

석사학위논문

베이지안 딥러닝을 활용한  
항만물동량 예측에 관한 연구

A Study On Bayesian Deep Learning Based  
Port Throughput Estimation

지도교수 임규건

이 논문을 경영학 석사학위논문으로 제출합니다.

2023년 2월

한양대학교 대학원  
비즈니스 인포매틱스학과

이준희

## 국문 초록

### 베이지안 딥러닝을 활용한 항만물동량 예측에 관한 연구

우리나라의 수출입화물 대부분은 해상운송으로 이루어지며 이는 해운항만물류 사업이 그만큼 우리나라의 경제와 매우 밀접한 관계를 가지는 것을 의미한다. 이 중 항만 컨테이너 물동량은 해운항만 경쟁력을 대표하는 요소 중 하나이다. 최근에는 글로벌 인플레이션 및 공급망 위기 등의 다양한 대내외적 거시경제의 불안정성으로 인하여 물동량 역시 점차 예측이 어려워지는 추세를 보인다고 할 수 있다. 이러한 불규칙적이고 비선형적 특징을 기존의 전통적인 시계열 방법론 대비 잘 반영할 수 있는 딥러닝을 활용한 예측 연구들이 이루어졌으나, 현재까지의 연구들은 모두 One-Step Prediction의 결과를 순차적으로 연결하는 기본적인 구조를 사용하는 데에 그친 상태이다. 이에 따라 본 연구에서는 딥러닝을 활용한 다 시점 예측(Multi-Step Prediction)이 가능한 4개의 Encoder-Decoder 기반 모델(LSTM,CNN-LSTM,TCN,Attention)과 베이지안 딥러닝 방법론 중 하나인 Monte-Carlo Dropout을 활용하여 물동량에 대한 예측의 불확실성을 함께 나타냈다. 데이터는 2000년 1월부터 2022년 6월까지의 총 10개의 다변량 변수를 구축했다. 이 중 누락된 자료는 별도의 추정과 유사한 자료를 통해서 보충했다. 예측 기간은 2017.07~2022.06으로 설정했으며 각각 1년,3년,5년 단위로 분할한 총 8개의 기간에 대한 예측을 수행했다. 먼저, 각 예측 기간당 MC-Dropout 기반으로 출력된 30개의 추정치들이 분포를 형성한다는 점에서 각 모델의 추정치들에 대한 MLE(최대우도추정) 기반의 확률밀도함수 추론을 MC-Dropout 비율 조정을 통해서 수행했다. MC-Dropout 비율은 각 모델마다 0.1에서 0.8까지 조정했으며 전반적으로 가장 높은 0.8의 비율에서 추정치들의 침도와 분산이 실제 물동량에 근사하는 특징을 확인했다. 여기서 MC-Dropout=0.8의 CNN-LSTM구조가 가장 예측에 있어서 적합한 모델로 판정되었으며 이는 4개 모델의 추정치들

과 실제 물동량의 분포에 대한 유사도 측정(KL-Divergence, Negative Log-Likelihood) 결과를 통해서 뒷받침했다. 나아가 각 시점당 추정치들의 분포의 특징을 반영한 평가지표인 CRPS(Continuous Ranked Probability Score)를 기준으로 4개의 예측 모델 간 구체적인 비교를 실시했다. 비교 결과, CRPS가 가장 낮은 CNN-LSTM구조에서 추정치들의 평균과 중앙값의 성능, 표준편차와 분위수 기준의 예측의 불확실성 구간의 적합도(Coverage, Sharpness), 그리고 실제 물동량의 추세 중심의 변동과의 일치 면에서 모두 종합적으로 안정적인 수치를 기록했다. 이 점을 통해서 불확실성 기반의 확률적 예측에 있어서 분포 중심의 성능 평가가 우선시 되어야 함을 확인했다. 나아가 단변량 CNN-LSTM 및 기존의 결정론적 관점에서의 여러 예측 방법론과의 종합 성능 비교, 각 기간당 출력 개수 증가를 실시하여 기존의 실험 결과의 안정성을 재차 검증했다. 이후, 해당 모델을 대상으로 물동량의 계절변동을 활용하여 각 기간당 30개의 MC-Dropout기반 추정치들 중 가장 발생 가능성이 높은 5개의 우선순위 추정치들을 선정했다. 이는 각 기간당 CNN-LSTM구조의 평균과 계절변동 예측값과의 곱셈과 MC-Dropout 추정치들 간의 유사도 측정(JS-Distance, Euclidean Distance, Dynamic Time Warping)을 통해서 수행했다. 실험 결과, 중장기 예측에서의 우선순위 추정치들은 실제로 가장 높은 정확도를 가진 5개의 MC-Dropout 개별 추정치들과 전반적으로 더 높은 조건부확률을 기록하여 교집합 형성에서 단기 예측 대비 더 안정적이었다. 하지만 단기 예측에서의 최고 성능을 가진 값들 역시 실제 물동량 대비 큰 차이를 보이지 않았으며 개별 예측 모델을 수행할 수준의 결과를 보였다. 이를 통해서 MC-Dropout을 통해서 컨테이너 물동량의 전반적인 움직임 중심의 예측의 불확실성에서 나아가 최고 성능을 가질 수 있는 우선순위 추정치들의 집합을 선정하는 것으로 구체적인 예측 계획 수립을 세우는 데에도 충분한 효율성을 가질 수 있는 점을 나타냈다.

