KIGAS Vol. 23, No. 5, pp 52~58 2019
(Journal of the Korean Institute of Gas)
https://doi.org/10.7842/kigas.2019.23.5.52

외재적 변수를 이용한 딥러닝 예측 기반의 도시가스 인수량 예측

†김지현 · 김지은 · 박상준 · 박운학
JB주식회사
(2019년 8월 22일 접수, 2019년 10월 25일 수정, 2019년 10월 26일 채택)

Deep Learning Forecast model for City-Gas Acceptance Using Extranoues variable

†Ji-Hyun Kim · Gee-Eun Kim · Sang-Jun Park · Woon-Hak Park
14, Sandongan-gil, Eumbong-myeon, Asan-si, Chungcheongnam-do, Republic of Korea
(Received August 22, 2019; Revised October 25, 2019; Accepted October 26, 2019)

요 약

본 연구에서는 국내 도시가스 인수량에 대한 예측 모델을 개발하였다. 국내의 도시가스 회사는 KOGAS에 차년도 수요를 예측하여 보고해야 하므로 도시가스 인수량 예측은 도시가스 회사에 중요한 사안이다. 도시가스 사용량에 영향을 미치는 요인은 용도구분에 따라 다소 상이하나, 인수량 데이터는 용도별 구분이 어렵기 때문에 특정 용도에 관계없이 영향을 주는 요인으로 외기온도를 고려하여 모델개발을 실시하였다.실험 및 검증은 JB주식회사의 2008년부터 2018년까지 총 11년 치 도시가스 인수량 데이터를 사용하였으며, 전통적인 시계열 분석 중 하나인 ARIMA(Auto-Regressive Integrated Moving Average)와 딥러닝 기법인 LSTM(Long Short-Term Memory)을 이용하여 각각 예측 모델을 구축하고 두 방법의 단점을 최소화하기 위하여 다양한 앙상블(Ensemble) 기법을 사용하였다. 본 연구에서 제안한 일별 예측의 오차율 절댓값 평균은 Ensemble LSTM 기준 0.48%, 월별 예측의 오차율 절댓값 평균은 2.46%, 1년 예측의 오차율 절댓값 평균은 5.24%임을 확인하였다.

Abstract - In this study, we have developed a forecasting model for city- gas acceptance. City-gas corporations have to report about city-gas sale volume next year to KOGAS. So it is a important thing to them. Factors influenced city-gas have differences corresponding to usage classification, however, in city-gas acceptence, it is hard to classificate. So we have considered tha outside temperature as factor that influence regardless of usage classification and the model development was carried out. ARIMA, one of the traditional time series analysis, and LSTM, a deep running technique, were used to construct forecasting models, and various Ensemble techniques were used to minimize the disadvantages of these two methods.Experiments and validation were conducted using data from JB Corp. from 2008 to 2018 for 11 years.The average of the error rate of the daily forecast was 0.48% for Ensemble LSTM, the average of the error rate of the monthly forecast was 2.46% for Ensemble LSTM, And the absolute value of the error rate is 5.24% for Ensemble LSTM.

Key words : city-gas, forecasting demand, LSTM, ARIMA, time-series, acceptance

I. 서 론

한국가스공사(Korea Gas Corporation, 이하 KOGAS)는 도시가스 회사와 수급계약을 체결하고 이를 기반으로 해외 LNG 원산지와 도시가스 수급에 대해 장기계약을 체결하고 있다. 그렇기 때문에 도시가스 수요량 예측은 도시가스 회사에 있어 중요한 사안 중 하나이다.

