creative commons
COMMONS DEED

저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

석사학위 논문

혼합 극단분포를 이용한 이상점 탐색 연구

A Study of Outlier Detection Using Mixture Extreme Distribution

2014년 12월

숭실대학교 대학원

정보통계보험수리학과

김 재 경

석사학위 논문

혼합 극단분포를 이용한 이상점 탐색 연구

A Study of Outlier Detection Using Mixture Extreme Distribution

2014년 12월

숭실대학교 대학원

정보통계보험수리학과

김 재 경

석사학위 논문

혼합 극단분포를 이용한 이상점 탐색 연구

지도교수 이 정 진

이 논문을 석사학위 논문으로 제출함

2014년 12월

숭실대학교 대학원

정보통계보험수리학과

김 재 경

김 재 경의 석사학위 논문을 인준함

심 사 위 원 장 _____ 인

심 사 위 원 _____ 인

심 사 위 원 _____ 인

2014년 12월

숭실대학교 대학원

목 차

표 목 차

그 림 목 차

국문초록

혼합 극단분포를 이용한 이상점 탐색 연구

김재경
정보통계보험수리학과
숭실대학교 대학원

남극해에서는 우리나라를 포함한 연안 강대국들의 원양어업이 활발히 성행하고 있다. 주인 없는 남극해의 생태계를 보호하기 위해 조업 국가들은 남극해양생물자원보존위원회를 만들고 협약을 맺어 일정한 어획량만 조업하고 조업 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. 자료에 적합한 혼합 극단분포 모형을 EM 알고리즘으로 추정한 후 로그 가능도함수 값을 이용한 이상점 탐색 알고리즘을 제안한다. 이 방법을 남극해 자료에 적용하여 시뮬레이션 한 결과 만족스러운 탐색결과를 보여주었다.

ABSTRACT

A Study of Outlier Detection Using Mixture Extreme Distribution

KIM, JAE-KYOUNG

Department of Statistics and Actuarial Science

Graduate School of Soongsil University

Deep-sea fishery around Antarctic Ocean has been active by the developed countries including Korea. In order to protect environmental destruction of Antarctic Ocean, related nations established the Commission for the Conservation of Antarctic Marine Living Resources(CCAMLR) and has watched illegal unreported or unregulated fishing. The fishery of Antarctic tooth fish, which is expensive, has been increased recently and so this has led to illegal unreported or unregulated fishing. Catch of a ship can be measured by the CPUE(catch per unit effort). The data of CPUEs in a fishing area show usually an extreme value distribution and/or a mixture of two extreme distributions. This paper proposes an algorithm to detect an outlier CPUE by using the mixture of two extreme distributions. The parameters of the mixture distributions are estimated by the EM algorithm. Log likelihood values are used to detect an outlier.

Simulation experiments shows that the proposed algorithm works well to detect outlier of the data.

제 1 장 서 론

1.1 연구배경 및 목적

우리나라의 수산업 발전과 수출에 큰 영향을 차지했던 원양어업은 그동안 우리나라의 경제성장과 발맞추어 빠르게 성장해왔다. 비록 지금은 예전보다 규모가 줄어들기는 했지만 여전히 우리나라는 세계 상위의 원양대국이다.

그러나 빠르게 성장한 한국의 원양어업은 연안 강대국들의 권리 행사로 인해 쇠락하기 시작하였고, 그에 따라 새로운 어장을 찾게 되었다. 그 중의 하나가 바로 남극에서 주로 잡히는 우리가 흔히 메로 라고 부르는 이빨고기를 잡는 것이었다. 현재 남극해 내의 이빨고기 조업국들은 연간 약 1만여 톤 이상 어획하고 있는 것으로 파악된다. 하지만 비싼 값 때문에 이빨고기의 어획량이 늘어나면서 불법조업이 성행하게 되었고, 이에 주인 없는 남극해의 생태계 파괴를 보호하고 유지하고자 남극해양생물자원보존위원회(Commission for the Conservation of Antarctic Marine Living Resources)가 생기게 되었다.

남극해양생물자원보존위원회는 남극해양생물자원에 대한 보존과 이용 방법 및 조사 연구, 불법조업 감시 등의 업무를 수행하고 있다. 이에 조업국들의 조업 현황 데이터를 수집, 분석하여 비슷한 상황의 다른 배들에 비해 비정상적으로 높은 어획량이 보이거나 조업금지구역에서의 불법조업여부를 판단하기 위해 노력하고 있다.

본 논문에서는 특정 시점, 특정 해역에서의 데이터가 극단적이고 혼합된 형태를 이룬다면 기존의 방법이 아닌 혼합 극단분포모형을 이상점 탐색에 활용하는 새로운 방법을 제시하고자 한다.

예를 들면, 현재 일반적으로 하나의 분포모형으로 적합 되는 자료에서

는 이상점을 탐색할 경우에 “상위 1%”, “상위 5%” 등으로 기준을 설정하는 방법을 사용하지만 혼합된 형태의 자료에서는 위와 같은 기준이 적절하지 않을 수 있다.

본 논문에서는 자료가 단일 또는 혼합 극단분포 형태를 가질 경우 혼합 극단분포모형을 이용하여 이상점을 탐색한다.

1.2 연구방법 및 구성

본 논문은 통계적 탐색방법을 중심으로 이상점 탐색을 진행하고자 한다. 이상점 탐색이란 전체 데이터 중에서 대부분의 데이터와는 속성이 많이 다른 데이터를 찾는 것을 의미한다.

이상점 데이터라도 잘못된 데이터가 아니라 정상적이지만 대부분의 데이터보다는 특별한 값을 가진다는 의미이므로 사용에 주의하여야 한다.

이러한 이상점 탐색문제는 통계학이나 기계학습, 데이터마이닝 등 여러 분야에서 연구되어 왔다.

실증적 사례연구는 CCAMLR의 2008년부터 5개년도 이빨고기 조업 자료를 이용하여 이루어 졌다.

본 논문은 총 5개의 장으로 구성되었다.

제1장 서론에서는 연구 배경 및 목적, 연구방법 및 구성을 기술하였다.

제2장에서는 앞서 이상점 탐색을 연구한 사례를 살펴보고자 하였다.

제3장에서는 실제 데이터를 이용하여 데이터를 탐색 및 전처리하고, 시각적 데이터 분석을 진행한 결과를 통해 혼합 극단분포 모형을 이상점 탐색에 적용하는 이유와 EM 알고리즘을 이용하여 혼합 극단분포 모형의 모수를 추정하는 방법에 대해 정리하였다.

제4장에서는 결정된 혼합 극단분포 모형의 모수를 추정한 결과와 실제 데이터와 적합 되었을 경우 얼마나 적합성이 뛰어나고 이상점 탐색률이

높은지에 대해 기술하였다.

마지막으로 제5장에서는 본 연구의 의의 및 활용방안과 향후 개선방안에 대해 제시함으로써 논문을 마무리하고자 한다.

제 2 장 선행연구

2.1 이상점 탐색의 의의 및 중요성

이상점 탐색이란 전체 데이터 중에서 대부분의 다른 데이터와는 속성이 불일치되거나 심하게 다른 데이터를 찾는 것을 의미하는데, 속성의 값들이 일반적인 값과 상당히 차이가 큰 값을 가지므로 편차 탐지(deviation detection)라고 하기도 하고, 예외적으로 나타난다는 의미에서 예외점 마이닝(exception mining)이라 부르기도 한다.

통계학자인 더글라스 호킨(Douglas Hawkin)은 이상점에 대해 다른 관찰들과 서로 다른 관찰로써 다른 방법을 사용하여 생성된 결과라는 의심을 불러일으키는 관찰이라고 정의를 내렸다.

이상점 탐색문제는 통계학, 기계학습, 데이터 마이닝 등 다양한 분야에서 연구되어 왔다. 이상점 데이터는 잘못된 데이터가 아니라 지극히 정상적이지만 다른 대부분의 데이터보다는 특별한 값을 가지고 중요한 사실을 반영하는 것일 수도 있다.

예를 들어 어느 회사 직원들의 월급을 조사하였는데 CEO의 월급은 다른 직원들에 비해 클 수 있다. 그렇다고 해서 데이터가 잘못된 것이라고 할 수는 없다.

또한 어느 신용카드를 사용하는 고객이 평소에 비하여 사용횟수와 빈도가 갑자기 많아졌다면 신용카드 회사는 이 고객이 평상시와 다르다는 점을 파악하거나 구매패턴을 조사하여 도난 신용카드인지 확인해 볼 수 있다.

이와 같이 사기 탐지, 침입 탐지, 질병의 발생 등 다양한 분야에서 이상점 탐색을 응용할 수 있다.

2.2 여러 논문에서 연구된 이상점 탐색 방안

Murray Aitkin(1980)의 논문 'Mixture Models, Outliers, and the EM Algorithm'에서는 이상점이 존재하는 데이터에서 혼합 모형을 유도하고, 혼합 모형의 MLE를 이용한 모수 추정에 대해 기술하였다.

또한 EM 알고리즘을 이용하여 Normal Mixture의 이상점 탐색 방법을 제시하였다.

서한손(2011)의 논문 '서포트 벡터 기계를 이용한 이상치 진단'에서는 $\mu-\epsilon-$서포트 벡터 회귀를 이용하여 이상점을 진단하는 방법을 제안하였다.

이 논문에서 제시한 $\mu-\epsilon-$서포트 벡터 회귀방법은 기존의 접근방법에 비하여 계산비용을 감소시키고 이상점 문턱에 대해 좀 더 정확하게 정의할 수 있다는 장점을 가진다고 하였다.

김승(2010)의 논문 '대용량 자료 분석을 위한 밀도기반 이상치 탐지'에서는 밀도기반 이상점 탐색 기법중 하나인 LOF 알고리즘을 개선하는 방법에 대하여 설명하였다.

LOF 알고리즘이란 각각의 전체 데이터 개체에 대해 개별적인 개체마다 이상점 정도를 나타내는 측정치를 계산하는 것으로 이 측정치를 Local Outlier Factor(LOF)라 한다.

이 LOF 알고리즘은 여러 가지 좋은 장점이 있으나 계산에 부하가 걸린다는 단점으로 인하여 제한적으로 활용되었다는 점을 지적하였다.

대용량 데이터에 적용하기 위하여 계산의 정확도 보다는 감수할 수 있을 만큼의 불확실성을 추가하여 계산의 질의 감소는 최소화하고 데이터 색인으로 얻을 수 있는 계산시간의 이점을 활용할 수 있는 방법으로 Approximated k-nearest Neighbor(ANN)을 사용하는 것을 제시하였다.

2.3 통계적 이상점 탐색방법

이상점 탐색에 적용되는 통계적 방법은 주어진 데이터 집단에 대해서 확률분포 모형을 가정하고, 각 데이터가 이 확률분포에서 발생할 가능성을 찾는다. 만일 어느 데이터가 가정된 확률분포에서 발생할 가능성이 매우 낮다면 이 데이터는 이상점으로 정의된다.

이 방법은 데이터에 대해 가장 적합한 확률분포가 무엇인지에 관한 지식을 요구한다. 그러나 많은 경우에 데이터의 확률분포는 알려지지 않는다. 그리고 몇 개의 변수를 이상점 탐색에 이용할지 결정해야 한다.

주어진 데이터가 특정한 확률분포에서 추출된 것인지는 Q-Q 그림, 카이제곱 적합성 검정, 콜모고로프-스미르노프(Kolmogorov-Smirnov) 검정 등을 이용할 수 있다.

또한 확률분포가 가정되었다면 이 확률분포에서 이상점에 대한 기준값을 무엇으로 정할지 결정해야 한다.

한 변수일 경우 확률분포 모형에 의한 이상점 탐색은 기준점(cutoff point) c를 정한 후 구간 $|X| < c$를 벗어나는 데이터를 찾는다. 이와 같은 방법은 2차원 이상의 데이터에도 확장하여 적용할 수 있다.

예를 들어 데이터가 m-차원 다변량정규분포 $N(\mu, \Sigma)$라면 다음과 같은 확률 식에 의해 이상점의 영역을 결정한다.

$$P(x-\mu)^{'}\Sigma^{-1}(x-\mu) < c = 1-\alpha$$

여기서 α는 다변량 정규분포에서 정의될 이상점의 확률인데 이 확률을 정하면 c가 결정된다. 이때 모수 μ와 Σ는 표본 데이터에서 추정한다.

혼합 분포모형을 이상점 탐색에 활용할 수도 있다. 혼합 분포모형으로 접근하는 방식은 얻어진 데이터가 혼합된 확률분포를 가지며 각 데이터는 이들 분포중의 하나로 식별된다고 가정하는 것이다.

$$f(x) = (1-\lambda) f_a(x) + \lambda f_b(x)$$

여기서 λ는 0과 1 사이의 값으로 이상점의 기대비율을 의미한다. 일반적으로 $f_a(x)$는 데이터를 통해 특정한 확률분포를 가정하고, 이상점들의 확률분포 $f_b(x)$는 균등분포(uniform distribution)를 가정하여 이상점을 탐색하게 된다.

2.4 유사성-기반 이상점 탐색방법

유사성-기반(proximity-based)의 이상점 탐색은 대부분의 데이터보다 거리가 멀리 떨어져 있는 데이터를 탐색하는 방법이다. 이 방법은 데이터에 적합한 유사성 측도만을 정의하면 되기 때문에 통계분포함수를 이용하는 방법보다 더 용이하게 응용되고 있다.

