RRB 90-03

A GENERAL DISTRIBUTIONAL MODEL FOR LAKE ALKALINITY WITH APPLICATIONS TO CANADIAN DATA

by
A.H. ElShaarawi and A. Naderisamani

January 1990
NWRI Contribution #90-55

ABSTRACT

A class of probability models is proposed for making inferences about the regional distribution of lake Alkalinity. This class includes the three-parameter lognormal distribution which is frequently used as a regional model for lake chemistry. The adequacy of the lognormal distribution can be assessed by testing the significance of a single parameter in the proposed model. It is shown that the three-parameter lognormal distribution does not fit the majority of data sets from Eastern Canada. Furthermore, the pH-alkalinity relationship developed by Small and Sutton is used to yield a class of probability models representing the derived distributions for pH.

RÉSUMÉ

Une classe de modèles de probabilité est proposée pour faire des inférences sur la distribution régionale de l'alcalinité d'un lac. Cette classe comprend une distribution log-normale à trois paramètres qui est fréquemment utilisée comme modèle régional pour la chimie des lacs. On peut déterminer si la distribution log-normale convient en vérifiant la signification d'un paramètre unique dans le modèle proposé. Il est démontré que la distribution log-normale à trois paramètres ne convient pas à la majorité des ensembles de données provenant de l'Est du Canada. De plus, la relation pH-alcalinité élaborée par Small et Sutton est utilisée pour donner une classe de modèles de probabilité représentant les distributions dérivées du pH.

MANAGEMENT PERSPECTIVE

When a probabilistic model is used to characterize the spatial variability of lake chemistry in a region of interest, it is essential to ensure that the chosen model is adequate so that valid inferences can be made. One approach to assessing model adequacy is to consider a more general model which includes the proposed model as a special case depending on the value taken by a specific parameter. Hence testing the adequacy of the model is reduced to testing a simple hypothesis about a single parameter. This approach is used to test the adequacy of the three-parameter lognormal model (Small & Sutton, 1986a), by defining a family of probability models using the Box and Cox (1964) transformation which includes the lognormal distribution as a special case. The results show that the Small and Sutton model is inadequate for representing lake Alkalinity in New Brunswick, Newfoundland and Nova Scotia while only the Quebec data sets supports the use of the lognormal model. Furthermore, the effects of using the incorrect model on predicting the proportion of lakes at risk is illustrated using data from Eastern Canada.

PERSPECTIVE-GESTION

Lorsqu'on utilise un modèle probabiliste pour caractériser la variabilité spatiale de la chimie d'un lac dans une région d'intérêt, il est essentiel de s'assurer que le modèle choisi est approprié, de sorte que les inférences faites soient valides. Une façon de vérifier si un modèle est approprié consiste à étudier un modèle plus général qui inclut le modèle proposé comme un cas particulier, selon la valeur prise par le paramètre particulier. De cette façon, la vérification du caractère approprié du modèle se réduit à la vérification d'une simple hypothèse portant sur un paramètre unique. Cette approche est utilisée pour vérifier la valeur du modèle log-normal à trois paramètres (Small et Sutton, 1986a), en définissant une famille de modèles de probabilité à l'aide de la transformation de Box et Cox (1964) qui inclut une distribution log-normale comme un cas particulier. Les résultats révèlent que le modèle de Small et Sutton ne convient pas pour la représentation de l'alcalinité des lacs du Nouveau-Brunswick, de Terre-Neuve et de Nouvelle-Écosse, et que seules les données du Québec justifient l'utilisation du modèle log-normal. De plus, les effets de l'utilisation d'un modèle inapproprié pour prévoir la proportion des lacs à risque sont illustrés à l'aide de données provenant de l'Est du Canada.

