

베이지안 혼합 정규 분포를 이용한 선박 재항시간 분포의 추정

Estimation of Distributions of the Ship's Time in a
Port through Bayesian Mixture Normal Distribution

이민규* · 김근섭⁺

Lee, Min-kyu · Kim, Geun-sub

〈목 차〉

- I. 서 론
 - II. 선행 연구 고찰
 - III. 모 형
 - IV. 분석 자료
 - V. 분석 결과
 - VI. 결 론
-

Abstract: This study adopted Bayesian mixture normal distribution to estimate distributions of the ship's time in a port. Data on the time in a port for full-container vessels which arrived at Kwangyang port and Incheon port with loading/unloading purpose in 2009 were used for empirical analysis. In the case of Kwangyang port data and Incheon port data, five and four were decided as the best component numbers of mixture normal distribution. Groups with the short vessel's time in a port showed a tendency that the time converged on the average of it, but groups with the long vessel's time in a port had marked variation in the time. Comparison of information-based criteria and fitted values of estimation results of mixture normal distribution and Erlang distribution revealed that the former distribution was superior to the latter distribution with regard to the fitting performance. This study will contribute to increasing explanatory power of the model by describing

* 한국해양수산개발원 책임연구원

† 교신저자, 한국해양수산개발원 책임연구원

the heterogeneity of the ship's time in a port. Finally, analysis results can be used for building a queueing model and simulation model.

Key Words : The Ship's Time in a Port, Mixture Normal Distribution, Bayesian Analysis, Information-based Criteria, Fitting Performance

I. 서 론

컨테이너선의 대형화, 선사 전용 터미널의 확산, 항만 물동량 창출을 위한 인센티브의 증가 및 효율의 인하 등 항만 산업의 여건이 변화하고 있다. 이러한 여건 변화에 대응하기 위해 항만 운영자들은 항만의 생산성 제고 방안을 도출하려고 노력하고 있으며, 이를 위해 항만 시설의 확충과 선박 재항시간¹⁾의 최소화를 모색하고 있다. 항만 시설을 확충하려면 오랜 시간과 많은 투자 재원이 요구되기 때문에 정부 예산의 부족 및 민간 기업의 참여 의지 약화 등에 의해 항만 시설의 확충이 어려워질 가능성이 높다. 따라서 항만의 생산성 제고를 위해 가장 우선적으로 추진할 전략은 선박 입출항 시스템과 하역 시스템의 효율적 관리를 통한 선박 재항시간의 최소화라고 할 수 있다. 특히, 글로벌 선사가 재항시간을 단축해 주는 항만에만 대형 선박을 투입하고 그렇지 않은 항만은 기항지에서 제외하는 전략을 채택하고 있기 때문에(김형태, 2005), 항만이 생존하기 위해서는 선박의 재항시간을 어떻게 단축시킬 수 있을지 반드시 고민해야 한다.

항만에 입항하는 선박의 재항시간을 줄이려면, 우선적으로 선박 재항시간의 분포를 파악할 필요가 있다. 선박 재항시간의 분포를 파악하면, 항만의 운영상 문제점을 발견할 수 있고 대기행렬모형 및 시뮬레이션을 적용·분석할 수 있다(장영태, 1994; 박병인, 1998). 선박 재항시간 분포 분석에 관한 대부분의 기존 연구는 선박의 도착간시간 및 서비스시간의 분포가 열랑 분포(Erlang distribution)를 따른다고 가정 한 후 분포 함수의 모수(parameter)를 추정하였다. 열랑 분포는 재항시간을 단지 두 개의 모수로 설명하기 때문에 단순하고 규칙적인 분포를 근사하는 데 적절하다. 하지만 최근 컨테이너선의 대형화, 항만 생산성의 향상 등으로 선박 재항시간 분포의 왜도(skewness)²⁾ 및 첨도(kurtosis)³⁾가 커지는 등 선박 재항시간의 분포가 불규칙해졌기

1) 본 연구에서 정의한 선박의 재항시간은 표박지의 대기시간, 항내 이동시간, 선석 서비스 시간의 합계이며, 선박의 출항시간과 입항시간의 차이로 계산한다.

2) 왜도는 자료의 분포가 대칭인지 아닌지를 측정해 주는 값이다.

3) 첨도는 자료의 분포가 어느 정도 뾰족한지를 나타내는 값이다.

때문에 이를 반영하기 위한 새로운 분포의 적용이 요구되고 있다.

본 연구에서는 선박 재항시간의 분포를 추정하기 위해 베이지안 혼합 정규 분포(Bayesian mixture normal distribution)를 적용한다. 즉, 선박 재항시간 분포의 불규칙적이고 복잡한 형태에 주목하여 단순한 형태의 일량 분포가 아닌 새로운 유연한(flexible) 분포의 적용을 시도한다. 모형의 추정에서 정규 분포 구성요소의 개수를 판별하고 구성요소 집단의 평균과 표준편차를 추정함으로써, 선박 재항시간 집단의 세분화(group segmentation)를 실시하고 각 구성요소 집단의 특징을 파악한다. 실증 분석에서는 2009년 한 해 동안 광양항 및 인천항에 입항한 외항선박 중 양·적하 목적의 풀컨테이너선의 재항시간 데이터를 이용한다. 본 연구는 복잡한 선박의 재항시간 분포를 혼합 정규 분포로 적절하게 근사시킴으로써 모형의 현실 설명력을 높이는 데 기여할 것이다. 또한 본 연구에서 제안하는 방법론은 부두별 적정 하역능력을 산정하는 데 적용할 수 있으며, 추정 결과는 효율적인 선박 입출항 시스템을 구축하기 위해서 실시하는 시뮬레이션의 기초 자료로 활용할 수 있다. 학술적 측면으로는 일량 분포로 선박 재항시간을 추정하고 있는 상황에서 혼합 정규 분포로 추정하는 새로운 접근 방법을 제시하고 있다.

I 장 이후의 본 논문은 다음과 같이 구성된다. II장에서는 선박 재항시간의 분석에 관한 선행 연구에 관해 살펴보고, III장에서는 본 연구의 실증 모형인 베이지안 혼합 정규 분포에 대해 상세하게 설명한다. IV장에서는 사용된 분석 자료를 제시하고, V장에서는 분석 결과를 설명한다. 특히, V장에서는 혼합 정규 분포와 일량 분포의 추정 결과로부터 각 분포의 적합성을 서로 비교한다. 마지막으로 VI장에서는 연구 내용을 요약하고 본 연구의 시사점을 제시한다.

II. 선행 연구 고찰

본 장에서는 선박 재항시간 분석에 관한 선행 연구에 대해 살펴본

다. 선박 재항시간 분석은 선박 재항시간의 요인 분석 및 선박 재항시간 분포 추정의 두 가지로 나누어질 수 있다.

