Journal of Korea Port Economic Association
Vol.37, No.3, 2021, pp.01-17.
https://doi.org/10.38121/kpea.2021.09.37.3.1

Datamining 기법을 활용한 단기 항만 물동량 예측

하준수* · 임채환** · 조광휘*** · 하헌구****

Forecasting the Daily Container Volumes Using Data Mining with CART Approach

Ha, Jun-Su · Lim Chae Hwan · Cho Kwang-Hee · Ha, Hun-Koo

Abstract

Forecasting the daily volume of container is important in many aspects of port operation. In this article, we utilized a machine-learning algorithm based on decision tree to predict future container throughput of Busan port. Accurate volume forecasting improves operational efficiency and service levels by reducing costs and shipowner latency. We showed that our method is capable of accurately and reliably predicting container throughput in short-term(days). Forecasting accuracy was improved by more than 22% over time series methods(ARIMA). We also demonstrated that the current method is assumption-free and not prone to human bias. We expect that such method could be useful in a broad range of fields.

Key words: port volume forecasting, data mining, CART, daily demand forecasting

▷ 논문접수: 2021. 08. 03. ▷ 심사완료: 2021. 09. 28. ▷ 게재확정: 2021. 10. 14.

* 인하대학교 물류전문대학원 석사과정, 제1저자, hajunsu93@gmail.com
** 인하대학교 물류전문대학원 통합과정, 공동저자, lch3289@gmail.com
*** 인하대학교 물류전문대학원 박사과정, 공동저자, jgh2065@naver.com
**** 인하대학교 물류전문대학원 교수, 교신저자, hkha@inha.ac.kr

Ⅰ. 서 론

빅데이터 분석을 통해 유의미한 통계적 규칙과 패턴을 찾아내는 과정을 의미하는 데이터마이닝(Datamining)은 4차 산업혁명과 디지털 기반 융합 산업의 발전에 힘입어 그 파급력이 점차 커질 것으로 예상된다. 빅데이터 시대에서 경쟁력을 유지하기 위해 선진 기업들은 데이터를 중심으로 비즈니스와 의사결정 방식을 재편하는 디지털 트랜스포메이션(Digital Transformation)에 주력하고 있다. 이러한 변화는 제조, 금융부터 의료, 국방까지 산업 전반에서 진행되고 있으며 물류 산업에서도 그 논의가 활발히 이루어지고 있다.

물류 산업에 데이터마이닝을 적용할 수 있는 방법은 다양하며 지금도 최신 방법론과 적용법들이 등장하고 있다. 그중에서도 데이터마이닝을 활용한 물동량 예측은 시계열 분석이나 회귀분석을 활용한 예측에 비해 모형 구축에 대한 제약이 적고 간결하지만 정확한 예측력을 가진다는 점에서 주목받고 있다. 정확한 물동량 예측은 물류 및 교통 산업의 공급사슬 위의 모든 경제주체의 비효율성과 기회비용을 감소시키고 이윤극대화를 추구하기 위한 필수조건이다. 따라서 정확한 물동량 예측 방법에 관한 연구는 매우 중요하다고 볼 수 있다.

기존의 물동량 예측에 관한 연구는 대부분 1년 이상의 기간의 물동량을 예측한 장기적 관점의 시계열 예측이 대부분이었다. 이는 인프라 투자에 대규모 비용과 시간이 소요되는 물류 산업의 특징을 고려할 때 장기적 관점에서의 물동량 예측이 중요하기 때문이며 동시에 지금까지는 단기 예측에 적합한 방법론과 빅데이터를 다룰 수 있는 컴퓨팅 파워가 부족했던 것을 원인으로 볼 수 있다. 따라서 이러한 제한조건이 해결된 시점에서 데이터마이닝 기법을 활용한 단기 물동량 예측 연구를 진행하는 것은 시의적절한 것으로 판단된다.

장기적 관점에서의 물동량 예측이 시설 확장 계획을 수립하기 위해 필요하다면 단기적 관점에서의 물동량 예측은 실질적인 시설 운영계획을 수립하기 위한 근거 자료로 사용될 수 있다. 항만의 원활한 선적과 하역을 위해서는 선적과 하역에 필요한 시설과 노동 인력이 충분히 배치되어야 한다. 충분한 시설과 인력이 확보되지 않으면 선적 및 하역이 지연되고 이로 인해 선주의 대기시간이 증가하여 항만의 서비스 수준이 하락할 수 있다. 반대로 시설과 인력이 과도하게 투입되면 비용이 증가하여 항만 운영의 비효율성이 증가한다. 따라서 정확한 단기 물동량 예측은 항만의 효율적인 운영계획을 수립하기 위한 필수 조건으로 볼 수 있다.

본 연구는 데이터마이닝 기법을 활용한 단기 물동량 예측 모형을 제시하였다. 제시한 예측 모형은 주별 예측과 일별 예측 두 가지 단계로 구성된다. 1단계에서는 CART(Classification And Regression Tree) 모형을 활용한 주별 예측을 진행한다. 설명변수로는 시계열 분석 주별 물동량 예측치와 주요국 주별 노동일 수를 사용하였다. 시계열 분석에는 ARIMA(Autoregressive Integrated Moving Average) 모형을 사용하였다. 2단계에서는 앞서 도출한 주별 예측치와 더불어 요일, 주요국 공휴일, 주요국 행사기간을 설명변수로 CART 모형을 구축하여 최종 일별 물동량 예측을 진행하였다. 이후 예측모형을 통해 미래 92일에 대한 부산항 일별 물동량 예측을 진행하고 실제 물동량과 비교하여 예측 정확도를 검증하였다. 최종적으로 ARIMA 모형으로 예측한 결과와 정확도 비교검증을 진행하여 제안한 예측 방법론의 효율성 및 정확도 측면에의 우수성을 확인하고 이를 바탕으로 결론 및 시사점을 도출하였다.

II. 선행 연구

지금까지 항만 물동량 예측에 관한 다양한 연구가 진행되었으며 그중에서도 거시적 관점의 장기 물동량 예측 연구가 특히 활발히 진행되었다. 예측을 위한 방법론으로는 ARIMA(Autoregressive Integrated Moving Average) 모형, ARIMA 모형에서 추가로 계절성까지 고려한 SARIMA(Seasonal Auto-regressive Integrated Moving Average) 모형 등 시계열 분석 모형이 주로 사용되었다. 비교적 최근에는 신경망 모형(Neural Network), 의사결정나무(Decision Tree)등 보다 다양한 기법이 점차 많이 활용되고 있다.

하준수 외(2021)는 ARIMA 모형을 활용하여 부산항 9개 부두의 물동량 예측을 진행하였다. 학습 데이터로는 2841일 동안의 부산항 일별 컨테이너 물동량 자료를 활용하였으며 추정한 모형을 바탕으로 미래 42일의 일별 물동량을 예측하였다. 또한, 예측치의 신뢰구간을 활용하여 부두 별로 물동량의 이상 징후를 사전에 감지하는 관리 방안을 제시하고 실제 데이터를 바탕으로 실증분석을 진행하였다. 마지막으로 연구의 한계점으로 일별 예측의 정확도 개선이 필요함을 강조하였다.

김두환, 이강배(2020)는 신경망 모형의 한 종류인 LSTM(Long Short Term Memory) 모형을 기반으로 부산항 컨테이너 물동량 예측을 진행하였다. 학습 데이터로는 180개월 동안의 부산항 월별 컨테이너 물동량 자료를 사용하였으며 추정한 모형을 바탕으로 미래 24개월의 월별 물동량을 예측하였다. 또한, 동일한 데이터를 ARIMA 모형에 적용하여 예측을 진행하고 두 모형의 예측 정확도를 비교하였다. 이를 통해 월별 항만 물동량 예측에 LSTM 모형이 ARIMA 모형보다 적합함을 보였다.