1. INTRODUCTION

A class of probability models is given for characterizing the alkalinity variability of a population of lakes in a given region and for predicting the corresponding pH distributions. This class includes the three-parameter lognormal model proposed by Small and Sutton (1986a) as a special case. This report shows that the Small and Sutton model is inadequate for most of the alkalinity data generated from Environment Canada monitoring networks for New Brunswick, Newfoundland, Nova Scotia (Howell et al., 1987) and Quebec. In the majority of cases, the data do not fit the lognormal distribution and hence using such a model will result in an inaccurate prediction of the number of lakes below or above a certain pH value. Thus, over-estimation or under-estimation of the resources at risk will result from using an inaccurate model. The proposed class of models avoids these difficulties and results in a more robust inference of the chemistry of the lake population.

2. THE GENERAL PROBABILITY MODEL FOR LAKE ALKALINITY

Consider the Box and Cox (1964) transformation g(x) defined by

$$g(x) = \begin{cases} \dfrac{(x + \lambda_2)^{\lambda_1}}{\lambda_1} & \lambda_1 \neq 0 \\ \log\,(x + \lambda_2) & \lambda_1 = 0,\ x > \lambda_2. \end{cases} \qquad (1)$$

Let X be a random variable. If g(X) is normal (μ, σ^2), then the probability density function (pdf) of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot (x + \lambda_2)^{\lambda_1 - 1} \cdot \exp -\{(\frac{g(x)-\mu)^2}{2\sigma^2}\}. \qquad (2)$$

This form generates a wide class of distributions for X depending on the values of λ_1 and λ_2. For example, the normal and lognormal distributions are obtained by setting the values of λ_1 and λ_2 to 1 and 0 respectively. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with Pdf f(x), and let x_i be the observed value of X_i $(i=1,\ldots, n)$. Then the parameters of f(x) are estimated as follows. The estimates of μ and σ^2 are obtained from the sample mean and variance of the transformed data:

$$\hat{\mu}_{\lambda_1,\lambda_2} = \sum_{i=1}^{n} g(x_i)/n \quad \text{and} \quad \hat{\sigma}^2_{\lambda_1,\lambda_2} = \sum_{i=1}^{n} (g(x_i)-\hat{\mu})^2 /n.$$

Substituting $\hat{\mu}_{\lambda_1,\lambda_2}$ and $\hat{\sigma}^2_{\lambda_1,\lambda_2}$ for μ and σ^2 respectively in the joint probability density of $X_1,\ldots, X_n$, yields the likelihood function for λ_1 and λ_2 which is

$$L(\lambda_1,\lambda_2) = c\,\hat{\sigma}^{-n} \prod_{i=1}^{n} (x_i + \lambda_2)^{\lambda_1 - 1} \exp[-\sum_{i=1}^{n} (g(x_i)-\hat{\mu})^2/(2\hat{\sigma}^2)], \qquad (3)$$

where c is a constant. The estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of λ_1 and λ_2 are obtained so that the likelihood function defined by (3) achives its maximum at $(\hat{\lambda}_1, \hat{\lambda}_2)$. Let

$$R = \ln L(\lambda_1,\lambda_2) - \ln L(\hat{\lambda}_1,\hat{\lambda}_2).$$

Then $-2R$ is assymptotically $\chi^2(2)$, and hence an approximate a % confidence region for λ_1, and λ_2 may be obtained by

$$\{(\lambda_1,\lambda_2) : -2R \leq C_a\} = \{(\lambda_1,\lambda_2) : R \geq -\frac{C_a}{2}\},$$

where C_a is determined from $\text{Prob}(\chi^2(2) \leq C_a) = a$.

