STATISTICAL ANALYSIS OF CLOUD DISTRIBUTION OBSERVED WITH A GROUND-BASED LIDAR

Mutsumi Takagiwa, Kunio Shimizu*
Keio University, Yokohama, Japan

Ichiro Matsui, Nobuo Sugimoto
National Institute for Environmental Studies, Tsukuba, Japan

1. INTRODUCTION

In this work, we apply the statistical analysis method to the ground-based lidar data (June 1996-March 1999) continuously observed in Tsukuba with the National Institute for Environmental Studies compact lidar. We analyze the vertical distribution by applying a distribution function and study seasonal variation. The vertical profile is recorded every 15 minutes and multiple cloud-base heights are allowed up to 10 for a single profile.

2. METHODS

It is clearly observed from the empirical distribution of the cloud-base height observed in June 1-30, 1996 (Fig. 1) that the distribution is not unimodal; it may be bimodal, trimodal or multimodal. Since there may be three cloud overlap in the vertical, mixtures of two or three distributions as theoretical distributions are fitted to empirical distributions of the cloud-base height in this study. Fitting methods are described below.

The normal (Gaussian) distribution with mean μ $(-\infty < \mu < \infty)$ and variance σ^2 $(\sigma > 0)$ whose probability density function is expressed as

$$\phi(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right], \quad -\infty < x < \infty$$

is denoted by $N(\mu,\sigma^2)$. The mixture of 2-component normal distributions $N(\mu_1,\sigma_1^2)$ and $N(\mu_2,\sigma_2^2)$ with weight p_1 (≥ 0) and $p_2 = 1-p_1$ (≥ 0) has a probability density function

$$f(x,\theta) = p_1\phi(x;\mu_1,\sigma_1) + p_2\phi(x;\mu_2,\sigma_2), \quad -\infty < x < \infty, \quad \theta = (p_1,\mu_1,\mu_2,\sigma_1,\sigma_2)'.$$

Similarly the mixture of 3-component normal distributions $N(\mu_1,\sigma_1^2)$, $N(\mu_2,\sigma_2^2)$ and $N(\mu_3,\sigma_3^2)$ with weight p_1 (≥ 0), p_2 (≥ 0) and $p_3 = 1-p_1-p_2$ (≥ 0) has a probability density function

$$f(x,\theta) = p_1\phi(x;\mu_1,\sigma_1) + p_2\phi(x;\mu_2,\sigma_2) + p_3\phi(x;\mu_3,\sigma_3), \quad -\infty < x < \infty.$$

Here the parameter of the distribution consists of a vector $\theta = (p_1,p_2,\mu_1,\mu_2,\mu_3,\sigma_1,\sigma_2,\sigma_3)'$. The frequency curve is fitted to the empirical distribution only over the interval $(0,\infty)$, because the cloud-base height conditional on cloud always takes positive values. Actually values smaller than $T_1 = 100$ (m) are not measured and from the empirical study the number of observations greater than $T_2 = 10{,}000$ (m) is very small. Thus, truncated 2- and 3-component normal distributions

$$g(x,\theta) = \frac{f(x,\theta)}{\int_{T_1}^{T_2} f(t,\theta)dt}, \quad T_1 < x \leq T_2$$

will be fitted to the empirical distribution.

The lognormal distribution with two parameters μ and σ^2, denoted by $LN(\mu,\sigma^2)$, is defined as the distribution of a random variable whose logarithm is distributed as $N(\mu,\sigma^2)$. It may be

* Corresponding author address: Kunio Shimizu, Department of Mathematics, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan; e-mail: shimizu@math.keio.ac.jp

adequate to consider lognormal distributions as a model distribution of the cloud-base height because it is known (Crow and Shimizu eds., 1988) that some cloud characteristics such as heights, horizontal sizes and durations are well fitted by lognormal distributions. The probability density function of $LN(\mu, \sigma^2)$ is

$$\xi(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{1}{2\sigma^2}(\log x - \mu)^2\right], \quad x > 0,$$