Rashed et al(2017)은 ARIMA Intervention 모형과 ARIMAX 모형을 활용하여 앤트워프(Antwerp)항의 월별 컨테이너 물동량을 예측하였다. 해당 연구에서는 특정한 경제 변화를 반영하는 핵심 지표들과 컨테이너 물동량의 상관관계를 분석하기 위해 ARIMAX 모형을 적용하였다. 연구에서는 과거 20년 동안의 월별 앤트워프항 물동량 데이터를 활용하였으며 이를 바탕으로 EU 산업 신뢰성지수(Industrial Confidence Index)가 앤트워프 항의 컨테이너 물동량에 가장 큰 연동성을 가진다고 주장하였다.

민경창, 하헌구(2014)는 SARIMA 모형을 활용하여 국내 전체 항만의 분기별 컨테이너 물동량을 예측하였다. 해당 연구에서는 84분기 동안의 분기별 물동량 데이터를 사용하여 시계열 모형을 추정하였으며 추정한 모형을 통해 미래 4분기 물동량을 분기 단위로 예측하였다. 또한, 동일한 데이터를 활용하여 계절성을 고려하지 않은 ARIMA 모형을 추정하고 마찬가지로 미래 물동량을 예측하여 정확도 비교를 진행하였다. 이를 바탕으로 분기 예측에 ARIMA 모형보다 SARIMA 모형이 적합함을 보였다.

선행 연구들은 대부분 거시적 관점에서의 장기 물동량 예측에 초점을 맞춰왔다. 비교적 최근 들어 반기, 분기 단위의 단기 수요예측이 진행되었지만, 아직도 일 단위 수요예측을 진행한 연구는 찾아보기 어렵다. 이에 본 연구에서는 대표적인 데이터마이닝 기법을 활용하여 단기 항만 물동량 예측에 적합한 모형을 제시하였다. 또한, 부산항 일별 데이터(2005년 ~2020년)를 활용하여 실증분석을 진행하고, 기존 시계열 모형으로 예측한 결과와 정확도를 비교하여 제시한 모형의 우수성을 증명하였다. 선행 연구에서 많이 다뤄지지 않은 일별 물동량 예측을 데이터마이닝 기법을 통해 수행한 본 연구는 충분한 가치가 있을 것으로 판단된다.

Ⅲ. 연구 모형

1. ARIMA 모형

본 연구에서는 2단계로 구성된 단기 예측 방법론을 제시하였으며 그중 주별 예측을 진행하는 1단계에서 시계열 예측치를 설명변수로 사용하였다. 본 연구에서 시계열 예측을 위해 사용한 ARIMA(Auto regressive Integrated Moving Average) 모형은 대표적인 시계열 예측 모형으로 데이터의 과거 시계열 수치와 오차를 기반으로 현재의 시계열 수치를 설명하는 모형이다. ARIMA 모형에서 종속변수는 독립변수에 의해 설명되는 것이 아니라 자신의 과거 데이터와 확률적 오차항에 의해 설명된다는 특징을 가진다. 일반적인 ARIMA(p,d,q) 모형의 형태는 (식 1)과 같다.

$$W_t = \varnothing_1 W_{t-1} + \cdots + \varnothing_t W_{t-p} \quad \text{-----(식 1)}$$
$$+ \epsilon_t - \theta_1 \epsilon_{t-1} - \cdots - \theta_q \epsilon_{t-q}$$

W_t: 원 시계열자료
t : 시간을 나타내는 연산자
ϵ : $N(0,\sigma^2)$을 따르는 오차항, 백색잡음
p : $AR(Auto-regressive)$항의 차수
q : $MA(Moving-average)$항의 차수

ARIMA 모형을 활용하여 시계열 예측을 진행하기 위해서는 시계열 자료의 안정성(Stationarity) 조건이 만족되어야 한다. 안정성 조건을 만족하지 않는 시계열 자료는 평균과 분산 값이 시점에 따라 변화하기 때문에 유의미한 예측치를 얻을 수 없다. 또한, 추정해야 하는 모수가 늘어나 모형 추정 자체에 어려움이 생긴다. 따라서 ARIMA 모형으로 시계열 예측을 진행하기 위해서는 반드시 안정성 조건을 만족하는지 확인해야 한다. 시계열 자료가 안정적이지 않다면 차분(Difference)의 과정을 통해 시계열의 추세를 제거함으로써 자료를 안정적으로 변환시킨 후 모형을 적용해야 한다. 추세 제거를 위한 차분 방법은 (식 2)와 같다.

$d=1$인경우, $\triangle Y_{t=} Y_t - Y_{t-1}$
$d=2$인경우, $\triangle^2 Y_t = \triangle Y_t - \triangle Y_{t-1}$ ———(식 2)

추정한 ARIMA 모형으로 미래의 수요를 예측하기 위해서는 5단계 과정을 거쳐야 한다. 구체적으로 데이터 준비(Data preparation), 모형선택(Model Selection), 추정(Estimation), 진단(Diagnostics)의 과정을 거쳐야 마지막 단계인 예측(Forecasting)을 실행할 수 있다. 본 연구에서는 위 5단계를 통해 추정한 ARIMA 모형을 활용하여 주별 물동량의 시계열 예측치를 도출하였다. 이후 도출한 예측치를 주별 예측을 위한 CART(Classification And Regression Tree) 모형의 설명변수로 사용하였다.

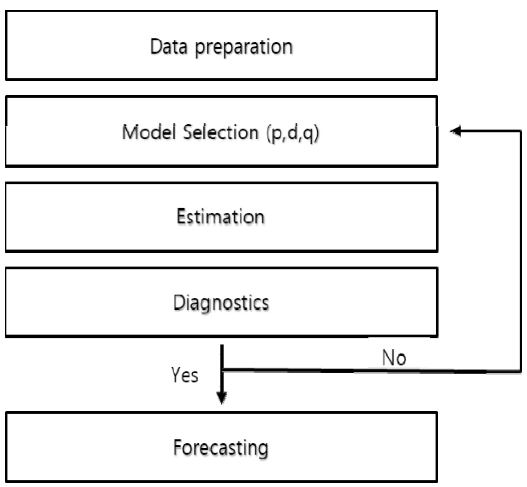


그림 1. ARIMA 모형 순서도

2. CART 모형

CART(Classification And Regression Tree)는 종속 변수가 범주형 혹은 연속형인지에 따라 분류나무와 회귀나무로 구분되는 비모수 의사결정나무 학습법이다. CART 모형은 대표적인 머신러닝(Machine Learning) 기법인 의사결정나무(Decision Tree)의 한 종류이며 직관적인 모형 구조와 해석이 쉽다는 장점으로 주목받고 있다. 실제로 의사결정나무는 자료 분류 및 세분화, 품질관리, 수요예측 등 다양한 분야에서 해석의 용이함과 강력한 분류 및 예측 성능을 인정받고 있다.

CART 모형의 추정 과정은 4단계로 구성된다. 구체적으로 '나무 형성', '가지치기', '타당성 평가', '해석 및 예측' 단계를 통해 적합한 모형을 추정할 수 있다. 첫 번째 단계인 '나무 형성' 단계는 '분리기준(split criterion)'을 이용하여 나무 구조를 형성하는 단계이다. 이때 분리기준이란 나무 구조를 구성하는 '부모마디(Parent Node)'에서 '자식마디(Child Node)'가 갈라져 나오는 기준을 의미하며 순수도(Purity)를 바탕으로 결정한다. 순수도는 일반적으로 지니계수(Gini Index)와 엔트로피 지수(Entropy Index)를 사용하여 측정한다. 두 지표 모두 분리된 그룹의 순수도가 높아질수록 그 값이 작아지며 분리된 그룹 내 모든 자료의 특성이 동일할 경우 계수 값이 0이 된다. 지니계수와 엔트로피 지수의 계산식은 (식 3)과 같다.

$$Gini\ Index(i) = \sum_{i=1}^{n} p(i)(1-p(i)) = 1 - \sum_{i=1}^{n} (\frac{n_i}{n})^2 \quad --(식3)$$

$$Entropy\ Index(i) = -\sum_{i=1}^{n} p(i)\log_2 p(i) = -\sum_{i=1}^{n} (\frac{n_i}{n})\log_2(\frac{n_i}{n})$$

