The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.847
28(5), 847–858

A Study of Outlier Detection Using the Mixture of Extreme Distributions Based on Deep-Sea Fishery Data

Jung Jin Lee[a,1] · Jae Kyoung Kim[a]

[a]Department of Statistics, Soongsil University

(Received December 26, 2014; Revised July 7, 2015; Accepted August 13, 2015)

Abstract

Deep-sea fishery in the Antarctic Ocean has been actively progressed by the developed countries including Korea. In order to prevent the environmental destruction of the Antarctic Ocean, related countries have established the Commission for the Conservation of Antarctic Marine Living Resources (CCAMLR) and have monitored any illegal unreported or unregulated fishing. Fishing of tooth fish, an expensive fish, in the Antarctic Ocean has increased recently and high catches per unit effort (CPUE) of fishing boats, which is suspicious for an illegal activity, have been frequently reported. The data of CPUEs in a fishing area of the Antarctic Ocean often show an extreme Distribution or a mixture of two extreme distributions. This paper proposes an algorithm to detect an outlier of CPUEs by using the mixture of two extreme distributions. The parameters of the mixture distribution are estimated by the EM algorithm. Log likelihood value and posterior probabilities are used to detect an outlier. Experiments show that the proposed algorithm to detect outlier of the data can be adopted instead of simple criteria such as a CPUE is greater than 1.

Keywords: outlier detection, mixture of extreme distributions

1. 서론

우리나라는 경제발전과 더불어 수산업도 발전하여 원양어업이 빠르게 성장해왔고 수출에서 큰 비중을 차지했었다. 최근 수출에서 원양어업의 비중이 줄어들기는 했지만 여전히 우리나라는 세계 상위의 원양어업 대국이다. 원양어업은 여러 바다에서 진행되고 있는데 남극해에서는 이빨고기(tooth fish), 크릴새우, 빙어 등을 잡는다. 이 중에서 흔히 '메로'라고 부르는 값비싼 이빨고기는 현재 남극해에서 연간 약 1만여 톤 이상 어획하고 있다. 주인 없는 남극해의 생태계를 보호하기 위해 조업 국가들은 남극해양생물자원보존위원회(Commission for the Conservation of Antarctic Marine Living Resources; CCAMLR)를 만들고 남극 해양생물에 대한 보존과 이용 방법 및 조사 연구, 불법조업 감시 등의 업무를 수행하고 있다. CCAMLR는 Figure 1.1과 같이 남극해를 여러 개의 조업 해역으로 나누어 각 원양어업의 조업 현황 데이터를 수집, 분석하여 다른 어선에 비해 비정상적으로 높은 어획량을 보이는 어선을 탐색하거나 조업 금지구역에서의 불법 조업여부를 판단한다.

CCAMLR는 한 배가 어느 해역에서 한 번 그물을 넣어 조업한 성과를 다음과 같은 CPUE(catch per

[1]Corresponding author: Department of Statistics, Soongsil University, Seoul 156-743, Korea.
E-mail: jjlee@ssu.ac.kr

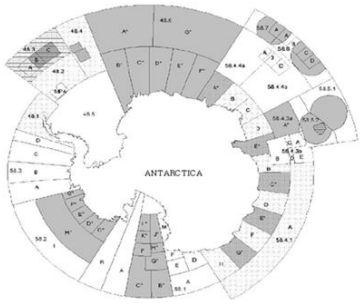

Figure 1.1. Partitioned fishing areas in Antarctic Ocean.

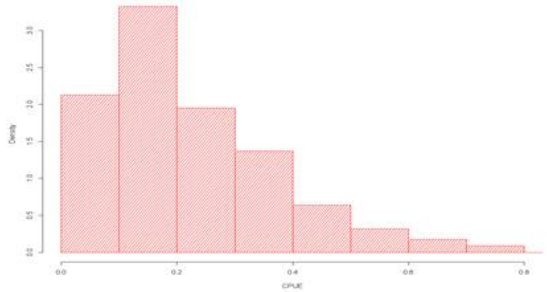

Figure 1.2. CPUE distribution of the fishing area 881 from 2008 to 2013.

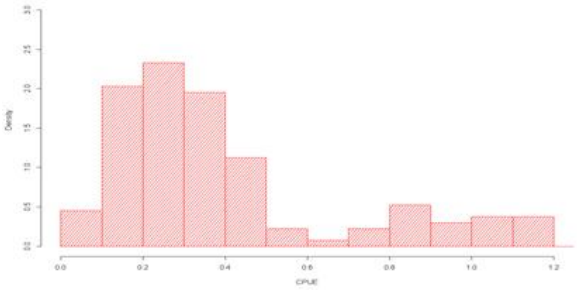

Figure 1.3. CPUE distribution of the fishing area 486E from 2008 to 2013.

unit effort)로 측정한다.

$$\text{CPUE} = \frac{\text{총어획량 (단위kg)}}{\text{사용한 낚시바늘수}}.$$

즉 CPUE는 낚시 바늘 하나에 얼마나 많은 양의 고기를 잡았는지를 측정하는 것이다. 현재 CCAMLR는 CPUE가 단순히 1보다 크면 많은 양의 고기를 잡은 것으로 생각하고 혹시 불법조업을 하지 않았는지 의심하여 세부 조사를 한다. 남극해의 각 조업 해역에서 2008–2013년도 기간 중 이빨고기를 조업한 모든 어선들의 CPUE에 대한 히스토그램을 그려보면 대개 Figure 1.2와 같은 극단분포나 Figure 1.3과 같은 두 분포함수의 혼합 형태를 가진다.

