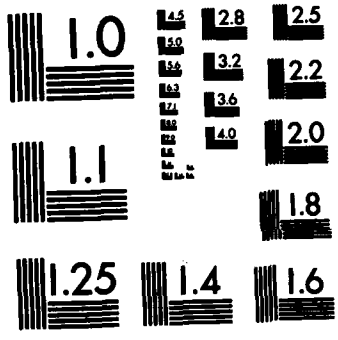
AD-A123 385 ESTIMATING A FAMILY OF DISTRIBUTIONS WITH APPLICATIONS TO THE EXTENSION O.. (U) UNIVERSITY OF CENTRAL FLORIDA ORLANDO DEPT OF MATHEMATICS AND.. S J BEAN ET AL. 1/1

UNCLASSIFIED 16 AUG 82 AFGL-TR-82-0239 F19628-82-K-0001 F/G 12/1 NL

END

FILMED

DTIC



AFGL-TR-82-0239

Estimating a Family of Distributions with Applications to the Extension of Climatology

S. J. Bean
P. N. Somerville

Department of Mathematics and Statistics
University of Central Florida
Orlando, FL 32816

Scientific Report No. 2

16 August 1982

Approved for Public Release; Distribution Unlimited

Air Force Geophysics Laboratory
Air Force Systems Command
United States Air Force
Hanscom AFB, Massachusetts 01731

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1 REPORT NUMBER: AFGL-TR-82-0239	2 GOVT ACCESSION NO: A123 385
	3 RECIPIENT'S CATALOG NUMBER
4 TITLE (and Subtitle): Estimating a Family of Distributions with Applications to the Extension of Climatology	5 TYPE OF REPORT & PERIOD COVERED: Scientific Report No. 2, 15 April - 1 Sept 82
	6 PERFORMING ORG. REPORT NUMBER
7 AUTHOR(s): S. J. Bean, P. N. Somerville	8 CONTRACT OR GRANT NUMBER(s): F19628-82K-0001
9 PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Mathematics and Statistics, University of Central Florida, Orlando, Florida 32816	10 PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62101 F, 667009 AK
11 CONTROLLING OFFICE NAME AND ADDRESS: Air Force Geophysics Laboratory, Hanscom AFB, MA 01731, Contract Monitor: D. Grantham, LYD	12 REPORT DATE: 16 August 1982
	13 NUMBER OF PAGES: 13
14 MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15 SECURITY CLASS. (of this report): Unclassified
	15a DECLASSIFICATION/DOWNGRADING SCHEDULE

16 DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17 DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18 SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Family of Distributions
Regression Model
Spreading
Least Squares Estimation
Maximum Likelihood Estimation

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

A family of distributions is used to model a climatic (or other) variable over a region. The entire family of distributions may be estimated by utilizing a sample of locations throughout the region and the dependence of the parameters of distributions on independent variables. Regression models are used to estimate the parameters for a location with given values of the independent variables.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Table of Contents

1.0 Introduction

One problem of concern in climatology studies is that of extending climatology information spatially. Somerville and Bean (1981) have given some techniques for developing probability models for visibility in West Germany for locations where no historical records exist. The current study is a consolidation and generalization of the basic theory for the above problem.

The problem may be stated in the following way. We wish to estimate the probability distribution of a climatic variable, such as visibility, for any location in a given region where historical records do not necessarily exist. However, we do assume that there are independent variables which may be measured at the location of interest that have some correlation with the parameters in the probability distribution to be estimated. These variables may be the elevation, average elevation of the surrounding area, or other geographical measures for example. Also, we assume that we have a sample of locations for which we have historical records. Further, we assume that the distribution at each location in the region is of the same form but the parameters change from location to location. The region may be considered as a collection or family of distributions. This family of distributions is indexed by a p-dimensional parameter θ. The parameter θ depends on an independent variable $Z = (Z_1, Z_2, \ldots, Z_k)$. The k components of Z may be measures of such attributes as elevation, average elevation of surround area, or others. Z_1 may be taken to be identically 1 to allow for a constant term. θ is assumed to depend on Z in the form

$$\theta = Z\beta$$

where

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} \cdots & \beta_{2p} \\ \vdots & & \\ \beta_{k1} & \beta_{k2} \cdots & \beta_{kp} \end{bmatrix} .$$