첫째, 선박 재항시간의 요인 분석 연구는 선박의 총중량, 선박의 종류, 항만의 유형, 하역 장비의 성능 등의 요인이 선박 재항시간에 미치는 영향을 분석한다. 이러한 분석을 통해 선박 재항시간을 예측하거나 선박 재항시간을 단축시키기 위한 세부적인 전략을 구축할 수 있다. 윤신휘(2009)는 의사결정나무 분석법을 이용하여 컨테이너의 수, 컨테이너의 분포, 하역장비의 성능 등이 선박 재항시간⁴⁾에 미치는 영향을 추정하고 선박 재항시간의 예측을 시도하였다. 사공훈·최석범(2009)은 벌크 선박의 접안 대기시간을 선박 적재능력, 항만의 특성 및 입지, 입항 시기, 화물처리량의 요인에 따라 t -검정과 분산 분석을 통해 실증적으로 분석하였다. 신강원·정장표(2010)는 생존분석 기법을 적용하여 부두의 서비스 용량, 선박의 총중량, 선박의 종류가 재항시간에 유의한 영향을 미친다는 것을 실증적으로 검증하였다.

둘째, 선박 재항시간 분포 추정에 관한 연구는 선박 재항시간이 특정한 분포 모형에 근사한지를 살펴보고 최적의 분포 함수를 추정한다. 추정 결과로부터 항만의 선박 입출항 패턴을 파악하고 대기행렬 모형과 시뮬레이션 모형을 구축할 수 있다. 장영태(1994)는 인천항 일반부두, 포항항 원료부두, 울산항 정유부두를 대상으로 선박 도착간시간 및 서비스시간의 분포가 얼랑 분포와 얼마나 일치하는지 검증하였다. 김창곤·홍동희·최종희(1997)는 포항항 원료전용부두에 입항한 선박에 대해서 선박 도착간시간과 서비스시간⁵⁾의 분포 함수를 추정하였다. 이를 위해 단계(phase)별 Coxian 분포 및 얼랑 분포를 추정함수로서 고려하였다. 백인흠(1998)은 인천항에 입항한 외항선을 대상으로 부두 서비스 시간을 분석한 결과 얼랑 분포를 따른다는 것을 규명하였다. 또한 추정치를 대기 행렬 이론에 적용한 분석 결과로부

4) 윤신휘(2009)는 선박 재항시간을 컨테이너 선박이 컨테이너 터미널 선석에 들어와서 야드 크레인으로부터 양·적하 서비스를 받는 시점부터 서비스가 끝나는 시점까지로 정의하였다.

5) 김창곤·홍동희·최종희(1997)는 선박이 선석에 접안하여 이안하기까지의 하역 시간을 서비스 시간으로 정의하였다.

터 인천항 부두 선석의 증설이 필요하다고 주장하였다. 박병인(1998)은 부산항 자성대 부두의 선박 입항 자료를 이용하여 선박의 도착간 시간 및 서비스시간의 분포가 각각 지수분포와 일랑 분포를 따른다는 것을 추정하였다.

선박 재항시간 분포 추정에 관한 기존 연구를 고찰한 결과, 대부분의 연구에서 추정 함수로 일랑 분포를 적용하였다. Page(1972)에 의하면, 일랑 분포의 모수를 재항시간 자료의 평균과 분산으로부터 간단하게 추정할 수 있다. 하지만 일랑 분포는 두 개의 모수만으로 분포 전체를 표현하기 때문에 실제 재항시간의 분포를 지나치게 단순화시킨다. 따라서 선박 재항시간이 불규칙적이고 복잡한 형태를 가진다면, 일랑 분포는 선박 재항시간을 적절하게 반영하기 어렵다. 일랑 분포를 적용한 기존 연구와는 달리, 본 연구는 혼합 정규 분포를 적용하여 선박 재항시간의 분포를 추정하고자 한다. 혼합 정규 분포는 일랑 분포로는 묘사하기 어려운 복잡한 형태의 데이터를 잘 근사할 수 있으며 추정 결과로부터 다양한 함의를 이끌어낼 수 있다.

Ⅲ. 모 형

다수의 정규 분포가 결합된 형태인 혼합 정규 분포는 많은 피크를 가지거나 왜도 및 첨도가 큰 데이터를 분석하는 데 유용하다(Park et al., 2010). 혼합 정규 분포는 생물학, 경제학, 경영학, 물리학, 천문학, 공학 등의 다양한 분야에서 불규칙적이고 복잡한 데이터를 분석하기 위한 용도로 활용되고 있다. 구체적으로는 다변량 데이터의 클러스터링, 이상치 데이터(outlier data)의 모델링, 확률 밀도 함수의 추정 등에 많이 적용된다(Frühwirth-Schnatter, 2006).

혼합 정규 분포 모형을 추정하기 위한 방법론으로는 EM(Expectation-Maximization) 알고리즘과 베이지안 추정 방법이 유명하다. Dempster et al.(1977)가 제안한 EM 알고리즘은 주어진 데이터를 손실된 값을 가진 불완전한 데이터로 간주하고 최적의 모수를 추정하는 방법이다.

EM 알고리즘은 우도 함수에서 전역 최적해(global optimum)보다는 국부 최적해(local optimum)를 찾을 가능성이 높고 특이한 형태를 가진 우도 함수의 표준 오차를 계산하기 어렵다는 단점을 가지고 있다 (McLachlan and Peel, 2000). 반면, Diebolt and Robert(1994)가 제시한 베이지안 추정 방법은 EM 알고리즘에 비해서 모수의 불확실성을 추정하기 쉽다는 장점이 있다. 따라서 본 연구에서는 선박 재항시간의 분포를 추정하기 위해 베이지안 혼합 정규 분포를 적용한다.

혼합 정규 분포를 따르는 선박 재항시간 관측변수(y)의 확률 밀도 함수(probability density function)는 식 (1)과 같이 표현할 수 있다.

$$p(y|\theta) = \sum_{k=1}^K \eta_k f_N(y|\mu_k, \sigma_k^2) \quad (1)$$

식 (1)에서 K 는 정규 분포 구성요소의 개수, η_k 는 구성요소 k 의 혼합 비율($0 \leq \eta_k \leq 1$, $\sum_{k=1}^K \eta_k = 1$)을 나타낸다. 그리고 $f_N(y|\mu_k, \sigma_k^2)$ 은 평균과 분산이 각각 μ_k 와 σ_k^2 인 정규 분포의 확률 밀도 함수를 의미한다.

$$f_N(y|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right\} \quad (2)$$

구성요소의 개수가 K 인 베이지안 혼합 정규 분포의 공액 사전 분포(conjugate prior distribution)는 식 (3)에서 식 (5)까지와 같다. η 는 디리슈레(Dirichlet) 분포를 따르며, 구성요소 k 집단의 평균과 분산인 μ_k 와 σ_k^2 는 각각 정규 분포와 역감마(Inverse Gamma) 분포를 따른다.