where $\log x$ stands for the natural logarithm of x. The parameter σ is a shape parameter and $\exp(\mu)$ is a scale parameter. When X is distributed as $LN(\mu, \sigma^2)$, the mean and the variance of X are $E(X) = \exp[\mu + \sigma^2/2]$ and $\mathrm{Var}(X) = \{E(X)\}^2(\exp[\sigma^2] - 1)$, respectively. The mixture of 2-component lognormal distributions $LN(\mu_1, \sigma_1^2)$ and $LN(\mu_2, \sigma_2^2)$ with weight p_1 (≥ 0) and $p_2 = 1 - p_1$ (≥ 0) and the mixture of 3-component lognormal distributions $LN(\mu_1, \sigma_1^2)$, $LN(\mu_2, \sigma_2^2)$ and $LN(\mu_3, \sigma_3^2)$ with weight p_1 (≥ 0), p_2 (≥ 0) and $p_3 = 1 - p_1 - p_2$ (≥ 0) may be defined in a similar way to the mixtures of 2- and 3-component normal distributions. Similarly to the case of normal distributions, truncated lognormal-mixture distributions will be fitted to the empirical distributions.

We explain below the fitting methods used in this paper. Let $x_1, x_2, ..., x_n$ denote a time series of the cloud-base height observed in a fixed time interval, for example in June in 1996, conditional on cloud. The size of the sample is n and the data set includes values greater than 100.

The usual maximum-likelihood (ML) estimate of θ is obtained by maximizing the likelihood

$$\ell(\theta) = \prod_{i=1}^{n} g(x_i, \theta)$$

or by minimizing the minus log-likelihood

$$S_1(\theta) = -\log \ell(\theta)$$

with respect to θ, where g is the probability density function of a candidate model.

We also employ estimation methods based on grouped data. First, histogram is produced as follows. The support of the histogram covers the values of all data $x_1, x_2, ..., x_n$ with range $h = 100$ (m) starting from $T_1 = 100$. Suppose that K is equal to $\lceil \max_{1\leq i\leq n}(x_i - T_1)/h \rceil$ with notation $\lceil a \rceil$, the least integer which is equal to or greater than a, and the interval $(T_1, Kh]$ is exclusively separated into $(T_1, T_1 + h]$, $(T_1 + h, T_1 + 2h]$,...,$(T_1 + (K-1)h, T_1 + Kh]$. Thus, K represents the total number of categories. From n observed data points, f_j ($j = 1, ..., K$), the frequency in the jth category, will be counted. Fig. 1 has been made in this way.

Second, the following fitting algorithm will be used to estimate the parameter θ of a candidate model whose probability density function is $g(x, \theta)$. Since the frequency in each category whose lower bound is greater than 10,000 m is very small, the theoretical probability of such a category would be very small, too, when a model distribution would be fitted to the empirical distribution. This may cause some problems of numerical computation when the fitting algorithm is used. Thus, we actually fit models to the empirical distribution only over the interval $(T_1, T_1 + kh]$ with $k = 99$ as in the ML estimate. The grouped data ML estimate is provided by the estimate that maximizes the likelihood

$$L(\theta) = \prod_{j=1}^{k} [\pi_j(\theta)]^{f_j}$$

or that minimizes the minus log-likelihood

$$S_2(\theta) = -\log L(\theta)$$

with respect to θ, where $\pi_j(\theta)$ denotes the probability of the jth category under the theoretical model, i.e.,

$$\pi_j(\theta) = \int_{T_1+(j-1)h}^{T_1+jh} g(x, \theta)dx.$$

The minimum χ^2 estimate, $\hat{\theta}$, of θ is provided by minimizing the χ^2 or goodness of fit statistic

$$S_3(\theta) = \sum_{j=1}^{k} \frac{(f_j - n\pi_j(\theta))^2}{n\pi_j(\theta)}$$

with respect to θ. The minimum χ^2 estimator is asymptotically equivalent to the grouped ML estimator.