θ must be restricted to a set, say $\Omega \in R^p$, on which the distribution F_θ is defined. Thus, for Z contained in set $\Gamma \in R^k$ we must restrict β to some set, say B, such that $\theta \in \Omega$. The family of distributions is characterized by

$$\mathcal{F} = \{F_\theta: \ \theta = Z\beta, \ Z \in \Gamma \text{ and } \beta \in B\}$$

Where F_θ is the cumulative probability distribution of the climatic (or other) variable of interest with parameter θ.

Our objective is to estimate the entire family of distributions $\mathcal{F}$, by utilizing the dependence of θ on Z. To accomplish this, a random sample of M distributions at M locations from the family of distributions is taken. Next, from these M distributions a sample of size N is taken from each distribution. That is, N observations of the variable of interest, say X, are taken at each of the sampled locations. Also, the independent variable Z is taken at each of the sampled locations. Thus at the j^{th} location the sample will consist of

$$Z_{j1}, Z_{j2}, \ldots, Z_{jk} \text{ and } x_{1j}, x_{1j}, \ldots, x_{Nj} .$$

The entire sample of size N, from M locations, using k attributes of a station is given by

$$\{(x_{ij}, Z_{j\ell};\ 1 \le i \le N,\ 1 \le j \le M,\ 1 \le \ell \le k\}$$

Note that Z is fixed at each location so that only one set of Z components are observed at each location. Using the above sample we may obtain estimators in the form

$$\hat{\theta} = Z\hat{\beta}.$$

2.0 Estimators for Parameters

Although we are primarily interested in θ, β is the global parameter that must be estimated. The estimate of β will then provide a model which will yield an estimate for θ at any location in the region.

2.1 Maximum Likelihood Estimators

The log-likelihood function is given by

$$L(\beta) = \sum_{j=1}^{M} \sum_{i=1}^{N} \ln [f(x_{ij};\ Z^{(j)}\ \beta)] \qquad (2.1.1)$$

where $Z^{(j)}$ refers to the vector $(Z_{j1}, Z_{j2}, \ldots, Z_{jk})$ and f is the probability density function.

Results for the Normal Distribution

Assuming X has a normal distribution with $\theta_{j1} = \mu_j$, $\theta_{j2} = \sigma_j^2$

$$\mu_j = \beta_{11}Z_{j1} + \beta_{21}Z_{j2} + \ldots + \beta_{k1}Z_{jk}$$

and

$$\sigma_j^2 = \beta_{12}Z_{j1} + \beta_{22}Z_{j2} + \ldots + \beta_{k2}Z_{jk} \quad .$$

(Note: the variables $Z_{j1}, Z_{j2}, \ldots, Z_{jk}$ are the same in the above expressions for μ_j and σ_j^2 for notational simplicity. Other variables could be used with only notational changes in the estimators).

For μ_j, we find that if we take partials of (2.1.1) with respect to $\beta_{11}, \beta_{21}, \ldots, \beta_{k1}$ we obtain the following system of equations.