3. APPLICATIONS

i) ALKALINITY DISTRIBUTION

Transformation (1) is used to develop a probability distribution model for the lake Alk data defined by

$$F_{Alk}(x) = \text{Prob}\ (Alk \leq x) = \int_{-\infty}^{x} f_{Alk}(y)dy, \tag{4}$$

where f_{Alk} is the Pdf of the Alk data defined by (2). F_{Alk} is called the cumulative distribution function (CDF) of the Alk data. The estimated parameters of the Canadian lakes Alk distribution are obtained using the International Mathematical and Statistical Libraries (IMSL (1987)) and given in Table 1. Figure 1 presents the 95% confidence regions for the Newfoundland and Quebec Alk data. The confidence region for the Quebec data includes $\lambda_1 = o$ and hence the lognormal distribution is adequate for this data set. This confidence region also indicates that the values of λ_1 and λ_2 for the Quebec data are more precisely determined than those of the Newfoundland data. Figures 2 and 3 show good agreement between the observed and fitted distributions of the Newfoundland and Quebec data sets.

ii) pH DISTRIBUTION

The pH distribution may be obtained directly by the methods described in the previous section. However, it is often desirable to allow predicted shifts in the Alk distribution to be directly translated into shifts in the pH distribution. Hence, the following pH-Alk relationship developed by Small and Sutton (1986b) may be used to derive the pH distribution directly from the Alk distribution defined by (4).

$$pH = a + \frac{1}{\ln 10} \text{ arc sinh } \left(\frac{Alk-d}{c}\right), \tag{5}$$

where a, c and d are unknown constants. Assuming (5), we have

$$F_{pH}(x) = \text{Prob}(pH \leq x) = \text{Prob}(Alk \leq d+c \sinh(\frac{x-a}{b})) = F_{Alk}(d+c \sinh (\frac{x-a}{b})).$$

The corresponding pdf may be obtained by

$$f_{pH}(x)=\frac{c}{b} \cosh \left(\frac{x-a}{b}\right) f_{Alk}(d+c \sinh (\frac{x-a}{b})).$$

The derived pH distributions using the optimal and log transformations as well as the corresponding observed distributions of the Newfoundland and New Brunswisk data are shown in Figures 4 and 5. The derived distributions using the optimal transformation show better agreement with the corresponding observed distributions than those obtained using the log transformation. These figures show that the automatic use of the log transformation could result in inadequate inference about the proportion of lakes in a specific pH class. This is especially clear from Figure 5.

4. CONCLUSIONS

The proposed general model indicates that the three-parameter lognormal model does not fit the majority of data sets from Eastern Canada. It is shown that using an incorrect model for representing the regional distribution of lake Alkalinity could result in unrealistic derived pH distributions and hence, incorrect predictions of the lake population characteristics. The model presented avoids such difficulties by adding a single parameter to the model. As a result, more flexibility is obtained and robust predictions can be made.

REFERENCES

Box G.E.P. and Cox D.R. (1964) An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society, Ser. B., 39, 211-252.

Howell G.D. and Brooksbank P. (1987) An assessment of LRTAP acidification of surface waters in Atlantic Canada, IW/L-AR-WQB-87-121, Environment Canada.

IMSL (1987) In MATH/LIBRARY and STAT/LIBRARY, IMSL, Inc., Houston, Texas.

Small, M.J. and Sutton, M.C. (1986a). A direct distribution model for regional aquatic acidification. Water Resour. Res., 22, 1749-1758.

Small M.J., Sutton M.C., (1986b). A regional pH-alkalinits relationship. Water Res., 20, 335-343.

Table 1. Estimates of the Alk distribution parameters.

Alk Data	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\mu}$	$\hat{\sigma}$
New Brunswick	-.7	3.95	1.030	.172
Newfoundland	-.7	2.55	.974	.156
Nova Scotia	- 1.9	24.8	.525	.0003
Quebec	.1	- .01	.061	1.362