**주요어:** Monte-Carlo Dropout, Encoder-Decoder, KL-Divergence, Negative Log-Likelihood, CRPS, 분포 추론, 분포 유사도, 예측의 불확실성, 우선순위 추정치 집합

# ABSTRACT

## A Study on Bayesian Deep Learning Based Port Throughput Estimation

Lee, Jun Hee

Dept. of Business Informatics

Graduate School of

Hanyang University

South Korea's import and export freight is mostly based on sea transportation, which means that maritime and port logistics has a vital relationship with its economy. Port throughput is one of the key elements which can represent both the shipping port and economic competitiveness of the country. Recently, it can be seen that various internal and external macroeconomic based uncertainty is making port throughput estimation more difficult. Deep Learning based studies which can more efficiently handle these kinds of irregular and nonlinear characteristics than traditional time series methods have been making a progress. But until now, deep learning based studies are mostly based on the basic One-Step prediction structure which connects single prediction results among the test period. So accordingly, this study applied four

multi-step based multivariate Encoder-Decoder models (LSTM, CNN-LSTM, TCN, Attention) and Monte-Carlo Dropout which is one of the Bayesian Deep Learning methods to not only just simply predict, but also to show prediction uncertainty as well. 10 multivariate macroeconomic data were collected from 2000.01 to 2022.06. Also, missing values were imputed with simple predictions and using other similar data. The forecasting period is from 2017.07 to 2022.06, and this was divided into 8 specific periods based upon 1, 3, and 5 year prediction units. Since 30 MC-Dropout based multiple prediction results per each period form probability distributions, maximum likelihood estimation of the variables was carried out by adjusting the MC-Dropout rates. By adjusting the rates from 0.1 to 0.8, it was seen that at rate 0.8, the kurtosis and variance of the predicted variables relatively matched with the distributions of the real port throughput values. The CNN-LSTM structure was chosen to be the most appropriate forecasting model and this was reinforced by distribution similarity tests (KL-Divergence, Negative Log-Likelihood) among all models. Further, CRPS (Continuous Ranked Probability Score) which can reflect the probability distribution based characteristic was used as the priority score to specifically compare the goodness among the four models. As a result, CNN-LSTM which has the lowest CRPS, showed the best overall stability of the mean and median values of the prediction results, coverage and sharpness measure of the prediction intervals, and the concordance with the trend of the real port throughput values. According to this, it can be said that in uncertainty based probabilistic forecasting, distribution based measures are to be prioritized. After choosing the best model, further verifications which are comparisons of univariate CNN-LSTM, models from previous studies, and adjusting the MC-Dropout output numbers were done to

double check the model stability. After this progress, 5 priority values were selected from the 30 prediction values per period with the standard of the multiplications of the mean values and the seasonal components of the port throughput data. The similarity measure of the multiplications and the MC-Dropout predictions was conducted with JS-Distance, Euclidean Distance and Dynamic Time Warping. As a result, the priority value sets of the mid and long term (3&5 years) showed more stability than the short term (1 year) prediction results. This was seen with the conditional probabilities of the similarity measure based priority sets which consists of 5 values given the 5 MC-Dropout individual predictions per period which have the least error compared to the real port throughput values. Nonetheless, the best values among the short term priority prediction sets didn't have much difference with the real port throughput values. And it could be seen that they are capable of making themselves a singular prediction model. As a result, this study can mention that not only MC-Dropout based predictions quantify prediction uncertainty in port throughput prediction, but also efficiently lead to a more concrete forest plan by selecting priority values from the entire uncertainty based results.

**Key Words:** Monte Carlo-Dropout, Encoder-Decoder, KL-Divergence, Negative Log-Likelihood, CRPS, Distribution Estimation, Distribution Similarity, Prediction Uncertainty, Priority Prediction Values