Figure 1.2와 같이 CPUE 데이터가 단일 극단분포일 경우 이상점 탐색은 상위 백분위수를 이용하든가, 아니면 많이 알려져 있는 거리를 이용한 이상점 탐색 모형을 사용하면 된다. 하지만 Figure 1.3과 같이

CPUE 데이터가 혼합 분포 형태인 경우 우측의 분포에 해당되는 이상점들을 찾는 적절한 통계적 모형을 적용하기가 쉽지 않다.

본 논문은 Figure 1.3과 같이 데이터가 혼합 분포 형태인 경우 이상점 탐색을 위한 통계적 방법을 연구하고자 한다. 2절에서는 데이터에 적합한 혼합 분포 모형을 추정한 후 로그 가능도함수(likelihood function)나 사후확률을 이용한 이상점 탐색 알고리즘을 제안한다. 3절에서는 이 알고리즘을 남극해 자료에 적용하여 이상점 탐색의 성과를 실험한다. 4절에서는 결론과 향후과제에 대한 제안을 한다.

2. 혼합 극단분포를 이용한 이상점 탐색

이상점 탐색이란 전체 데이터 중에서 대부분의 다른 데이터와는 속성이 불일치되거나 심하게 다른 데이터를 찾는 것을 의미하는데 (Lee, 2011; Kang 등, 2007), 속성의 값들이 일반적인 값과 상당히 차이가 큰 값을 가져 편차 탐지(deviation, detection)라고 하기도 하고, 예외적으로 나타난다는 의미에서 예외점 마이닝(exception mining)이라 부르기도 한다 (Yong 등, 2007). 많이 이용되는 간단한 방법은 데이터 간의 k-인접이웃거리(k-nearest neighbor distance)를 계산하여 거리가 현저히 높은 데이터는 이상점으로 간주한다 (Kim 등, 2010). 이밖에도 데이터의 밀도나 군집을 정의하여 이상점 탐색을 하기도 하고 서포트벡터기계(support vector machine)를 이용하기도 한다 (Seo와 Yoon, 2011). 하지만 이러한 방법은 대개 거리 개념을 이용한 것이고 통계적 모형에 근거한 방법이 아니다.

Aitkin과 Wilson (1980)은 혼합 정규분포 모형을 이용한 이상점 탐색 방법을 제시하였다. 이 모형은 데이터를 다음과 같은 두 확률분포의 혼합모형으로 가정하는 것이다.

$$f(x) = (1-\lambda) f_a(x) + \lambda f_b(x), \quad 0 < \lambda < 1, \tag{2.1}$$

여기서 $f_a(x)$는 정상적인 데이터를 의미하는 정규분포와 같은 특정한 확률분포를 가정하고, $f_b(x)$는 이상 데이터들의 확률분포로서 균등분포(uniform distribution) 등을 가정한다. 하지만 Figure 1.3과 같은 남극해의 이빨고기 CPUE 데이터는 $f_a(x)$와 $f_b(x)$ 모두 극단적인 분포 형태로서 혼합 극단분포 모형을 이용한 이상점 탐색은 아직 적용된 사례가 없다. 본 논문에서는 다음과 같은 네 종류의 확률분포 혼합 모형으로 이상점 탐색을 연구하였다.

1) 와이불(Weibull) 분포. α: 형상(shape) 모수, β: 척도(scale)모수

$$f(x) = \begin{cases} \dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

2) 정규(Normal) 분포

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

3) 감마(Gamma) 분포. α: 형상(shape) 모수, β: 척도(scale) 모수

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}.$$

4) 로그정규(Log-Normal) 분포. μ: 평균로그(meanlog), σ: 표준편차로그(sdlog)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}.$$

혼합 확률분포 모형에서 주어진 표본 데이터 $x_1, x_2, \ldots, x_n$에 대한 가능도함수는 다음과 같다

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \{(1-\lambda) f_a(x_i) + \lambda f_b(x_i)\}. \tag{2.2}$$

최대우도법으로 모수를 추정하기 위해 이와 같은 혼합 확률분포의 가능도함수가 최대가 되는 모수를 추정하는 것은 함수의 형태가 복잡하여 쉽지 않다. 하지만 이상점 탐색을 위한 혼합 확률분포 모형에서는 한 데이터가 한쪽 분포에 속할 경우 다른 분포에서 관측될 가능성이 거의 없다고 가정할 수 있다. 확률분포 $f_a(x)$와 $f_b(x)$를 따르는 데이터 집합을 각각 A와 B로 표시하고 이 집합에 속하는 데이터 수를 각각 n_a, n_b라고 하면 위의 식은 다음과 같이 근사하여 표현할 수 있다.

