기상 데이터를 활용한 배달음식 주문 수요 예측

- Seasonal ARIMA와 VAR을 중심으로

2014310249 오희준
2013312161 김민구

요약

우리나라는 배달 음식의 문화와 산업이 특히 발달된 나라이며 그 산업의 규모도 점점 더 커지고 있다. 배달음식 스마트폰 어플리케이션이 더욱 활성화됨에 따라 시장은 계속 성장과 발전을 하고 있다. 하지만 배달음식에 대한 연구나 예측은 이루어진 바가 많지 않았다. 이에 따라 주문량에 영향을 주는 기상 데이터를 반영하여 Seasonal ARIMA 모형과 Vector Auto-Regression (VAR) 모형으로 모델을 수립하고 예측해보고자 한다. 실제로, 기상 변수를 이용하여 Seasonal ARIMA 모형을 이용하여 예측하였을 때 예측력이 더 좋아졌음을 확인하였다. 이 연구를 통해 날씨에 따른 주문량을 파악하여 자영업자들의 효율적인 재고 및 인력 관리와 효과적인 마케팅 전략 수립 등 여러 방면으로 공헌할 것으로 기대된다. 분석에는 R과 Eviews가 사용되었다.

1. 서론

국내 식품, 외식산업 매출 규모가 2015년 기준으로 200조원에 육박할 정도로 해마다 꾸준한 성장을 보여주고 있다. 특히, 그 중에서도 외식업의 매출 규모는 2015년 기준으로 100조원을 넘어선 108조원에 달한다 (농림축산식품부, 2017).

<표 1-1. 식품산업 성장 추이>

단위: 조원

구 분	2012	2013	2014	2015	**연평균 성장률**
제조, 외식(A+B)	152.4	156.9	163.7	192.0	**7.9%**
음식료품제조업(A)	75.1	77.3	79.9	83.9	**6.8%**
음식점업(B)	77.3	79.5	83.8	108.0	**8.9%**

출처: 농림축산식품부, 외식 트렌드 보고서, 2017.

이 중 치킨, 피자, 중국음식으로 대표되는 배달 요식업의 매출규모는 2001년 6,000억원 수준에서 매년 빠르게 성장하여 2015년에 16배 성장한 12조원에 이르고 있는 것으로 나타났다 (신한금융투자, 2014).

이렇듯 외식업 산업 중 특히 배달 외식업 산업의 성장규모가 증가하게 된 것에는 여러 가지 이유를 들 수 있다. 먼저, 해마다 증가하는 1인가구를 첫 번째 원인으로 꼽을 수 있다. 1인가구가 점점 증가함에 따라 '혼밥' 또한 하나의 트렌드로 자리잡게 되면서 외식 빈도도 늘고 있다. 실제로 월 평균 혼자 외식한 횟수도 2015년에는 2.8회에서 2016년에는 3.7회, 2017년에는 4.1회로 증가하는 경향을 보여주고 있다. (농림축산식품부, 2017).

<표 1-2. 1인가구 비중 추이>

단위: 명, %

	2013	2014	2015	2016	2017
1인가구 수	4,756,220	4,961,662	5,179,573	5,381,719	5,562,048
1인가구 비율	**25.9%**	**26.5%**	**27.2%**	**27.9%**	**28.5%**

출처: 통계청, 장래가구추계, 2017.