한 데이터에 대해 유사성에 근거한 이상점 점수(outlier score)의 정의는 여러 가지 방법이 있으나 가장 간단한 방법으로 k-인접이웃거리(k-nearest neighbor distance)가 많이 이용된다.

즉, 모든 데이터에 대해 k-번째 인접한 데이터까지의 거리를 계산하여, 이 거리가 현저히 높은 데이터는 이상점으로 간주한다. 이상점 점수의 최저값은 0이며, 최고값은 거리 함수의 최대 가능값으로 보통 무한대가 된다.

이 방법에서는 k를 몇 번째로 하느냐에 따라 결과가 민감하게 달라질 수 있다. k가 너무 작으면 이상점을 탐색 못할 가능성이 있고, k가 너무 크면 많은 수의 데이터가 이상점으로 분류될 위험이 있다. 이러한 문제점을 방지하기 위하여 k-인접이웃까지의 평균 거리를 사용하기도 한다.

유사성-기반 이상점 탐색방법은 간편하지만 여러 개의 군집이 있을 경우에 이상점 탐색이 힘들 수도 있다.

2.5 밀도-기반 이상점 탐색방법

밀도-기반(density-based)의 이상점 탐색은 한 지역의 데이터 밀도가 낮을 때 이상점으로 간주한다. 데이터의 밀도는 보통 유사성에 의해서 정의되기 때문에 이 방법은 유사성-기반 이상점 탐색방법과 밀접한 연관이 있다.

밀도의 정의로는 k-인접이웃거리(k-nearest neighbor distance) 평균과 반비례 하도록 하는 방법이 많이 이용된다.

즉, k-인접이웃들의 평균 거리가 작다면 밀도는 높고 평균 거리가 크면 밀도는 낮게 된다. 이러한 밀도-기반 이상점 탐색방법도 k를 어떻게 정의하느냐에 따라 결과가 민감하게 달라질 수 있다.

주어진 m차원 데이터 x에 대하여 k-인접이웃이 $y_1, y_2, \cdots, y_n$라면 x의 밀도는 아래와 같이 정의할 수 있다.

$$density(x) = \left(\frac{\sum_{i=1}^{k} d(x, y_i)}{k} \right)^{-1}$$

만일 거리가 작은 경우 많은 데이터들이 적은 밀도를 가지게 되어 높은 이상점 점수를 갖게 된다. 그러나 거리가 큰 경우에는 많은 이상점들이 일반적인 데이터와 비슷한 이상점 점수를 가지게 되어 탐색이 용이하지 않을 수 있다.

이러한 밀도의 변형으로 아래와 같은 상대밀도(relative density)를 이용할 수도 있다.

$$density(x) = \frac{d(x, y_k)}{\frac{1}{k} \sum_{i=1}^{k} d(x, y_i)}$$

이 밖에도 밀도를 정의하는 방법은 여러 가지가 있을 수 있다.

2.6 군집-기반 이상점 탐색방법

군집 분석은 서로 강하게 관련된 데이터들의 군집을 탐색하는 것이고, 군집-기반(cluster-based)의 이상점 탐색은 기존 군집분석과는 반대로 군집과 강하게 관련이 없는 데이터를 탐색하는 방법이다. 즉, 기존의 군집분석 방법을 역으로 이상점 탐색에 이용할 수 있다.

군집분석의 몇 가지 모형은 군집의 크기가 기준 값보다 작은 경우 해당 군집을 버렸는데, 이렇게 버리는 군집에 적당한 이상점 점수(outlier score) 기준을 적용하면 이상점을 찾을 수 있다.

모든 데이터에 대한 군집화를 수행한 후에 각 데이터가 임의의 군집에 속하는 정도를 평가할 때 각 데이터의 군집 중심까지의 거리를 이상점 점수로 이용할 수 있다.

목적함수를 최대 또는 최소로 하는 군집방법에서는 목적함수 값을 이상점 점수로 이용할 수 있다.

예를 들어 k-means에서 한 데이터를 제거하였을 때 그 군집의 오차제곱합(SSE)을 많이 개선하게 된다면 이 데이터는 이상점이라고 분류할 수 있다.

즉, 군집의 중심에서 멀리 떨어진 데이터를 제거하게 되면 그 군집의 오차제곱합(SSE)은 많이 줄어들 것이다. 다시 말하면 군집은 데이터의 모델을 생성하며 이상점은 그 모델을 왜곡시킨다는 것이다.

군집분석과 이상점 탐색은 서로 독립적으로 이루어지는 것이 아니라 일반적으로 동시에 이루어지게 된다. 어떠한 군집분석이라도 단 한 번의 시도로서 원하는 군집을 모두 찾아내기는 쉽지 않다.

이러한 경우 데이터를 군집화한 후에 이상점을 제거하고 다시 데이터들을 군집화 한다. 비슷한 절차를 여러 번 반복하여 군집화 과정을 진행하면 최종적인 결과는 원하는 군집의 형성과 더불어 이상점 집합을 얻게

된다.

그러나 이와 같은 군집분석과 이상점의 동시 탐색은 k-means의 경우 군집의 개수를 자동으로 결정하지 못하므로 군집의 수 k를 어떻게 정하느냐에 따라 결과가 달라질 수 있다.

그러므로 군집분석에 근거한 이상점 탐색은 k를 변화시켜 가며 분석을 여러 번 반복해야 만족스러운 결과를 얻을 수 있다.

제 3 장 혼합 극단분포를 이용한 이상점 탐색

3.1 연구 의의 및 방향

일반적인 이상점 탐색 절차는 다음과 같다.

[그림 3-1] 이상점 탐색 과정

본 논문에서는 먼저 데이터 탐색 및 전처리를 통해 중복된 데이터나 잘못 입력된 데이터를 제거하고, 시각화 데이터 분석을 통해 단일 또는 혼합 극단분포 형태를 가지는 해역을 탐색한다.

탐색된 해역에 대한 혼합 극단분포 모형을 추정하고 χ^2 적합성 검정을 통해 가장 적합한 분포 모형을 찾아 이상점 탐색에 활용한다.

원양어선에서 고기를 잡았을 때 각 배의 크기가 차이가 있고, 사용하는 미끼나 낚싯바늘의 개수가 다르므로 어획량에 대한 객관적인 평가를 위해 CPUE라는 지표를 사용한다.

CPUE(Catch Per Unit Effort)는 1개의 낚싯바늘에 걸리는 고기의 무게이며, 하나의 낚싯바늘에서 평균 1kg 이상의 고기가 잡히는 경우를 HIGH CPUE로 판단하게 된다.

그러나 데이터의 형태가 극단적인 모습으로 나타나지만 CPUE 값이 1 보다 작으면 HIGH CPUE로 취급하지 않는다.

만약 특정 해역에서 대부분의 데이터와는 속성이 다른 지속적인 HIGH CPUE가 발생한다면 불법조업을 한 것으로 의심하게 된다.

혼합 극단분포 모형을 활용하여 이상점 탐색을 하게 된다면 불법조업 여부에 대해 좀 더 과학적인 판단을 할 수 있을 것이라 생각된다.

3.2 데이터 탐색 및 전처리

데이터 탐색 및 전처리는 데이터를 탐색하고 분석에 맞게 적당한 형태로 변환시키는 것이다. 이를 통해 중복된 데이터나 잡음이 있는 데이터를 제거한다. 데이터가 중복되거나 잡음이 존재하게 되는 이유는 여러 가지가 존재할 수 있다. 시간적으로 차이가 나는 조업 데이터인데 매번 데이터가 수집될 때마다 입력하는 것이 아니라 여러 시점에 조업한 결과를 한꺼번에 입력하면서 실수가 발생될 수 있고, 조업을 하다가 불가피한 상황에서 그물이 끊어지게 되는 경우 결측값이 생길 수 있다. 이런 데이터를 포함시켜 분석을 하게 될 경우에 믿을 수 없는 결과가 도출될 가능성이 있으므로 데이터 탐색 및 전처리 과정을 진행하였다.

3.2.1 데이터 구성

본 논문에서는 CCAMLR이 관리하고 있는 수역 내에서 2008/09 어기 ~ 2012/13 어기(5년) 동안의 이빨고기 조업 자료 7,738개를 이용하였다. 변수는 총 37개가 존재하며 크게 기초정보 관련 변수, 시간 및 어구 관련 변수, 어획량 관련 변수, 데이터 수집 목적 관련 변수, 부가 설명 변수 5가지로 구성되어 있다. [표 3-1]에 세부 변수에 대한 설명이 되어 있으며, [표 3-2]에서는 세부 변수를 요약한 자료가 정리되어 있다.

[표 3-1] 변수 설명

번호	변수명	내용
1	ID	기록번호
2	AssignedASD	해역
3	AssignedSSRU	소해구
4	CCAMLRSeason	어기
5	CalendarYear	조사년
6	Month	조사월
7	NATIONALITY_CODE	국가
8	SHIP_CODE	선박번호
9	SHIP_NAME	선박명
10	DEPTH_START_SET	어구투승해역수심
11	SET_START_DATE	투승시작시간
12	SET_END_DATE	투승마무리시간
13	DEPTH_END_SET	어구양승해역수심
14	HAUL_START_DATE	양승시작시간
15	HAUL_END_DATE	양승마무리시간
16	LONGLINE_CODE	어구종류
17	LINE_LENGTH	어구길이
18	HOOK_SIZE	낚시바늘크기
19	HOOK_COUNT	낚시바늘수
20	PERCENT_BAITED	미끼비율
21	BAIT_SPECIES_CODE	미끼
22	VME_INDICATOR_TOTAL_NUMBER_UNITS	VME 지표종 수
23	FishingLatitude	어획위치위도
24	FishingLongitude	어획위치경도
25	FishingDepthM	어획수심
26	SetTime	어구투승시간
27	HaulTime	어구양승시간
28	SoakTimeH	침적시간
29	CalculatedTotalCaughtKG	목표종어획량(이빨고기)
30	CalculatedTotalCaughtN	목표종어획개체수(이빨고기)
31	CalculatedTotalCaughtKG	부수어획종어획량(이빨고기 제외 전 어종)
32	CalculatedTotalCaughtN	부수어획종어획개체수(이빨고기 제외 전 어종)
33	CPUE(kg/hook)	단위노력당어획량(이빨고기)
34	FISHING_CODE	어획형태 (C-상업조사, R-과학조사)
35	Note1	비고1 - 자료 상태표시
36	Note5	비고5 - 어획시간 없음
37	Note6	비고6 - 어획자료 없음

[표 3-2] 변수 요약

구분	변수
기초정보 관련 변수 : 9개	기록번호, 해역, 소해구, 어기, 조사년, 조사월, 국가, 선박번호, 선박명
시간 및 어구 관련 변수 : 19개	어구투승해역수심, 투승시작시간, 어구양승해역수심, 어구종류, 어구길이, 낚시바늘수, 미끼, 어획수심, 어구투승시간, 어구양승시간, 침적시간 등
어획량 관련 변수 : 5개	목표종어획량, 부수어획종어획량, CPUE(단위노력당어획량) 등
수집 목적 관련 변수 : 1개	어획형태(상업조사, 과학조사)
부가 설명 변수 : 3개	자료상태, 어획시간 유무, 어획자료 유무
총합계	37개

3.2.2 데이터 전처리

아래와 같은 데이터 전처리 과정을 통해 중복되거나 잡음이 있을 수 있는 데이터를 제거하여 7,738개 중에서 7,394개의 유효한 데이터가 결정되었다. 그에 대한 결과를 (연도*해역)별 교차표로 표현한 것이 [표 3-3]과 같다.

1) 자료의 상태가 Best Quality인 것만 선택

2) CPUE가 0이 아닌 자료만 선택

3) 목표종 어획개체수가 0이 아닌 자료만 선택

[표 3-3] (연도*해역)별 유효 데이터

	486A	486B	486C	486D	486E	486G	4861C	4861E	4861G	4862A	4862E	4881B	4881C	4881H	4881I	4881J	4881K	4881L	4882A	4882C	4882D	4882E	4882F	4882G	4882H	합계
2008/09	0	0	0	0	37	81	66	35	92	5	33	49	113	162	266	35	313	4	0	0	24	24	98	0	361	1798
2009/10	92	0	0	40	18	144	58	13	72	5	47	1	74	413	342	113	59	26	0	0	0	0	0	9	191	1717
2010/11	39	32	50	0	0	95	98	4	99	0	21	61	67	236	281	55	228	25	8	4	53	12	44	0	306	1818
2011/12	6	0	0	35	78	150	91	42	37	20	73	90	99	203	55	140	349	45	0	0	0	0	26	0	318	1857
2012/13	0	0	0	0	0	39	0	0	0	0	0	0	23	12	69	0	23	0	0	0	0	0	0	0	38	204
합계	137	32	50	75	133	509	313	94	300	30	174	201	376	1026	1013	343	972	100	8	4	77	36	168	9	1214	7394

3.2.3 최종 변수 선택

37개의 변수 중 본 논문의 분석에서 활용되는 것으로 기초정보 관련 변수와 어획량 관련 변수에서 각 1개의 변수를 선택하였고 그 결과가 [표 3-4]에 나와있다.