나무 구조를 형성한 후에는 모형의 과적합 문제(Over-fitting Problem)를 해결하고 오차(Error)를 방지하기 위해 '가지치기(Pruning)'을 진행한다. 과적합이란 모형이 학습데이터를 과도하게 학습한 경우를 의미한다. 과적합 모형은 실제 데이터의 부분집합인 학습데이터에서는 오차율이 매우 낮지만 실제 데이터에서는 오히려 오차율이 증가한다. CART 모형은 과적합 문제를 방지하기 위해 모형의 오차율과 복잡도(Complexity)를 동시에 고려한 가지치기를 진행하고 최종 나무 구조의 깊이(Depth)를 결정한다. 일반적으로 가지치기는 비용함수를 통해 비용이 최소화되는 수준까지 진행한다. 비용함수의 계산식은 (식 4)와 같다.

$$Cost(T) = Err(T) + \alpha \cdot N(T) \quad ---- (식4)$$
$$= \sum_{i=1}^{N(T)} (y_i - \hat{y}_i)^2 + \alpha \cdot N(T)$$

$Err(T)$: 오분류율
$N(T)$: 최종 가지 ($Terminal\ Node$) 수
α : 가중치

가지치기를 마친 나무모형은 분류 및 예측에 활용하기 전에 타당성 평가를 거친다. 타당성 평가란 이익도표(Gains Chart), 위험도표(Risk Chart) 또는 검증용 자료(Test Data)에 의한 교차타당성(Cross Validation) 등을 활용하여 의사결정나무를 평가하는 과정을 의미한다. 최종적으로 타당성 평가를 통과한 모형은 분석에 사용할 수 있다. 본 연구에서도 마찬가지로 주단위 예측과 일단위 예측을 위한 모형을 추정하기위해 나무 형성, 가지치기, 타당성 평가를 시행하였고 최종적으로 추정한 CART 모형을 적용하여 예측을 진행하였다.

3. 최종 단기 물동량 예측 모형

본 연구에서는 2단계로 구성된 일별 물동량 예측 모형을 제시하였다. 1단계는 주별 물동량 예측 모형을 추정하는 단계이다. 주별 예측 모형은 앞서 언급한 CART 모형을 활용하여 추정하였으며 설명변수로 ARIMA 예측치와 주요 교역국의 주당 근로일수를 함께 사용하였다. 주요 교역국은 2021년 부

산항 물동량을 기준으로 최대 교역국인 중국, 미국, 일본을 선정하였다. 다음으로 2단계에서는 1단계에서 도출한 예측치와 요일 정보, 주요국 공휴일 정보, 주요국 행사 기간 정보를 설명변수로 활용하여 최종적인 단기 항만 물동량 예측 모형을 추정하였다. 본 연구에서 제시하는 단기 항만 물동량 예측 모형을 그림으로 단순화하여 표시하면 [그림 2]와 같다.

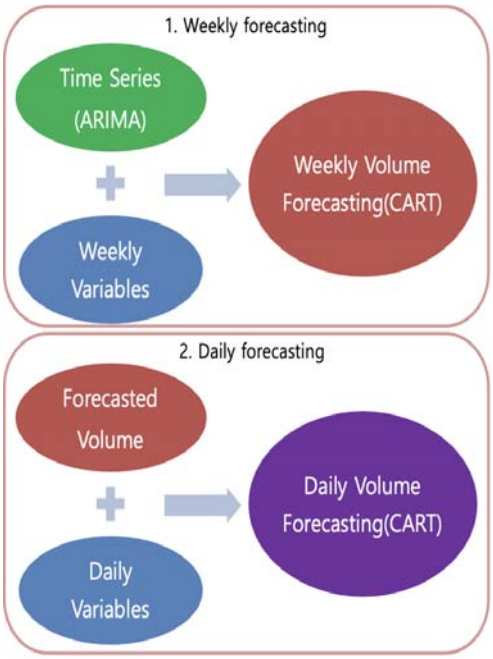


그림 2. 단기 항만 물동량 예측 모형

Ⅳ. 실증분석 및 결과

본 연구에서는 제시한 단기 물동량 예측 모형의 우수성을 검증하기 위해 실제 부산항 데이터를 활용한 물동량 예측을 진행하였다. 활용한 데이터는 2005년 1월 1일부터 2020년 12월 31일까지 집계한 16년치 부산항 일별 물동량 자료이다. 이중 2005년 1월 1일부터 2020년 9월까지의 데이터를 모형 추정을 위한 학습데이터(Training Data)로 사용하였고 추정한 모형을 바탕으로 2020년 10월부터 12월까지 총 92일의 물동량을 예측하였다. 앞서 언급한대로 제시 모형의 추정 및 예측은 시계열 예측(ARIMA), 주별 물동량 예측(CART 1), 일별 물동량 예측(CART 2)으로 구성된다. 본 연구에서는 제시한 모형의 우수성을 입증하기 위해 제시 모형으로 예측한 예측치와 ARIMA 모형으로 예측한 예측치의 정확도를 비교 검증하고 이를 바탕으로 결론 및 시사점을 제시하였다.

1. 시계열 예측치 도출

본 연구에서 제시한 단기 물동량 예측 모형은 주별 물동량 예측, 일별 물동량 예측 2단계로 구성된다. 주별 물동량 예측은 시계열 예측치와 주요국 주당 근로일수를 설명변수로 가진다. 시계열 예측을 위해서는 ARIMA 모형을 사용했으며 데이터 준비(Data preparation), 모형선택(Model Selection), 추정(Estimation), 진단(Diagnostics) 과정을 거쳐 최종 ARIMA 모형을 추정하였다. 시계열 예측치를 얻기 위해 사용한 데이터는 부산항의 2005년 1월 1일부터 2020년 9월까지의 주단위 시계열 데이터이다. 자료는 1기 분기 차분을 통해 시계열의 안정성 조건을 만족시켰다. 안정화된 시계열 데이터의 형태는 [그림 4]와 같다.