또 다른 이유로는 '배달의 민족', '요기요', '배달통'과 같은 배달 어플리케이션의 등장 때문이라고 할 수 있다. 현재 치열한 외식시장에서 살아남기 위해 외식업소들은 각 매장마다 배달 서비스를 병행함에 따라, 소비자 입장에서는 배달 음식 메뉴가 보다 다채로워지고 있다. 게다가, 1인가구의 증가로 인한 '혼밥족'이 늘면서, 배달 주문이 쉽고 간편한 배달 어플리케이션을 통해 다양한 메뉴로 끼니를 해결하는 상황이 맞물리면서 배달 어플리케이션은 배달 외식업 산업의 성장을 이끌고 있다. 2016년 기준 배달 어플리케이션 매출은 2조원에 달하고 있으며, 전체 음식 매출액 10조 원 가운데 배달 어플리케이션을 통한 배달 주문이 20%에 달한다 (박길수, 2017). 실제로 배달 어플리케이션 가맹점의 연간 매출액은 2015년 대비 2016년에 504만원 증가했다 (우아한형제들 & RPG코리아, 2016). 이처럼 배달 어플리케이션 가맹 여부는 외식업소의 매출에 직접적인 영향을 끼침으로써 외식 산업에 빠져서는 안 될 필수품이 되었다.

눈부시게 성장 중인 외식업계에서 살아남기 위해서는 꾸준한 매출이 필요하다. 꾸준한 매출은 직접적으로 손님이 매장에 얼마나 많이 와서 음식을 얼마나 시키는지에 달려있다. 그러나 배달의 비중이 높은 외식업의 경우, 음식 주문이 많을수록 매출이 높다. 그렇다면 사전에 음식주문 수를 예측할 수 있다면 매출의 규모도 어느 정도 예측할 수 있고, 그에 따라 사업 전략을 구상할 수 있을 것이다. 또한, 배달의 비중이 높은 업소의 특성상 주문과 동시에 조리를 시작하기 때문에 주문수요를 예측함으로써 매장 운영을 효율적으로 할 수 있을 것이다. 예측된 주문수요에 따라 재료를 준비하고 배송인력과 조리인력을 효율적으로 나눔으로써 매장의 운영비용을 절감할 수 있기 때문이다.

그러나 배달 외식업과 관련된 연구는 그리 활발하지 않다. 대부분이 설문조사를 기반으로 한 연구였고 단순한 설문지 문항들의 해석에 그치고 있다. 그 중, 눈에 띄는 것은 계절 ARIMA 모형을 이용하여 배달음식의 주문수요를 예측한 (윤현준, 2016)의 선행연구가 있었다. 하지만 요일변수만을 사용함으로써 연구를 진행했다는 점에서 아쉬운 내용이 있었다. 또 다른 연구로는 날씨가 배달음식 매출에 미치는 영향을 살펴본 연구가 있었다(정수미, 2017). 위 선행연구는 체감 기상

변수와 계절 간 차이를 변수로 활용하여 데이터마이닝 기법을 통해 부스팅(Boosting) 모형이 가장 우수하였음을 입증했다. 또한, 김다영·김대룡·변수지(2016)의 연구에서도 날씨데이터인 기온과 치킨 배달과의 연관성이 있음을 보여주면서 주문 수요를 부스팅(Boosting) 모형을 활용해 예측하였다.

하지만, 위의 두 선행연구는 데이터마이닝 기법을 활용했기 때문에 시계열 자료의 특성을 잘 반영해주는 ARMA, GARCH, VAR등 과 같은 모형을 활용하지 않았다는 아쉬움이 있었다. 따라서 이번 분석에서는 요일을 넘어, 기온, 습도와 같은 기상데이터를 활용하고 추가로 미세먼지 데이터를 가져와 계절성을 고려하는 SARIMA를 넘어선 VAR모형을 활용하여 분석을 진행할 것이다.

2. 연구대상

2-1. 자료 출처 및 변수 설명

본 연구에서 사용되는 배달음식 주문량 자료는 SK텔레콤의 데이터 개방 서비스인 빅데이터허브(http://www.bigdatahub.co.kr)에서 제공되는 '배달 업종 이용 현황 분석' 중 2016년 11월부터 2017년 10월까지의 데이터만을 추출한 것이다. 데이터는 서울지역의 치킨음식점, 피자음식점, 중국음식점, 족발/보쌈음식점으로 구별된 일별, 시간 별 주문량 데이터로 SK텔레콤의 이동통신 가입자가 해당 음식업종의 사업장에 전화 연결을 요청하여 성공한 건수를 기록한 것이다. 통화량이 5건 미만인 데이터는 5건으로 표시되어 있으며 연구에는 중국음식점 자료만을 사용했다. 네 분류 중 중국음식점을 사용한 이유는 음식점 분류 중 치킨과 중국음식의 이용률이 높았고 그 중 중국음식의 주문 건수에 결측값이 적었기 때문이다.