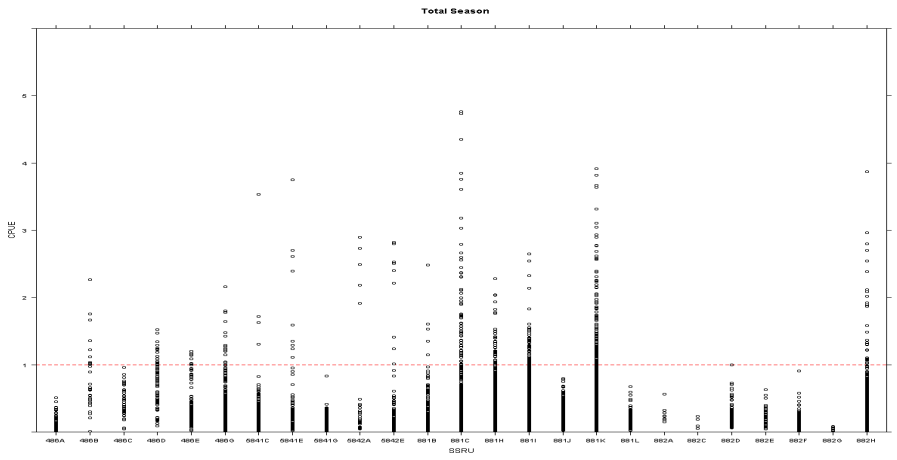
[표 3-4] 최종 변수 선택

구분	변수
기초정보 관련 변수 : 1개	소해구
어획량 관련 변수 : 1개	CPUE(단위노력당어획량)

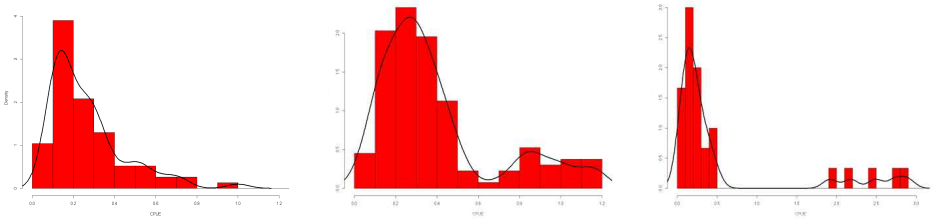
3.3 시각화 데이터 분석

전처리 과정이 완료된 데이터를 이용하여 가지고 있는 정보의 특성을 쉽게 이해하고 본 논문의 목적인 혼합 극단분포 모형을 적용시킬 수 있는 해역이 존재하는지 알아볼 수 있도록 그래프를 활용하였다.

기초정보 관련 변수 중 소해구는 총 25개의 해역이 존재한다. 시각적으로 해역별로 하나의 분포로 적합이 될 것인지 아니면 혼합 분포로 적합이 되는지 판단하기 위해 dot plot으로 표현한 것이 [그림 3-2]와 같다.



[그림 3-2] 해역별 CPUE 분포



<882D> <486E> <5842A>

[그림 3-3] 단일 또는 혼합 극단분포 형태 해역

해역 중에는 단일 또는 혼합 극단분포 형태를 가지는 해역과 그렇지 않은 해역들이 섞여 있다. [부록 3]에 25개 해역의 분포 형태에 대해 histogram과 density plot으로 정리하였다.

위의 [그림 3-3]은 25개의 해역 중 데이터가 단일 또는 혼합 극단분포 형태를 가지는 3개 해역을 탐색한 모습이다. 882D 해역은 단일 극단분포 형태, 486E 해역과 5842A 해역은 혼합 극단분포 형태를 가지는 것을 확인할 수 있다.

3.4 일반 혼합 분포 모형

우선 전체 데이터 중에서 대부분의 데이터가 확률분포 $f_a(x)$을 따르고, 일부의 이상점들은 확률분포 $f_b(x)$를 따른다면 전체 데이터의 확률분포 $f(x)$는 다음과 같은 혼합 확률분포로 표시될 수 있다.

$$f(x) = (1-\lambda) f_a(x) + \lambda f_b(x)$$

혼합 확률분포 모형에서 주어진 표본 데이터 $x_1, x_2, \cdots, x_n$에 대한 가능도 함수는 다음과 같이 표현할 수 있다.

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

본 논문에서는 아래와 같은 네 종류의 확률분포 혼합 모형을 고려하여 이상점 탐색을 연구하였다.

1) Weibull $\alpha : shape \quad \beta : scale$

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

2) Normal

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

3) Gamma $\alpha : shape \quad \beta : scale$

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

4) Log-Normal $\mu : meanlog \quad \sigma : sdlog$

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

3.5 EM 알고리즘을 이용한 혼합 극단분포 모형 탐색 방법

3.5.1 EM 알고리즘

혼합 확률분포 모형의 경우 추정해야 하는 모수의 수가 많아지므로 최대가능도 추정법으로 모수를 추정하기가 쉽지 않아 EM 알고리즘을 이용한다.

EM 알고리즘은 추정해야하는 모수들에 대한 초기값을 설정하는 것으로 시작한다. 그리고 각 데이터가 각각의 확률분포에 속할 확률을 계산한 후 최대가능도 추정법으로 모수 추정을 다시 한다. 이와 같이 재 추정된 모수의 값이 거의 변하지 않을 때까지 반복한다.

[표 3-5]에 EM 알고리즘 절차에 대하여 정리하였다.

[표 3-5] 혼합 모형 EM 알고리즘 절차

모수 초기값 설정
평균 혹은 중심을 나타내는 k개의 집단을 임의로 선정하고 추가 될 모수들에 대해 랜덤 또는 다양한 방법으로 초기값 추정

↓

Expectation step
추정된 모수를 이용하여 각 데이터가 각 확률분포에 속할 확률을 계산

↓

Maximization step
E step에서 계산된 확률을 이용하여 최대가능도 추정법으로 확률분포 모형의 모수를 다시 추정

↓

repeat
모수의 변화가 없을 때까지 반복하여 최종 모수 선택

3.5.2 혼합 극단분포 모형 이상점 탐색 절차

정상적인 데이터의 경우 두 분포에 나올 가능성이 모두 있다. 예를 들면, 혼합 정규분포일 경우 $f_a(x)$에 나올 가능성이 70%라고 한다면 $f_b(x)$에 나올 가능성은 30%가 된다.

그러나 이상점 탐색을 위해 혼합 분포모형을 활용할 경우에는 한쪽 분포에 데이터가 관측되면 나머지 다른 한쪽은 데이터가 관측될 가능성이 거의 없다고 생각할 수 있다.

이제 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.

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

계산의 용이를 위해서 위 식에 대하여 로그 가능도 함수를 구하게 되면 아래와 같이 표현할 수 있다.

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

전체 데이터를 정상적 데이터로 하였을 때의 로그 가능도함수 값과 각 데이터를 이상점으로 간주하고 혼합 분포 모형의 로그 가능도함수 값을 계산하여 비교하면 이상점을 구별할 수 있다.

위의 식에서 λ는 이상 데이터의 비율로 0과 1사이의 값이다.

혼합 극단분포 모형의 이상점 탐색은 [표 3-6]와 같은 단계로 진행할 수 있다.

[표 3-6] 혼합 극단분포 모형 이상점 탐색 절차

확률분포의 분할
정상적 데이터의 확률분포 $f_a(x)$와 이상 데이터의 확률분포 $f_b(x)$를 가정

↓

초기 로그 가능도 값 계산
모든 데이터를 정상적인 데이터 집합 A에 속한다고 가정하고 초기 로그 가능도함수 값 L_0 계산

↓

이상 데이터로 가정시 로그 가능도 값 계산
각 데이터를 하나씩 이상 데이터 집합 B로 옮기고 로그 가능도함수 값 L_i 계산

↓

이상점 분류
이상 데이터 가정시 로그 가능도 값 L_i와 초기로그 가능도 값 L_0를 비교했을 시 매우 차이가 크면 이상점으로 분류

3.5.3 혼합 극단분포 모형 로그 가능도함수

혼합 극단분포모형의 로그 가능도 값을 계산할 때에는 아래와 같은 식을 고려할 수 있다.

[표 3-7] 16가지 혼합 확률분포 모형의 로그 가능도 함수

No	$f_a(x)$	$f_b(x)$	로그 가능도 함수
1	Weibull	Weibull	$n_a\ln(1-\lambda)+n_a\ln\alpha_1+(\alpha_1-1)\sum_{x_i\in A}\ln x_i-n_a\alpha_1\ln\beta_1-\sum_{x_i\in A}(\frac{x_i}{\beta_1})^{\alpha_1}$ + $n_b\ln\lambda+n_b\ln\alpha_2+(\alpha_2-1)\sum_{x_i\in B}\ln x_i-n_b\alpha_2\ln\beta_2-\sum_{x_i\in B}(\frac{x_i}{\beta_2})^{\alpha_2}$
2	Normal	Normal	$n_a\ln(1-\lambda)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma_1-(\frac{1}{2\sigma_1^2})\sum_{x_i\in A}(x_i-\mu_1)^2$ + $n_b\ln\lambda-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma_2-(\frac{1}{2\sigma_2^2})\sum_{x_i\in B}(x_i-\mu_2)^2$
3	Gamma	Gamma	$n_a\ln(1-\lambda)-n_a\ln(\Gamma(\alpha_1))-n_a\ln(\beta_1^{\alpha_1})+(\alpha_1-1)\sum_{x_i\in A}\ln(x_i)-\sum_{x_i\in A}(\frac{x_i}{\beta_1})$ + $n_b\ln\lambda-n_b\ln(\Gamma(\alpha_2))-n_b\ln(\beta_2^{\alpha_2})+(\alpha_2-1)\sum_{x_i\in B}\ln(x_i)-\sum_{x_i\in B}(\frac{x_i}{\beta_2})$
4	Log Normal	Log Normal	$n_a\ln(1-\lambda)-\sum_{x_i\in A}\ln(x_i)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma_1-\frac{\sum_{x_i\in A}(\ln x_i-\mu_1)^2}{2\sigma_1^2}$ + $n_b\ln\lambda-\sum_{x_i\in B}\ln(x_i)-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma_2-\frac{\sum_{x_i\in B}(\ln x_i-\mu_2)^2}{2\sigma_2^2}$
5	Weibull	Normal	$n_a\ln(1-\lambda)+n_a\ln\alpha+(\alpha-1)\sum_{x_i\in A}\ln x_i-n_a\alpha\ln\beta-\sum_{x_i\in A}(\frac{x_i}{\beta})^{\alpha}$ + $n_b\ln\lambda-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-(\frac{1}{2\sigma^2})\sum_{x_i\in B}(x_i-\mu)^2$
	Normal	Weibull	$n_a\ln(1-\lambda)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma_1-(\frac{1}{2\sigma_1^2})\sum_{x_i\in A}(x_i-\mu_1)^2$ + $n_b\ln\lambda+n_b\ln\alpha+(\alpha-1)\sum_{x_i\in B}\ln x_i-n_b\alpha\ln\beta-\sum_{x_i\in B}(\frac{x_i}{\beta})^{\alpha}$
6	Weibull	Gamma	$n_a\ln(1-\lambda)+n_a\ln\alpha_1+(\alpha_1-1)\sum_{x_i\in A}\ln x_i-n_a\alpha_1\ln\beta_1-\sum_{x_i\in A}(\frac{x_i}{\beta_1})^{\alpha_1}$ + $n_b\ln\lambda-n_b\ln(\Gamma(\alpha_2))-n_b\ln(\beta_2^{\alpha_2})+(\alpha_2-1)\sum_{x_i\in B}\ln(x_i)-\sum_{x_i\in B}(\frac{x_i}{\beta_2})$
	Gamma	Weibull	$n_a\ln(1-\lambda)-n_a\ln(\Gamma(\alpha_1))-n_a\ln(\beta_1^{\alpha_1})+(\alpha_1-1)\sum_{x_i\in A}\ln(x_i)-\sum_{x_i\in A}(\frac{x_i}{\beta_1})$ + $n_b\ln\lambda+n_b\ln\alpha_2+(\alpha_2-1)\sum_{x_i\in B}\ln x_i-n_b\alpha_2\ln\beta_2-\sum_{x_i\in B}(\frac{x_i}{\beta_2})^{\alpha_2}$