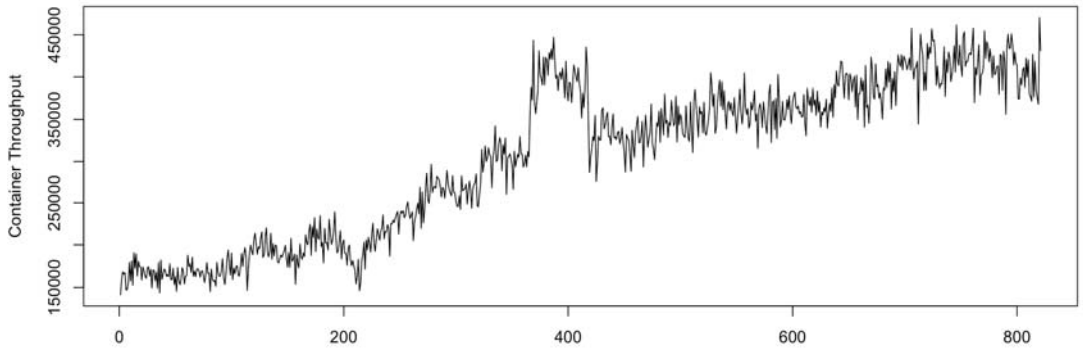


그림 3. 부산항 물동량 원시계열 그래프

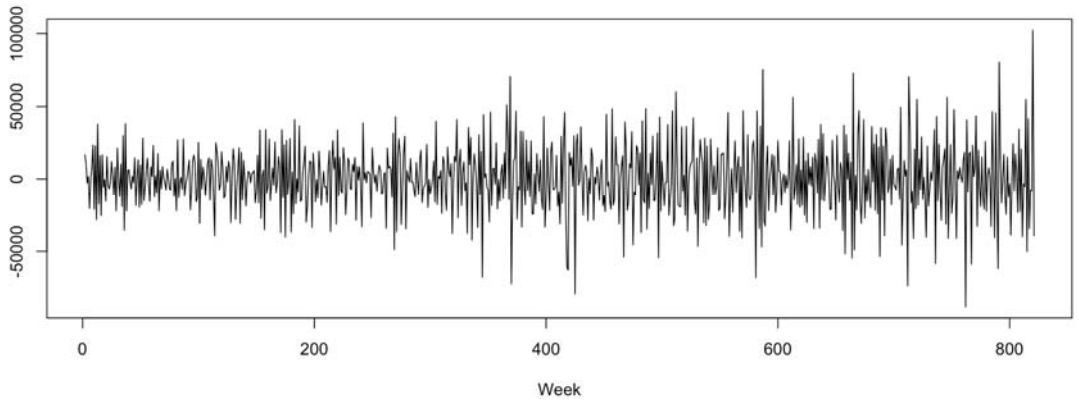


그림 4. 분기 차분한 부산항 물동량 시계열 그래프

자료를 살펴보면 [그림 3]의 원시계열은 시간에 따른 증가 추세가 확연했지만 [그림 4]의 시계열은 평균이 일정한 안정적인 시계열임을 확인할 수 있다. 정확한 판단을 위해 시계열의 안정성을 검증하는 대표적인 단위근(unit root) 검정법인 DF(Dickey Fuller) 검정을 시행한 결과 p-value가 0.01로 도출되어 '시계열 자료가 불안정하다'는 귀무가설을 기각하였다. ARIMA(p,d,q) 모형의 최종 형태는 각 항의 차수인 'p', 'd', 'q' 의 값에 따라 결정된다. 단위근 검정을 통해 1기 분기 차분된 시계열 자료가 안정성 조건을 만족함을 확인하였으므로 'd' 값은 '1'임을 알 수 있다.

AR 항의 차수인 'p'와 MA 항의 차수인 'q' 는 시계열 자료의 자기상관(Autocorrelation)과 편자기상관(Partial-Autocorrelation)을 기준으로 결정한다. [표 1]은 안정화된 시계열의 AC와 PAC를 나타낸 표이다. 표를 보면 AC는 2차(0.005)부터 0에 가까운 값을 가지므로 AR 항의 차수인 'p'는 2 이하의 값을 가진다. PAC는 5차(-0.054)에 값이 큰 폭으로 감소하는 절단 점을 가진다. 따라서 'q'는 5 이하의 값을 가진다고 추정할 수 있다. 따라서 'p'와 'q'의 범위를 [식 5]와 같이 정의할 수 있다.

$$\begin{matrix} p: 0 \le p \le 2 \\ q: 0 \le q \le 5 \end{matrix} \quad \text{———————} \quad (\text{식 } 5)$$

표 1. 안정화된 시계열의 AC 와 PAC

Lag	AC	PAC
0	1.000	-
1	-0.444	-0.044
2	0.005	-0.240
3	-0.100	-0.269
4	0.072	-0.153
5	0.027	-0.054
6	-0.026	-0.054
7	-0.061	-0.019
8	0.064	-0.039
9	0.047	-0.053

위의 (식 5)를 만족하는 ARIMA 모형은 총 18개가 존재한다. 18개의 후보 모형들은 모두 수요예측에 활용할 수 있는 모형이지만 모형별로 담고 있는 정보의 양이 다르기 때문에 정보 기준(Information criteria)을 활용하여 최적의 모형을 선택한다. 정보 기준은 모형의 잔차 크기가 작을수록 모형의 설명력이 뛰어나다는 점을 이용한 선택 기준으로 그 값이 작을수록 활용 가능한 정보의 양이 많다는 것을 의미한다. 일반적으로 가장 많이 사용되는 정보 기준은 AIC(Akaike information Criteria)와 BIC(Bayesian Information Criterion)가 있다. AIC 와 BIC의 추정식은 [식 6]과 같다.

$$AIC = -2 \times \ln(likelihood) + 2 \times k$$
$$BIC = -2 \times \ln(likelihood) + \ln(N) \times k \quad \text{---- (식 6)}$$
k : 모수의 수
N: 관측 수

본 연구에서는 AIC값을 기준으로 후보 모형들을 비교하여 ARIMA(1,1,1)을 선택하였다. 이후 마지막으로 모형의 잔차항(Residual)이 자기상관(Autocorrelation)을 가지는지를 확인하기 위해 잔차검증을 수행하였다. 검증 결과 ARIMA(1,1,1)의 잔차항들은 자기상관을 갖지 않는 백색잡음임을 확인할 수 있었으며, 따라서 최종 예측 모형으로 결정하고 예측을 진행하였다. 예측한 시계열 예측치는 앞서 언급한 대로 다음 단계인 CART 모형을 활용한 주별 물동량 예측의 설명변수로 활용한다.

표 2. 주별 물동량 예측모형 설명변수

설명변수	비고
시계열 예측치	ARIMA 모형 활용
한국 주당 근로일수	신정, 설날, 3·1절, 석가탄신일, 어린이날, 현충일, 광복절, 추석, 개천절, 한글날, 크리스마스, 선거일
중국 주당 근로일수	신정, 춘절, 청명절, 노동절, 현충일, 단오절, 중추절, 국경절, 대체공휴일
일본 주당 근로일수	설날, 성인의 날, 건국기념일, 천황탄생일, 춘분의 날, 쇼와의 날, 헌법 기념일, 녹색의 날, 어린이날, 바다의 날, 산의 날, 경로의 날, 국민의 휴일, 추분의 날, 체육의 날, 문화의 날, 근로감사의 날
미국 주당 근로일수	신정, 마틴 루터 킹 주니어 탄생일, 대통령 취임식, 워싱턴 탄생일, 메모리얼 데이, 노예해방기념일, 독립기념일, 노동절, 콜럼버스의 날, 군인의 날, 추수감사절, 크리스마스

2. 주별 물동량 예측

다음 단계인 주별 물동량 예측에서는 앞서 예측한 시계열 예측치와 더불어 한국, 중국, 미국, 일본의 주당 근로일수를 설명변수로 가지는 CART 모형을 추정한다. 주당 근로일수는 국가별 16년 치 공휴일 일자를 정리하여 반영하였다. 주당 근로일수에 반영한 국가별 공휴일 정보는 [표 2]와 같다.

CART 모형은 그룹의 순수도(Purity)를 기준으로 학습데이터를 반복적으로 분리하는 작업을 통해 나무 구조를 형성한다. 이때 중요한 것은 나무 구조의 복잡도(Complexity)와 비례하는 끝마디(Terminal Node)의 수를 적절하게 조절하는 것이다. 나무 구조의 복잡도가 증가할수록 모형의 오류가 감소하여 정확도가 증가한다. 하지만 과도하게 학습시킨 모형은 학습데이터(Training Data)에서는 성능이 매우 뛰어나지만 실제 데이터를 투입할 때 오히려 정확도가 현저히 떨어지는 과적합(Overfitting) 문제가 발생할 수 있다. 따라서 의사결정나무 분석에서는 복잡도와 모형의 오류를 동시에 고려하는 비용함수를 활용해서 나무의 크기를 조절하는 가지치기(Pruning)를 수행한다.

대표적인 가지치기 방법은 2가지가 있다. 구체적으로 나무 구조를 형성하는 과정 중에 부모노드에서 자식노드가 나누어지는 분기마다 비용함수를 계산하는 방식과 먼저 나무 구조를 최대크기로 형성한 뒤에 다시 끝마디(Terminal Node)부터 뿌리마디(Root Node)까지 거슬러 오르며 비용함수를 계산하여 가지치기를 진행하는 방법이 있다. 본 연구에서는 두 번째 방식을 사용하여 가지치기를 진행하였다.

[그림 5]는 주별 예측모형을 추정하기 위해 나무 구조를 최대로 성장시키고 다시 끝마디부터 뿌리마디까지 분기별로 비용함수를 계산한 결과를 나타낸 그래프이다. 주별 예측모형의 나무 구조를 최대로 성장시키면 끝마디는 총 807개이다. 이는 비용함수에서 나무 구조의 복잡성에 대한 페널티를 의미하는 'α' 값을 '0'으로 설정하는 것과 같은 의미이다. 끝마디가 807개인 나무 구조에 학습데이터를 입력하면 모형의 오류는 0에 가깝게 나와 예측 정확도가 100%에 수렴함을 알 수 있다. 이 경우에는 앞서 언급한 과적합 문제가 발생하므로 모형의 정확도와 복잡성을 동시에 고려하여 가지치기를 진행한다. 그래프를 보면 나무의 크기(Size of tree)가 13일 때를 기점으로 모형의 복잡성이 증가해도 더 이상 모형의 오류가 감소하지 않고 오히려 증가하는 모습을 확인할 수 있다. 따라서 적합한 나무 크기는 끝마디가 13개라고 판단할 수 있다. 최종적인 주별 예측 모형은 [그림 6]과 같다.