설명변수로는 기상청에서 제공하는 기상자료개방포털(https://data.kma.go.kr)에서 종관기상관측 자료를 사용했다. 종관기상관측이란 정해진 시각의 대기 상태를 파악하기 위해 모든 관측소에서 같은 시각에 실시하는 지상관측을 말한다. 또한 미세먼지농도 변수를 추가하기 위해 부유분진 측정 데이터를 활용하였다. 부유분진측정기(PM10) 관측이란 대기 중의 부유하는 공기를 흡입하여 직경이 10㎛ 이하인 먼지(황사 포함)가 필터에 침적되고, 동위원소 C-14 에서 방출되는 베타선을 필터 여지에 쏘아 감쇄된 베타선을 검출기로 측정하여 황사의 농도를 산출한 것을 뜻한다. 대기 상태를 나타내는 데이터와 미세먼지 농도를 측정한 데이터 모두 시간 별로 관측한 값을 가져와 다음과 같은 변수들을 분석에 적용하였다. 추가적으로, 설날과 추석과 같은 공휴일은 일요일로 코딩하였다.

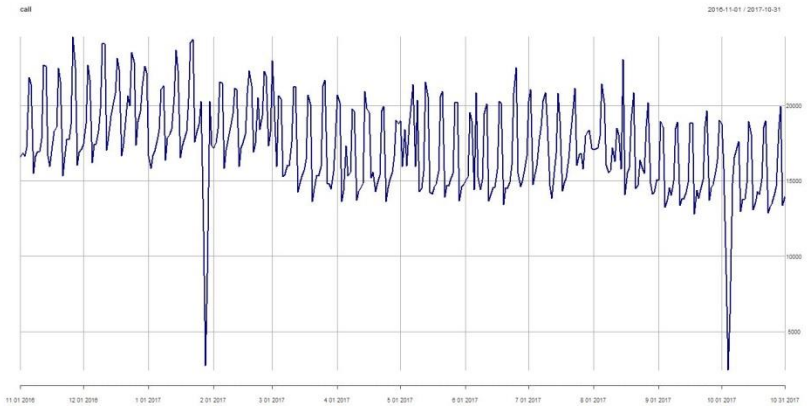
call	배달 주문 건수(개)
day	요일
max.temp	일 최고기온(℃)
min.temp	일 최저기온(℃)
temp.diff	일교차(℃)

avg.temp	일 평균기온(℃)
wind	풍속(m/s)
water	강수량(mm)
visibility	가시거리(100m)
dust	미세먼지(㎛/㎥)

2-2. 데이터 탐색

다음은 분석에서 사용된 변수를 시계열 타임플랏으로 나타낸 것이다.