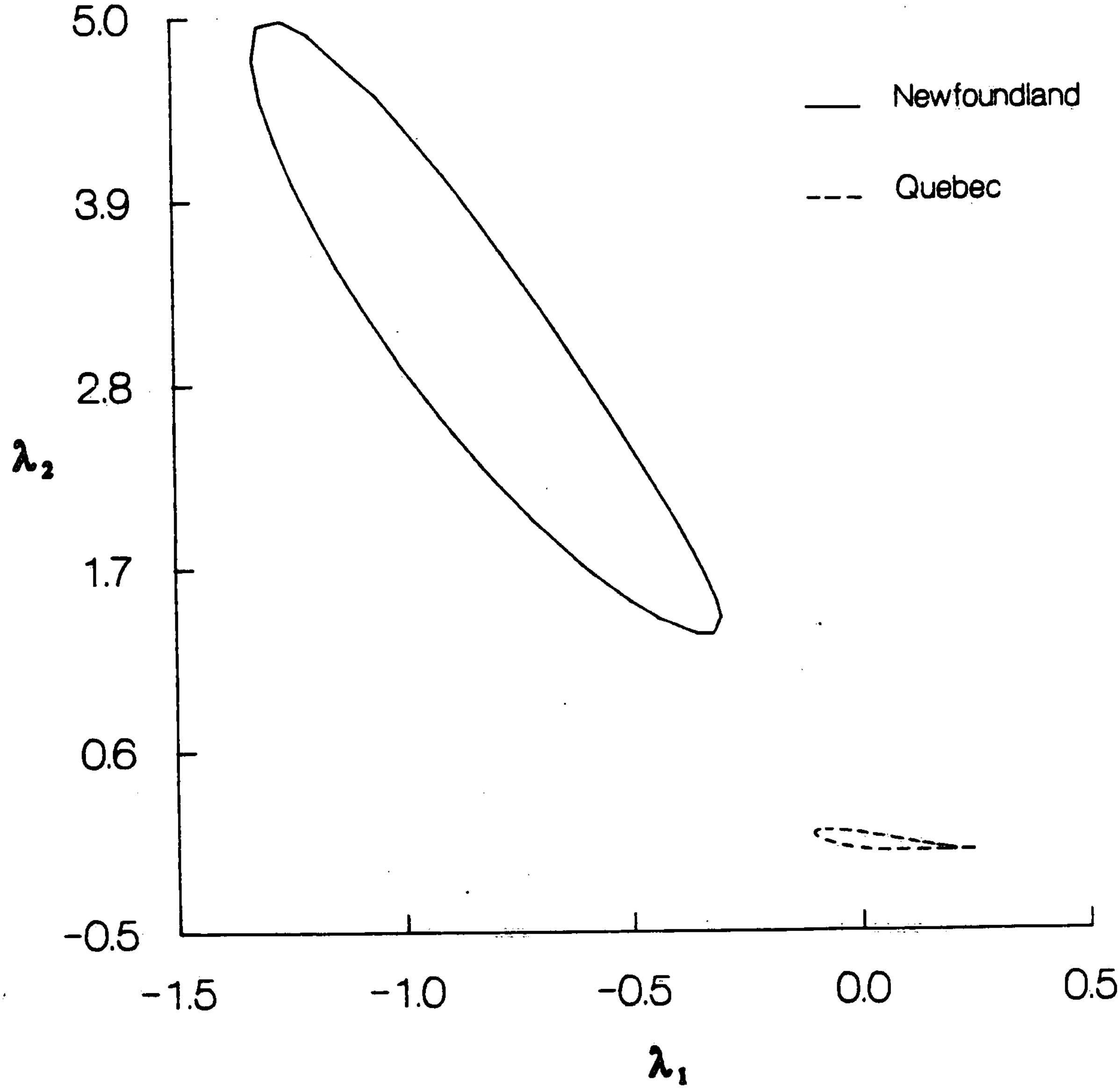

Figure 1. 95% confidence regions for the transformation parameters of the Newfoundland and Quebec data.