하지만 도시가스 판매량 데이터는 한 달에

†Corresponding author:kjihyun@jbcorporation.com

한 번 검침을 통해 확인할 수 있으며, 검침 주기에 따라 도시가스 판매량의 변동 가능성이 존재한다. 예를 들면 1월의 검침 일자는 25일, 2월은 28일, 3월은 27일인 경우에 1월과 2월 사이의 검침 주기는 34일이고, 2월과 3월 사이의 검침 주기는 27일이 되므로 월별 도시가스 판매량 데이터의 일관성이 떨어진다고 할 수 있다. 또한, 가스계량기가 고객 주거지 내부에 있어 고객이 직접 계량기 지침 값을 입력하는 경우 임의검침, 미 검침 등의 사유로 고객이 기록한 지침 값이 신뢰도가 있는 데이터라 판단하기 어려우며, 기체인 도시가스의 특성상 온도와 압력에 의해 부피가 증감하는 등 고려해야 하는 변수가 상당수 존재한다. 이에 반해 KOGAS에서 제공하는 도시가스 인수량 데이터는 지능형 검침 인프라(AMI, Advanced Metering Infrastructure)를 통해 일 단위로 기록되며 KOGAS로부터 도시가스를 공급받은 지점의 지침 값을 기록하기 때문에 데이터의 신뢰도와 적합성을 고려해보았을 때 도시가스 인수량 데이터가 더욱 적절할 것으로 판단하였으며, 충남지역 9개 시·군에 도시가스를 공급하고 있는 JB주식회사의 11년치의 도시가스 인수량 데이터를 제공받아 사용하였다.

도시가스 판매량 예측에 대한 기존 연구를 살펴보면 대부분 계절성(seasonality), 온도, 요일을 반영하여 예측을 시행하거나[1], 도시가스 시간대별 판매량은 직전 시간(24시간 전)의 데이터 기반으로 회귀분석을 실시한 것을 볼 수 있었다[2]. 이를 참고하여 JB주식회사로부터 받은 도시가스 인수량 데이터를 분석한 결과, 판매량 데이터와 마찬가지로 계절적 요인으로 인한 1년의 주기성을 보이는 것으로 확인되어(Fig. 1), 기존 연구에서 검토한 것과 같이 온도 변수를 주된 변수로 고려하였다.

또한, 기존 도시가스 판매량 예측에 대해 선행된 연구들은 온도를 독립변수로 설정한 회귀분석(regression) 모형, 누적 자기 회귀이동평균(auto-regressive integrated moving average, ARIMA)모형이나 순환 인공신경망(recurrent neural networks, RNN) 중 하나인 LSTM(long short memory networks) 모형으로 도시가스 판매량 예측을 시행하였다. 그러나 위 모형들은

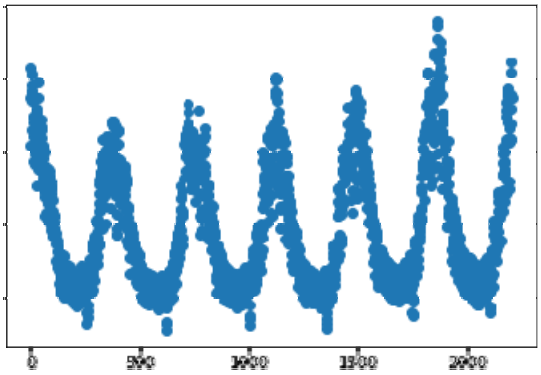

Fig 1. The graph of city-gas acceptance during 6years

추세 또는 온도에 절대적으로 의존하기 때문에 예상 밖의 과거 흐름(일정하지 않은 데이터의 패턴)과 기온 변화에 취약하다는 한계점이 존재한다. 따라서 본 연구에서는 JB 주식회사의 도시가스 인수량을 일(1일), 월(1개월), 1년 단위로 예측하는 것을 목표로하여 과거 추세를 따라 예측하는 것에 중점을 두는 ARIMA 기법과 온도 변수를 반영한 LSTM 기법을 앙상블하여 두 기법의 단점을 최소화함과 동시에 평균 오차율을 줄이는 성능을 보이는 모델을 소개하고자 한다.

Ⅱ. 도시가스 인수량 예측 모형 개발

이 장에서는 도시가스 인수량과 온도로 구성한 데이터 셋을 사용하여 도시가스 인수량 예측 모형을 구현한다. 모형을 구현하기 위해 전통적 통계 기법인 ARIMA와 딥러닝 기법인 LSTM을 사용한다.

2.1. ARIMA

ARIMA(Auto-Regressive Integrated Moving Average) 모형은 과거 데이터 기반으로 현재를 설명하는 시계열 데이터 분석 기법으로, 주로 주가 예측, 수요 예측 등에 사용되는 모델이다. ARMA 모형이 안정적 시계열 (Stationary Series)에만 사용 가능한 것과 비교하면 ARIMA는 비안정적 시계열 (Non Stationary Series)에도 사용할 수 있다는 장점이 있다. ARIMA의 일반

적인 형태는 다음과 같다. Xt는 ARIMA를 통해 예측하고자 하는 데이터 이다.

$$Y_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_t - p \quad (1)$$

2.2. LSTM

딥러닝(Deep Learning)이란 대용량 데이터에서 핵심적인 내용이나 기능을 시도하는 머신러닝(Machine Learning)의 한 분야이다.