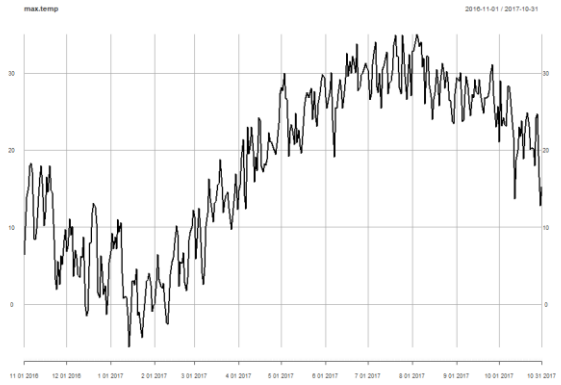
- 반응변수



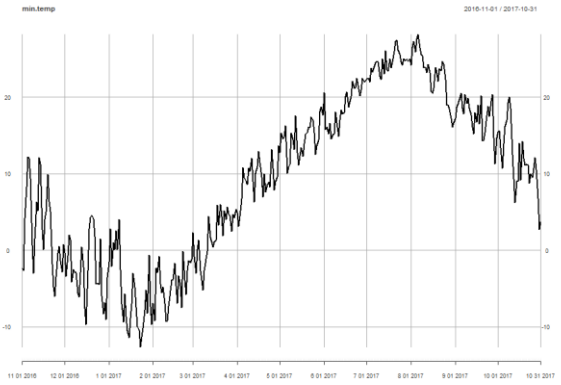
<그림 2-1> 2016 년 10 월부터 2017 년 11 월까지의 중국음식 배달 주문량

배달 주문량이 가장 적었던 날은 2017 년 10 월 4 일로 주문수가 2447 건이었고 그 다음으로 적은 날은 2017 년 1 월 28 일로 주문량은 2726 건이었다. 이 두 날의 공통점은 각각 추석 당일과 설날 당일이었다. 중국음식은 설과 추석에 주문량이 급격하게 감소하는 특징을 보였다.

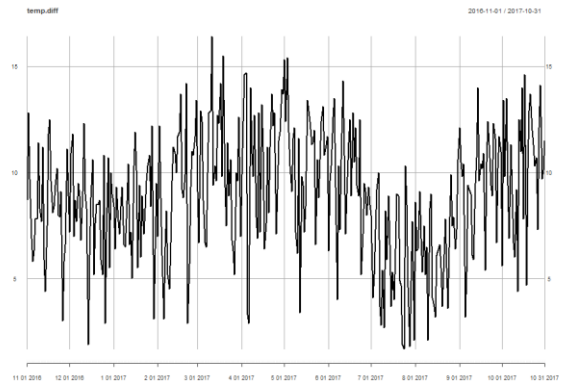
- 설명변수



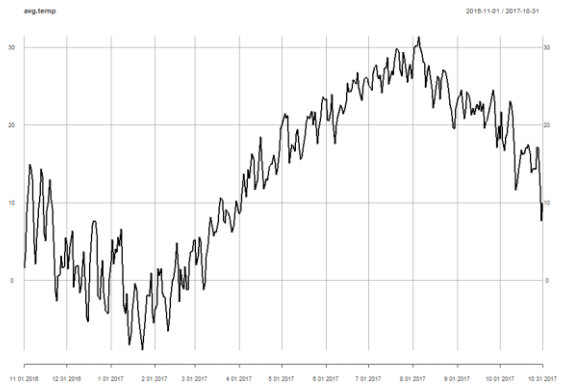
<그림 2-2> max.temp (일 최고기온)



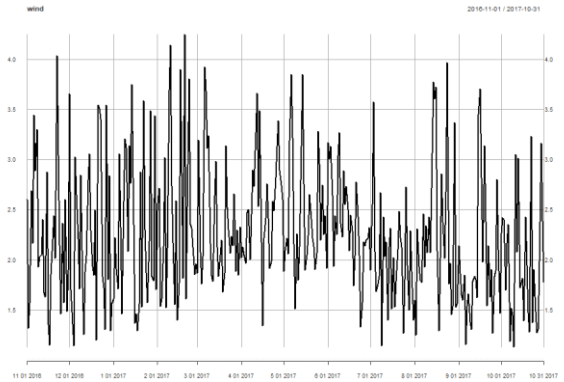
<그림 2-3> min.temp (일 최저기온)



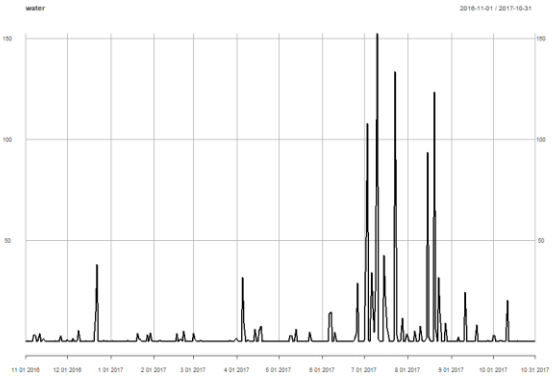
<그림 2-4> temp.diff (일교차)