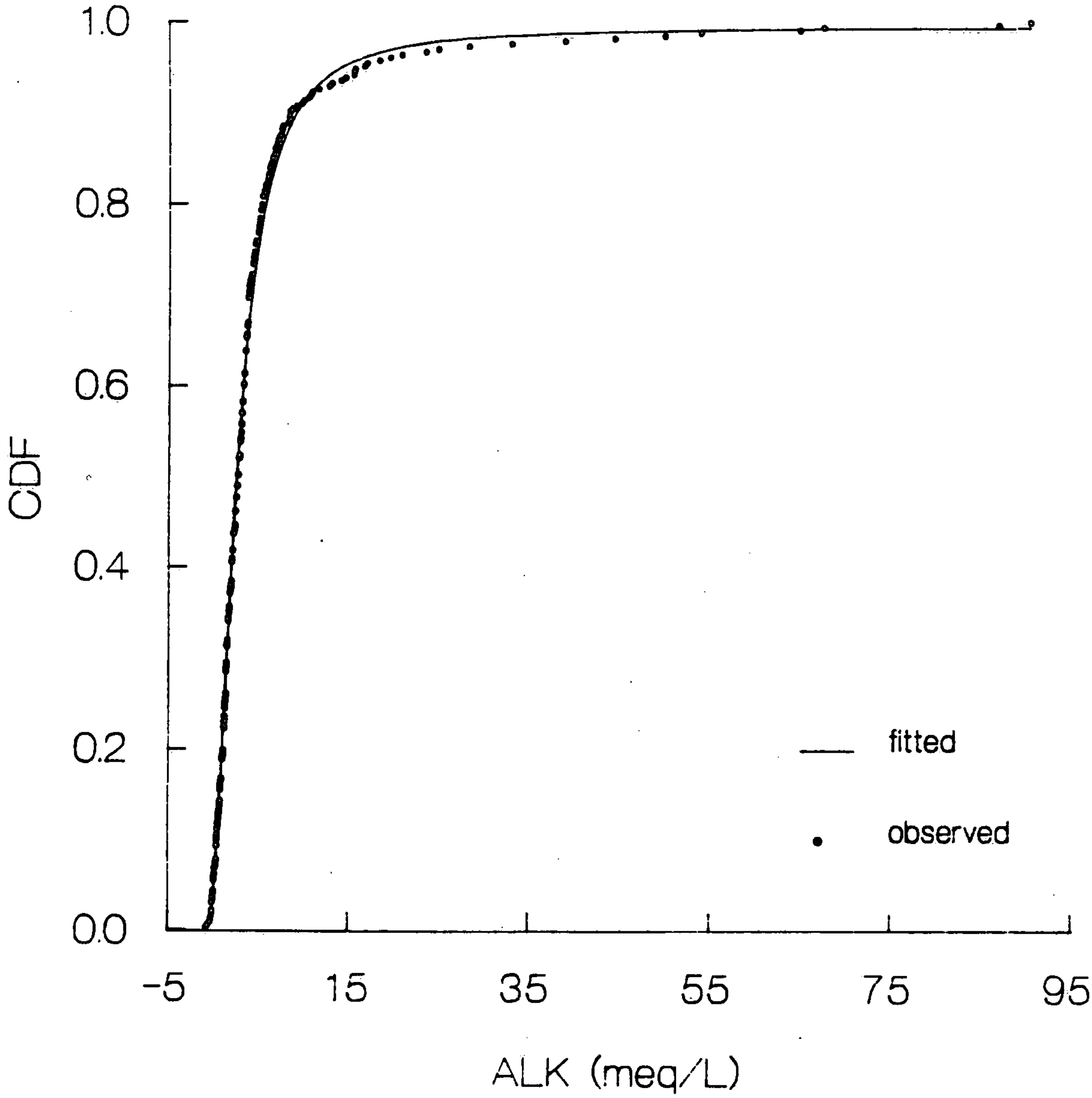

Figure 2. Observed and fitted distribution functions of the Newfoundland alkalinity data.