$$f(x_1, x_2, \ldots, x_n) \approx \left\{ (1-\lambda)^{n_a} \prod_{x_i \in A} f_a(x_i) \right\} \lambda^{n_b} \prod_{x_i \in B} f_b(x_i). \tag{2.3}$$

이 경우 로그 가능도함수는 다음과 같다.

$$\ln f(x_1, x_2, \ldots, x_n) = n_a \ln(1-\lambda) + \sum_{x_i \in A} \ln f_a(x_i) + n_b \ln \lambda + \sum_{x_i \in B} \ln f_b(x_i). \tag{2.4}$$

본 논문에서 고려하는 네 가지 분포함수에 대한 16가지 혼합 확률분포 모형의 로그 가능도함수 목록은 다음과 같다.

1) $f_a(x)$: Weibull, $f_b(x)$: Weibull

$$n_a \ln(1-\lambda) + n_a \ln \alpha_1 + (\alpha_1 - 1) \sum_{x_i \in A} \ln x_i - n_a \alpha_1 \ln \beta_1 - \sum_{x_i \in A} \left(\frac{x_i}{\beta_1} \right)^{\alpha_1}$$
$$+ n_b \ln \lambda + n_b \ln \alpha_2 + (\alpha_2 - 1) \sum_{x_i \in B} \ln x_i - n_b \alpha_2 \ln \beta_2 - \sum_{x_i \in B} \left(\frac{x_i}{\beta_2} \right)^{\alpha_2}.$$

2) $f_a(x)$: Weibull, $f_b(x)$: Normal

$$n_a \ln(1-\lambda) + n_a \ln \alpha + (\alpha - 1) \sum_{x_i \in A} \ln x_i - n_a \alpha \ln \beta - \sum_{x_i \in A} \left(\frac{x_i}{\beta} \right)^{\alpha}$$
$$+ n_b \ln \lambda - \left(\frac{n_b}{2} \right) \ln 2\pi - n_b \ln \sigma - \left(\frac{1}{2\sigma^2} \right) \sum_{x_i \in B} (x_i - \mu)^2.$$

3) $f_a(x)$: Weibull, $f_b(x)$: Gamma

$$n_a \ln(1-\lambda) + n_a \ln \alpha_1 + (\alpha_1 - 1) \sum_{x_i \in A} \ln x_i - n_a \alpha_1 \ln \beta_1 - \sum_{x_i \in A} \left(\frac{x_i}{\beta_1} \right)^{\alpha_1}$$
$$+ n_b \ln \lambda - n_b \ln(\Gamma(\alpha_2)) - n_b \ln({\beta_2}^{\alpha_2}) + (\alpha_2 - 1) \sum_{x_i \in B} \ln(x_i) - \sum_{x_i \in B} \left(\frac{x_i}{\beta_2} \right).$$

4) $f_a(x)$: Weibull, $f_b(x)$: LogNormal

$$n_a \ln(1-\lambda) + n_a \ln \alpha + (\alpha - 1) \sum_{x_i \in A} \ln x_i - n_a \alpha \ln \beta - \sum_{x_i \in A} \left(\frac{x_i}{\beta} \right)^{\alpha}$$
$$+ n_b \ln \lambda - \sum_{x_i \in B} \ln(x_i) - \left(\frac{n_b}{2} \right) \ln 2\pi - n_b \ln \sigma - \frac{\sum_{x_i \in B} (\ln x_i - \mu)^2}{2\sigma^2}.$$

5) $f_a(x)$: Normal, $f_b(x)$: Weibull

$$n_a \ln(1-\lambda) - \left(\frac{n_a}{2}\right)\ln 2\pi - n_a \ln\sigma - \left(\frac{1}{2\sigma^2}\right)\sum_{x_i\in A}(x_i-\mu)^2$$
$$+ n_b \ln\lambda + n_b \ln\alpha + (\alpha-1)\sum_{x_i\in B}\ln x_i - n_b\alpha\ln\beta - \sum_{x_i\in B}\left(\frac{x_i}{\beta}\right)^{\alpha} c.$$

6) $f_a(x)$: Normal, $f_b(x)$: Normal

$$n_a \ln(1-\lambda) - \left(\frac{n_a}{2}\right)\ln 2\pi - n_a \ln\sigma_1 - \left(\frac{1}{2{\sigma_1}^2}\right)\sum_{x_i\in A}(x_i-\mu_1)^2$$
$$+ n_b \ln\lambda - \left(\frac{n_b}{2}\right)\ln 2\pi - n_b \ln\sigma_2 - \left(\frac{1}{2{\sigma_2}^2}\right)\sum_{x_i\in B}(x_i-\mu_2)^2.$$

7) $f_a(x)$: Normal, $f_b(x)$: Gamma

$$n_a \ln(1-\lambda) - \left(\frac{n_a}{2}\right)\ln 2\pi - n_a \ln\sigma - \left(\frac{1}{2\sigma^2}\right)\sum_{x_i\in A}(x_i-\mu)^2$$
$$+ n_b \ln\lambda - n_b \ln(\Gamma(\alpha)) - n_b \ln(\beta^{\alpha}) + (\alpha-1)\sum_{x_i\in B}\ln(x_i) - \sum_{x_i\in B}\left(\frac{x_i}{\beta}\right).$$