S-PLUS (1995) has a function **nlminb** (Nonlinear Minimization with Box Constraints) to find

the local minimum of a multivariate function. There are two required arguments to **nlminb**: **objective** (functions $S_1(\theta)$, $S_2(\theta)$ and $S_3(\theta)$ to be minimized in our case) and **start** (a vector of starting values for the minimization). Since, by default, there are no boundary constraints on the parameters, it is impossible to give the constraint $p_1 + p_2 \leq 1$ for instance. However, we can find an estimate by moving p step by step under the conditions $p_1 \in [0, p]$ and $p_2 \in [0, 1-p]$ because the **nlminb** function also takes the optional arguments **lower** and **upper** that specify the bounds on the parameters. The range $[0, \infty]$ is used for the range of σ's because model parameters σ's move from zero to plus infinity. Model parameters μ's move from minus infinity to plus infinity, but it should be noted that the range of μ's is also constrained as $[100, \infty]$ because modality occurs on that interval.

An optimal model between 2- and 3-component normal and lognormal mixtures is chosen by maximizing approximate p-values $P(\chi^2 \geq S_3(\hat{\theta}))$, where χ^2 denotes a random variable whose distribution is the χ^2 distribution with $k-1-r$ degrees of freedom. Here r is the number of parameters in the hypothesized model; $r = 5$ for 2-component normal and lognormal mixtures, and $r = 8$ for 3-component normal and lognormal mixtures.

3. RESULTS

Three estimation methods, exact ML, grouped ML and minimum χ^2, were compared. The difference between the results (not cited) was very small. Because it is theoretically known that these three are asymptotically equivalent, this is a valid conclusion (in our case the size of the sample is very large). Thus, it is enough to use the exact ML method for the parameter estimation of a model.

Table 1 shows the estimated parameter values and the values of $S_3(\hat{\theta})$ for the method of minimum χ^2 when the hypothesized model is a truncated 3-component lognormal mixture. From Table 1 we see that the fit of truncated 3-component lognormal mixture is poor in the sense that the hypothesized model is rejected at the 0.01 level of significance in many cases. The fit is worse for truncated 2- and 3-component normal mixtures and for truncated 2-component lognormal mixture except several cases. However, it may be enough to choose a 2-component normal model for interpreting the data set. Fitted distributions can be seen in Figs. 1 and 2.

Roughly speaking the 3-component lognormal mixture is an adequate model for data in summer (Fig. 2), while for data in winter (Fig. 3) the 3-component normal mixture is better. This reflects clear seasonal variation. The proposed model is useful in the sense that it gives some information on the distribution such as modality, weight and local maximum. When the 2-component normal mixtures are fitted to the empirical distribution, the estimated values of μ's are shown in Fig. 4. Yearly variation can be seen because of continuous observation from June 1996 to March 1999.

Table 1. Minimum χ^2 estimation (truncated 3-component lognormal mixture)

Fig. 1. Empirical and fitted distributions of cloud-base heights (96/06) (truncated 2-component normal and lognormal mixtures)

Fig. 2. Empirical and fitted distributions of cloud-base heights (96/06) (truncated 3-component normal and lognormal mixtures)

Fig. 3. Empirical and fitted distributions of cloud-base heights (96/12) (truncated 3-component normal and lognormal mixtures)

Fig. 4. Variation of estimated μ's (truncated 2-component normal mixtures)

REFERENCES

Crow, E. L. and Shimizu, K. eds. (1988), Lognormal Distributions: Theory and Applications, pp. 387, Marcel Dekker, Inc., New York.

S-PLUS (1995), S-PLUS Documentation, Statistical Sciences, Inc., Seattle.

Table 1: Minimum χ^2 estimation (truncated 3-component lognormal mixture)