딥러닝 구조는 인간의 뇌구조에서 영감을 얻은 인공신경망(ANN, Arificial Neural Networks)에 기반하여 설계되는데 주로 데이터에서 학습을 하여 규칙기반으로 풀기 어려운 컴퓨터 비전이나 음성인식과 같은 다양한 범위에 사용되고 있다.

시계열 데이터처럼 시퀀스가 있는 데이터에 적합한 신경망의 일종으로 순환 인공 신경망(RNN, Recurrent Neural Networks)이 있다. 일반적인 신경망을 Feed-forward neural networks (FFNets)라고 하는데 FFNets은 연산이 입력층(Input Layer)에서 은닉층(Hidden Layer), 출력층(Output Layer)까지 단계적으로 진행되며 데이터들은 모든 노드(Node)를 한번 씩 지나가게 된다. 하지만 RNN은 그 이름의 의미처럼 루프의 구조를 갖고 있고 은닉층의 결과가 다시 같은 은닉층의 입력으로 들어가도록 재귀적으로 움직이는 신경망이다. 이렇듯 데이터의 순서를 고려할 수 있는 RNN은 주로 손글씨, 주가와 같은 시계열 데이터를 처리하는데 자주 활용된다. 하지만 과거의 정보를 가지고 미래 또는 현재의 문제를 다루는 RNN은 관련되어 있는 정보와 그 정보를 사용하려는 지점 사이

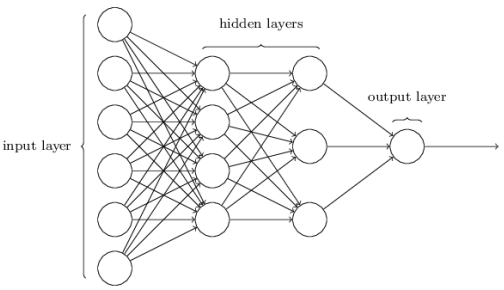

Fig 2. Structure of an ANN

거리가 멀어질수록 성능이 저하되는 장기 의존성(Long Term Dependency) 문제를 가지고 있다. 이 문제를 극복하기 위해서 고안된 알고리즘이 LSTM(Long Short Term Memory)이다. LSTM의 경우 단일 네트워크 레이어를 가지지 않고 RNN의 은닉층에 셀 스테이트(Cell-state)를 추가하여 정보들을 선택적으로 사용할 수 있기 때문에 RNN에 비해 비교적 긴 시퀀스의 입력데이터를 처리하는데 탁월한 성능을 보인다. 본 연구에서는 도시가스 인수량과 같이 비교적 대용량의 데이터에 딥러닝 학습이 효과적일 수 있다는 판단을 하여 LSTM을 활용한 도시가스 인수량 예측을 시도하였다.

2.3. Ensemble LSTM

본 연구에서는 기본 LSTM 모형을 사용하지 않고 LSTM을 여러 번 학습하여 나온 결과들을 Ensemble 하여 예측값을 찾아내는 방법을 제안한다. 본 연구에서 사용한 데이터는 2008년부터 2018년까지 총 11년 치의 자료로, 학습을 시키기에는 다소 데이터의 양이 적었다. 이렇게 적은 양의 데이터로 LSTM 모델을 사용하는 경우 오차값이 현저하게 높아질 수 있기 때문에, 약한 학습(weak learner)을 강한 학습(strong learner)으로 바꿔줄 방법으로 데이터 양이 적을수록 오히려 학습 데이터의 양을 적게 하고 테스트 데이터의 양을 늘려가며 앙상블(Ensemble) 하는 방법을 도입하였다. 단계는 다음과 같다.

STEP1. epoch, batch size 등 하이퍼 파라미터(Hyper parameter)를 조절하여 LSTM 모델링을 실시한다.

STEP2. STEP1의 LSTM 모델을 N 회 반복하여 예측값을 N개 추출한다. 이때, 학습데이터와 테스트 데이터의 비율은 5:5 또는 4:6으로 설정한다.

STEP3. 강한 학습(strong learner)으로 만들기 위해 학습을 N 회 반복하여 예측된 값을 다양한 Ensemble 방법에 적용하여 가장 좋은 예측값을 찾는다.