8) $f_a(x)$: Normal, $f_b(x)$: LogNormal

$$n_a \ln(1-\lambda) - \left(\frac{n_a}{2}\right)\ln 2\pi - n_a \ln\sigma_1 - \left(\frac{1}{2{\sigma_1}^2}\right)\sum_{x_i\in A}(x_i-\mu_1)^2$$
$$+ n_b \ln\lambda - \sum_{x_i\in B}\ln(x_i) - \left(\frac{n_b}{2}\right)\ln 2\pi - n_b \ln\sigma_2 - \frac{\sum_{x_i\in B}(\ln x_i-\mu_2)^2}{2{\sigma_2}^2}.$$

9) $f_a(x)$: Gamma, $f_b(x)$: Weibull

$$n_a \ln(1-\lambda) - n_a \ln(\Gamma(\alpha_1)) - n_a \ln({\beta_1}^{\alpha_1}) + (\alpha_1-1)\sum_{x_i\in A}\ln(x_i) - \sum_{x_i\in A}\left(\frac{x_i}{\beta_1}\right)$$
$$+ n_b \ln\lambda + n_b \ln\alpha_2 + (\alpha_2-1)\sum_{x_i\in B}\ln x_i - n_b\alpha_2\ln\beta_2 - \sum_{x_i\in B}\left(\frac{x_i}{\beta_2}\right)^{\alpha_2}.$$

10) $f_a(x)$: Gamma, $f_b(x)$: Normal

$$n_a \ln(1-\lambda) - n_a \ln(\Gamma(\alpha)) - n_a \ln(\beta^{\alpha}) + (\alpha-1)\sum_{x_i\in A}\ln(x_i) - \sum_{x_i\in A}\left(\frac{x_i}{\beta}\right)$$
$$+ n_b \ln\lambda - \left(\frac{n_b}{2}\right)\ln 2\pi - n_b \ln\sigma - \left(\frac{1}{2\sigma^2}\right)\sum_{x_i\in B}(x_i-\mu)^2.$$

11) $f_a(x)$: Gamma, $f_b(x)$: Gamma

$$n_a \ln(1-\lambda) - n_a \ln(\Gamma(\alpha_1)) - n_a \ln({\beta_1}^{\alpha_1}) + (\alpha_1-1)\sum_{x_i\in A}\ln(x_i) - \sum_{x_i\in A}\left(\frac{x_i}{\beta_1}\right)$$
$$+ n_b \ln\lambda - n_b \ln(\Gamma(\alpha_2)) - n_b \ln({\beta_2}^{\alpha_2}) + (\alpha_2-1)\sum_{x_i\in B}\ln(x_i) - \sum_{x_i\in B}\left(\frac{x_i}{\beta_2}\right).$$

12) $f_a(x)$: Gamma, $f_b(x)$: LogNormal

$$n_a \ln(1-\lambda) - n_a \ln(\Gamma(\alpha)) - n_a \ln(\beta^\alpha) + (\alpha-1)\sum_{x_i \in A} \ln(x_i) - \sum_{x_i \in A} \left(\frac{x_i}{\beta}\right)$$
$$+ n_b \ln \lambda - \sum_{x_i \in B} \ln(x_i) - \left(\frac{n_b}{2}\right) \ln 2\pi - n_b \ln \sigma - \frac{\sum_{x_i \in B} (\ln x_i - \mu)^2}{2\sigma^2}.$$

13) $f_a(x)$: LogNormal, $f_b(x)$: Weibull

$$n_a \ln(1-\lambda) - \sum_{x_i \in A} \ln(x_i) - \left(\frac{n_a}{2}\right) \ln 2\pi - n_a \ln \sigma - \frac{\sum_{x_i \in A} (\ln x_i - \mu)^2}{2\sigma^2}$$
$$+ n_b \ln \lambda + n_b \ln \alpha + (\alpha-1) \sum_{x_i \in B} \ln x_i - n_b \alpha \ln \beta - \sum_{x_i \in B} \left(\frac{x_i}{\beta}\right)^\alpha.$$

14) $f_a(x)$: LogNormal, $f_b(x)$: Normal

$$n_a \ln(1-\lambda) - \sum_{x_i \in A} \ln(x_i) - \left(\frac{n_a}{2}\right) \ln 2\pi - n_a \ln \sigma_1 - \frac{\sum_{x_i \in A} (\ln x_i - \mu_1)^2}{2{\sigma_1}^2}$$
$$+ n_b \ln \lambda - \left(\frac{n_b}{2}\right) \ln 2\pi - n_b \ln \sigma_2 - \left(\frac{1}{2{\sigma_2}^2}\right) \sum_{x_i \in B} (x_i - \mu_2)^2.$$