$$\eta \sim D(\alpha_1, \dots, \alpha_K) \quad (3)$$

$$\sigma_k^2 \sim IG(c_0, C_0) \quad (4)$$

$$\mu_k | \sigma_k^2 \sim N(b_0, \sigma_k^2/B_0) \quad (5)$$

관측변수 y 가 K 개의 혼합 정규 분포 중에서 어떤 구성요소로부터 도출된 것인지 알 수 없기 때문에 y 가 속한 구성요소의 표시 벡터 (component indicator vector)를 $S = (S_1, \dots, S_N)$ 로 정의한다. S 가 결정 되면 사후 분포(posterior distribution)는 다음과 같이 전개된다.

$$\eta | S \sim D(\alpha_1 + n_1, \dots, \alpha_K + n_K)^{6)} \quad (6)$$

$$\sigma_k^2 | S, y \sim IG(c_k^*, C_k^*)^{7)} \quad (7)$$

$$\mu_k | \sigma_k^2, S, y \sim N(b_k^*, \sigma_k^2 / B_k^*)^{8)} \quad (8)$$

$$p(S_i = k | \mu, \sigma^2, \eta, y_i) \sim \eta_k N(y_i | \mu_k, \sigma_k^2) \quad (9)$$

깁스 샘플링(Gibbs sampling) r 번째 단계($r = 1, \dots, R_0, \dots, R$)는 $S^{(r-1)}$ 로부터 식 (6)부터 식 (8)까지 순서대로 $\eta^{(r)}$, $\sigma_k^{2(r)}$, $\mu_k^{(r)}$ 를 생성한다. 생성한 모수와 관측치 y_i 로부터 식 (9)에서 $S^{(r)}$ 을 도출할 수 있다. 이후 $S^{(r)}$ 은 식 (6)에 다시 적용되면서 $(r+1)$ 번째 단계의 모수 추정을 위해 활용된다. 즉, 각 단계별로 식 (6)에서 식 (9)까지 연속적인 과정을 거치면서 사후 분포의 모수 표본을 추출할 수 있다. 추출한 모수 표본 집단에서 안정적인 상태로 수렴하는 값들을 얻기 위해서 첫 번째 단계부터 R_0 번째 단계까지 추출한 값들은 burn-in 구간으로 처리하였다. burn-in 구간을 제외한 추정 모수의 기대값은 식 (10)과 같이 계산된다.

$$E(\theta | y) = \frac{1}{(R - R_0)} \sum_{r=R_0+1}^R \theta^{(r)} \quad (10)$$

6) n_k 는 S 에서 구성요소 k 로 배분된 관측치의 개수를 의미한다.

7) $c_k^* = c_0 + \frac{1}{2}n_k$, $C_k^* = C_0 + \frac{1}{2}\left\{n_k s_{y,k}^2 + \frac{n_k B_0}{n_k + B_0}(\bar{y}_k - b_0)^2\right\}$, \bar{y}_k 와 $s_{y,k}^2$ 는 구성요소 k 인 관측치의 평균과 분산을 나타낸다.

8) $b_k^* = \frac{B_0}{n_k + B_0}b_0 + \frac{n_k}{n_k + B_0}\bar{y}_k$, $B_k^* = \frac{1}{n_k + B_0}\sigma_k^2$

베이저안 혼합 정규 분포에서 구성요소의 개수를 결정하려면 구성 요소의 개수를 순차적으로 증가시키면서 추정 결과가 모형 선택 기준(model selection criteria)에 부합하는지 검증하는 절차를 따라야 한다(Lo et al., 2001; Nylund et al., 2007). 본 연구에서는 모형 선택 기준으로서 AIC(Akaike Information Criteria), BIC(Bayesian Information Criteria) 및 검정 도표(diagnostic plot)를 살펴본다. 식 (11)과 식 (12)에서 L , p , n 은 각각 우도 값, 추정 모수의 개수, 전체 관측치의 개수를 뜻한다.

$$AIC = -2\ln L + 2p \quad (11)$$

$$BIC = -2\ln L + p \ln n \quad (12)$$

검정 도표는 각 구성요소의 모수가 정확하게 식별되는지 검증하기 위해 실시한다. 깃스 샘플링 추출 값의 도표 $(\mu_k^{(r)}, \sigma_k^{(r)})$ 와 $(\mu_k^{(r)}, \mu_l^{(r)})$ ⁹⁾를 통해 모수의 식별 여부를 판별한다.

IV. 분석 자료

본 연구에서는 2009년 한 해 동안 광양항 및 인천항에 입항한 외항선박의 재항시간 데이터를 이용하였다. 입항 목적과 선박 종류에 따라 선박 재항시간이 달라질 수 있기 때문에, 입항한 외항선박 중에서 양·적하 목적을 가진 폴컨테이너선으로 범위를 한정하였다. 외항선박의 입항 목적별 선박 척수와 비중은 <표-1>에 제시되어 있다. 광양항과 인천항에 입항한 외항선박 중 양·적하 목적을 가진 선박의 비중은 각각 53.3%와 40.1%이다.

9) 단, $k \neq l$ 이다.

<표-1> 외항선박의 입항 목적별 선박 척수와 비중

(단위 : 척, %)

입항 목적	광양항	인천항
양·적하	2,831 (53.3)	3,500 (40.1)
양하	838 (15.8)	4,078 (46.8)
적하	1,579 (29.7)	899 (10.3)
선박수리	31 (0.6)	2 (0.0)
급유	2 (0.0)	167 (1.9)
기타	31 (0.6)	75 (0.9)
합계	5,312 (100.0)	8,721 (100.0)

주 : 1) ()의 값은 입항 목적에 따른 선박 척수가 입항한 외항선박의 척수에서 차지하는 비중을 나타냄

2) 기타 목적으로는 단순경유, 선용품적재, 승무원교대, 여객상륙 등이 있음

자료 : 1) 여수지방해양항만청 PORT-MIS(yeosu.mltm.go.kr)

2) 인천지방해양항만청 PORT-MIS(www.portincheon.go.kr/portmis)

<표-2> 양·적하 목적의 선박 종류별 선박 척수와 비중

(단위 : 척, %)

선박 종류	광양항	인천항
폴컨테이너선	2,583 (91.2)	1,714 (49.0)
일반화물선	180 (6.4)	241 (6.9)
자동차운반선	59 (2.1)	58 (1.7)
세미(혼재)컨테이너선	6 (0.2)	121 (3.5)
케미컬운반선	1 (0.0)	1 (0.0)
기타	2 (0.1)	1,365 (39.0)
합계	2,831 (100.0)	3,500 (100.0)

주 : 1) ()의 값은 선박 종류에 따른 선박 척수가 양·적하 목적의 외항선박 척수에서 차지하는 비중을 나타냄

2) 기타 선박으로는 여객선, 화객선, 산물선(벌크선), 국제카페리 등이 있음

자료 : 1) 여수지방해양항만청 PORT-MIS(yeosu.mltm.go.kr)

2) 인천지방해양항만청 PORT-MIS(www.portincheon.go.kr/portmis)

양·적하 목적으로 광양항과 인천항에 입항한 선박의 종류별 척수와 비중은 <표-2>와 같다. 광양항의 경우 폴컨테이너의 비중이 91.2%로 상당히 높지만, 인천항의 폴컨테이너선 비중은 49.0%로 절반 가까이

차지하고 있다. 인천항은 양·적하 목적으로 입항한 선박 중에서 여객선과 화객선의 비중이 높은 것이 특징이다.