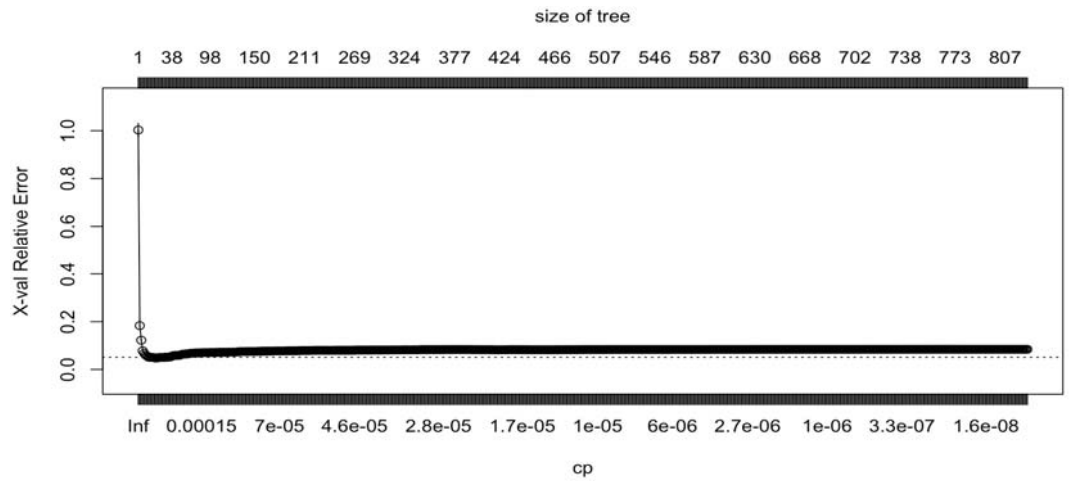


그림 5. 주별 물동량 예측 모형 비용함수 그래프

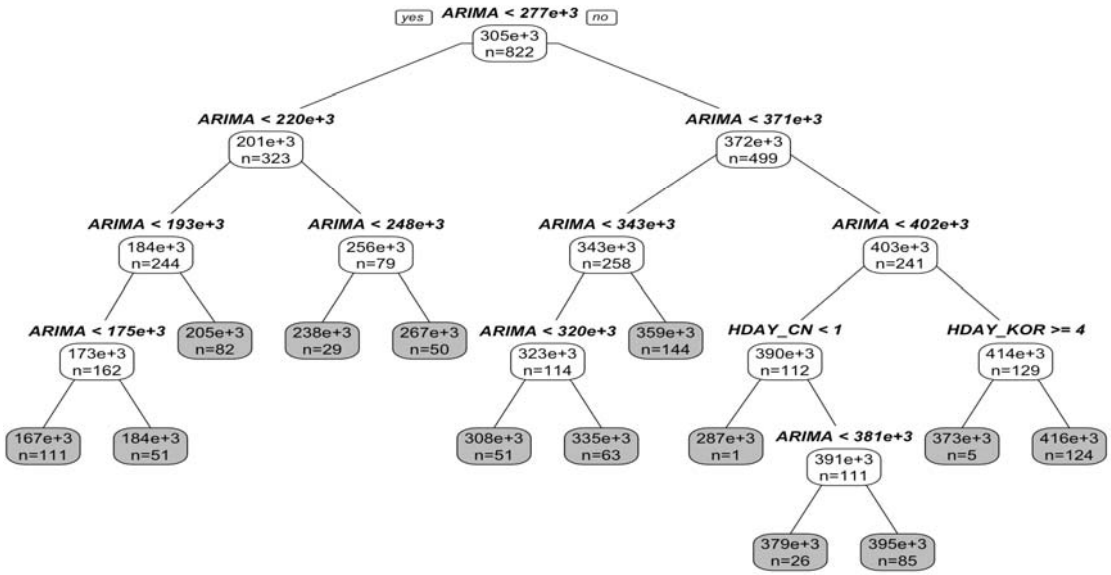


그림 6. 주별 물동량 예측모형 나무 구조

3. 일별 물동량 예측

다음 단계에서는 최종적인 일별 물동량 예측을 위한 두 번째 CART 모형을 추정한다. 2단계에서는 1단계에서 도출한 주별 예측치와 요일 정보, 주요국 공휴일 정보, 주요국 행사 기간 정보를 설명변수로 사용한다. 요일 정보는 요일에 따른 물동량의 증감 추이를 반영하기 위해 월요일부터 일요일까지 7개 명목변수로 반영하였다. 공휴일 정보는 1단계와 마찬가지로 한국, 중국, 미국, 일본의 공휴일을 고려하였다. 1단계에서는 주요국의 공휴일을 주별로 계산하여 주당 노동 일수를 구하고 이를 변수로 사용했지만 2단계에서는 해당 일자의 공휴일 여부를 설명변수로 사용하였다.

주요국 행사 기간 정보는 대규모 할인 행사가 진행되는 기간을 의미한다. 일반적으로 크리스마스와 같은 행사 기간에는 대규모 구매가 발생하여 항만 물동량이 증가하는 경향이 있다. 본 연구에서는 국가별 대표 행사 기간을 측정하여 설명변수로 사용하였다. 구체적으로 크리스마스(Christmas), 블랙프라이데이(Black Friday), 광군제(光棍節), 골든위크(Golden Week) 기간을 연도별로 측정하여 국가별 설명변수로 반영하였다. 일별 예측 모형에서 고려한 설명변수는 [표 3]과 같다.

표 3. 일별 물동량 예측모형 설명변수

설명변수	비고
1단계 예측치	CART 모형 활용
요일 정보	월, 화, 수, 목, 금, 토, 일
주요국 공휴일 정보	한국, 중국, 미국, 일본
주요국 행사 기간	크리스마스(Christmas) 블랙프라이데이(Black Friday) 광군제(光棍節) 골든위크(Golden Week)

앞서 설명한 설명변수를 활용하여 추정한 일별 예측을 위한 나무 구조는 [그림 7]과 같다. 일별 예측 모형 역시 앞서 주별 예측 모형 추정 시 진행했