15) $f_a(x)$: LogNormal, $f_b(x)$: Gamma

$$n_a \ln(1-\lambda) - \sum_{x_i \in A} \ln(x_i) - \left(\frac{n_a}{2}\right) \ln 2\pi - n_a \ln \sigma - \frac{\sum_{x_i \in A} (\ln x_i - \mu)^2}{2\sigma^2}$$
$$+ n_b \ln \lambda - n_b \ln(\Gamma(\alpha)) - n_b \ln(\beta^\alpha) + (\alpha-1) \sum_{x_i \in B} \ln(x_i) - \sum_{x_i \in B} \left(\frac{x_i}{\beta}\right).$$

16) $f_a(x)$: LogNormal, $f_b(x)$: LogNormal

$$n_a \ln(1-\lambda) - \sum_{x_i \in A} \ln(x_i) - \left(\frac{n_a}{2}\right) \ln 2\pi - n_a \ln \sigma_1 - \frac{\sum_{x_i \in A} (\ln x_i - \mu_1)^2}{2{\sigma_1}^2}$$
$$+ n_b \ln \lambda - \sum_{x_i \in B} \ln(x_i) - \left(\frac{n_b}{2}\right) \ln 2\pi - n_b \ln \sigma_2 - \frac{\sum_{x_i \in B} (\ln x_i - \mu_2)^2}{2{\sigma_2}^2}.$$

이와 같은 혼합 극단분포 모형을 이용한 이상점 탐색은 가능도 확률의 차이를 이용하는 알고리즘과 사후 확률을 이용한 방법을 생각할 수 있다.

2.1. 가능도 확률의 차이를 이용한 이상점 탐색 알고리즘

데이터 $x_1, x_2, \ldots, x_n$의 혼합분포 모형(식 (1.1))에서 $f_a(x)$를 정상 데이터의 분포, $f_b(x)$를 이상 데이터의 분포라 하자. 만일 데이터 x_1이 이상 데이터이고, $x_2, \ldots, x_n$이 정상 데이터라고 가정하면 x_1의 $f_a(x)$에서 추출되었을 확률은 매우 작을 것이고 $f_b(x)$에서 추출되었을 확률은 상대적으로 클 것이다. 이 경우 모든 데이터 $x_1, x_2, \ldots, x_n$이 $f_a(x)$분포에서 추출되었다고 가정하고 가능도 확률을 구했을 때와 식 (1.3)과 같은 혼합 확률분포의 가능도 확률을 비교하면 차이가 많이 나게 된다. 이와 같은 점을 이용하여 다음과 같은 이상점 탐색 알고리즘을 제안할 수 있다.

[가능도 확률의 차이를 이용한 이상점 탐색 알고리즘]

(단계 1) 모든 데이터를 정상적인 데이터 집합 A에 속한다고 가정하고 초기 로그 가능도함수 값 L_0 계산, $i = 1$.

(단계 2) 데이터 x_i를 이상 데이터 집합 B로 하고, 나머지 데이터를 집합 A로 하여 로그 가능도함수 값 L_i 계산.

(단계 3) 로그 가능도함수 값 L_i와 초기 로그 가능도함수 값 L_0를 비교했을 때 매우 차이가 크면 이상점으로 분류.

(단계 4) 데이터가 끝날 때까지 $i = i + 1$로 하여 (단계 2)와 (단계 3)을 반복.

하지만 이 알고리즘은 가능도함수 값의 차이가 어느 정도일 때 이상점으로 간주해야 하는지 기준 값을 설정해야 하는 문제가 있을 수 있다.

2.2. 사후 확률을 이용한 이상점 탐색

데이터 $x_1, x_2, \ldots, x_n$의 혼합분포 모형 추정식이 다음과 같다고 하자.

$$\hat{f}(x) = (1 - \hat{\lambda})\hat{f}_a(x) + \hat{\lambda}\hat{f}_b(x), \quad 0 < \lambda < 1. \tag{2.5}$$

관측치 x_i가 정상 데이터 분포인 $f_a(x)$에서 나온 데이터인지, 아니면 이상 데이터 분포 $f_b(x)$로부터 나온 데이터인지 판정하는 방법은 추정된 확률분포를 이용하여 다음과 같은 사후확률을 구해보는 것이다.

$$\widehat{\Pr}(x_i \in A \mid x_i) = \frac{\left(1 - \hat{\lambda}\right)\hat{f}_a(x_i)}{\left(1 - \hat{\lambda}\right)\hat{f}_a(x_i) + \hat{\lambda}\hat{f}_b(x_i)}, \tag{2.6}$$

$$\widehat{\Pr}(x_i \in B \mid x_i) = \frac{\left(1 - \hat{\lambda}\right)\hat{f}_b(x_i)}{\left(1 - \hat{\lambda}\right)\hat{f}_a(x_i) + \hat{\lambda}\hat{f}_b(x_i)}. \tag{2.7}$$

만일 B에 속할 가능성이 크면 이상점으로 판정한다.

$$\widehat{\Pr}(x_i \in A \mid x_i) < \widehat{\Pr}(x_i \in B \mid x_i). \tag{2.8}$$

식 (2.8)이면 x_i를 이상점으로 판정

각각의 데이터에 대해 A에서 나왔을 사후 확률과 B에서 나왔을 사후 확률을 구하여 이상점 여부를 판정한다.