$$\begin{aligned}
\sum_{j=1}^{M} Z_{j1}(\beta_{11}Z_{j1} + \ldots + \beta_{k1}Z_{jk}) &= \sum_{j=1}^{M} Z_{j1}\bar{X}_{\cdot j} \\
\sum_{j=1}^{M} Z_{j2}(\beta_{11}Z_{j1} + \ldots + \beta_{k1}Z_{jk}) &= \sum_{j=1}^{M} Z_{j2}\bar{X}_{\cdot j} \\
&\vdots \\
\sum_{j=1}^{M} Z_{jk}(\beta_{11}Z_{j1} + \ldots + \beta_{k1}Z_{jk}) &= \sum_{j=1}^{M} Z_{jk}\bar{X}_{\cdot j}
\end{aligned} \qquad (2.1.2)$$

where

$$\bar{X}_{\cdot j} = \sum_{i=1}^{N} X_{ij}/N \quad .$$

Let

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1k} \\ Z_{21} & Z_{22} & \cdots & Z_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ Z_{M1} & Z_{M2} & \cdots & Z_{Mk} \end{bmatrix} \quad , \quad \bar{X} = \begin{bmatrix} \bar{X}_{\cdot 1} \\ \bar{X}_{\cdot 2} \\ \cdot \\ \cdot \\ \cdot \\ \bar{X}_{\cdot M} \end{bmatrix}$$

and

$$\beta_{\cdot 1} = \begin{bmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{k1} \end{bmatrix} .$$

The system of normal equations given in (2.1.2) may be written as

$$Z'Z\beta_{\cdot 1} = Z'\bar{X} \qquad (2.1.3)$$

Solving (2.1.3) we obtain

$$\hat{\beta}_{\cdot 1} = (Z'Z)^{-1} Z'\bar{X} , \quad \text{and thus,}$$

based on the previous assumption

$$\hat{\mu}_j = Z^{(j)} \hat{\beta}_{\cdot 1} .$$

We can obtain the normal equations to estimate the parameters in the model for σ_j^2 in a similar way. If we let

$$S^2_{\cdot j} = \sum_{i=1}^{N} (x_{ij} - \hat{\mu}_j)^2/N,$$

$$S^2 = \begin{bmatrix} S^2_{\cdot 1} \\ S^2_{\cdot 2} \\ \vdots \\ S^2_{\cdot M} \end{bmatrix} \qquad \text{and} \quad \beta_{\cdot 2} = \begin{bmatrix} \beta_{12} \\ \beta_{22} \\ \vdots \\ \beta_{k2} \end{bmatrix}$$

then

$$\hat{\beta}_{\cdot 2} = (Z'Z)^{-1} Z' S^2 \quad ,$$

and thus

$$\hat{\sigma}_j^2 = Z^{(j)} \hat{\beta}_{\cdot 2} \quad .$$

The above estimates are equivalent to estimating (μ_j, σ_j^2) with the usual MLE's, $(\bar{X}_{\cdot j}, S_j^2)$. Then to obtain the overall model, $\beta_{\cdot 1}$ and $\beta_{\cdot 2}$ are estimated by regressing $\bar{X}_{\cdot j}$ and S_j^2 on $Z_{j1}, Z_{j2}, \ldots, Z_{jk}$ for $j = 1, 2, \ldots, M$.

2.2 Least Squares Estimators

Another type of estimator which we will refer to as the least squares estimator (LSE) is an alternative to the MLE. This type (the LSE) is useful for its robustness qualities. Somerville and Bean (1982) give a simulation study to illustrate the robustness of the LSE. Parr and Schucany (1980) give a discussion of minimum distance (MD) estimation which cover a number of estimators with robustness qualities. One of the most useful MD estimators is one which is obtained by minimizing the Cramér-von Mises statistic. This particular MD estimator is equivalent to the LSE. A form of the Cramér-von Mises distance or discrepancy between the empirical distributions and the model distributions we define to be

$$d_F(\beta) = \sum_{j=1}^{M} \sum_{i=1}^{N} \left[F(x_{(i)j} \; ; \; Z^{(j)}\beta) - \frac{2i-1}{2N} \right]^2 \Big/ NM \quad .$$

The notation $x_{(i)j}$ refers to the i^{th} largest observation at the j^{th} distribution (location).