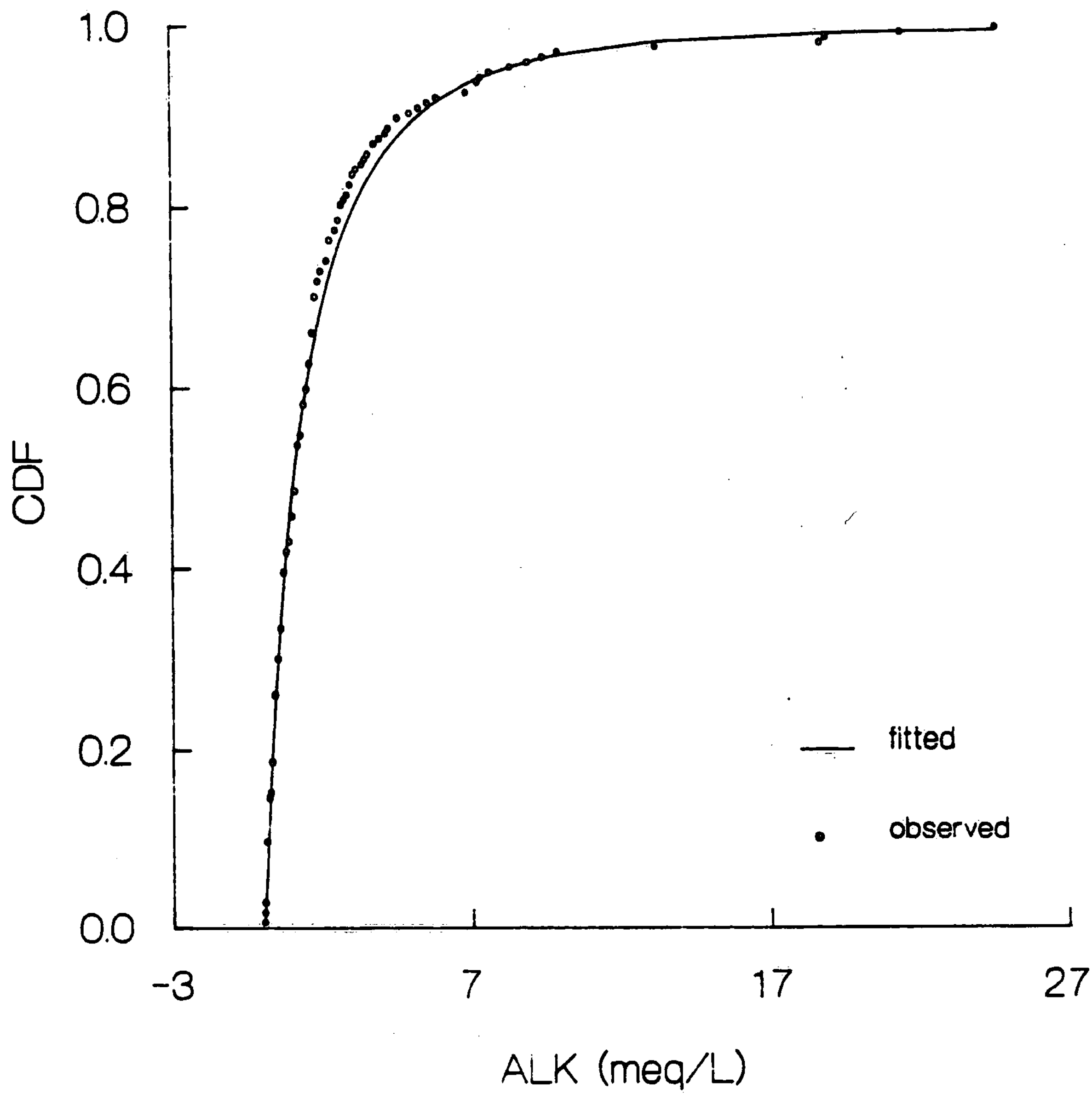

Figure 3. Observed and fitted distribution functions of the Quebec alkalinity data.