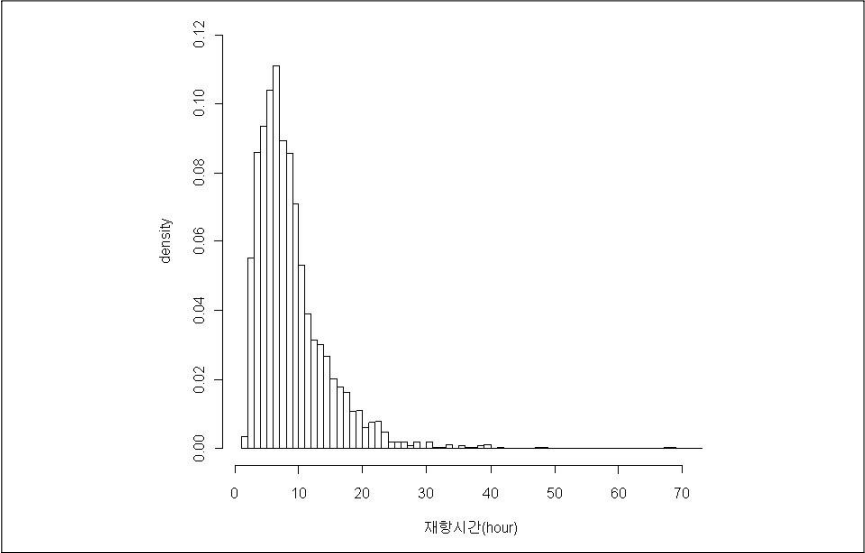
광양항과 인천항에 입항한 양·적하 목적의 풀컨테이너선에서 재항시간이 2시간 미만이거나 72시간 초과인 경우는 분석에서 제외하였다.¹⁰⁾ 최종적으로 2009년 한 해 동안 광양항에 입항한 양·적하 목적의 외항 풀컨테이너선 2,521척 선박과 인천항에 입항한 1,692척을 분석 자료로 선택하였다. <표-3>은 분석에 사용된 선박 재항시간 변수의 요약 통계량을 나타내고 있다. 광양항에서는 평균 선박 재항시간이 8.949시간, 표준편차 5.742시간이며, 인천항에서는 선박 재항시간의 평균과 표준편차가 각각 13.624시간, 8.777시간이다. <그림-1>과 <그림-2>는 광양항 및 인천항의 선박 재항시간에 대한 히스토그램이며, 이를 통해 선박 재항시간의 분포가 높은 첨도를 가지고 있고 히스토그램의 꼬리가 오른쪽으로 늘어진 모양을 띠고 있다는 것을 확인할 수 있다.

<표-3> 분석에 사용된 선박 재항시간 변수의 요약 통계량

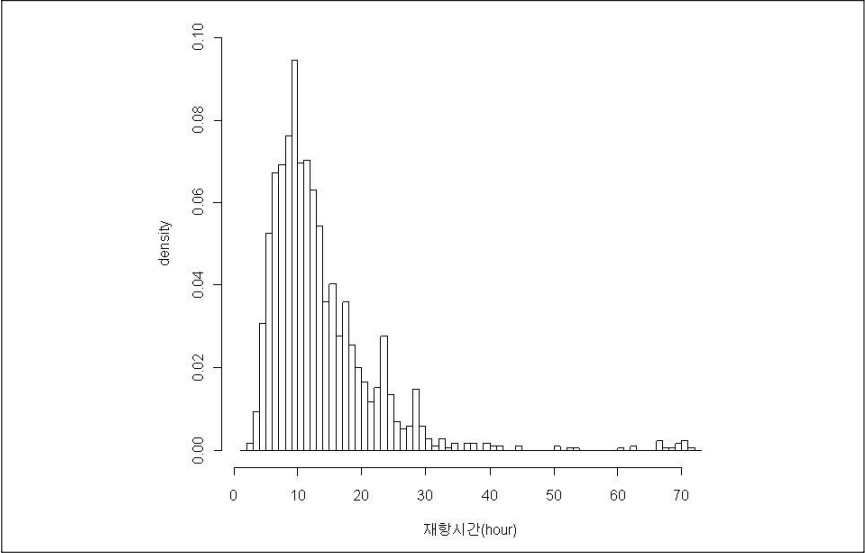
구분	광양항 (관측치 개수=2,521)	인천항 (관측치 개수=1,692)
평균	8.949	13.624
중앙값	7.600	11.333
최빈값	4.000	10.167
표준편차	5.742	8.777
분산	32.968	77.028
첨도	13.288	14.040
왜도	2.515	2.990
최소값	2.000	2.417
최대값	68.083	71.583

10) 재항시간이 2시간 미만인 경우는 실제 하역이 이루어졌을 가능성이 거의 희박하다고 판단했다. 또한, 재항시간이 72시간 초과인 경우는 선박 입항 목적이 양·적하뿐만 아니라 선박 수리 등의 다른 목적이거나 기상 상태 악화 등으로 선박의 출항이 지연된 예외적인 상황이 발생한 것으로 간주하였다.

<그림-1> 광양항 선박 재항시간에 대한 히스토그램



<그림-2> 인천항 선박 재항시간에 대한 히스토그램



V. 분석 결과

1. 혼합 정규 분포의 추정 결과

본 절에서는 선박 재항시간 분포의 추정을 위해서 베이지안 혼합 정규 분포를 적용하였다.¹¹⁾ Rossi and McCulloch(2007)에 의거하여 사전 분포의 모수 α_k , b_0 , B_0 , c_0 , C_0 를 $\alpha_k = 5$, $b_0 = 0$, $B_0 = 0.01$, $c_0 = 3$, $C_0 = 3$ 으로 설정하였다. 추정 과정에서 모두 5천 번의 깃스 샘플링 추출을 실시하였으며, 처음의 1천 번의 추출은 burn-in 구간으로 버려지고 나머지 4천 번의 추출 값은 모수 추정에 사용되었다. 깃스 샘플링 추정에서 구성요소의 라벨 스위칭(label switching) 및 과적합(overfitting)이 발생하면, 모형의 식별이 어렵고 추정 결과를 전혀 신뢰할 수 없게 된다(Frühwirth-Schnatter, 2006). 본 연구에서는 모형의 식별을 위해 $\mu_1 < \mu_2 < \dots < \mu_K$ 의 제약 조건을 부여한다.