The LSE for β is that value of $\hat{\beta} \in B$ such that

$$d_F(\hat{\beta}) = \inf_{\beta \in B} d_F(\beta) \quad . \tag{2.2.1}$$

This estimator is found using non-linear regression methods. Somerville and Bean (1981) illustrate the methods for solving the non-linear regression problem in this context.

3.0 An Example Using the LSE

The Weibull distribution has been used extensively by Somerville and Bean (1979) for various climatic variables including visibility. The probability that visibility is less than x miles, using the Weibull distribution is given by

$$F(x) = 1 - e^{-\alpha x^{\beta}} \quad , \quad \alpha, \beta > 0 \quad .$$

The parameters (α,β) vary from location to location. Note that the use of β in this context differs from the previous sections.

A number of different variables including some climatic and geographical variables were investigated, and the LSE for (α,β) using these variables was developed by Somerville and Bean (1981). A future report which uses more data than the above study will show that the LSE for (α,β) based on the cube of elevation and average elevation of the surrounding area provides a practical model which gives reasonable fits to the data.

Let

Z_1 = (elevation in feet)$^3 \cdot 10^{-9}$,

Z_2 = (average elevation in feet)$^3 \cdot 10^{-9}$ of 20 locations equally spaced a distance of 20 km from the location of interest,

and

$$\alpha = b_{01} + b_{11} Z_1 + b_{21} Z_2$$
$$\beta = b_{02} + b_{12} Z_1 + b_{22} Z_2 \quad .$$

Sixty locations throughout West Germany were used to obtain the LSE, $\hat{b}$, for $b = (b_{01}, b_{11}, b_{21}, b_{02}, b_{12}, b_{22})$ for various times of day and each month of the year.

For example, the LSE for the above parameters were obtined by the appropriate form of (2.2.1) for April for hours 10-12. The resulting equations are given by

$$\hat{\alpha} = .0365 + .0049 Z_1 - .0009 Z_2$$
$$\hat{\beta} = 1.32 - .0092 Z_1 - .0160 Z_2 \quad .$$

The RMS defined by

$$RMS = \left(d_F(\hat{b}) \right)^{1/2}$$

is .058 based on the sixty locations for April hours 1000 - 1200.

Now suppose we wish to have the probability distribution at Konstanz, Germany at the above mentioned month and time. The elevation of 1368 feet yields $Z_1 = 2.56$, and the average elevation of 1725 yields $Z_2 = 4.71$. This gives

$$\hat{\alpha} = .045$$
$$\hat{\beta} = 1.22$$

and

$$F(x) = 1 - e^{-.045 x^{1.22}} \quad , \quad x > 0 \quad .$$

4.0 Summary and Conclusions

A general theory for the estimation of a family of distributions was considered. To estimate a family of distributions, a form of maximum likelihood estimation and least squares estimation was investigated. It was found that if the family of distributions is a family of normal distributions, then a linear regression model results for the estimation of parameters using the MLE.

On the other hand, the LSE was found to be based on non-linear regression methods. The LSE has been found to be practical because of its robustness, and even though non-linear regression is required, it has been found to be no more costly or difficult to calculate than the MLE.

It has been found that the concept of estimating a family of distributions is applicable to extending probability distributions for climatic variables to data-void regions. Also, these concepts should prove to be useful in other areas of research.

5.0 References

1. Parr, William C. and William R. Schucany (1980), "Minimum Distance and Robust Estimation," Journal of the American Statistical Association, 75, 616-624.

2. Somerville, P.N. and S. J. Bean (1979), "Statistical Modeling of Climatic Probabilities," AFGL-TR-79-0222, AD A080559.

3. Somerville, P.N. and S. J. Bean (1981), "Modeling Visibility for Locations in Germany where No Records Exist," AFGL-TR-81-0313, AD A111890.

4. Somerville, P.N. and S. J. Bean (1982), "A Comparison of Maximum Likelihood and Least Squares for the Estimation of a Cumulative Distribution," Journal of Statistical Computation and Simulation, 14, 229-239.