	$\hat{p}_1$	$\hat{p}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$S_3(\hat{\theta})$
96/06	0.6182	0.1380	7.0391	7.0407	8.8182	1.3915	0.3216	0.1976	163.66
96/07	0.4908	0.0980	6.1980	7.1504	8.9676	0.9332	0.2804	0.4007	175.29
96/08	0.6540	0.1927	6.9494	7.3378	8.8746	1.3992	0.3309	0.2076	128.79
96/09	0.0340	0.7695	7.3231	7.9155	8.8154	0.0948	1.2657	0.1886	172.48
96/10	0.0266	0.8287	5.6081	7.5219	8.8364	0.3808	0.6079	0.1845	187.76
96/11	0.0310	0.2363	5.7421	7.3575	8.9715	0.1922	0.2286	1.5706	310.68
96/12	0.0000	0.4209	4.9013	7.2975	8.0814	1.0843	0.2319	1.1603	175.54
97/01	0.2634	0.1834	7.1262	7.3951	8.3707	1.3605	0.2095	0.3964	115.97
97/02	0.3995	0.5655	7.3447	8.3765	8.5635	0.3651	1.3164	0.0538	151.60
97/03	0.2702	0.5224	7.4163	7.4364	8.5035	0.2347	1.0446	0.1693	144.04
97/04	0.1943	0.0357	7.4615	8.5135	10.529	0.5235	0.1049	9.7963	190.68
97/05	0.3831	0.4298	6.1677	7.5498	8.7533	0.7597	0.4813	0.1938	134.27
97/06	0.6841	0.2075	7.0459	8.4712	8.8483	1.2159	0.3002	0.0822	160.50
97/07	0.0011	0.0089	6.5505	8.6293	31.228	0.0068	0.2799	15.111	168.19
97/08	0.6001	0.1093	6.5688	7.1957	8.6963	0.7840	0.1124	0.3971	173.43
97/09	0.5685	0.2497	6.7610	7.6074	8.5935	1.0010	0.3138	0.2985	161.92
97/10	0.0000	0.1848	5.0197	7.3790	9.0953	1.2134	0.2939	1.6359	248.63
97/11	0.3604	0.4983	6.2711	7.5226	8.7191	0.6389	0.4291	0.1971	115.39
97/12	0.1710	0.7036	6.0798	7.2947	8.6265	0.4286	0.3644	0.2052	108.42
98/01	0.0688	0.5359	6.8871	7.3507	8.4439	0.8887	0.3157	0.3073	158.92
98/02	0.1930	0.6411	6.2873	7.2483	8.3664	0.4303	0.4304	0.2728	138.22
98/03	0.1886	0.5532	6.6360	7.4413	8.5500	0.8070	0.3703	0.2006	146.30
98/04	0.6821	0.1096	6.1990	7.3263	8.6180	1.5954	0.1833	0.2988	181.28
98/05	0.4099	0.2632	6.3574	7.1000	8.5999	1.1161	0.4536	0.4338	120.13
98/06	0.3409	0.3305	4.9949	7.0186	8.7597	1.1908	0.5138	0.3177	241.28
98/07	0.0100	0.0100	6.5731	8.8213	22.666	0.3423	0.1047	11.191	145.64
98/08	0.9300	0.0370	1.5039	6.6097	8.6634	1.3236	0.6973	0.4468	143.85
98/09	0.8676	0.0952	4.1763	7.0246	8.8251	0.1647	0.9765	0.2838	170.09
98/10	0.3025	0.2880	6.6811	7.3125	8.7327	1.2629	0.4259	0.2450	153.15
98/11	0.0909	0.3610	6.1302	7.4232	8.7740	0.4573	0.3156	0.2453	164.65
98/12	0.2972	0.4579	6.9914	7.4760	8.7511	0.6899	0.2964	0.1981	161.12
99/01	0.0000	0.3017	5.1678	7.2871	8.3732	0.7583	0.4405	1.1204	218.09
99/02	0.1037	0.7691	4.5742	7.7638	8.4683	2.0699	0.7718	0.2427	110.03
99/03	0.3837	0.3340	6.5826	7.3426	8.6866	0.8725	0.2806	0.2624	165.37

