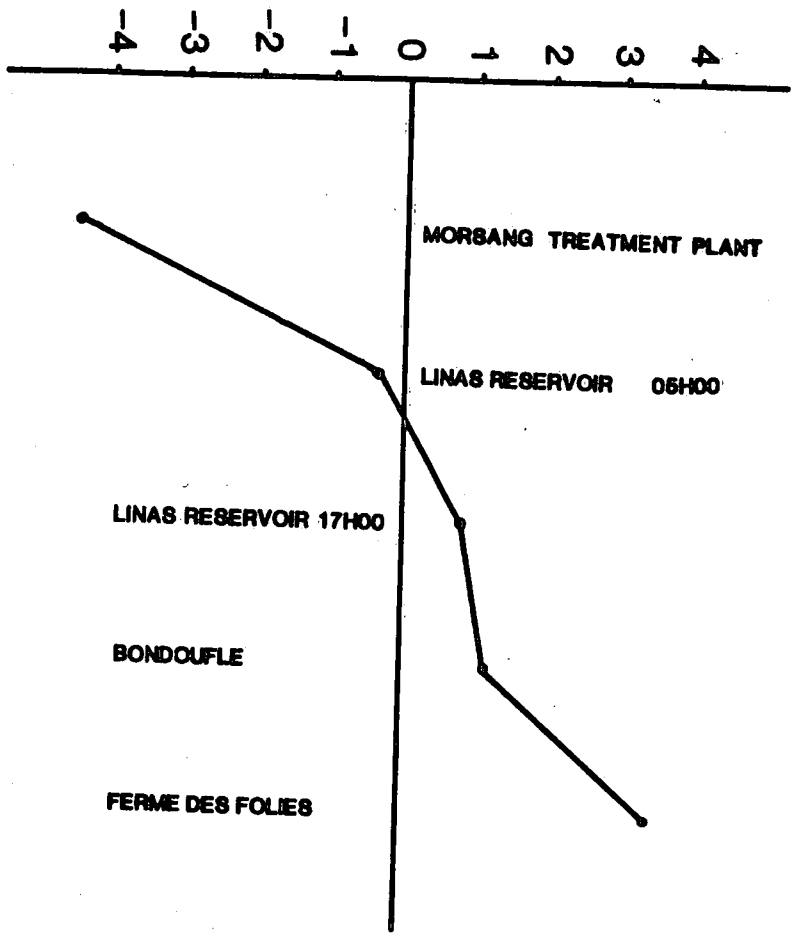
Société Lyonnaise des Eaux, Étude de l'évolution de la qualité de l'eau dans les réseaux, Le Pecq, 1983.

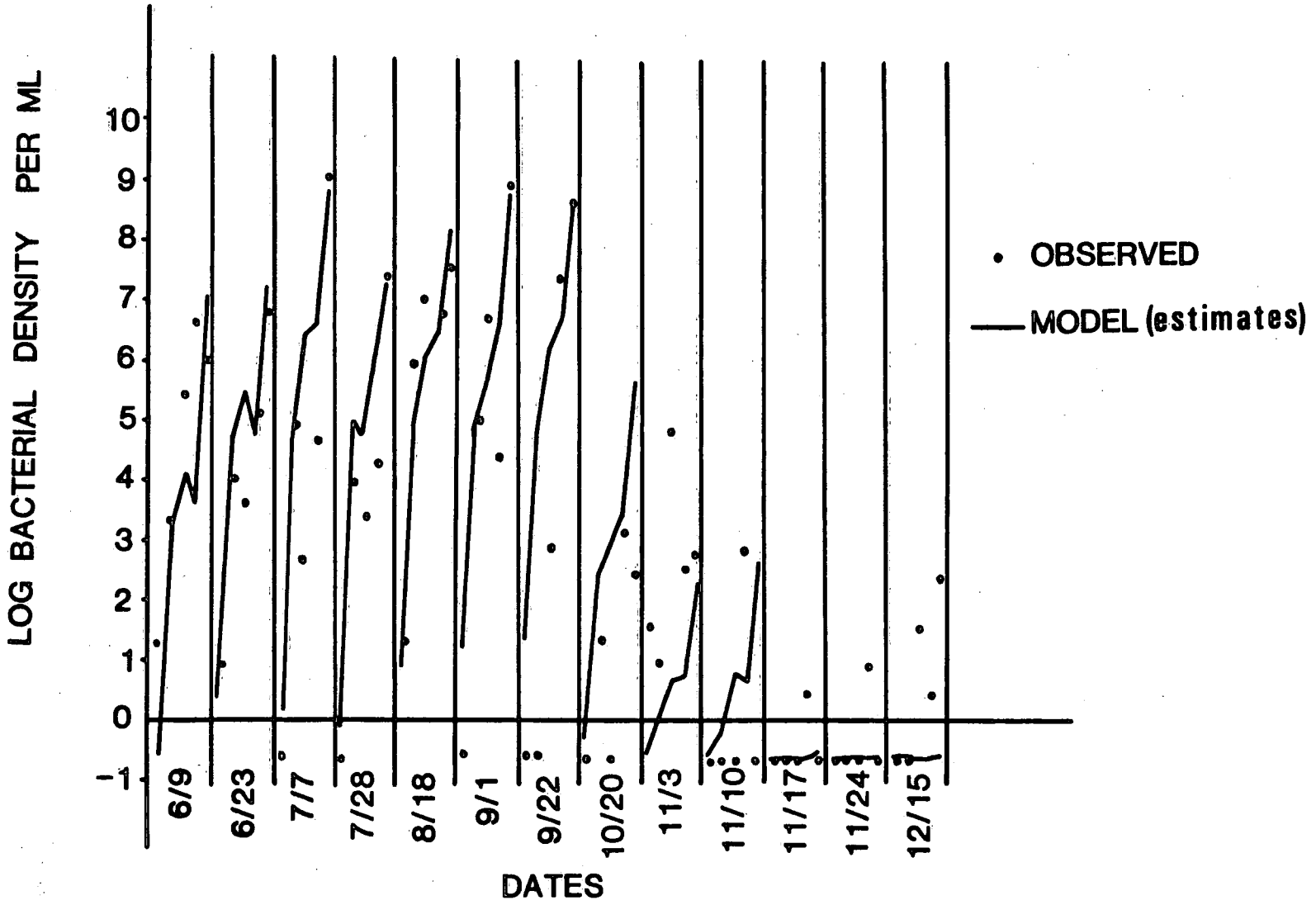
FIGURES


DISTRIBUTION SYSTEM 2
Ferme des folies
+ 1 month
DISTRIBUTION SYSTEM 1
Bondoufle
+ 1 day
LINAS RESERVOIR
+12 hours
+5 hours
MORSANG
TREATMENT PLANT
* sampling station


TIME EFFECT β̂
5
4
3
2
1
0
−1
−2
−3
−4
−5
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
NOVEMBER
DECEMBER

SPATIAL EFFECT $\hat{\alpha}$




LOG BACTERIAL DENSITY PER ML
10
9
8
7
6
5
4
3
2
1
0
-1
6/9
6/23
7/7
7/28
8/18
9/1
9/22
10/20
11/3
11/10
11/17
11/24
12/15
DATES
OBSERVED
MODEL (estimates)