2.4. Ensemble

본 연구에서는 네 가지의 Ensemble 방법을 이용하였다. 단순 평균인 SA(Simple Average), 검증 세트(Validation set)에서 더 우수한 성능을 낸 모델에 가중치를 주는 OP(Out Performance), MAPE를 기준으로 더 작은 오차 값을 가진 모델의 가중치를 높게 주는 EB(Error-Based), MSE (Mean Squared Error)를 기준으로 더 작은 오차 값 분산을 가진 모델에 더 많은 가중치를 주는 VB(Variance-Based)를 사용하여 성능을 비교하였다.

Ⅲ. 도시가스 데이터 분석

3.1. 회귀분석

본 연구에서는 2008년 1월 1일부터 2018년 12월 31일까지 총 2,192일(11년)의 도시가스 인수량 데이터를 사용하여 일별, 월별, 1년 인수량 예측을 시행하였다.

본 실험에 들어가기 전에 지역별 평균온도를 대상으로 회귀분석을 실시하여 도시가스 인수량과의 상관관계 정도를 나타내는 R Square 값을 계산하였다. JB주식회사 공급권역 내 KOGAS Gas Station이 위치한 지역을 기준으로 '1. 천안-아산, 2. 세종-공주, 3. 논산-부여, 4. 금산, 5. 보령-대천'으로 구분하였으며, 지역별로 단일 회귀분석을 시행하였다. 회귀분석에 대한 결과는 아래와 같다.

실험 결과, 지역별로 평균온도와 도시가스 인수량의 상관관계가 높다는 결과는 도출되었으나, 온도 이외에 도시가스 사용자의 생활패턴이나 주말/공휴일 등 다른 변수들의 영향도 일정 부분 존재함을 JB주식회사의 도시가스 수요량 기록을 통해 확인할 수 있었다. 또한, 지역별로 도시가스 사용 용도(주택용/일반용/산업용 등)의 비중에 따라 온도 변수의 영향을 크게 받지 않는 곳도 존재하였다.

3.2. 자기 회귀(AutoRegressive; AR) 분석

과거의 도시가스 인수량 데이터가 현재 데이터에 어떻게 영향을 주고 있는지 살펴보기 위한 자기 회귀를 시행하였다. 여기서 Order는 자기 회귀 차수, Var. pred는 자기 회귀 모델에 의해 설명되지 않는 시계열의 분산 부분 추정값이다. Table 2에서 확인 할 수 있는 지역별 도시가스 수요의 특징은 다음과 같다.

Table 1. R square for each area

Area	R square
1. Cheonan-Asan	-0.8633
2. Sejong-Gongju	-0.7823
3. Nonsan-Buyeo	-0.7216
4. Geumsan	-0.5986
5. Boryeong-Daecheon	-0.8616

Table 2. Result of AR analysis for each area

Area	Order	Var.pred
1. Cheonan-Asan	29	42012894
2. Sejong-Gongju	30	778103766
3. Nonsan-Buyeo	29	67588154
4. Geumsan	15	11961592
5. Boryeong-Daecheon	29	24786719

(1) R square값이 -0.5986인 금산을 제외한 모든 지사의 도시가스 일일 인수량은 약 한 달 전 (30일)의 수요까지 영향을 준다.

(2) 각 지역의 Var. pred 값을 보았을 때 자기 회귀 모형만으로 설명되지 않는 분산 추정값이 상당히 높다.

Ⅳ. 모형 적용 결과 및 분석

본 연구에서는 ARIMA, Ensemble LSTM, 그리고 두 가지 모델을 앙상블하여 총 세 가지의 방법으로 실험하였다. 온도는 기상청에서 제공하는 지역별 최저온도, 최고온도, 평균온도를 사용하였다.

일별, 월별, 1년 예측 오차율을 평가하기 위해 실제 인수량과 예측값의 차이를 실제 인수량으로 나눈 평균값 MAPE(mean absolute percentage error)를 사용하였다. A_t는 실제 값이며 F_t는 예측값이다.