3. 혼합 극단분포 모형을 이용한 이상점 탐색 실험

3.1. 혼합 극단분포 모형의 추정

남극해에서 조업한 이빨고기의 CPUE 데이터 중에서 혼합 극단분포 형태를 보이는 조업지역은 486E (Figure 3.1)와 5842A (Figure 3.2) 해역이다.

각 해역의 데이터에 대해 R을 이용하여 EM 알고리즘으로 혼합 확률분포의 모수를 추정하는 프로그램을 작성한 후 Table 3.1에서 제시된 16가지 혼합 확률분포 모형의 모수를 추정하였다. 16가지 모형

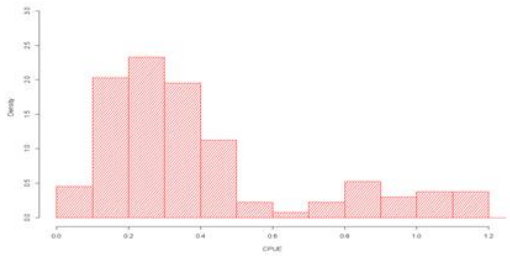

Figure 3.1. CPUE distribution of the fishing area 486E from 2008 to 2013.

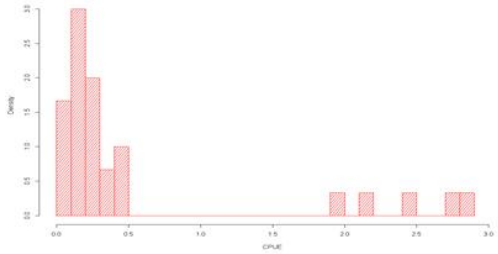

Figure 3.2. CPUE distribution of the fishing area 5842A from 2008 to 2013.

Table 3.1. χ^2 goodness of fit tests for the mixture distribution models at each fishing area

해역	혼합 분포 모형	p-값
486E	Weibull-Weibull	0.9335
	Weibull-Normal	0.9184
	Weibull-LogNormal	0.8900
	Normal-Normal	0.8412
	Normal-LogNormal	0.7908
	Gamma-Weibull	0.6847
	Gamma-Normal	0.6711
	Gamma-Gamma	0.6545
	Gamma-LogNormal	0.6376
	LogNormal-Normal	0.2179
5842A	Weibull-LogNormal	0.9961
	Weibull-Normal	0.9960
	Gamma-Gamma	0.9953
	Gamma-LogNormal	0.9952
	Gamma-Normal	0.9950
	Weibull-Weibull	0.9949
	Gamma-Weibull	0.9938
	LogNormal-LogNormal	0.9921
	LogNormal-Normal	0.9920
	Normal-Normal	0.9889

중 χ^2 적합성 검정을 한 결과 p-값이 큰 순으로 각 해역에 적합한 10개의 혼합 확률분포 모형이 Table 3.1과 같다.

χ^2 적합성 검정 결과 486E 해역에서는 Weibull-Weibull 모형의 p-값이 가장 높고, 5842A 해역에서

Table 3.2. Parameter estimation for a mixture distribution of the fishing area 486E

혼합모형	모수	추정값	모수	추정값	lambda
Weibull-Weibull	α_1	2.3355	α_2	6.8264	0.1894
	β_1	0.3048	β_2	1.0022	

Table 3.3. Parameter estimation for a mixture distribution of the fishing area 5842A

혼합모형	모수	추정값	모수	추정값	lambda
Weibull-LogNormal	α	1.7953	μ	0.8820	0.1667
	β	0.2275	σ	0.1502	

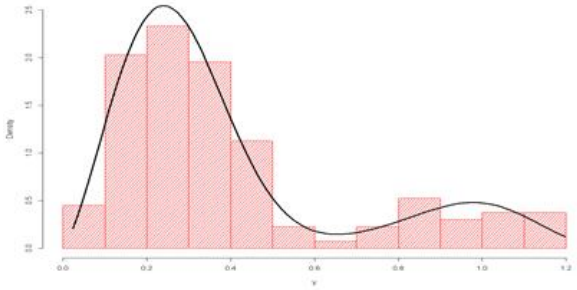

Figure 3.3. A mixture of extreme distribution model for the CPUE data of the fishing area 486E.

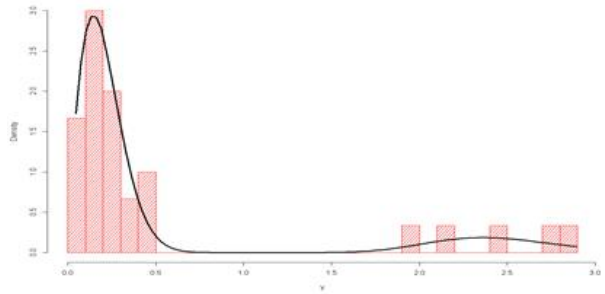

Figure 3.4. A mixture of extreme distribution model for the CPUE data of the fishing area 5842A.