<표-4>는 혼합 정규 분포의 구성요소의 개수별 추정 결과에 따른 AIC와 BIC를 나타낸다.¹²⁾ <표-4>에서 AIC 기준에 의하면 광양항과 인천항의 선박 재항시간에 대해서 각각 구성요소가 6개와 5개인 혼합 정규 분포가 가장 적합하다. 하지만 BIC 기준으로는 5개와 4개의 구성요소를 가진 혼합 정규 분포가 광양항과 인천항의 선박 재항시간을 가장 잘 근사하고 있다.

11) 베이지안 혼합 정규 분포의 추정을 위한 프로그램으로는 R을 사용하였다. R은 다양한 통계적 분석과 우수한 그래픽 방법을 제공하며 뛰어난 프로그래밍 기능이 있어서 사용자가 새로운 함수를 작성하여 추가할 수 있다(김달호, 2005).

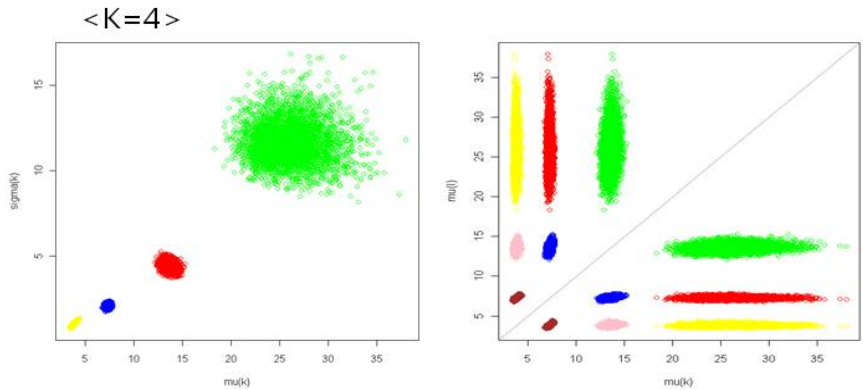
12) AIC와 BIC가 작은 모형일수록 모형의 적합성이 우수하다는 것을 의미한다.

<표-4> 혼합 정규 분포의 구성요소 개수의 선택 기준

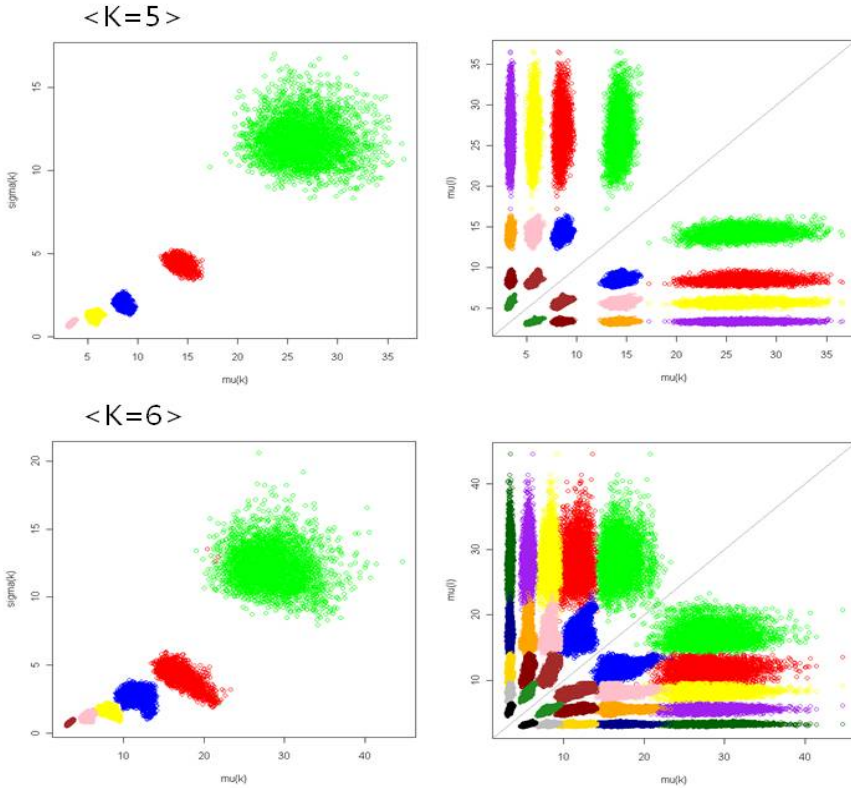
구성요소의 개수	추정모수의 개수	광양항		인천항	
		AIC	BIC	AIC	BIC
2	5	14,910.4	14,939.5	11,244.6	11,271.8
3	8	14,695.3	14,741.9	10,949.9	10,993.3
4	11	14,589.8	14,653.9	10,919.7	10,979.5
5	14	14,560.7	14,642.3	10,906.4	10,982.5
6	17	14,553.2	14,652.3	10,908.7	11,001.0

<그림-3>과 <그림-4>에서 구성요소 개수를 정하는 세 번째 기준으로 깃스 샘플링 추출 값의 도표를 살펴본다. 구성요소별 깃스 샘플링 추출 값이 서로 중첩되는 경우, 모형이 과적합될 수 있다.¹³⁾ 세 가지 선택 기준을 모두 검토한 결과, 광양항의 선박 재항시간은 5개의 구성요소, 인천항의 경우는 4개의 구성요소를 가진 혼합 정규 분포를 추정 모형으로 결정한다.