[표 3-7 (계속)] 16가지 혼합 확률분포 모형의 로그 가능도 함수

No	$f_a(x)$	$f_b(x)$	로그 가능도 함수
7	Weibull	Log Normal	$n_a\ln(1-\lambda)+n_a\ln\alpha+(\alpha-1)\sum_{x_i\in A}\ln x_i-n_a\alpha\ln\beta-\sum_{x_i\in A}(\frac{x_i}{\beta})^\alpha$ $+\ n_b\ln\lambda-\sum_{x_i\in B}\ln(x_i)-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-\frac{\sum_{x_i\in B}(\ln x_i-\mu)^2}{2\sigma^2}$
	Log Normal	Weibull	$n_a\ln(1-\lambda)-\sum_{x_i\in A}\ln(x_i)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma-\frac{\sum_{x_i\in A}(\ln x_i-\mu)^2}{2\sigma^2}$ $+\ n_b\ln\lambda+n_b\ln\alpha+(\alpha-1)\sum_{x_i\in B}\ln x_i-n_b\alpha\ln\beta-\sum_{x_i\in B}(\frac{x_i}{\beta})^\alpha$
8	Normal	Gamma	$n_a\ln(1-\lambda)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma-(\frac{1}{2\sigma^2})\sum_{x_i\in A}(x_i-\mu)^2$ $+\ n_b\ln\lambda-n_b\ln(\Gamma(\alpha))-n_b\ln(\beta^\alpha)+(\alpha-1)\sum_{x_i\in B}\ln(x_i)-\sum_{x_i\in B}(\frac{x_i}{\beta})$
	Gamma	Normal	$n_a\ln(1-\lambda)-n_a\ln(\Gamma(\alpha))-n_a\ln(\beta^\alpha)+(\alpha-1)\sum_{x_i\in A}\ln(x_i)-\sum_{x_i\in A}(\frac{x_i}{\beta})$ $+\ n_b\ln\lambda-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-(\frac{1}{2\sigma^2})\sum_{x_i\in B}(x_i-\mu)^2$
9	Normal	Log Normal	$n_a\ln(1-\lambda)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma_1-(\frac{1}{2\sigma_1^2})\sum_{x_i\in A}(x_i-\mu_1)^2$ $+\ n_b\ln\lambda-\sum_{x_i\in B}\ln(x_i)-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-\frac{\sum_{x_i\in B}(\ln x_i-\mu)^2}{2\sigma^2}$
	Log Normal	Normal	$n_a\ln(1-\lambda)-\sum_{x_i\in A}\ln(x_i)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma-\frac{\sum_{x_i\in A}(\ln x_i-\mu)^2}{2\sigma^2}$ $+\ n_b\ln\lambda-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-(\frac{1}{2\sigma^2})\sum_{x_i\in B}(x_i-\mu)^2$
10	Gamma	Log Normal	$n_a\ln(1-\lambda)-n_a\ln(\Gamma(\alpha))-n_a\ln(\beta^\alpha)+(\alpha-1)\sum_{x_i\in A}\ln(x_i)-\sum_{x_i\in A}(\frac{x_i}{\beta})$ $+\ n_b\ln\lambda-\sum_{x_i\in B}\ln(x_i)-(\frac{n_b}{2})\ln2\pi-n_b\ln\sigma-\frac{\sum_{x_i\in B}(\ln x_i-\mu)^2}{2\sigma^2}$
	Log Normal	Gamma	$n_a\ln(1-\lambda)-\sum_{x_i\in A}\ln(x_i)-(\frac{n_a}{2})\ln2\pi-n_a\ln\sigma-\frac{\sum_{x_i\in A}(\ln x_i-\mu)^2}{2\sigma^2}$ $+\ n_b\ln\lambda-n_b\ln(\Gamma(\alpha))-n_b\ln(\beta^\alpha)+(\alpha-1)\sum_{x_i\in B}\ln(x_i)-\sum_{x_i\in B}(\frac{x_i}{\beta})$

EM 알고리즘을 실행하게 되면 $f_a(x)$를 좌측의 분포, $f_b(x)$를 우측의 분포로 강제로 설정하는 것이 아니라 데이터에 가장 잘 맞게 $f_a(x)$와 $f_b(x)$가 자동으로 할당되게 된다.

즉 위에서 제시된 16가지 모형 중 5-10번 까지는 두 개의 모형 중 가능도 값이 큰 모형으로 추정된다. 결과적으로 10개의 모형이 얻어진다.

n_a는 정상적 데이터의 수이다. 초기 로그 가능도함수 값의 경우에는 이상 데이터가 없이 모든 데이터가 정상적 데이터라고 가정하므로 위의 식에서 n_a가 포함되는 부분만 계산하면 된다.

관측데이터가 이상 데이터의 확률분포 $f_b(x)$를 따른다고 한다면 위의 식을 이용하여 혼합 로그 가능도함수 값을 계산할 수 있다.

다시말해 데이터가 100개가 있다면 초기 로그 가능도함수 값은 100개가 모두 정상적 확률분포를 따른다고 가정하고 계산을 하면 된다.

i번째 값이 이상점인지 판단을 하기 위해서는 100개 중 99개는 정상적 데이터의 확률분포를 따른다고 가정을 하고 로그 가능도함수 값을 계산하고, 1개는 이상 데이터의 확률분포를 따른다고 가정한 로그 가능도함수 값을 계산하여 더해주면 된다.

제 4 장 모형의 실험

4.1 단일 극단분포

4.1.1 단일 극단분포 모형의 모수 추정

데이터가 단일 극단분포 형태를 가질 경우 MLE를 이용하여 $f_a(x)$에 해당되는 분포 모형에 대한 모수 추정을 할 수 있다. [표 4-1]은 882D 해역에서 적합 가능한 분포 모형에 대해 모수 추정한 결과이다.

[표 4-1] 단일 극단분포 해역 모수 추정 값

해역	분포 모형	모수	추정 값
882D	Weibull	shape	1.6062
		scale	0.2970
	Normal	mean	0.2638
		sd	0.1801
	Gamma	shape	2.6639
		scale	0.0990
	LogNormal	meanlog	-1.5318
		sdlog	0.6227

4.1.2 χ^2 적합성 검정

통계분석을 하다보면 자주 당면하게 되는 의문은 관측된 데이터가 어떤 확률분포에 적합하고 있는가 하는 문제이다. 이와 같은 경우의 검정이 적합성 검정(the goodness-of-fit test)이다. 이 검정을 할 때 χ^2분포가 사용되기 때문에 χ^2 적합성 검정이라 한다. 882D 해역에 적합 가능한 분포 모형들에 관해 χ^2 적합성 검정 한 결과가 [표 4-2]와 같다.

[표 4-2] 단일 극단분포 해역 χ^2 적합성 검정 결과

해역	분포모형	p-value
882D	Weibull	0.0739
	Normal	0.0000
	Gamma	0.2620
	LogNormal	0.8529

그 결과 882D 해역에서는 LogNormal 모형이 가장 잘 적합되는 것으로 나타났다.

4.1.3 혼합 극단분포 모형 추정

단일 극단분포 형태의 데이터를 혼합 극단분포 모형으로 적합하기 위해 $f_b(x)$에 해당되는 분포를 가정할 필요가 있다.

CPUE가 1 이상인 데이터가 현재 없으므로 향후 이상 데이터의 분포 $f_b(x)$가 Weibull 분포를 따른다고 가정을 하고 χ^2 적합성 검정 결과 가장 잘 적합한 LogNormal 모형과 결합하여 LogNormal-Weibull 혼합 극단분포 모형을 만들었다. 이때 lambda는 0.15로 가정하였다.

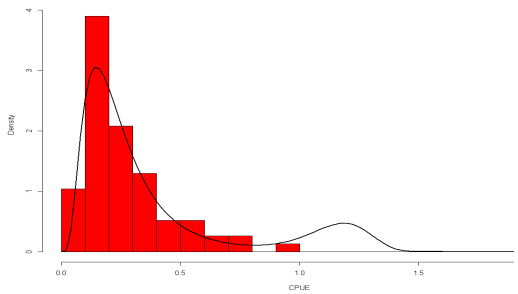
혼합 극단분포 모형의 모수에 대한 추정 값이 [표 4-3]에 있다.

[표 4-3] 단일 극단분포 해역의 혼합 극단분포 모형 모수 추정 값

해역	혼합 모형	모수	추정 값	모수	추정 값	lambda
882D	LogNormal-Weibull	meanlog	-1.5318	shape	10.0	0.15
		sdlog	0.6227	scale	1.2	

4.1.4 모형 검증

882D 해역의 데이터를 히스토그램으로 표현한 후 추정된 혼합 극단분포 모형을 겹쳐 표현한 것이 [그림 4-1]이다.



<882D>

[그림 4-1] 혼합 극단분포 모형

추정된 혼합 극단분포 모형의 이상점 탐색률을 알아보기 위해 기존 데이터에 CPUE가 1보다 큰 데이터 10개를 추가하였다.

모든 데이터를 정상 데이터라고 가정했을 시의 초기 로그 가능도 값과 각 관측값들이 이상점이라고 가정했을 시의 혼합 극단분포 모형의 로그 가능도 값을 계산하여 두 값의 차이가 5% 이상 크게 나타나는 경우를 정리한 결과가 [표 4-4]와 같다.

[표 4-4] 단일 극단분포 해역의 로그 가능도 비교

해역	혼합 모형	초기 로그 가능도	혼합 모형 로그 가능도	CPUE
882D	LogNormal - Weibull	13.2749	322.7210	2.1231
			174.4840	1.9847
			24.7897	1.4579
			19.8124	1.3391
			17.0818	1.2267
			16.0980	1.1708
			15.3069	1.1187
			15.3032	1.1184
			14.8386	1.0850
			14.5166	1.0608

882D 해역에서는 모든 관측데이터를 정상데이터로 가정했을 시 초기 로그 가능도 값과 각 관측값을 이상데이터로 가정했을 시 혼합 극단모형의 로그 가능도 값 차이가 5% 이상 크게 나타나기 시작할 때 CPUE 값이 1.0608로 나타났다.

예상했던 이상점(CPUE가 1 이상인 값)들이 초기 로그 가능도 값과 혼합 분포모형의 로그 가능도 값의 차이를 어떻게 설정함에 따라 실제로 이상점으로 판단되는 비율이 변화하는지 파악하기 위하여 탐색률을 비교한 결과가 [표 4-5]과 같다.

[표 4-5] 단일 극단분포 해역 혼합 극단분포 모형의 탐색률 비교

해역	로그 가능도 값 차이	예상 이상점 (CPUE ≥1)	실제 탐색된 이상점	탐색률
882D	1%	11	11	100%
	5%		10	90.9%
	10%		9	81.8%

초기 로그 가능도 값과 혼합 극단분포 모형의 로그 가능도 값 차이가 1%일 때 탐색률이 가장 좋고, 5%일 때도 90% 이상으로 높게 나타났다.

4.2 혼합 극단분포

4.2.1 혼합 극단분포 모형의 모수 추정

EM 알고리즘을 이용하여 486E와 5842A 해역의 혼합 분포모형에 대하여 모수 추정 한 결과가 [표 4-6], [표 4-7]과 같다.

[표 4-6] EM 알고리즘을 이용한 모수추정 결과(486E 해역)

해역	혼합모형	모수	추정값	모수	추정값	lambda
486E	Weibull-Weibull	shape1	2.3355	shape2	6.8264	0.1894
		scale1	0.3048	scale2	1.0022	
	Weibull-Normal	shape	2.3232	mean	0.9455	0.1855
		scale	0.3062	sd	0.1451	
	Weibull-LogNormal	shape	2.3119	meanlog	-0.0640	0.1839
		scale	0.3070	sdlog	0.1497	
	Normal-Normal	mean1	0.2705	mean2	0.9418	0.1880
		sd1	0.1221	sd2	0.1480	
	Normal-LogNormal	mean	0.2710	meanlog	-0.0684	0.1865
		sd	0.1227	sdlog	0.1535	
	Gamma-Weibull	shape1	3.4952	shape2	7.3226	0.1730
		scale1	0.0802	scale2	1.0165	
	Gamma-Normal	shape	3.4968	mean	0.9597	0.1709
		scale	0.0803	sd	0.1385	
	Gamma-Gamma	shape1	3.5015	shape2	49.5797	0.1706
		scale1	0.0802	scale2	0.0194	
	Gamma-LogNormal	shape	3.5000	meanlog	-0.0495	0.1704
		scale	0.0802	sdlog	0.1415	
	LogNormal-LogNormal	meanlog1	-1.3696	meanlog2	-0.0391	0.1376
		sdlog1	0.6460	sdlog2	0.1349	

[표 4-7] EM 알고리즘을 이용한 모수추정 결과(5842A 해역)

해역	혼합모형	모수	추정값	모수	추정값	lambda
5842A	Weibull-LogNormal	shape	1.7953	meanlog	0.8820	0.1667
		scale	0.2275	sdlog	0.1502	
	Weibull-Normal	shape	1.7954	mean	2.4426	0.1667
		scale	0.2275	sd	0.3566	
	Gamma-Gamma	shape1	2.8293	shape2	45.3337	0.1667
		scale1	0.0712	scale2	0.0539	
	Gamma-LogNormal	shape	2.8293	meanlog	0.8820	0.1667
		scale	0.0712	sdlog	0.1502	
	Gamma-Normal	shape	2.8292	mean	2.4427	0.1667
		scale	0.0712	sd	0.3566	
	Weibull-Weibull	shape1	1.7953	shape2	8.2258	0.1667
		scale1	0.2275	scale2	2.5960	
	Gamma-Weibull	shape1	2.8291	shape2	8.2255	0.1667
		scale1	0.0712	scale2	2.5960	
	LogNormal-LogNormal	meanlog1	-1.7893	meanlog2	0.8821	0.1666
		sdlog1	0.6373	sdlog2	0.1502	
	LogNormal-Normal	meanlog	-1.7893	mean	2.4430	0.1665
		sdlog	0.6373	sd	0.3565	
	Normal-Normal	mean1	0.2014	mean2	2.4427	0.1667
		sd1	0.1189	sd2	0.3566	

4.2.2 χ^2 적합성 검정

위에서 계산된 모수를 이용한 혼합 분포 모형에 대하여 χ^2 적합성 검정을 실시해 본 결과가 [표 4-8]과 같았다.