는 Weibull-LogNormal 모형이 p-값이 가장 높게 나타났다. 두 모형들의 모수 추정 결과는 Table 3.2 및 Table 3.3과 같고, 각 해역 데이터의 히스토그램 위에 추정된 혼합 극단분포 함수를 겹쳐 표현하면 Figure 3.3 및 Figure 3.4과 같다. 해역 486E의 경우 정상과 비정상 데이터가 겹치는 부분이 존재하는 혼합 극단분포 형태이다. 해역 5842A의 경우 정상 데이터와 비정상 데이터가 확연히 구분되어 두 개별적인 분포의 결합으로 보이는 형태이다.

3.2. 가능도 확률의 차이를 이용한 이상점 탐색

추정된 혼합 극단분포 모형을 이용하여 이상점을 탐색하기 위해 먼저 모든 관측값들을 정상 데이터라고 가정하고 초기 로그 가능도함수 값(L_0)을 계산한다. 그리고 각 관측 값 x_i가 이상점이라고 가정했을 때의 로그 가능도함수 값(L_i)을 계산하여 두 값의 차이가 L_0의 5% 이상 크게 나타나는 경우를 찾는다. 여기서 5%는 임의로 정한 기준이고 다른 적절한 기준 값을 찾기 위한 실험을 계속한다. 이 알고리즘으로 찾은 CPUE 값을 조사한 결과가 Table 3.4와 같다. 두 해역 모두 CPUE 값이 1보다 큰 이상점을 잘

Table 3.4. Comparison of the initial log-likelihood and log-likelihood of the mixture distribution

해역	혼합모형	초기 로그 가능도	혼합모형 로그 가능도	CPUE
486E	Weibull-Weibull	−226.2602	−207.5280	1.1989
			−208.3893	1.1725
			−208.7972	1.1603
			−209.4216	1.1418
			−209.4735	1.1403
			−211.2065	1.0902
			−213.5336	1.0242
			−213.8737	1.0145
			−213.8807	1.0143
			−214.2873	1.0028
5842A	Weibull-LogNormal	−315.4898	−235.6047	2.8940
			−244.4709	2.7336
			−257.4627	2.4883
			−272.8878	2.1810
			−285.6048	1.9164

Table 3.5. Outlier detection rate by using the difference of log-likelihood ratios

해역	데이터수	로그 가능도함수 값 차이	탐색 이상점 수(CPUE ≥ 1)	탐색률(CPUE ≥ 1)
486E	133	1%	10	100%
		5%	10	100%
		10%	0	0%
5842A	30	1%	5	100%
		5%	5	100%
		10%	4	80%

Table 3.6. Outlier detection rate by using the posterior probabilities

해역	데이터수	이상점 기준값	탐색 이상점 수	이상점 비율	현행 이상점 탐색률(CPUE ≥ 1)
486E	133	0.6651	25	18.7%	100%
5842A	30	1.9163	5	17.7%	100%

탐색하였다. 즉 현재 관례적으로 사용되고 있는 CPUE 값이 1보다 큰 값을 이상점으로 하는 것은 L_0와 L_i의 차이가 L_0의 5%를 기준으로 하면 '적어도' 찾을 수 있다는 의미이다.

Table 3.5는 초기 로그 가능도함수값과 혼합 분포모형의 로그 가능도함수 값의 차이가 L_0의 1%, 5%, 그리고 10%인 경우의 현행 이상점(CPUE ≥ 1) 탐색 결과이다.

당연히 1%인 경우는 현행 이상점을 잘 탐색해 내지만 10%일 경우는 모두 탐색하지는 못한다. 과연 10% 기준일 때 탐색된 이상점 데이터가 실제로 이상점인지는 이 수치를 보이는 배의 항적 검사 등 불법 조업여부를 조사하여야 하는데 본 연구에서는 이와 같은 조사를 할 수 없다. 향후 현장에서 어느 기준일 때가 가장 적절한지에 대한 연구가 더 필요하다.

3.3. 사후 확률을 이용한 이상점 탐색

추정된 혼합 극단분포 모형을 이용하여 사후 확률을 이용한 이상점 탐색(2.2절)을 한 결과가 Table 3.6과 같다.

사후 확률을 이용하여 이상점을 탐색한 결과 486E 해역에서는 25개의 데이터가 이상점으로 분류되었고 그 기준값은 CPUE $\geq$ 0.6651이다. 이 해역의 데이터는 두 극단분포가 혼합된 형태이어서 이상 데이터가 많이 분류되었다. 이 방법도 현행 이상점 기준(CPUE $\geq$ 1)에 맞는 데이터는 잘 탐색한 것으로 나타났다. 하지만 이상점으로 분류된 데이터가 과연 불법조업을 하였는지는 본 연구에서 결론내릴 수 없고 현장 조사를 해 보아야 알 수 있을 것이다. 5842A 해역에서는 5개의 데이터가 이상점으로 분류되었고 그 기준값은 CPUE $\geq$ 1.9163이다. 현행 기준의 이상점은 모두 탐색하였지만 사후확률을 이용하면 현재 사용되는 기준보다 더욱 강화된 기준이 되어 앞으로 추가 연구가 역시 필요하다.