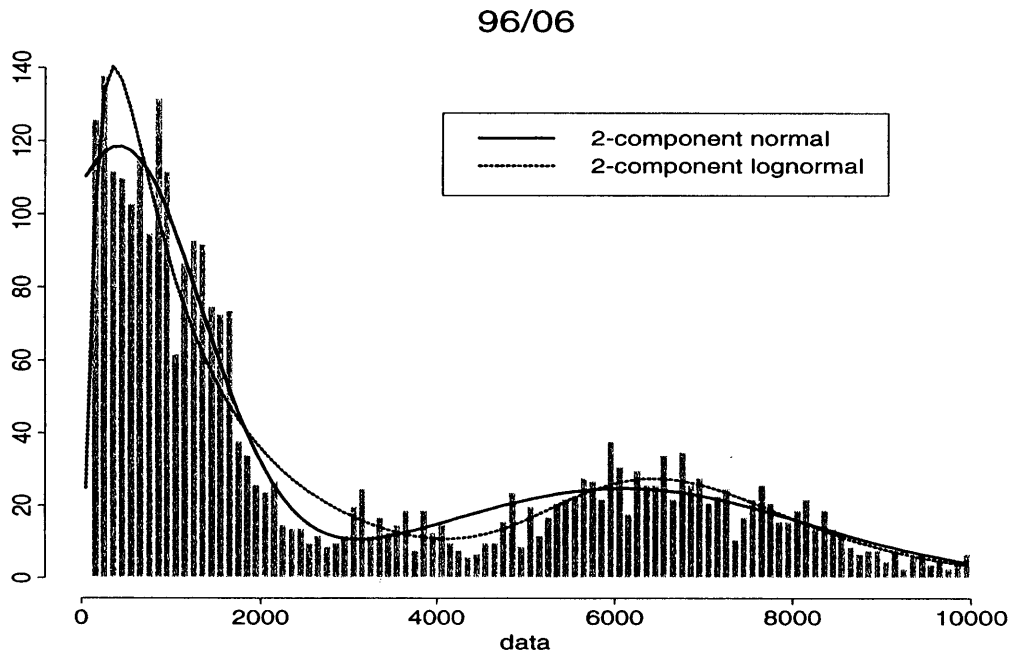

Fig. 1: Empirical and fitted distributions of cloud-base heights (96/06) (truncated 2-component normal and lognormal mixtures)

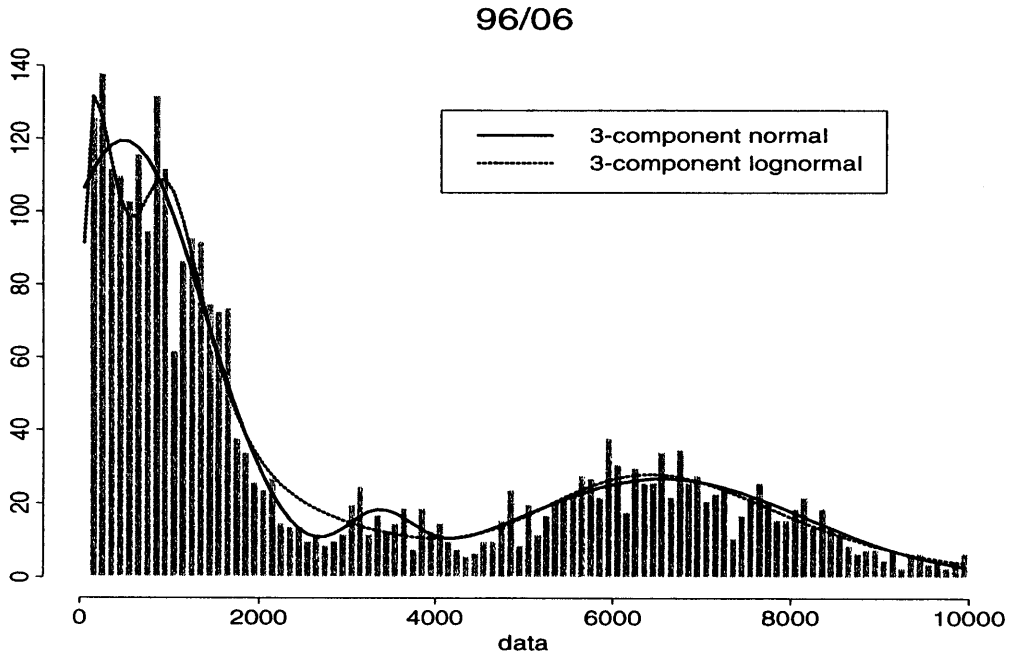

Fig. 2: Empirical and fitted distributions of cloud-base heights (96/06) (truncated 3-component normal and lognormal mixtures)