<그림-3> 광양항 선박 재항시간의 깃스 샘플링 추출 값의 도표

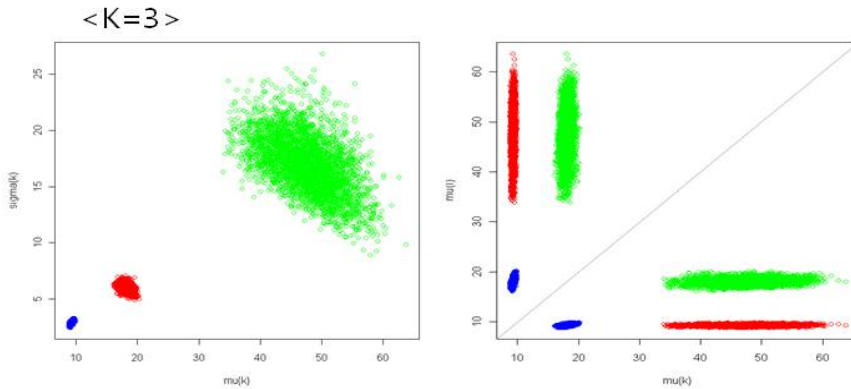


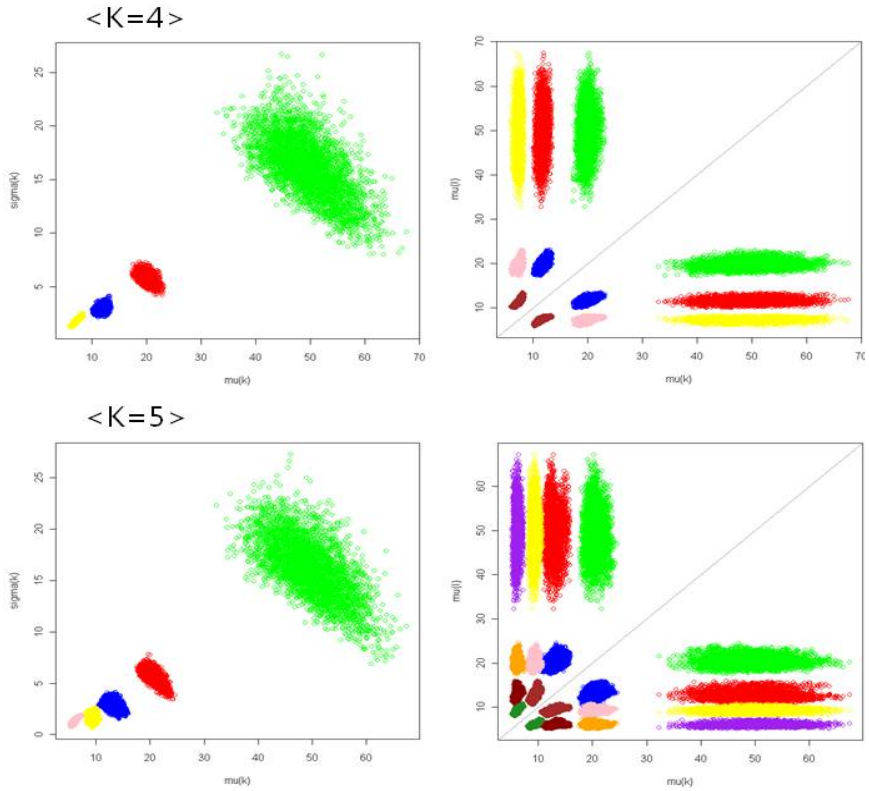
13) 광양항의 경우 구성요소의 개수가 6개일 때 깃스 샘플링 추출 값이 중첩되며, 인천항의 경우는 구성요소의 개수가 5개일 때 중첩이 발생한다.



주 : 왼쪽 도표는 $(\mu_k^{(r)}, \sigma_k^{(r)})$, 오른쪽 도표는 $(\mu_k^{(r)}, \mu_l^{(r)})$ 을 나타냄(단, $k \neq l$ 임)

<그림-4> 인천항 선박 재항시간의 깃스 샘플링 추출 값의 도표





주 : 왼쪽 도표는 $(\mu_k^{(r)}, \sigma_k^{(r)})$, 오른쪽 도표는 $(\mu_k^{(r)}, \mu_l^{(r)})$ 을 나타냄(단, $k \neq l$ 임)

<표-5>는 베이저안 혼합 정규 분포의 추정 결과이며, 광양항과 인천항의 선박 재항시간에 대한 구성요소 집단별 모수의 평균, 표준편차, 2.5% 분위수, 97.5% 분위수 및 전체 집단의 평균, 분산, 집단 간 분산의 결정 계수가 제시되어 있다.

<표-5>

혼합 정규 분포의 추정 결과

구성 요소	모수	광양항(K=5)				인천항(K=4)			
		평균	표준 편차	2.5% 분위수	97.5% 분위수	평균	표준 편차	2.5% 분위수	97.5% 분위수
1	η_1	0.146	0.026	0.094	0.195	0.294	0.064	0.162	0.413
	μ_1	3.383	0.136	3.106	3.640	7.425	0.398	6.511	8.095
	σ_1	0.781	0.071	0.643	0.923	1.967	0.200	1.519	2.318
2	η_2	0.239	0.052	0.137	0.345	0.397	0.056	0.293	0.509
	μ_2	5.700	0.257	5.199	6.209	11.650	0.593	10.563	12.911
	σ_2	1.224	0.154	0.888	1.498	2.977	0.263	2.484	3.537
3	η_3	0.339	0.056	0.216	0.441	0.281	0.039	0.199	0.357
	μ_3	8.446	0.351	7.829	9.294	19.832	0.910	18.202	21.954
	σ_3	1.995	0.222	1.567	2.425	5.801	0.438	4.928	6.649
4	η_4	0.246	0.028	0.187	0.299	0.028	0.006	0.017	0.042
	μ_4	14.207	0.603	13.154	15.478	49.493	5.240	39.563	59.992
	σ_4	4.400	0.270	3.834	4.881	16.217	2.751	10.813	21.382
5	η_5	0.030	0.008	0.018	0.047	-	-	-	-
	μ_5	26.476	2.724	21.665	32.405	-	-	-	-
	σ_5	11.795	1.203	9.730	14.457	-	-	-	-
μ		9.011				13.772			
σ^2		33.899				81.315			
R_d^2		0.683				0.736			

주 : 1) $\mu = \sum_{k=1}^K \eta_k \mu_k$

2) $\sigma^2 = \sum_{k=1}^K \eta_k \sigma_k^2 + \sum_{k=1}^K \eta_k (\mu_k - \mu)^2$, 전자는 집단 내 분산을, 후자는 집단 간 분산을 나타냄

3) $R_d^2 = \sum_{k=1}^K \eta_k (\mu_k - \mu)^2 / \sigma^2$

혼합 정규 분포의 구성요소별 추정 모수로부터 전체 집단의 평균(μ)과 분산(σ^2)을 계산할 수 있으며, 이들의 값은 <표-3>의 평균, 분산과 거의 일치한다. 평균은 각 구성요소 집단의 평균(μ_k)을 구성요소

의 혼합 비율(η_k)로 가중 평균한 값이며, 분산은 집단 내 분산(within-group variance)과 집단 간 분산(between-group variance)의 합계이다. 집단 간 분산이 전체 분산에서 차지하는 비율을 집단 간 분산의 결정 계수(R_d^2)라고 하면, 결정 계수의 크기가 1에 가까울수록 집단이 서로 잘 분리되었음을 나타낸다(Frühwirth-Schnatter, 2006). 광양항과 인천항의 R_d^2 가 모두 0.5보다 크기 때문에 구성요소 집단의 분리가 대체로 잘 이루어졌다. 이는 혼합 정규 분포의 적용이 적절하다는 것을 의미한다.