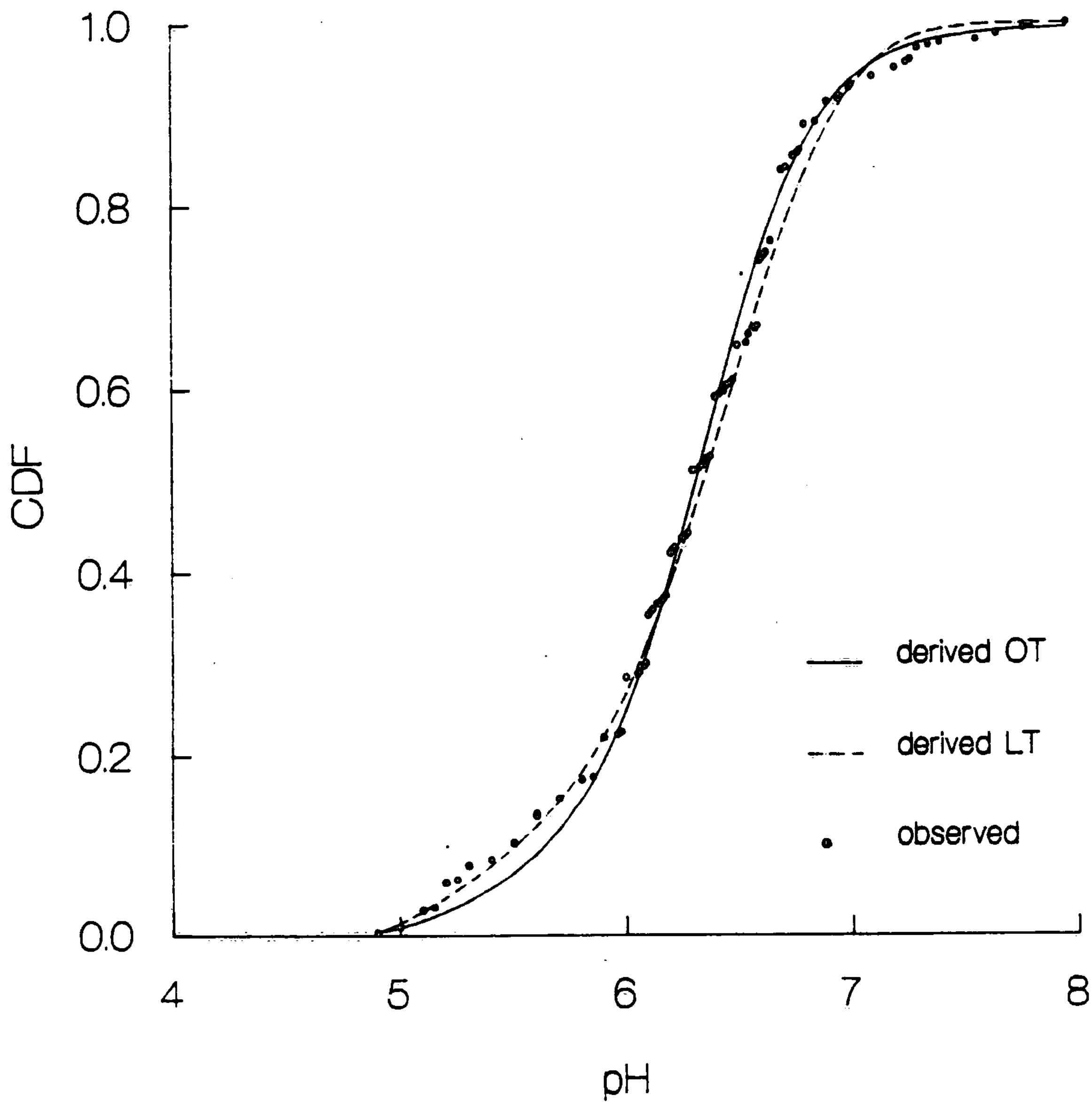

Figure 4. Observed and derived distribution functions of the Newfoundland pH data. The derived distributions are obtained by the optimal transformation (OT) and the log transformation (LT) of the data.