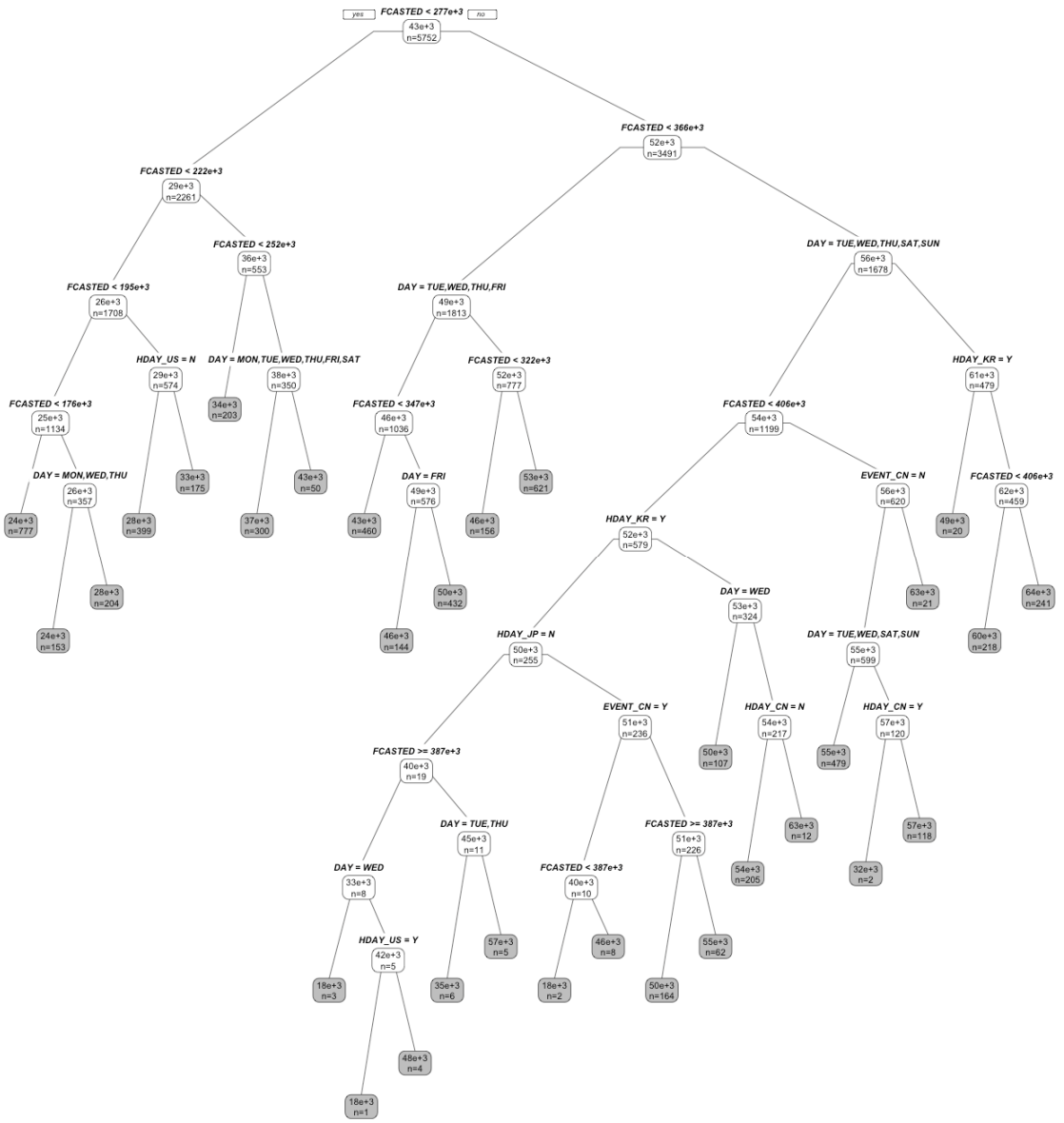


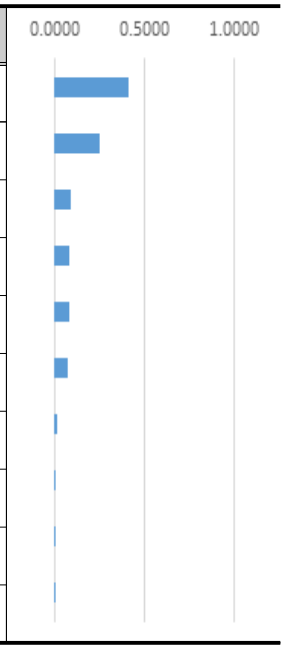
그림 7. 일별 물동량 예측모형 나무 구조

던 절차와 마찬가지로 나무 구조의 복잡성에 대한 페널티인 'α' 값을 '0'으로 설정하여 나무를 최대로 성장시킨 뒤, 비용함수를 통해 모형의 복잡도와 모형의 정확도를 동시에 고려하여 가지치기를 진행하였다. 그 결과 32개의 끝마디를 가진 최종 나무 구조를 추정하였고 최종적으로 예측에 사용하였다.

최종 나무 구조는 학습데이터의 개별 일자 데이터가 어떤 특성을 가지는지 각각 확인하고 그 특성을 분석하여 예측을 수행한다. 구체적으로 해당 일자의 요일, 주요국 공휴일 여부, 특정 행사 기간 포함 여부와 더불어 해당 일자가 포함된 주차의 예측 물동량을 기준으로 학습데이터를 분석하고 그 결과를 예측 대상의 특성에 대입하여 결과를 도출한다. 나무 구조는 중요도가 높은 변수를 뿌리노드에 가깝게 배치하고 중요도가 적은 변수를 끝마디에 가깝게 배치하여 효율적으로 모형을 추정한다. 최종 나무 구조를 구성하는 변수의 중요도는 [표 4]와 같다.

표 4. 일별 물동량 예측모형 설명변수 중요도

변수	중요도
주별 예측치	0.4145
요일	0.2466
공휴일(US)	0.0858
공휴일(KR)	0.0845
공휴일(JP)	0.0796
공휴일(CN)	0.0748
행사(CN)	0.0134
행사(JP)	0.0003
행사(KR)	0.0003
행사(US)	0.0003



설명변수들에 대한 변수 중요도를 살펴보면 주별 예측치의 중요도가 0.4145로 가장 높게 나타났고 그 뒤로 요일(0.2466), 미국 공휴일 여부(0.0858), 한국 공휴일 여부(0.0845), 일본 공휴일 여부(0.0796), 중국 공휴일 여부(0.0784), 중국 행사 기간(0.0134), 일본 행사 기간(0.0003), 한국 행사 기간(0.003), 미국 행사 기간(0.003) 순서로 나타났다. 주별 예측치는 해당 시점의 전체적인 물동량 추이를 결정하는 변수이므로 가장 높은 중요도를 가지는 것으로 판단된다. 주별 예측치를 바탕으로 해당 시점의 물동량이 정해지면 이후 개별적인 날짜의 요일 및 주요국 공휴일 여부에 따라 물동량이 결정되는 것으로 판단된다. 행사 기간 여부는 상대적으로 낮은 중요도를 가진 것으로 나타났으며 유일하게 중국의 행사 기간이 유효한 영향력을 가지는 것으로 나타났다. 이는 실증분석에서 예측한 기간인 4분기(10월, 11월, 12월)에 중국의 대규모 행사 기간인 광군제(11.11)가 포함되어 있기 때문으로 판단된다. 따라서 크리스마스가 포함된 1분기나 골든위크가 포함된 2분기를 예측한다면 다른 주요국의 행사 기간의 영향력이 증가할 것으로 예상된다.

본 연구에서는 추정한 최종 나무 구조를 사용하여 미래 92일에 대한 부산항 일별 물동량 예측을 진행하였다. 또한, 제시한 모형의 우수성을 검증하기 위해 전통적인 시계열 예측 방법으로 동일 기간을 예측하여 예측 정확도 비교검증을 시행하였다. 일별 시계열 예측은 대표적인 예측 방법론인 ARIMA 모형을 적용하였다. 예측을 위해 앞서 언급한 모형 추정을 위한 단계인 데이터 준비(Data preparation), 모형선택(Model Selection), 추정(Estimation), 진단(Diagnostics)을 거쳐 최종적으로 ARIMA(2,1,1) 모형을 추정하고 일별 예측을 진행하였다. 본 연구에서 제시한 예측모형으로 예측한 결과와 ARIMA(Auto regressive Integrated Moving Average) 모형의 예측 결과, 그리고 해당 기간의 실제 부산항 일별 물동량을 비교한 결과는 [표 5]와 같다.