혼합 정규 분포는 전체 집단을 동질적인(homogeneous) K 개의 구성요소 집단으로 나누는 것이다(Park et al., 2010). 이러한 관점에서 볼 때, 광양항과 인천항의 선박 재항시간은 각각 5개와 4개의 동질적인 재항시간 집단으로 분류될 수 있다. 각 구성요소 집단을 비교하면 선박 재항시간의 평균이 큰 집단일수록 표준편차가 크다. 즉, 선박의 재항시간이 짧은 경우 재항시간이 평균값 주위로 일정하게 수렴하는 경향이 강하지만, 재항시간이 길 때는 재항시간의 변동성이 크다. 구성요소의 혼합 비율(η_k)은 선박 재항시간이 개별 구성요소 집단에 속할 확률을 의미한다. 광양항의 경우, 집단 3의 혼합 비율이 33.9%로 가장 높지만, 집단 5의 혼합 비율은 3.0%에 불과하다. 인천항은 집단 2의 혼합 비율(39.7%)이 가장 높지만, 집단 4에서 가장 낮은 혼합 비율(2.8%)을 보인다. 두 항만 모두 공통적으로 평균이 가장 큰 집단일수록 혼합 비율이 5% 미만으로 매우 낮다. 이와 같이 혼합 비율의 추정 결과로부터 낮은 혼합 비율을 가진 집단을 식별할 수 있다.

2. 열량 분포의 추정 결과

본 절에서는 기존 연구에서 많이 활용한 열량 분포를 적용하여 선박 재항시간의 분포를 추정하였다. 열량 분포는 감마 분포의 일종으로 형태 모수(shape parameter)가 정수 값을 가진다는 제약 조건을 가지고 있다. 감마 분포의 확률 밀도 함수는 다음과 같다.

$$f_G(y|\lambda, k) = \lambda^k y^{k-1} \frac{e^{-\lambda y}}{(k-1)!} \quad (13)$$

여기서 y 는 선박 재항시간, k 는 형태 모수, λ 는 크기 모수를 나타낸다. 감마 분포의 평균과 분산은 식 (14), (15)와 같다.

$$E(y) = \frac{k}{\lambda} \quad (14)$$

$$Var(y) = \frac{k}{\lambda^2} \quad (15)$$

식 (14)와 (15)로부터 다음과 같이 모수를 간단하게 추정할 수 있다. 이후 얼랑 분포의 제약 조건에 의거하여 추정한 실수 값 k 를 가장 가까운 정수 값으로 대체한다(Page, 1972).

$$\lambda = \frac{E(y)}{Var(y)} \quad (16)$$

$$k = \frac{\{E(y)\}^2}{Var(y)} = \lambda \times E(y) \quad (17)$$

자료의 평균과 분산으로부터 감마 분포의 모수를 추정하는 방법 이외에 모수를 추정하는 다른 방법으로 최우추정법(Maximum Likelihood Estimator : MLE)을 고려한다. 식 (13)에 최우추정법을 적용하여 우도 함수를 최대화 하는 모수를 추정할 수 있다. 두 가지 방법에 의한 추정 결과가 <표-6>에 제시되어 있다. 형태 모수(k)의 추정 값은 서로 같지만, 크기 모수(λ)는 다른 값을 가진다.

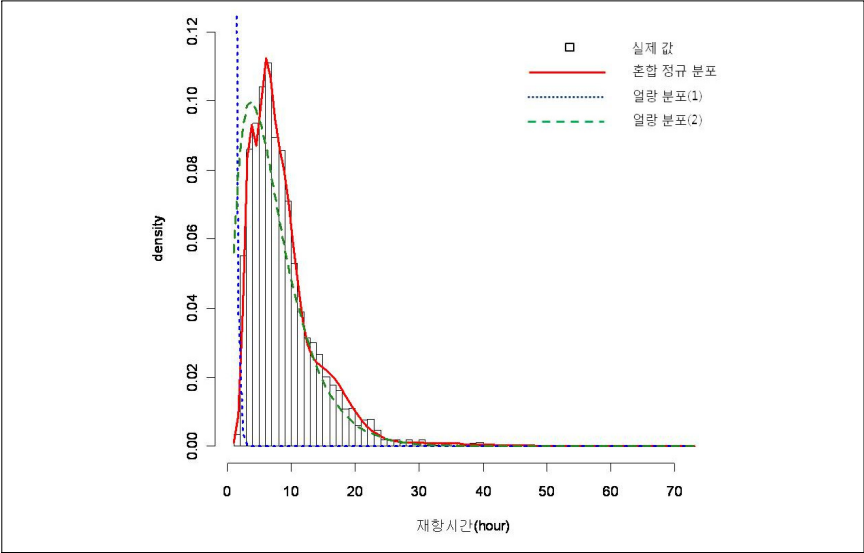
<표-6>

얼랑 분포의 추정 결과

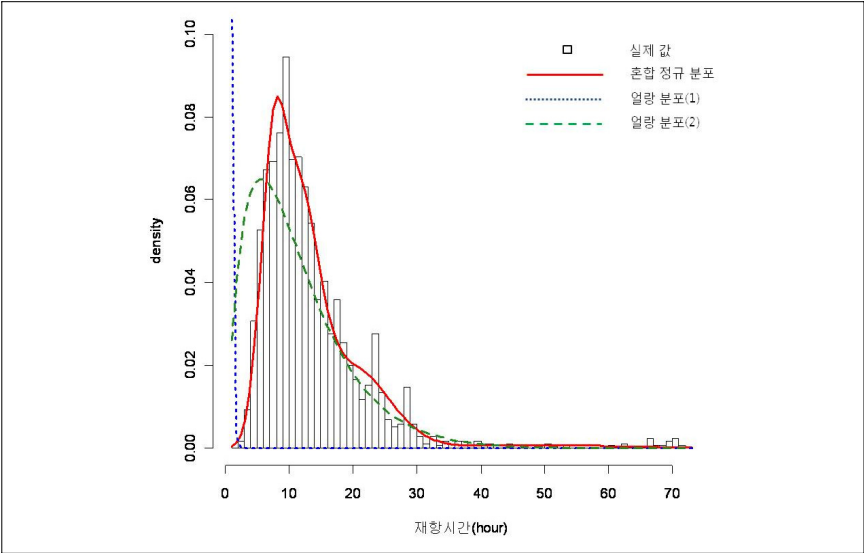
구분		광양항	인천항
Page(1972)	λ	0.271	0.177
	k	2.429	2.410
최우추정법	λ	3.684	5.654
	k	2.429	2.410