Table 3. Result of ARIMA MAPE(%) for each area

Area	Daily MAPE	Monthly MAPE	Yearly MAPE
1. Cheonan-Asan v	2.51	7.28	3.9
2. Sejong-Gongju	3.26	0.9	12.38
3. Nonsan-Buyeo	0.88	6.94	9.31
4. Geumsan	8.68	4.93	7.92
5. Boryeong-Daecheon	14.44	7.17	6.95

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (2)$$

Table 3, Table 4는 각각 최적의 하이퍼 파라미터값으로 ARIMA, Ensemble LSTM을 실험하여 일별,월별,1년을 예측한 실험결과이다. Table 5는 ARIMA와 LSTM을 앙상블한 비율을 나타내고 있다. 일별 예측은 오차 값이 3% 이내로 나타나지만 4. 금산 지역의 경우 2011년 이후에 신설되어 데이터의 양이 적기 때문에 다른 지역에 비해 오차 값이 큼을 확인할 수 있다. 또한 비교적 기온의 영향을 적게 받는 산업용의 비율이 높은 지역일수록 오차값이 증가하는 것을 볼 수 있다. 추세만을 반영하는 ARIMA와 외기온도를 함께 반영한 Ensemble LSTM의 오차 값을 비교한 결과, 앞선 회귀분석 결과와 동일하게 도시가스 사용량에 외기온도가 큰 영향을 미친다는 것을 확인하였다. 하지만 1년 예측을 시행하게 되면 5~8월에 해당하는 여름에 대한 오차 값이 ARIMA보다 Ensemble LSTM의 오차 값이 더 커진다. 이는 인수량이 용도 구분(주택/일반/산업용)이 되어 있지 않아 여름철 주택용 도시가스의 낮은 사용량에 대하여 모델의 학습이 정확하게 이루어지지 않는다는 것이다. 이러한 문제를 해결하기 위해 ARIMA와 LSTM의 Ensemble 방법을 시도하였다.

Table 5에서 LSTM Ratio(1)과 Ensemble Method(1)는 월 단위 예측 모델에 대한 Ensemble 옵션을 나타내며, LSTM Ratio(2)과 Ensemble Method(2)는 1년 단위 예측 모델에 대한 옵션을 의미한다.

Table 4. Result of Ensemble LSTM MAPE(%) for each area

Area	Daily MAPE	Monthly MAPE	Yearly MAPE
1. Cheonan-Asan	0.14	1.8	3.8
2. Sejong-Gongju	0.2	3.1	6.5
3. Nonsan-Buyeo	0.8	3.1	5.6
4. Geumsan	1.1	2.1	4.4
5. Boryeong-Daecheon	0.1	2.2	5.9

Table 5. LSTM Ratio(%) and Ensemble met hod for each Ensemble

Area	LSTM Ratio(1)	Ensemble Method(1)	LSTM Ratio(2)	Ensemble Method(2)
1. Cheonan-Asan	13.03	VB	13.03	VB
2. Sejong-Gongju	33.7	VB	54.84	EB
3. Nonsan-Buyeo	43.42	EB	99.99	VB
4. Geumsan	41.95	VB	85.41	VB
5. Boryeong-Daecheon	50.0	SA	83.0	OP

Table 6. Result of Ensemble MAPE(%) for each area

Area	Yearly MAPE
1. Cheonan-Asan	**3.68**
2. Sejong-Gongju	**8.34**
3. Nonsan-Buyeo	**7.30**
4. Geumsan	**6.23**
5. Boryeong-Daecheon	**4.76**

각각의 지역을 1년 예측을 기준으로 4가지 앙상블을 적용한 결과는 Table 6 과 같다. 1년 예측 앙상블 결과는 지역별로 0.2%~1% 정도의 성능 향상을 끌어냄을 확인할 수 있다.

Ⅳ. 결론 및 향후 연구

본 연구는 JB주식회사에서 도시가스를 공급하는 9개 시·군의 과거 도시가스 인수량 실적 데이터 (2008~2018년)를 바탕으로 외기온도와 계절별 특성을 고려하여 일별, 월별, 1년 인수량 예측을 시행하였다. 과거 일정 기간별 도시가스 인수량은 매우 높은 자기 상관성을 갖는데 일별 단위의 경우 주로 직전 과거 24시간 전, 7일 전, 10일 전까지에 영향을 주며, 월별 단위의 경우 1년전 까지 영향을 미치는 것을 확인할 수 있었다.도시가스 인수량은 용도별 구분이 어렵기 때문에 용도구분 없이 인수량 자체에 영향을 주는 주된 요인으로 외기온도를 고려하였다. 본 연구는 일별 인수량 예측에 일별 최저온도, 평균온도, 최고온도를 사용하였으며 월별, 1년 인수량 예측에는 월별 최저온도, 평균온도, 최고온도를 사용하였으며 회귀분석을 통해 외기온도와의 상관관계가 매우 높음을 확인하였다.