표 5. 부산항 일별 물동량 예측 결과

기간	실측치	ARIMA	CART	기간	실측치	ARIMA	CART
D+1	25094	60554	35278	D+47	61388	58231	63899
D+2	39886	60209	49180	D+48	55015	58231	62980
D+3	46610	60590	36895	D+49	70775	58231	62980
D+4	44837	61425	37910	D+50	47745	58231	62980
D+5	33333	58895	33849	D+51	64582	58231	63899
D+6	27776	56928	33849	D+52	68379	58231	68648
D+7	24085	57349	33849	D+53	67345	58231	70538
D+8	17622	57775	33849	D+54	54878	58231	63899
D+9	43580	57924	33849	D+55	63665	58231	62980
D+10	59004	58380	58296	D+56	59910	58231	62980
D+11	58891	58605	59901	D+57	65582	58231	57381
D+12	45690	58407	53483	D+58	54635	58231	63899
D+13	55740	58250	49566	D+59	74160	58231	59954
D+14	56814	58207	49566	D+60	54519	58231	61604
D+15	46387	58159	49566	D+61	64060	58231	63899
D+16	57273	58159	46433	D+62	61434	58231	55003
D+17	57234	58216	59954	D+63	57041	58231	55003
D+18	72945	58248	61604	D+64	50886	58231	57381
D+19	58047	58249	63899	D+65	67931	58231	63899
D+20	57819	58246	55003	D+66	60542	58231	59954
D+21	70876	58239	55003	D+67	76598	58231	61604
D+22	54249	58228	57381	D+68	53612	58231	63899
D+23	64710	58224	63899	D+69	66359	58231	55003
D+24	64369	58227	59954	D+70	57691	58231	55003
D+25	56590	58229	61604	D+71	61024	58231	57381
D+26	62693	58231	63899	D+72	48399	58231	63899
D+27	55845	58233	55003	D+73	72394	58231	59954
D+28	60230	58232	55003	D+74	68562	58231	61604
D+29	69790	58231	57381	D+75	61827	58231	63899
D+30	59055	58231	63899	D+76	47414	58231	55003
D+31	73199	58231	59954	D+77	63524	58231	55003
D+32	62875	58231	61604	D+78	57952	58231	57381
D+33	67314	58231	63899	D+79	55241	58231	63899
D+34	65931	58231	55003	D+80	59540	58231	59954
D+35	67414	58231	55003	D+81	68521	58231	61604
D+36	59725	58231	57381	D+82	64615	58231	63899
D+37	75537	58231	63899	D+83	55216	58231	55003
D+38	61336	58231	59954	D+84	59090	58231	55003
D+39	58001	58231	61604	D+85	52288	58231	57381
D+40	66722	58231	63899	D+86	62851	58231	49180
D+41	59527	58231	55003	D+87	53274	58231	54052
D+42	57822	58231	62980	D+88	68882	58231	55540
D+43	65058	58231	62980	D+89	50023	58231	59690
D+44	68033	58231	63899	D+90	56227	58231	53582
D+45	69871	58231	68648	D+91	39621	58231	50475
D+46	63339	58231	70538	D+92	58207	58231	53582

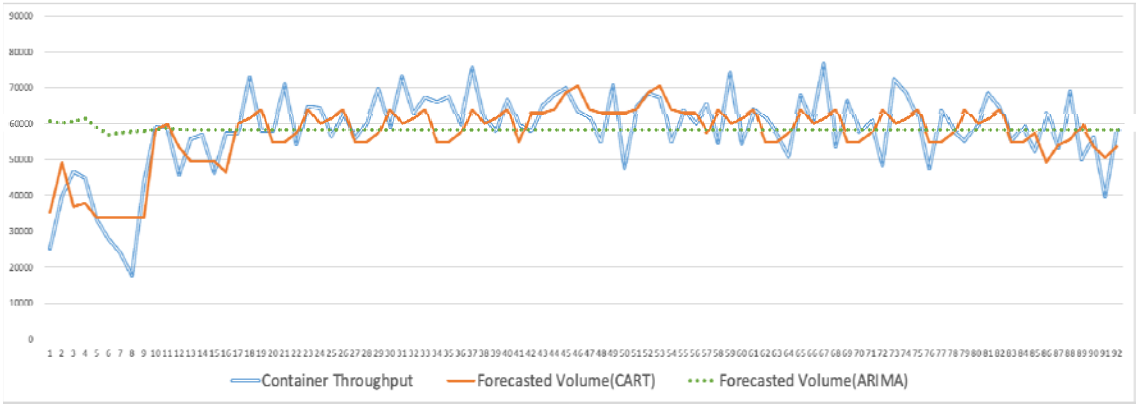


그림 8. 부산항 일별 물동량 예측 결과 비교

[그림 8]은 2020년 10월 01일부터 12월 31일까지 총 92일의 부산항 일별 물동량을 예측한 결과이다. 청색 복선은 해당 기간의 부산항 물동량 실측치를 나타내며 황색 실선은 CART 모형으로 예측한 예측치, 녹색 점선은 ARIMA로 예측한 예측치를 나타낸다. 예측한 기간의 부산항 실제 물동량은 날짜에 따라 최저 20000(TEU)에서 최대 80000(TEU)의 실측치를 기록하였다. 일별로 편차가 심한 부산항의 물동량을 정확하게 예측하는 것은 부산항의 운영효율성과 직결된다. 전통적인 시계열 모형인 ARIMA 모형은 일별 예측의 정확도가 매우 떨어지며 물동량의 증감 추이를 제대로 반영하지 못하는 모습을 확인할 수 있다. ARIMA 모형은 실제 물동량의 평균값에 근접한 예측치를 도출했을 뿐 일별 물동량의 변화 추이에 민감하게 반응하지 못했다. 이는 ARIMA 모형이 과거 시계열의 정보를 바탕으로 추세에 기반한 예측을 진행하는 모형이라는 특징 때문으로 판단할 수 있다.

반면에 CART 모형으로 예측한 예측치는 실측치에 상당히 근접한 정확한 예측치임을 알 수 있다. 실측치와 CART 예측치의 증감 추이를 보면 대부분 구간에서 유사한 증감 추세를 나타냄을 확인할 수 있다. 특히 예측 대상 기간의 초반인 10월 1주차(D+1~D+7) 구간을 살펴보면 갑작스러운 물동량의 하락을 전혀 따라가지 못한 ARIMA 모형에 비해 CART 모형은 물동량의 급격한 감소까지도 예측하는 모습을 보인다.

표 6. 일별 물동량 예측 정확도 비교

지표	ARIMA	CART	개선율
MSE	75040996	58120253	22.5%
RMSE	8663	7624	12%
MAPE	18.08%	11.85%	6.23%

[표 6]은 예측 기간 전반의 예측 정확도를 확인하기 위해 대표적인 정확도 측정 지표인 MSE(Mean squared error), RMSE(Root mean square error), MAPE(Mean absolute percentage error)를 ARIMA 모형의 예측치와 CART 모형의 예측치를 대상으로 측정한 결과이다. ARIMA 모형과 CART 모형의 RMSE를 비교해본 결과 CART 모형을 활용했을 때 RMSE가 22.5% 증가하는 것을 확인할 수 있었다. 또한, 두 모형의 예측 오차를 컨테이너 단위(TEU) 단위로 비교해본 결과 CART 모형으로 물동량 예측을 진행하면 ARIMA 모형으로 예측할

때 비해 166,504(TEU)의 오차를 줄일 수 있는 것으로 나타났다. 이러한 결과로 보았을 때 본 연구에서 제시한 모형을 일별 항만 물동량 예측에 적용 시 예측 오차 감소 및 운영의 효율성 증대에 도움이 될 것으로 기대된다.

V. 결 론

본 연구에서는 항만의 단기 물동량을 예측하기 위해 ARIMA 모형과 CART 모형을 활용한 단기 수요예측 모형을 제시하였다. 제시한 모형은 2단계로 구성된다. 1단계에서는 시계열 예측치와 주요 교역국의 주당 근로일수를 변수로 사용하여 CART 모형을 추정하고 주별 물동량 예측을 진행한다. 2단계에서는 1단계에서 도출한 예측치와 요일 정보, 주요국 공휴일 정보, 주요국 행사 기간 정보를 설명변수로 활용하여 최종적인 일별 물동량 예측 모형을 추정한다. 제시한 수요예측 모형을 활용하여 2020년 10월 1일부터 12월 31일까지 92일의 부산항 물동량을 예측한 결과 제시한 모형의 평균 정확도(MSE)가 기존 시계열 모형보다 '22.5%' 높은 것으로 나타났다. 제시 모형은 일별 물동량의 추세뿐만 아니라 물동량이 급락하는 지점에서도 높은 정확도를 보였다. 또한, 제시 모형을 활용하면 시계열 예측 모형을 사용했을 때 비해 총 166,504(TEU)의 오차를 줄일 수 있는 것으로 나타났다.