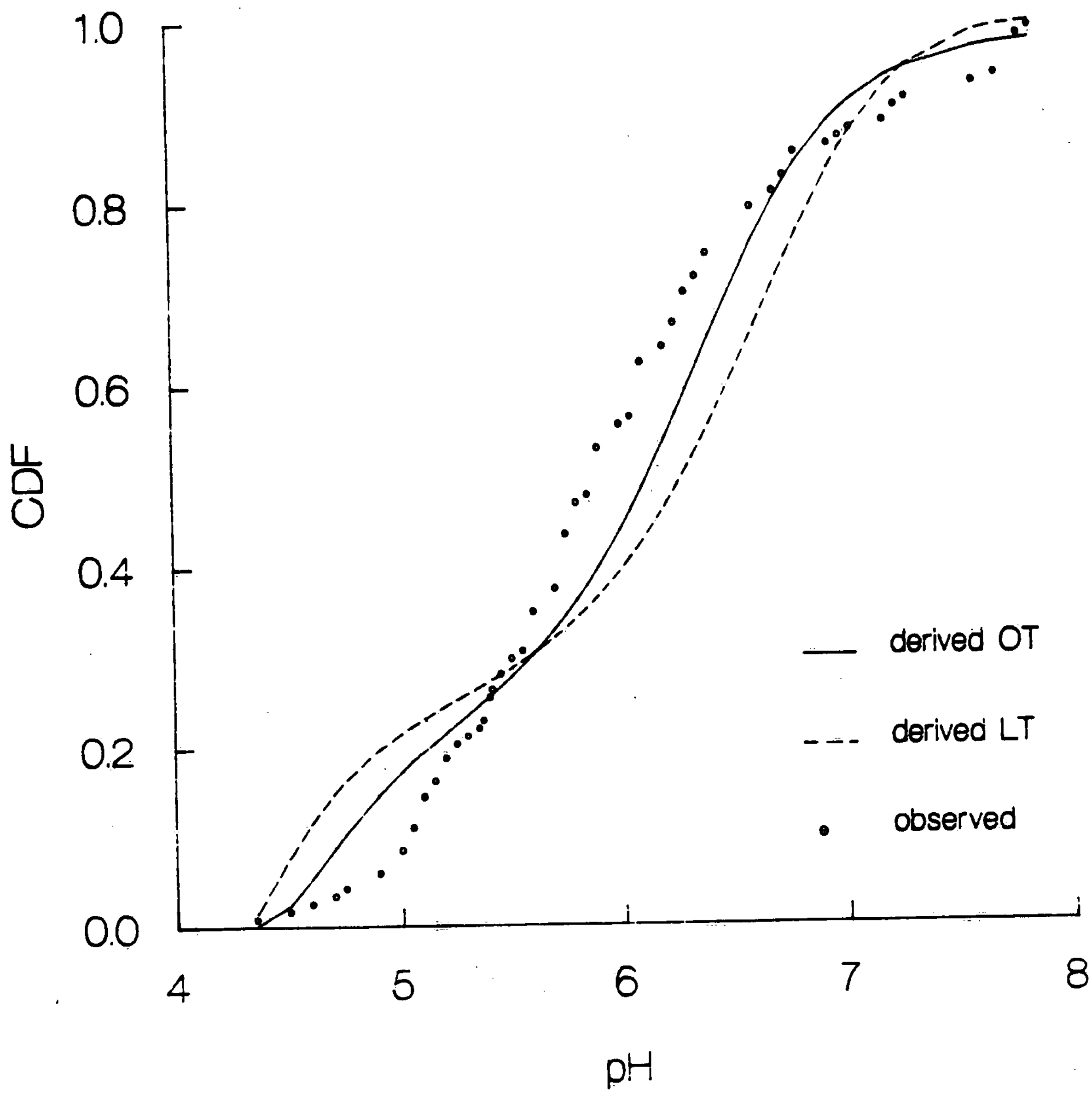

Figure 5. Observed and derived distribution functions of the New Brunswick pH data. The derived distributions are obtained by the optimal transformation (OT) and the log transformation (LT) of the data.