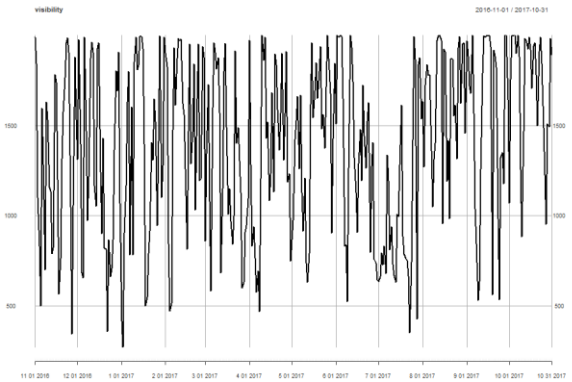
<그림 2-5> avg.temp (평균기온)



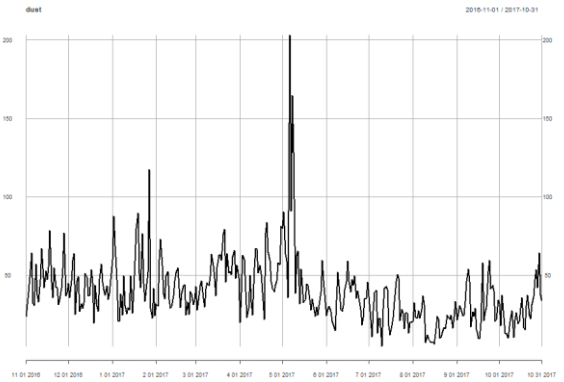
<그림 2-6> wind (풍속)



<그림 2-7> water (강수량)

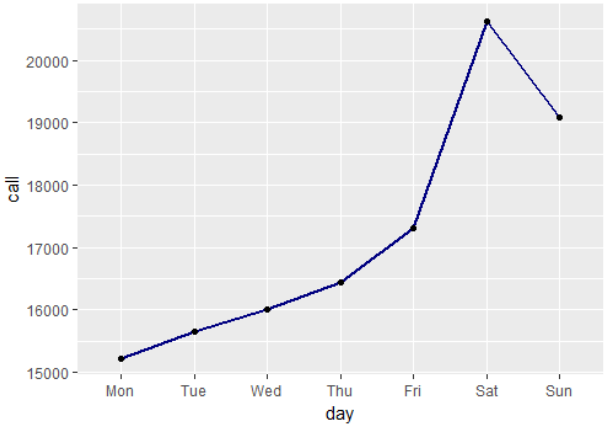


<그림 2-8> visibility (가시거리)



<그림 2-9> dust (미세먼지농도)

마지막으로 '요일' 변수에 대해서는 요일 별 주문량의 합계를 선 그래프로 표현하였다. 평일에 비해 주말인 토요일과 일요일에 주문량이 급격하게 증가하는 것을 알 수 있고 토요일과 일요일 중에는 토요일의 주문 수가 더 많았다.



<그림 2-3> 요일 별 주문량 합계

3. 연구 과정

3-1. 기초통계량과 정상성 검정

본격적인 연구에 들어가기 앞서, 중국음식 배달 주문량에 대한 기초통계량을 구하였다.