일반적으로 수요예측은 예측 시점으로부터 시차가 길어질수록 정확도가 떨어지는 경향을 보인다. 본 연구에서 제시한 단기 물동량 예측 모형을 활용한 실증분석 결과를 보면 예측 시점으로부터 '92'기 떨어진 시점까지도 정확도가 높게 유지됨을 확인할 수 있다. 따라서 제시 모형은 전통적인 수요예측 모형보다 일별 예측에 적합한 것으로 판단된다. 또한, 제시 모형은 외생변수를 설명변수로 포함하지 않는다는 장점이 있다. 외생변수를 설명변수로 포함하는 예측 모형은 설명변수의 정확도에 따라 모형 전체의 정확도가 결정되기 때문에 외생변수의 정확한 예측이 매우 중요하다. 본 연구에서 제시한 예측 모형은 외생변수를 포함하지 않기 때문에 외부 변수에 대한 별도의 예측이 필요하지 않다. 따라서 모형 추정이 복잡하지 않으면서도 안정적인 예측 정확도를 보인다.

항만의 선적과 하역이 원활히 이루어지기 위해서는 필요한 장비와 인력이 적절한 수준으로 배치되어야 한다. 장비와 인력을 적절히 투입하기 위해서는 정확한 일별 물동량 예측이 선행되어야 한다. 시설 및 인력 투입이 부족할 경우 선주의 대기시간 증가로 항만 서비스 수준이 하락하며 반대로 과도하게 투입되면 비용이 증가하여 운영효율성이 떨어진다. 하지만 지금까지는 빅데이터를 다룰 수 있는 방법론과 컴퓨터 연산 능력의 한계로 단기 수요예측을 다룬 연구가 많지 않았다. 이러한 문제가 해결된 시점에서 데이터마이닝 기법을 활용한 일별 수요예측 모형을 제시한 본 연구는 시의적절하며 충분한 활용 가치가 있다고 판단된다.

물동량에 영향을 미치는 변수는 매우 다양하다. 선박 운임, 유가, 교역량 등 다양한 변수가 항만 물동량에 영향을 미친다는 사실이 다양한 선행 연구를 통해 확인되었다. 본 연구에서는 시계열 예측 결과와 공휴일 정보와 같은 내생변수를 사용했지만, 이외에도 다양한 외생변수들을 함께 고려한다면 더욱 정확한 물동량 예측이 가능할 것으로 기대된다. 하지만 다양한 외생변수를 고려하기 위해서는 변수에 대한 추가적인 예측이 필요할 가능성이 있으므로 외생변수 선정 및 정확한 예측에 대한 충분한 근거가 필요할 것으로 판단된다. 또한, 본 연구에서는 예측 대상을 부산항 전체 물동량으로 선정하였다. 제시한 예측 방법을 적용할 때 항만의 각 터미널별 특성을 고려하여 변수를 조정한다면 더욱 정확한 예측 모형을 추정할 수 있을 것으로 판단된다.

참고문헌

김두환 · 이강배(2020), LSTM 을 활용한 부산항 컨테이너 물동량 예측, 한국항만경제학회지, 제 36집 제2호, 53-62.

김창범(2015), 개입 승법계절 ARIMA와 인공신경망모형을 이용한 해상운송 물동량의 예측, 한국항만경제학회지, 제31집 제1호, 69-84.

김종길 · 박지영 · 왕영 · 박성일 · 여기태(2011), Study on forecasting container volume of port using SD and ARIMA(2011), 한국항해항만학회지 제35집 제4호, 343-349.

민경창 · 하헌구(2014), SARIMA 모형을 이용한 우리나라 항만 컨테이너 물동량 예측, 대한교통학회지, 제32집 제6호, 600-614.

손용정 · 김현덕(2012), 의사결정나무분석을 이용한 컨테이너 수출입 물동량 예측, 한국항만경제학회지, 제28집 제4호, 193-207.

이충배 · 노진호(2018), 우리나라와 동아시아 항만간의 수출 컨테이너 물동량 추이 분석, 한국항만경제학회지, 제34집 제2호, 97-113.

여기태 · 정현재(2011), SD 기법에 의한 한 · 중 · 일 환적 물동량 변화량 추정에 관한 연구, 한국항만경제학회지, 제27집, 제4호, 165-185.

하준수 · 나준호 · 조광휘 · 하헌구(2021), 시계열 분석 기반 신뢰구간 추정을 활용한 항만 물동량 이상감지 방안. 한국항만경제학회지, 제37집, 제1호, 179-196.

Chan, H. K., Xu, S., and Qi, X. (2019), A comparison of time series methods for forecasting container throughput, International Journal of Logistics Research and Applications, 22(3), 294-303.

Chen, S., Goo, Y. J. J., & Shen, Z. D. (2014). A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements. The Scientific World Journal, 2014.

Chen, S. H., & Chen, J. N. (2010). Forecasting container throughputs at ports using genetic programming. Expert Systems with Applications, 37(3), 2054-2058.

Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140-1154.

Diaz, R., Talley, W., and Tulpule, M.(2011), Forecasting empty container volumes, The Asian Journal of Shipping and Logistics, 27(2), 217-236.

Farhan, J., and Ong, G. P.(2018), Forecasting seasonal container throughput at international ports using SARIMA models, Maritime Economics & Logistics, 20(1), 131-148.

Liu, C., Hu, Z., Li, Y., & Liu, S. (2017). Forecasting copper prices by decision tree learning. Resources Policy, 52, 427-434.

Patcha, A., & Park, J. M(2007), An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer networks, 51(12), 3448-3470.

Patcha, A., & Park, J. M(2007), An overview of anomaly detection techniques: Existing solutions and latest technological trends, Computer networks, 51(12), 3448-3470.

Rashed, Y., Meersman, H., Van de Voorde, E., and Vanelslander, T.(2017), Short-term forecast of container throughout: an ARIMA-intervention model for the port of Antwerp, Maritime Economics & Logistics, 19(4), 749-764.

Rahmawati, D., & Sarno, R.(2019), Anomaly detection using control flow pattern and fuzzy regression in port container handling, Journal of King Saud University-Computer and Information Sciences.

Schulze, P. M., and Prinz, A.(2009), Forecasting container transshipment in Germany, Applied Economics, 41(22), 2809-2815.

Viglioni, G., Cury, M. V. Q., & da Silva, P. A. L. (2007). Methodology for railway demand forecasting using data mining. In SAS global forum (Vol. 161, No. 2007, pp. 1-8).

Datamining 기법을 활용한 단기 항만 물동량 예측

하준수 · 임채환 · 조광휘 · 하헌구

국문요약

본 연구에서는 항만의 단기 물동량을 예측하기 위해 ARIMA 모형과 CART 모형을 활용한 단기 수요예측 모형을 제시하였다. 제시한 모형은 2단계로 구성된다. 1단계에서는 시계열 예측치와 주요 교역국의 주당 근로일수를 변수로 사용하여 CART 모형을 추정하고 주별 물동량 예측을 진행한다. 2단계에서는 1단계에서 도출한 예측치와 요일 정보, 주요국 공휴일 정보, 주요국 행사 기간 정보를 설명변수로 활용하여 최종적인 일별 물동량 예측 모형을 추정한다. 제시한 수요예측 모형을 활용하여 2020년 10월 1일부터 12월 31일까지 92일의 부산항 물동량을 예측한 결과 제시한 모형의 평균 정확도가 기존 시계열 모형보다 '22.5%' 높은 것으로 나타났다. 제시 모형은 일별 물동량의 추세뿐만 아니라 물동량이 급등락하는 지점에서도 높은 정확도를 보였으며 시계열 예측 모형을 사용했을 때 비해 총 166,504(TEU)의 오차를 줄일 수 있는 것으로 나타났다. 항만의 효율적인 운영을 위해 필수적인 단기 물동량 예측에 적합한 예측 모형을 제시한 본 연구는 충분한 활용 가치가 있을 것으로 판단된다.

주제어: 항만 물동량 예측, 데이터마이닝, CART, 일별 수요예측