4. 결론 및 향후과제

현재 CCAMLR는 어느 해역에서나 조업한 배의 CPUE가 1보다 클 경우 불법조업 가능성이 있는지 조사를 한다. 본 논문에서는 모든 해역에 공통적으로 적용하는 단순한 이상점 탐색보다 각 해역의 데이터의 확률분포를 고려한 이상점 탐색 가능성을 제시하였다. 통계적 모형을 이용하면 향후 남극해에서 불법 조업하는 배의 탐색을 잘 할 수 있을 것으로 기대한다.

본 논문에서는 데이터가 일반적으로 많이 관찰되지 않는 혼합 극단분포 모형을 따를 때 이상점을 탐색하는 알고리즘을 제안하였다. 혼합분포의 초기 가능도 확률값 L_0와 각 데이터의 로그 가능도함수 값의 차이를 이용하여 이상점을 탐색하면 현행 이상점 기준(CPUE $\geq$ 1)에 맞는 데이터는 잘 탐색한 것으로 나타났다. 또 다른 방법으로 사후확률을 이용한 이상점 탐색 방법을 제안하였다. 이 방법도 현행 이상점 기준(CPUE $\geq$ 1)에 맞는 데이터는 잘 탐색한 것으로 나타났다. 하지만 탐색된 이상점 데이터가 실제로 이상점인지는 이 값을 가진 배의 항적 검사 등 불법조업여부를 조사하여야 하는데 본 연구에서는 이와 같은 조사를 할 수 없어 한계가 있다. 향후 현장에서 어느 기준일 때가 가장 적절한지에 대한 연구가 더 필요하다. 그리고 CPUE에 영향을 미치는 요인에는 해역, 조업방법, 미끼, 어구 등 여러 가지 요인이 있어 일반화된 탐색기준을 찾으려면 더 많은 연구가 필요하다.

본 논문에서는 카이제곱 검정을 이용하여 혼합분포의 적합성을 조사하였다. 과연 혼합분포의 적합성 검정으로 카이제곱 검정이 이론적으로 타당한지, 아니면 기타 방법을 이용하여야 하는지 향후 연구해 볼만한 과제이다.

감사의 글

본 논문의 심사를 맡아 세밀하게 이론적 검증을 하여주신 심사위원들에게 감사드립니다. 특히 혼합분포의 사후확률을 이용한 이상점 탐색을 제안해 주신 심사위원께 감사드립니다.

References

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm, *Technometrics*, **22**, 325–331.

Kang, C. W., Kang, H. C., Park, S. H., Seung, H. W., Yong, H. S., Lee, D. H., Lee, S. K., Lee, Y. S., Jin, S. H., Choi, J. H. and Han, S. T. (2007). *Data Mining Concepts and Methods*, Cyprus.

Kim, S., Cho, N. W. and Kang, S. H. (2010). Density based outlier detection for massive data analysis, *Korean Journal of Management Science*, **15**, 71–88.

Lee, J. J. (2011). *Data Mining Using R, SAS and MS-SQL*, Freedom Academy.

Seo, H. S. and Yoon, M. (2011). Outlier detection using support vector machines, *Communications for Statistical Applications and Methods*, **18**, 171–177

Yong, H. S., Na, Y. M., Seung, H. W., Lee, M. S., Lee, S. J. and Choi, L. (2007). *Data Mining*, Infinity Books.

원양어선 조업 데이터의 혼합 극단분포를 이용한 이상점 탐색 연구

이정진[a,1] · 김재경[a]

[a]숭실대학교 정보통계보험수리학과

(2014년 12월 26일 접수, 2015년 7월 7일 수정, 2015년 8월 13일 채택)

요 약

남극해에서는 우리나라를 포함한 연안 강대국들의 원양어업이 활발히 성행하고 있다. 주인 없는 남극해의 생태계를 보호하기 위해 조업 국가들은 남극해양생물자원보존위원회를 만들고 협약을 맺어 일정한 어획량만 조업하고 금지기간과 금지구역을 설정하여 불법조업을 방지하고 있다. 남극해에서 조업하는 어종 중의 하나가 이빨고기(tooth fish)인데 비싼 값 때문에 불법조업이 있는 경우가 많다. 한 배의 조업성과는 CPUE(catch per unit effort)로 나타낼 수 있고, 한 지역에서 조업한 배들의 CPUE는 단일 또는 혼합 극단분포 형태를 가진다. 단일 극단분포일 경우 이상점 탐색은 상위 백분위수를 이용하면 된다. 본 논문은 자료가 혼합 극단분포인 경우 이상점 탐색을 위한 통계적 방법을 연구하고자 한다. 본 연구에서는 자료에 적합한 혼합 극단분포 모형을 EM 알고리즘으로 추정한 후 로그 가능도함수 값을 이용하거나 사후 확률을 이용한 이상점 탐색 알고리즘을 제안한다. 이 방법을 남극해 조업 데이터에 적용하여 시뮬레이션 한 결과 통계적 방법 적용의 가능성을 보여주었다.

주요용어: 이상점 탐색, 혼합 극단분포

[1]교신저자: (156-743) 서울시 동작구 상도로 369, 숭실대학교 정보통계보험수리학과. E-mail: jjlee@ssu.ac.kr