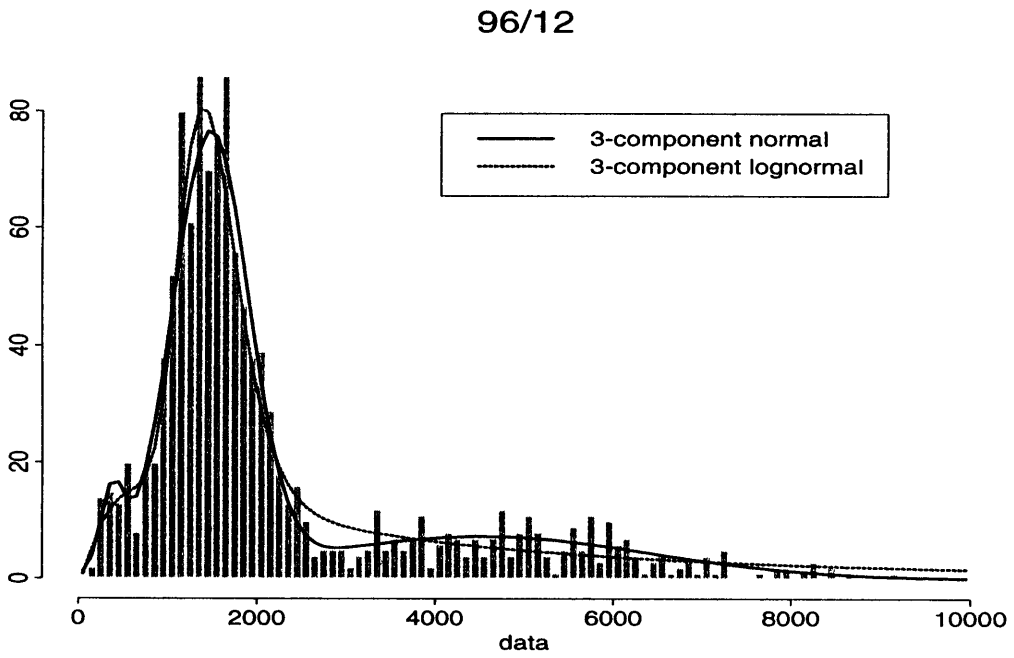

Fig. 3: Empirical and fitted distributions of cloud-base heights (96/12) (truncated 2-component normal and lognormal mixtures)

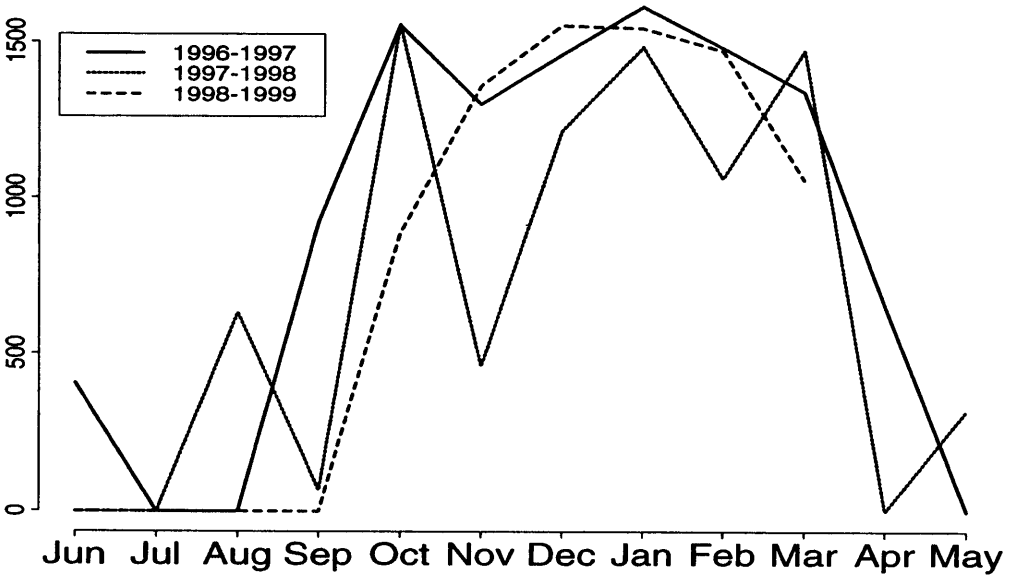

Fig. 4: Variation of estimated μ's (truncated 2-component normal mixtures)