[표 4-8] χ^2 적합성 검정 결과

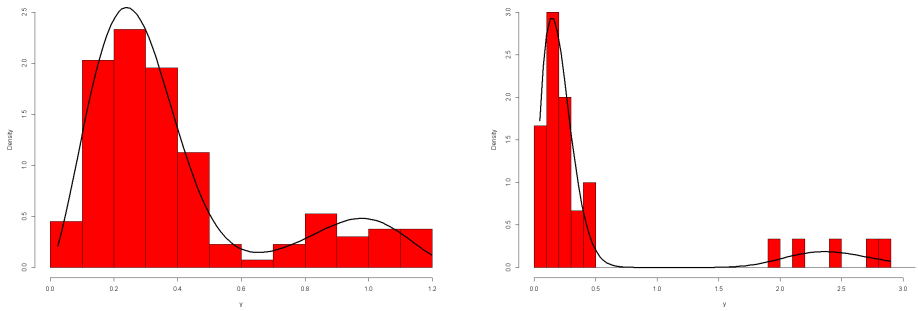
해역	혼합모형	p-value
486E	Weibull-Weibull	0.9335
	Weibull-Normal	0.9184
	Weibull-LogNormal	0.8900
	Normal-Normal	0.8412
	Normal-LogNormal	0.7908
	Gamma-Weibull	0.6847
	Gamma-Normal	0.6711
	Gamma-Gamma	0.6545
	Gamma-LogNormal	0.6376
	LogNormal-LogNormal	0.2179

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

χ^2 적합성 검정 결과 두 해역의 가장 적합한 혼합 극단분포 모형이 서로 다르게 나왔다. 486E 해역에서는 Weibull-Weibull 모형, 5842A 해역에서는 Weibull-LogNormal 모형이 χ^2 적합성 검정 결과 p-value가 가장 높게 나타났다.

4.2.3 모형 검증

486E와 5842A 해역의 데이터를 히스토그램으로 표현한 후 가장 적합하다고 판단된 혼합 극단분포 모형을 겹쳐 표현한 것이 [그림 4-2]이다.



<486E> <5842A>

[그림 4-2] 적합된 혼합 극단분포 모형

이상점 파악을 위해 각 해역의 데이터에 혼합 극단분포 모형을 적용시킨 후 모든 관측값들을 정상 데이터라고 가정했을 시의 초기 로그 가능도 값과 각 관측값들이 이상점이라고 가정했을 시의 로그 가능도 값을 계산한다.

두 값의 차이가 5%를 넘는 기준으로 한 이상점 탐색 결과가 [표 4-9]와 같다.

[표 4-9] 초기 로그 가능도와 혼합모형 로그 가능도 비교

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

486E 해역에서는 모든 관측데이터를 정상데이터로 가정했을 시 초기 로그 가능도 값과 각 관측값을 이상데이터로 가정했을 시 혼합 극단모형의 로그 가능도 값 차이가 5% 이상 크게 나타나기 시작할 때 CPUE 값이 1.0028로 나타났다. 5842A 해역에서는 로그 가능도 값 차이가 5% 이상 크게 나타날 때 CPUE 값이 1.9164였다.

예상했던 이상점(CPUE가 1 이상인 값)들이 초기 로그 가능도 값과 혼합 분포모형의 로그 가능도 값의 차이를 어떻게 설정함에 따라 실제로

이상점으로 판단되는 비율이 변화하는지 파악하기 위하여 탐색률을 비교한 결과가 [표 4-10]과 같다.

[표 4-10] 혼합 극단분포 모형의 탐색률 비교

해역	로그 가능도 값 차이	예상 이상점 (CPUE ≥1)	실제 탐색된 이상점	탐색률
486E	1%	10	24	240%
	5%		10	100%
	10%		0	0%
5842A	1%	5	5	100%
	5%		5	100%
	10%		4	80%

두 해역에서 초기 로그 가능도 값과 혼합 모형의 로그 가능도 값 차이가 5% 정도일 때 탐색률이 가장 좋은 것으로 나타났다.

제 5 장 결 론

본 논문은 과거의 데이터가 극단적인 혼합 분포모형을 이룰 때 이상점을 탐색해야 하는 경우 혼합 극단분포 모형을 이용하면 이상점 탐색률이 얼마나 좋아지는지 실험을 하기위한 알고리즘을 구현하였다. 만약 데이터의 분포가 혼합 분포 형태로 표현할 수 있다면 정상적 데이터의 분포와 이상 데이터의 분포를 모두 알고 있으므로 정상적 데이터에서 나오기 힘든 데이터일 경우에는 이상 데이터 분포에 속할 것이라고 생각하고 이상점으로 판단할 수 있을 것이다.

혼합 극단분포 모형을 활용하여 내년도 조업과 관련된 데이터가 들어오게 되면 어떠한 값이 이상점인지 아닌지 판명을 할 수 있을 것이다. 분포함수를 알고 있으므로 해당 해역의 데이터가 추가 된다면 추가된 데이터에 대해서 제일 처음에는 $f_a(x)$로 가정하고 로그 가능도함수 값을 계산하고 하나씩 $f_b(x)$로 이동시켜 혼합 극단분포 모형의 로그 가능도함수 값을 계산하여 차이를 가지고 비교하면 될 것이다.

이제 본 논문의 결과를 바탕으로 하나의 분포로 적합되는 다른 해역도 혼합 극단분포 모형으로 적합이 가능할 것이다.

CPUE가 대부분 1.0 미만인 데이터로 이루어진 해역은 주어진 데이터를 $f_a(x)$로 판단하고 로그 가능도함수 값을 계산할 수 있다. $f_b(x)$의 경우는 CPUE가 1.0 이상으로 나오는 데이터의 분포를 가정을 하여 혼합 극단분포 모형으로 생각해볼 수 있다.

이 경우 새로운 데이터가 추가되면 가정된 혼합 극단분포 모형을 활용하여 로그 가능도함수 값을 구할 수 있다. 예를 들어 새로운 데이터 50개가 들어왔을 때 가정된 혼합 극단분포 모형을 이용하여 50개가 $f_a(x)$에 속할 때 초기 로그 가능도함수 값을 구한다. 이후에 49개는 $f_a(x)$에

포함되고 1개는 $f_b(x)$에 포함된다고 여기고 혼합 극단분포 모형의 로그 가능도함수 값을 구하여 초기로그 가능도함수 값과 혼합 극단분포 모형의 로그 가능도함수 값의 차이가 크게 되면 이상값으로 판단할 수 있을 것이다.

따라서 내년부터는 두 개의 분포로 적합이 되는 해역뿐만 아니라 하나의 분포로 적합이 되는 해역에서도 혼합 극단분포 모형을 이용하여 이상점을 탐색하는 방법을 적용시키는 방안을 고려할 수 있을 것이다.

본 논문에서 혼합 극단분포 모형의 로그 가능도 값의 차이가 L_0의 5% 이상일 때 이상점 탐색을 잘 하는 것으로 나타났지만 이는 데이터에 좌우될 수 있어 향후 일반화 시키기 위한 연구가 더 필요하다.

참고문헌

[1] 이정진(2011), R, SAS, MS-SQL을 활용한 데이터마이닝, 자유아카데미

[2] 강창완, 강현철, 박우창, 승현우, 용환승, 이동희, 이성건, 이영섭, 진서훈, 최종후, 한상태(2007), 데이터 마이닝-개념과 기법, 사이플러스

[3] 용환승, 나연묵, 박종수, 승현우, 이민수, 이상준, 최린(2007), 데이터 마이닝, 인피니티북스

[4] 김홍준, 김진수(2013), 통계학 강의, 도서출판 명진

[5] Murray Aitkin and Granville Tunnicliffe Wilson(1980), Mixture Models, Outliers, and the EM Algorithm, Technometrics

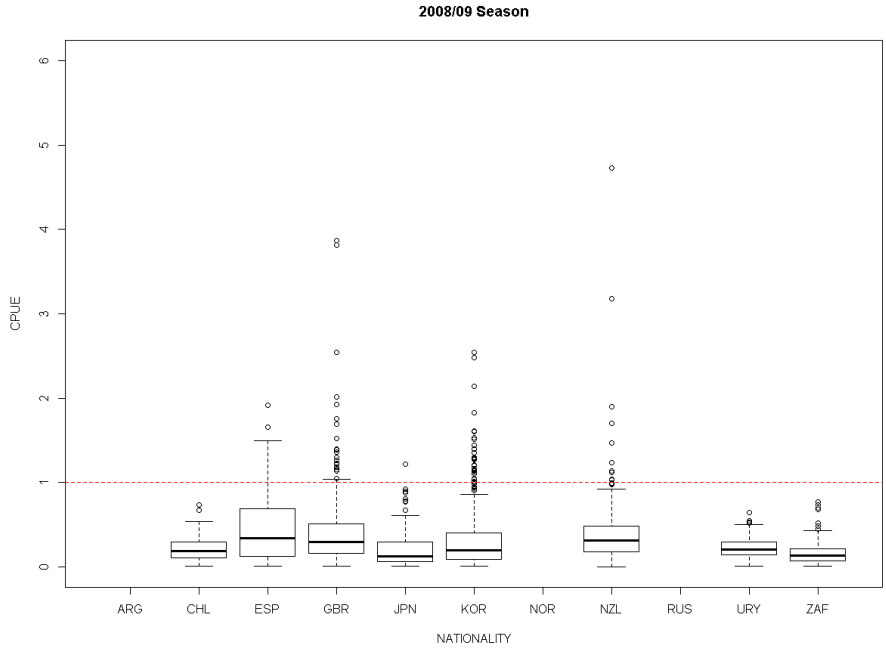
[6] 김승, 조남욱, 강석호(2010), 대용량 자료 분석을 위한 밀도기반 이상치 탐지, 한국경영과학회지

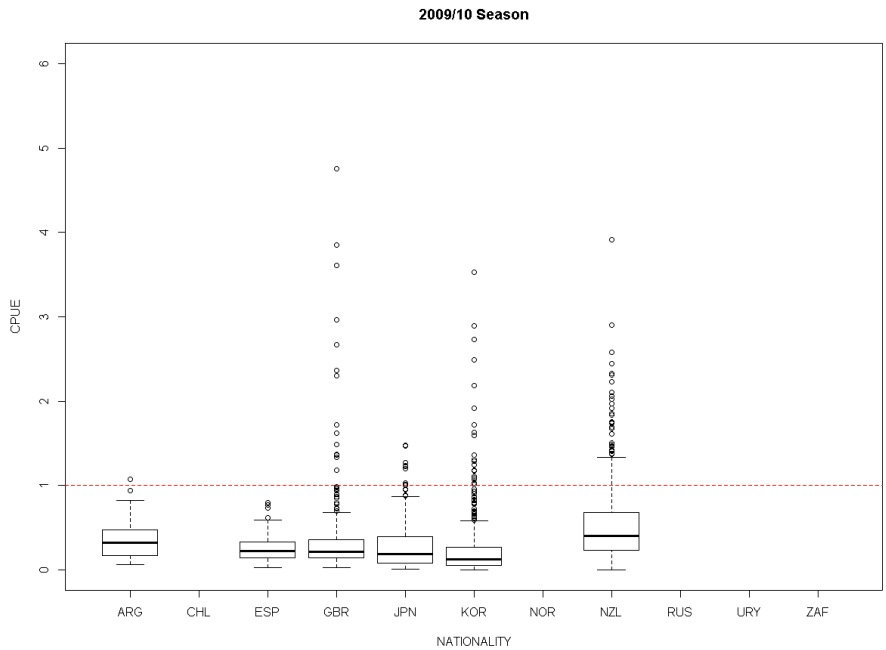
[7] 서한손, 윤민(2011) 서포트 벡터 기계를 이용한 이상치 진단, 한국통계학회논문집