앞선 데이터 분석 결과를 토대로 시계열 분석 방법인 ARIMA와 시계열 데이터에 탁월한 성능을 보이는 딥러닝(deep learning) 기반 LSTM(long short-term memory) 두 가지를 이용하여 예측을 시행하였다. ARIMA는 도시가스 인수량을 단일 변수로 하여 모델을 만들었고, LSTM은 기온과 도시가스 인수량을 함께 반영하되, 단일모델링보다 수차례 모델링한 결과값을 단순평균(Simple Average)하여 예측모델을 만들었다. ARIMA의 경우 초기 예측 정확도는 높은 편이지만 추세를 정확하게 따라감에 무리가 있고 LSTM은 특성상 추세 반영 정도는 높은 편이나 시점의 편차를 줄이는데 한계가 있어 두 모델을 앙상블(ensemble) 하였다. 또한 전반적으로 일별이나 월별은 LSTM의 성능이 월등히 높았지만, 1년 예측의 경우 지역별로 Ensemble 의 MAPE 값이 0.2% ~1% 낮음을 확인할 수 있었다. LSTM이 1년 예측 평균이 ARIMA 와 Ensemble LSTM에 비해 좋은 편이나, 6~8월에는 도시가스의 사용이 적거나 주택/산업용의 비율을 반영하지 못하여 MAPE 평균값이 높게 계산되는 것을 확인할 수 있었다. 따라서 사용자는 세 가지 방법론에 따른 수요량 예측모델의 특성을 각각 이해해야 하며, 유동성있게 모델을 선택하여 사용해야 한다.

본 연구는 최초로 인수량을 이용하여 일별, 월별, 1년 예측의 결과를 보여주며 전통적인 시계열 분석 방법뿐 만 아니라 인공지능 알고리즘인 딥러닝 모형을 함께 사용하여 제안하였으며 추후 발전되는 딥러닝 방법을 활용하여 인수량뿐 아니라 판매량 예측에 관한 심도있는 연구를 진행할 수 있는 발판을 마련했다는데에 그 의의가 있다.

본 연구의 한계점은 도시가스의 용도별(주택용, 일반용, 산업용)로 모델을 수립하지 않아 여름철 주택용의 계절성을 반영이 충분히 이루어지지 않았다. 그렇기 때문에 여름철 보정 방법을 개발하거나 용도별로 나누어 예측모델을 수립해야 할 필요가 있다. 또한 예측 정확도에 영향을 미칠 수 있는 과거 시점 등과 같은 하이퍼 파라미터를 조절함에 따라 결과가 달라질 수 있다. 4 지역인 금산의 경우 2011년 4월부터 신설된 지역이기에 데이터가 턱없이 부족하여 오차값(MAPE)이 높았다. 이러한 이슈를 해결하기 위하여 데이터 수집에 대한 중요성을 알아야 하며 오차값을 개선하기 위한 발전된 딥러닝 방법을 테스트해 볼 필요성이 있을 것으로 보인다.

REFERENCES

[1] Park, J. S., Kim, Y. B., and Jung, C. W., “Short-Term Forecasting of City Gas Daily Demand”, *Journal of the Korean Institute of Industrial Engineers*, 39(4), 247-252, (2013)

[2] Han, J. H., and Lee, G. C., “Forcasting Hourly Demand of City Gas in Korea”, *Journal of the Korea Academia-Industrial*, 17(2), 87-95, (2016)

[3] Kweather Corp., “Analysis and Forecast Solution for City-Gas Service Providers”, KMA, (2012)

[4] Park, C. W., and Park, C. H., “Estimation of the Demand Function for City Gas Based on Characteristics of Its Uses in Korea”,

Korean Energy Economic Review, 17(2), 1-29, (2018)

[5] Ratnadip Adhikari, and R. K. Agrawal, "A Novel Weighted Ensemble Technique for Time Series Forecasting", *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Berlin, Geidelberg, (2012)

[6] S. Krstanovic, and H. Paulheim, "Ensembles of Recurrent Neural Networks for Robust Time Series Forecasting", *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Spinger, Cham, 34-46, (2017)