주 : 얼랑 분포의 제약 조건에 따라서 k 는 가장 가까운 정수 값인 2로 대체됨

<그림-5> 광양항 선박 재항시간의 분포별 추정 값의 비교



<그림-6> 인천항 선박 재항시간의 분포별 추정 값의 비교



VI. 결 론

항만 운영자들이 항만의 생산성 제고 방안을 모색하고 있는 상황에서, 현실 설명력이 뛰어난 새로운 선박 재항시간 분포 모형을 제안하는 것은 선박 재항시간을 줄이고 항만의 운영상 문제점을 파악하는 데 큰 도움이 된다. 이러한 맥락에서 본 연구는 복잡한 데이터를 분석하기 위해 활용되고 있는 혼합 정규 분포를 선박 재항시간 분포의 추정에 적용하였다. 실증 분석으로 2009년 한 해 동안 광양항과 인천항에 양·적하 목적으로 입항한 폴컨테이너선의 선박 재항시간 자료를 이용하여 베이지안 혼합 정규 분포를 추정하였다.

혼합 정규 분포의 구성요소 개수를 판별하기 위해 구성요소 개수별 모형 적합도 및 검정 도표를 살펴봤으며, 광양항과 인천항에 대해서 각각 5개와 4개의 구성요소의 혼합 정규 분포를 최적의 추정 모형으로 선택하였다. 추정 결과에서 다음과 같은 시사점을 얻을 수 있었다. 첫째, 집단 간 분산의 결정 계수 값으로부터 혼합 정규 분포의 적용이 적절했다고 판단할 수 있었다. 둘째, 선박의 재항시간이 짧은 집단일수록 재항시간이 평균 주위로 수렴하지만, 그렇지 않은 집단에서는 재항시간의 변동성이 컸다. 셋째, 평균이 가장 큰 집단의 구성요소의 혼합 비율이 가장 낮았다. 넷째, 혼합 정규 분포와 일랑 분포의 적합성을 서로 비교한 결과, 혼합 정규 분포의 현실 설명력이 일랑 분포에 비해 훨씬 뛰어난 것으로 나타났다.

본 연구에서 혼합 정규 분포를 이용하여 선박 재항시간 분포를 추정함으로써 일랑 분포를 대체할 수 있는 새로운 분포의 적용 가능성을 확인하였다. 또한 혼합 정규 분포의 추정으로부터 선박 재항시간 집단을 세분화하고 구성요소 집단의 특징을 알 수 있었다. 본 연구는 복잡한 선박의 재항시간 분포를 적절하게 근사시킴으로써 모형의 현실 설명력을 높이는 데 기여할 것이다. 아울러, 본 연구에서 제안한 방법론은 부두별 적정 하역능력 산정에 적용할 수 있으며, 추정 결과는 시뮬레이션을 위한 기초 자료로 활용할 수 있다. 학술적 측면에서 본 연구는 선박 재항시간의 추정 연구 가운데 최초로 혼합 정규 분포

의 적용을 시도했다는 점에서 큰 의의가 있다.

본 연구에서 제안한 혼합 정규 분포의 적용 가능성을 높이기 위해서는 분석 대상 항만의 범위를 넓힐 필요가 있다. 그리고 연도별로 특정 항만의 선박 재항시간 분포를 추정하면, 선박 재항시간의 변화 패턴을 자세하게 파악할 수 있을 것이다. 본 연구에서 제시한 혼합 정규 분포로 추정한 재항 시간 분포를 대기이론에 적용하기 위해서는 재항 시간 분포가 기억 상실 성질(memoryless property) 조건을 만족해야 한다. 혼합 정규 분포를 대기이론으로 전개하기 위한 방법의 모색은 추후 연구 과제로 남겨 둔다.

투고일(2010년 10월 4일)

심사일(2010년 11월 12일)

게재확정일(2010년 11월 26일)

참고문헌

1. 김달호, 「R과 WINBUGS를 이용한 베이지안 통계학」, 자유아카데미, 2005.
2. 김창곤 · 홍동희 · 최종희, “항만 대기시스템에서 시간분포의 통계적 검증에 대한 사례연구”, 「해양정책연구」, 제12권, 1997.
3. 김형태, “항만하역장비 현대화자금 지원제도 도입 절실하다”, 「월간 해양수산」, 제247호, 2005. 4.
4. 박병인, “이질적 복수서버를 갖는 혼잡 컨테이너터미널의 선박관련 시간분포 추정”, 「해양정책연구」, 제13권 제1호, 1998.
5. 백인흠, “선박재항시간에 대한 분석연구: 인천항의 경우”, 「수산해양교육연구」, 제10권 제1호, 1998.
6. 사공훈 · 최석범, “국내 벌크선박의 체선원인에 관한 실증적 분석: 접안 대기시간을 중심으로”, 「해운물류학회」, 제25권 제2호, 2009.
7. 신강원 · 정장표, “생존분석모형을 이용한 선박의 재항시간 및 온실가스 배출량 분석”, 「대한토목학회논문집」, 제30권 제4D호, 2010.
8. 윤신휘, 「기계학습 기법을 이용한 본선의 재항시간 예측」, 부산대학교 석사학위논문, 2009.
9. 장영태, “우리나라 주요 수출입 항만에서의 선박 입출항시간 분포 추정에 관한 연구”, 「한국해운학회지」, 제19권, 1994.
10. Dempster, A. P., N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, *Journal of the Royal Statistical Society Series B*, 39, 1977.
11. Diebolt, J. and C. P. Robert, “Estimation of Finite Mixture Distributions through Bayesian Sampling”, *Journal of the Royal Statistical Society Series B*, 56, 1994.
12. Frühwirth-Schnatter, S., *Finite Mixture and Markov Switching Models*, New York, Springer, 2006.
13. Lo, Y., N. R. Mendell and D. B. Rubin, “Testing the Number of Components in a Normal Mixture”, *Biometrika*, 88, 2001.
14. McLachlan, G. and D. Peel, *Finite Mixture Model*, New York, John

Wiley & Sons, 2000.

15. Nylund, K. L., T. Asparouhov and B. O. Munthén, “Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study”, *Structural Equation Modeling*, 14, 2007.
16. Page, E., *Queueing Theory in OR*, London, Butterworths, 1972.
17. Park, B. J., Y. Zhang and D. Lord, “Bayesian Mixture Modeling Approach to Account for Heterogeneity in Speed Data”, *Transportation Research Part B*, 44, 2010.
18. Rossi, P. E. and R. McCulloch, *bayesm: Bayesian Inference for Marketing/Micro-econometrics*, R package version 2.1-3, 2007.
19. 여수지방해양항만청 PORT-MIS(yeosu.mltm.go.kr)
20. 인천지방해양항만청 PORT-MIS(www.portincheon.go.kr/portmis)