[8] 송규문, 문지은, 박철용(2011), R을 이용한 이상점 탐지 알고리즘의 구현, 한국데이터정보과학회지

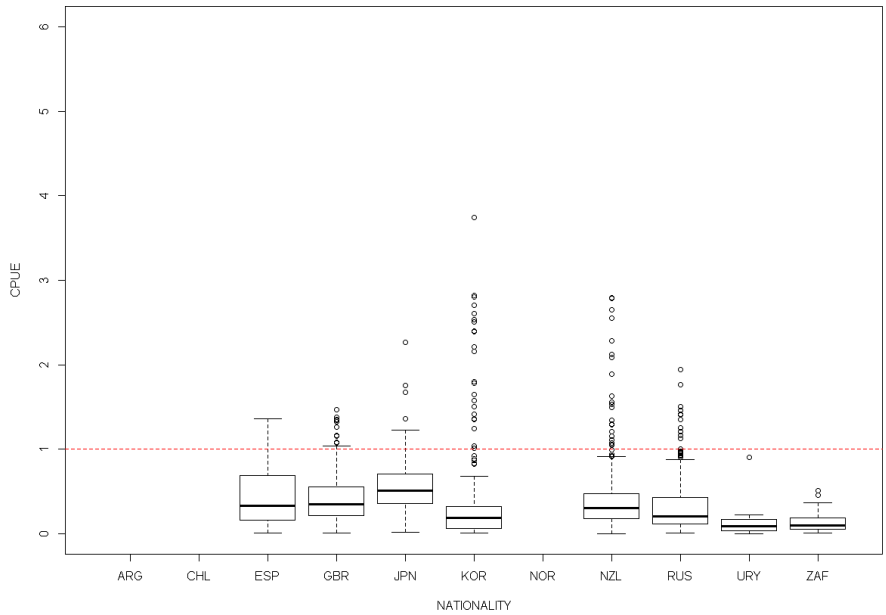
부 록

[부록 1] (연도*국가)별 CPUE의 box-whisker plot

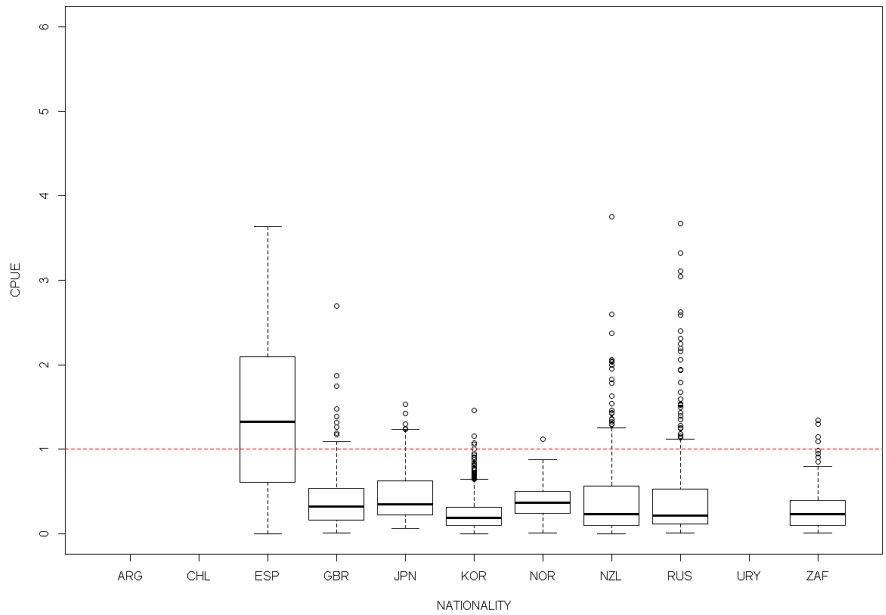


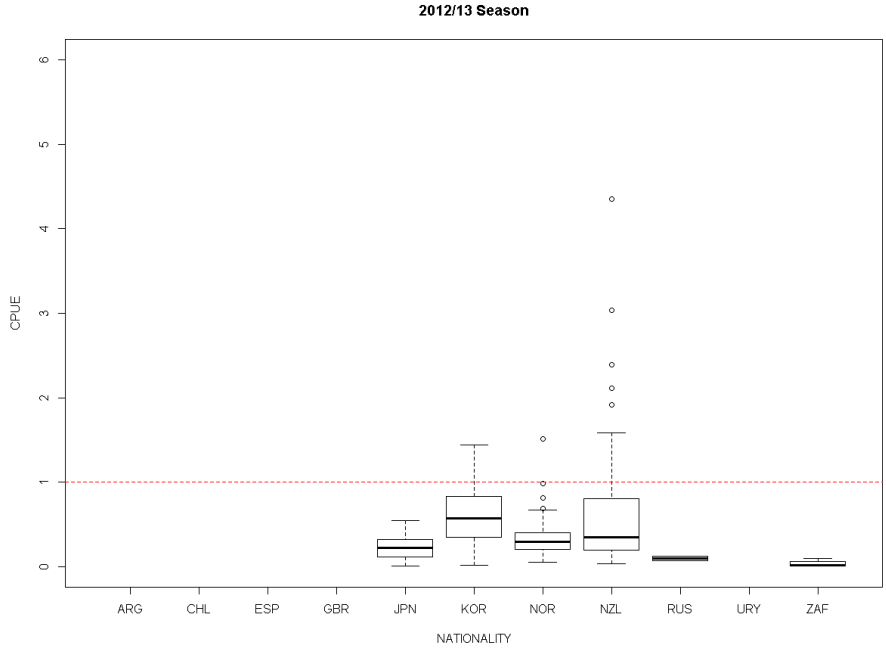


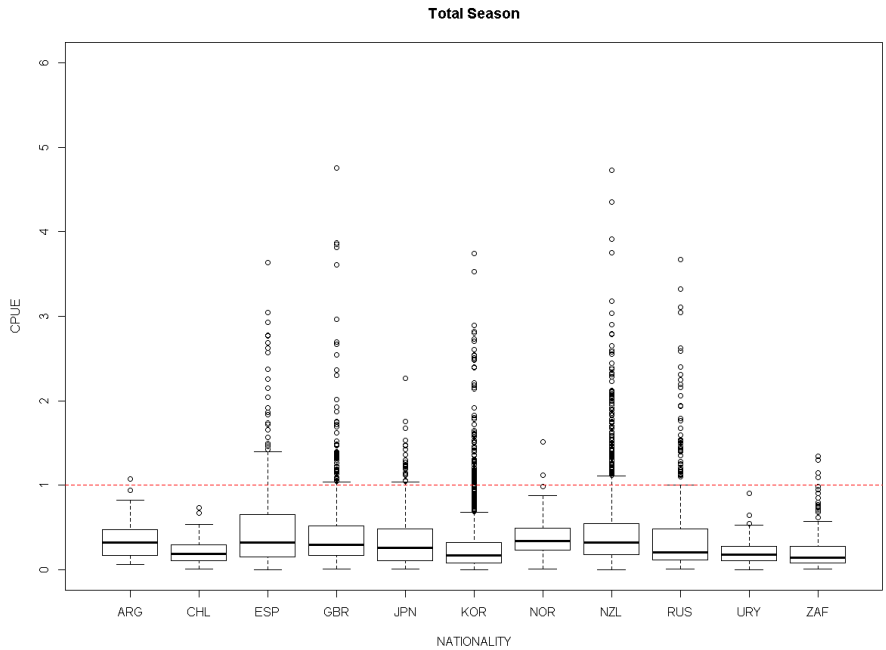
2010/11 Season



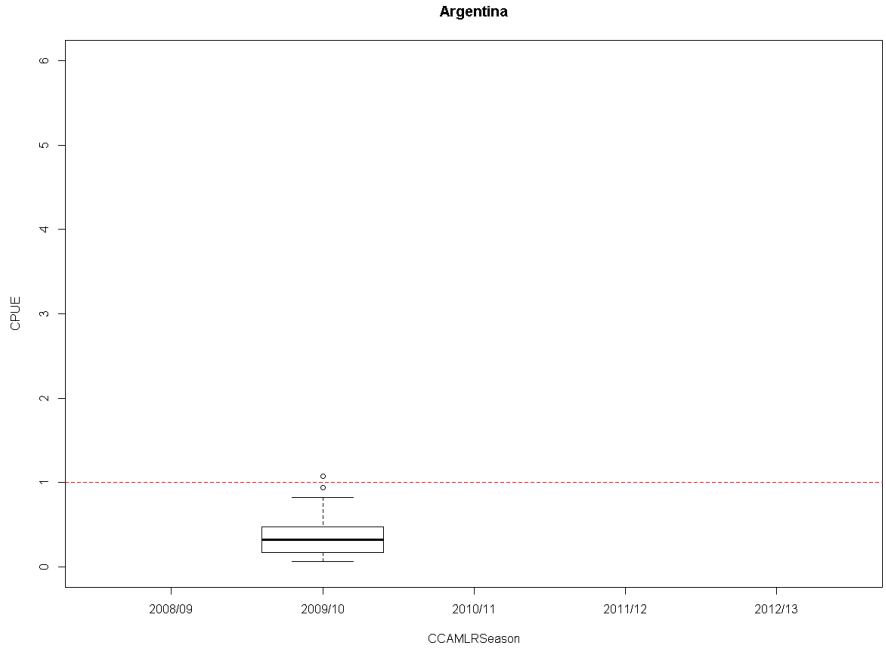
2011/12 Season

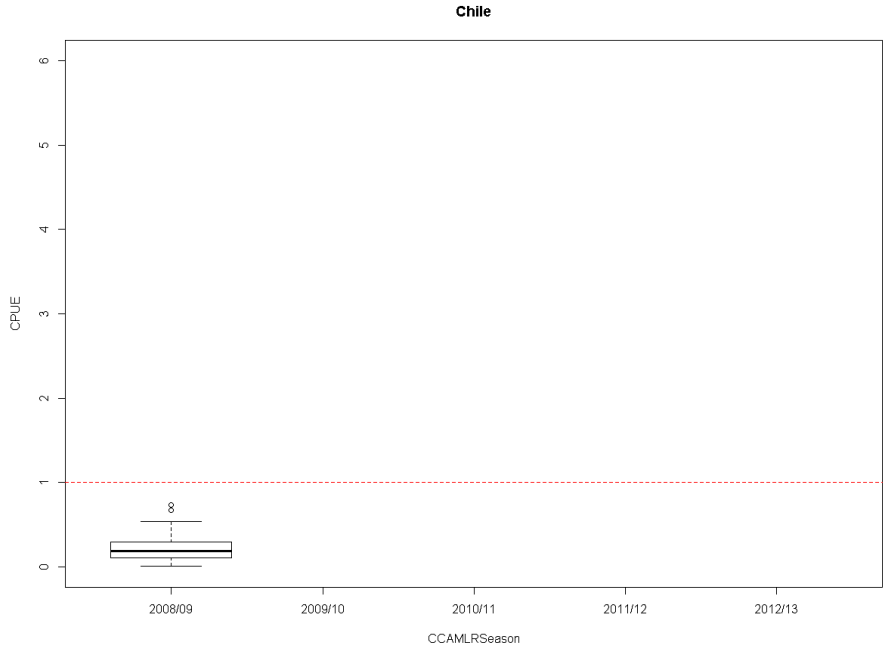


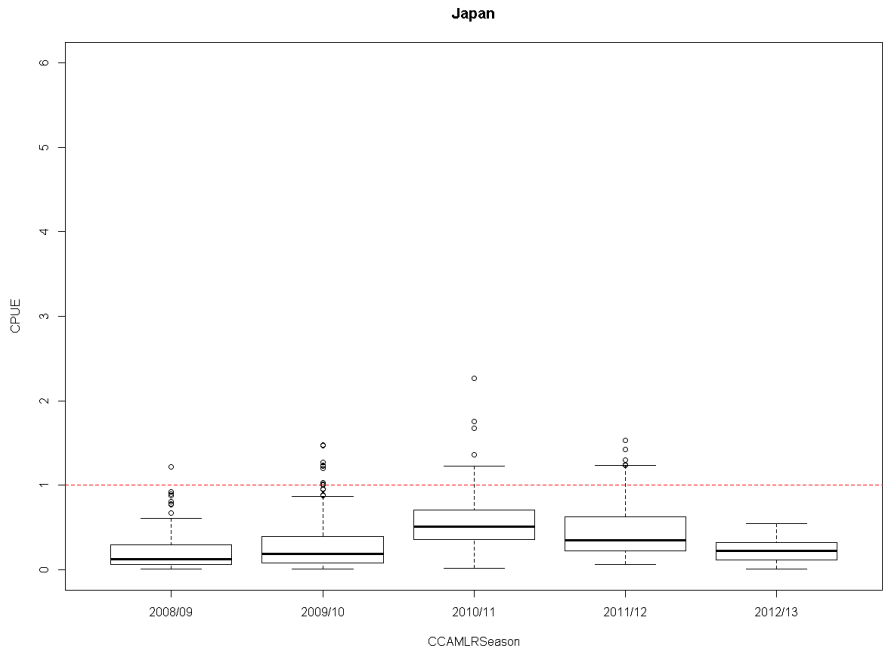

2012/13 Season
CPUE
0
1
2
3
4
5
6
ARG
CHL
ESP
GBR
JPN
KOR
NOR
NZL
RUS
URY
ZAF
NATIONALITY

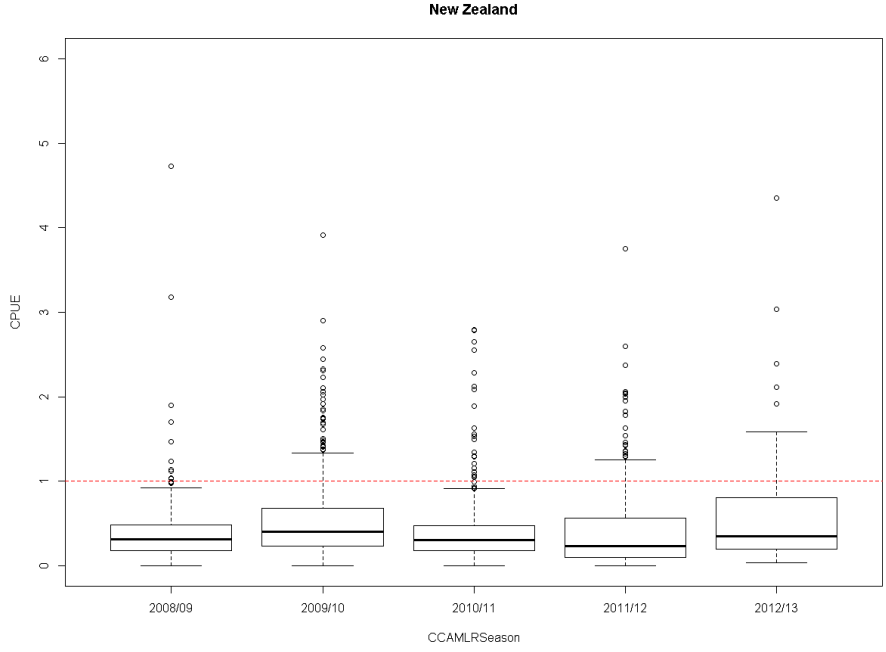

Total Season
CPUE
0
1
2
3
4
5
6
ARG
CHL
ESP
GBR
JPN
KOR
NOR
NZL
RUS
URY
ZAF
NATIONALITY

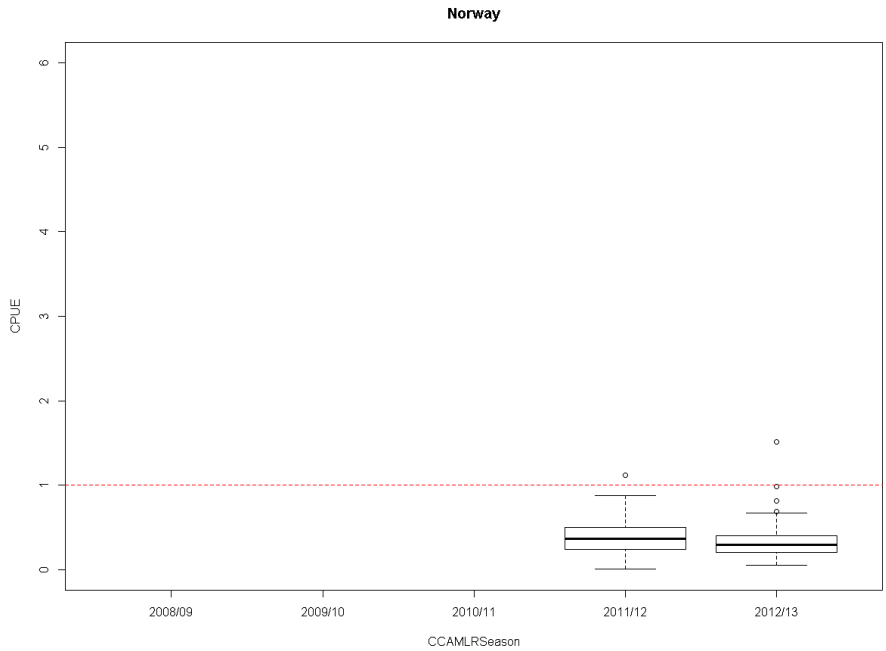
[부록 2] (국가*연도)별 CPUE의 box-whisker plot

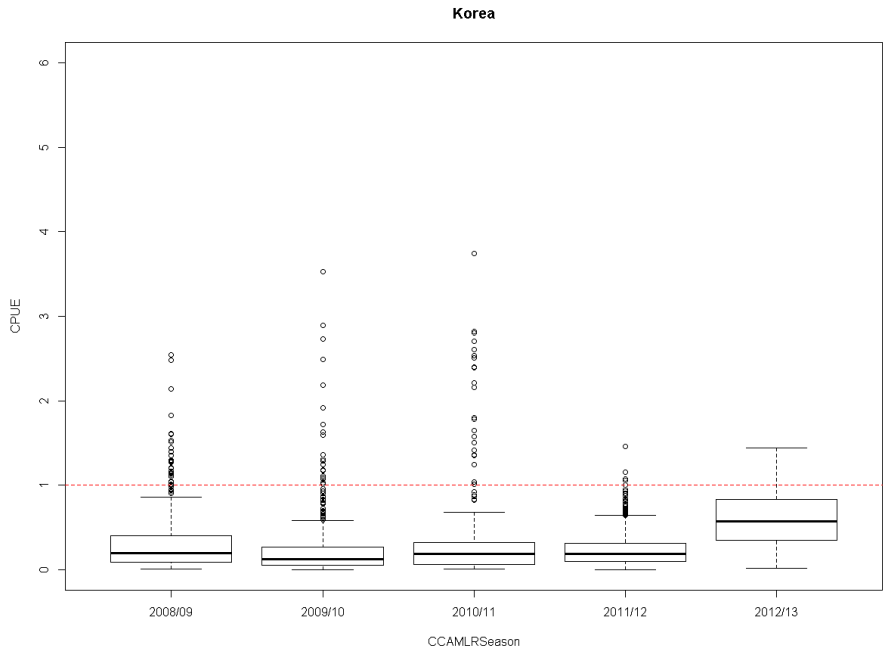


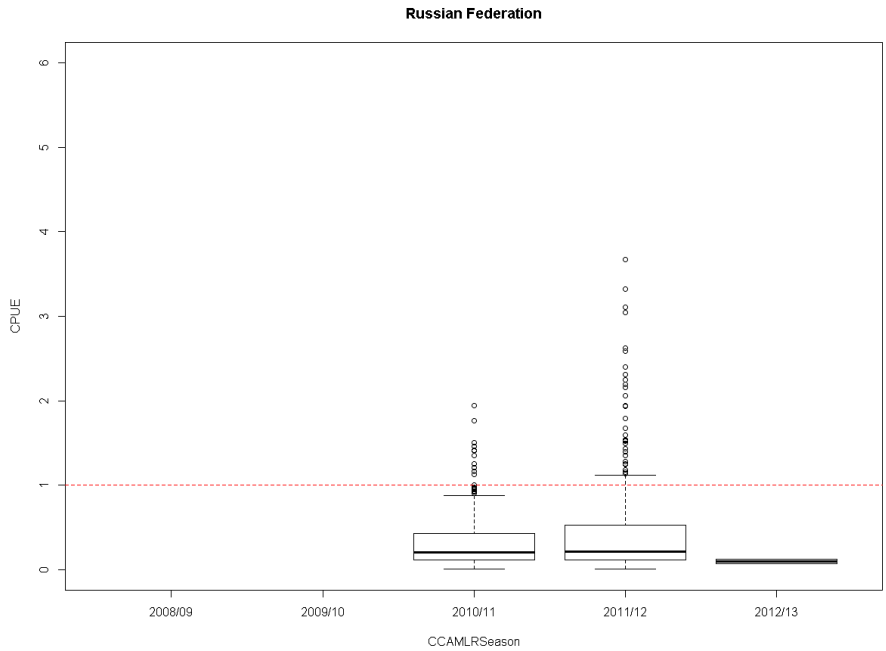


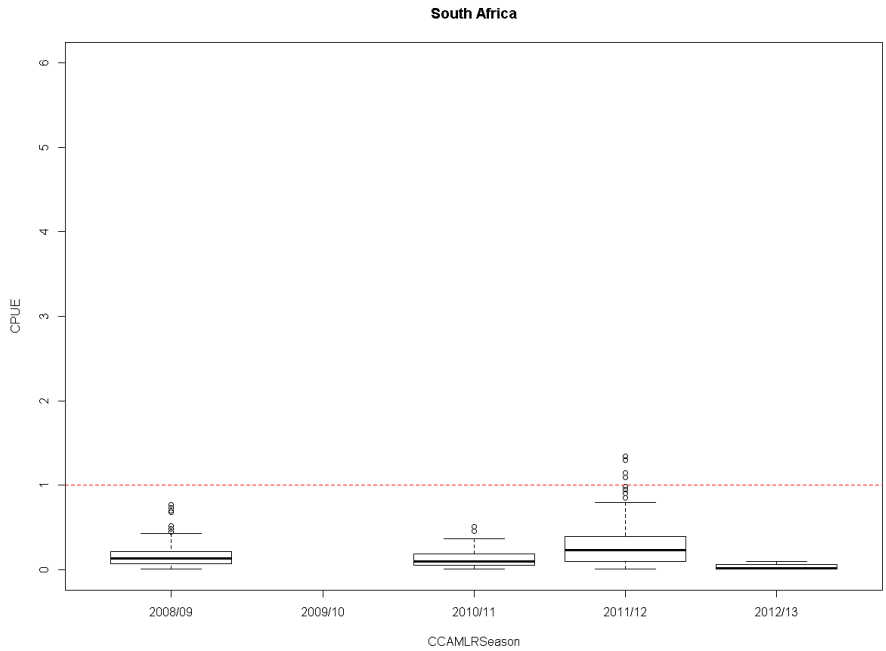

Japan
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

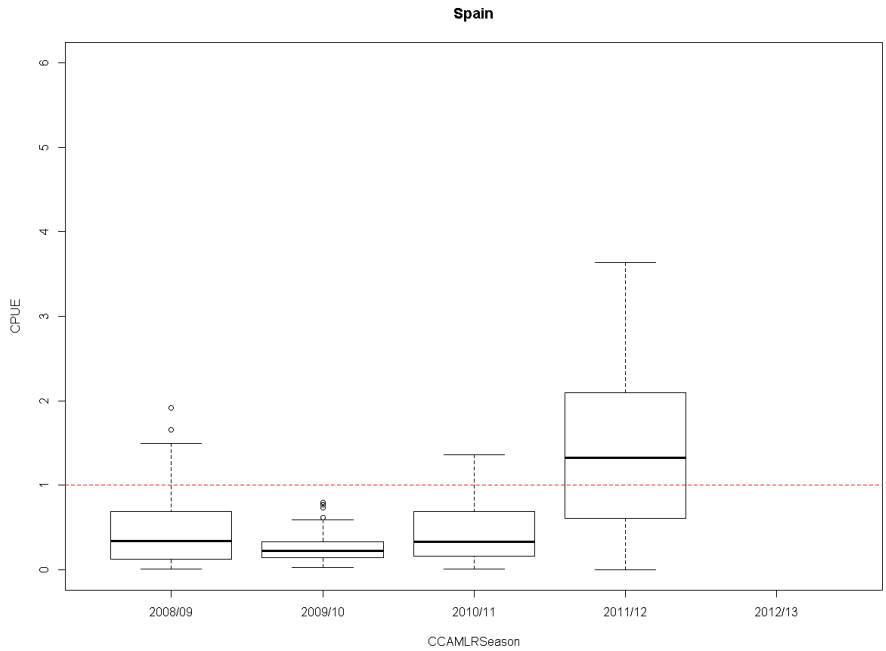

New Zealand
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

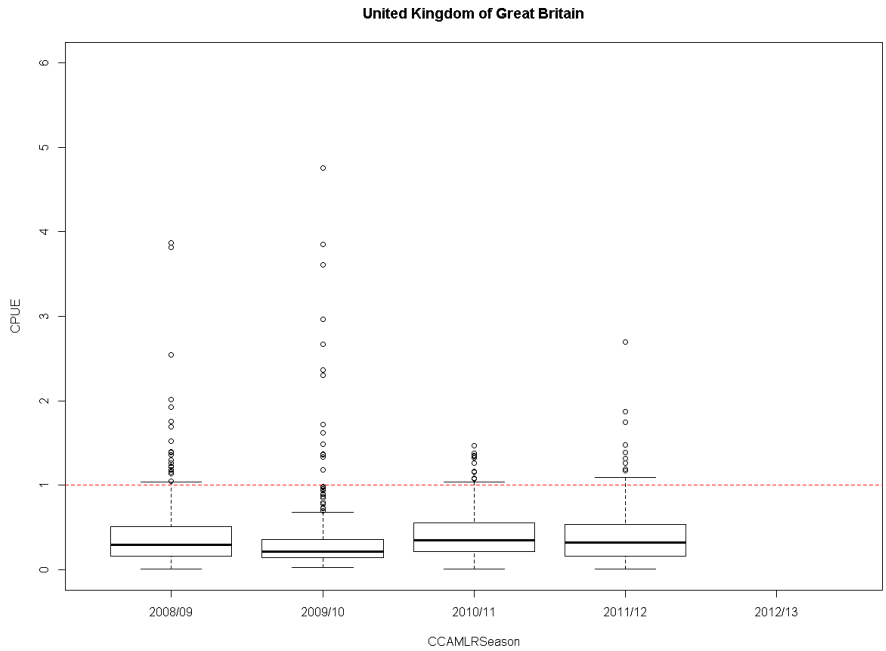

Norway
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

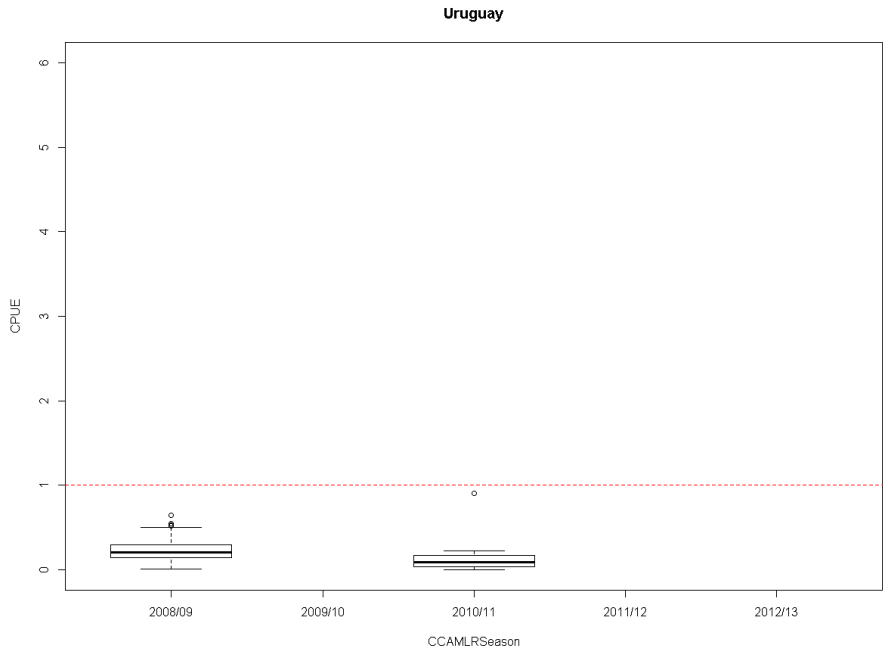

Korea
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason


Russian Federation
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

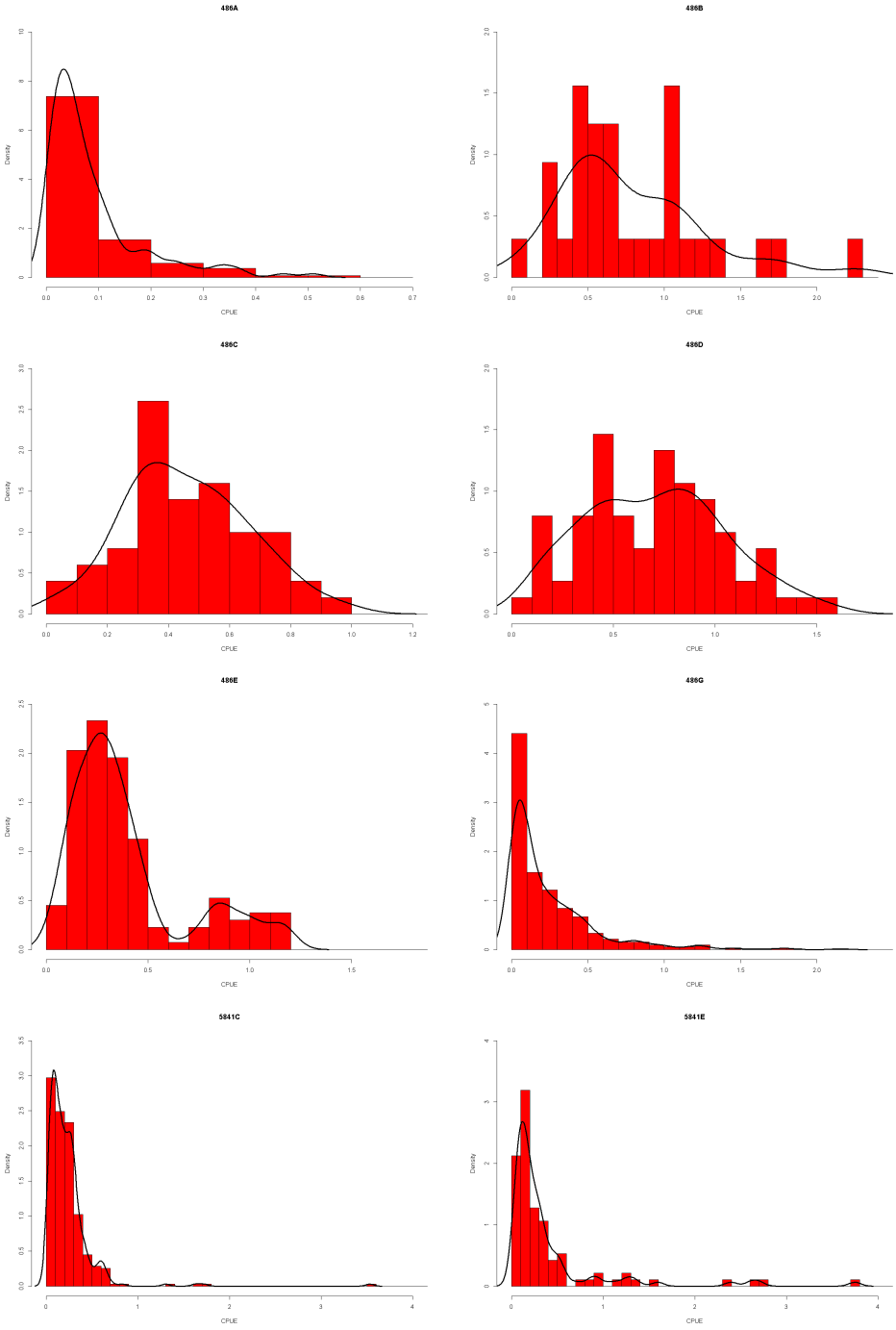

South Africa
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

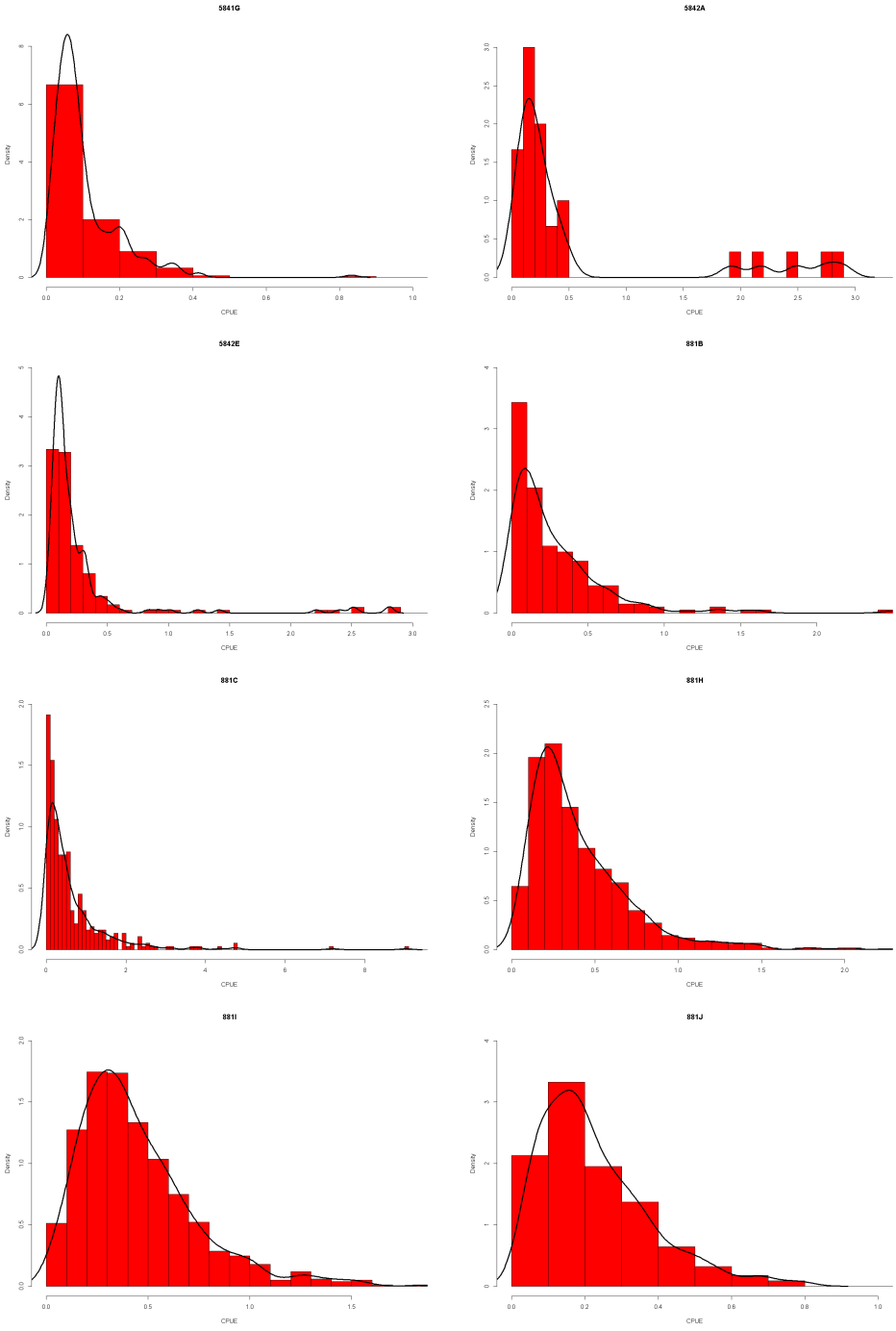

Spain
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

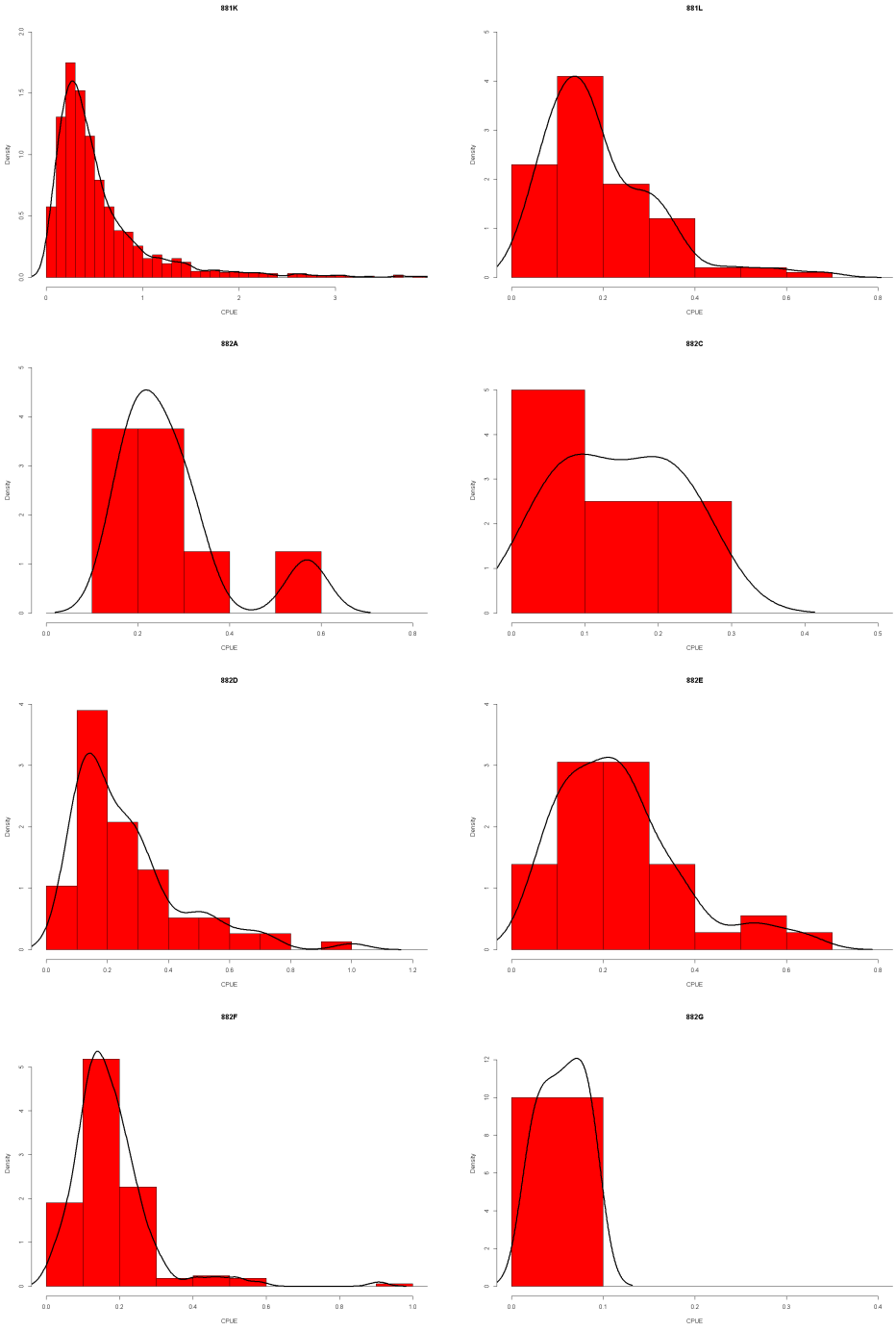

United Kingdom of Great Britain
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

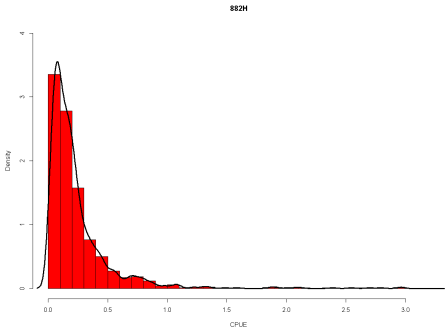

Uruguay
CPUE
0
1
2
3
4
5
6
2008/09
2009/10
2010/11
2011/12
2012/13
CCAMLRSeason

[부록 3] 해역별 histogram & density plot




5841G
5842A
5842E
881B
881C
881H
881I
881J
Density
CPUE


881K
881L
882A
882C
882D
882E
882F
882G
Density
CPUE


882H
Density
0
1
2
3
4
0.0
0.5
1.0
1.5
2.0
2.5
3.0
CPUE