<표 3-1. 중국음식 배달 주문량 기초통계량>

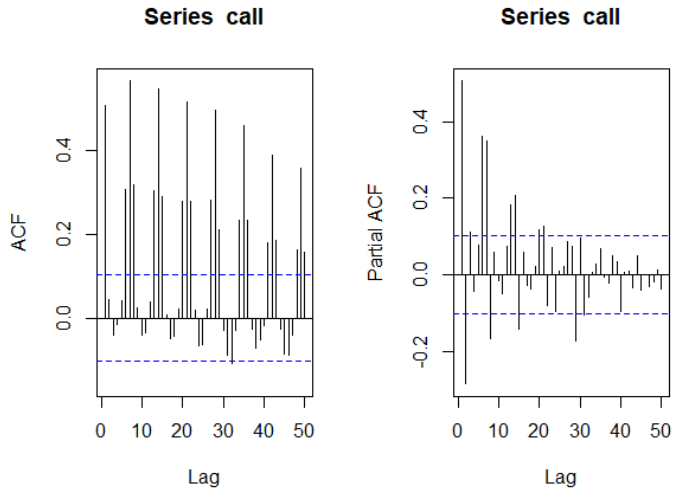
	평균	분산	왜도	첨도
기초통계량 값	17,303.88	9,494,313	-0.425428	2.184978

다음으로, 반응변수를 포함한 변수들의 정상성 검정을 시행하였다. 정상성 여부 판단은 Augmented Dickey Fuller (ADF) 검정, Phillips-Perron (PP) 검정, Kwiatkowski–Phillips–Schmidt–Shin (KPSS) 검정을 통해 이루어졌다.

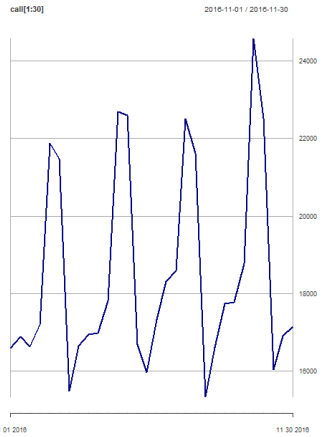
<표 3-2. 각 변수들의 정상성 검정 결과>

Variables	ADF		PP		KPSS
	none	constant	none	constant	constant
call	-0.594630	-2.163745	-1.028753	-13.80727	2.935175***
max.temp	-0.685422	-1.512223	-1.238264	-2.771879*	1.804960***
min.temp	-1.026428	-1.307850	-2.132992**	-2.790525*	1.736568***
temp.diff	-0.606021	-4.087481***	-2.460266	-14.90852***	0.208026*
avg.temp	-0.890950	-1.537205	-1.265147	-2.311837	1.781845***
wind	-0.888935	-14.70226***	-2.423035**	-14.70226***	0.763972***
water	-3.645544***	-4.058383***	-17.94020***	-17.88970***	0.482903**
visibility	-0.853044	-10.44349***	-2.605217***	-10.15237***	0.470862**
dust	-1.728871*	-7.360543***	-3.231190***	-11.20483***	0.957629***

결과표에서 *이 하나인 경우는 유의수준 10%에서 귀무가설을 기각한 것이고, **이 두 개인 경우는 유의수준 5%, 세 개인 경우는 유의수준 1%에서 기각하였음을 의미한다. ADF 검정과 PP 검정의 귀무가설은 '시계열에 단위근이 존재한다'는 것이고, KPSS 검정의 귀무가설은 '시계열이 정상성을 만족한다'이다. ADF, PP, KPSS 검정 결과 다수의 변수에서 정상성을 만족하지 않는 것으로 나타났다. 변수들이 정상성을 만족하지 않는 이유는 변수들이 가지는 계절성(seasonality) 때문이다. 하나의 예로 <그림 3-2>에서 주문량의 자기상관함수 (Autocorrelation Function, ACF)와 부분자기상관함수 (Partial Autocorrelation Function, PACF)를 나타내고 있는데 7 의 주기로 스파이크(spike)가 보이는 것을 알 수 있다. 이를 통해서 주기가 7 인 계절성이 존재함을 유추할 수 있다. 또한 <그림 3-3>은 2016 년 11 월 한 달 동안의 주문량의 그래프를 나타낸 것인데 여기서도 7 일의 주기가 있음을 확인할 수 있다.

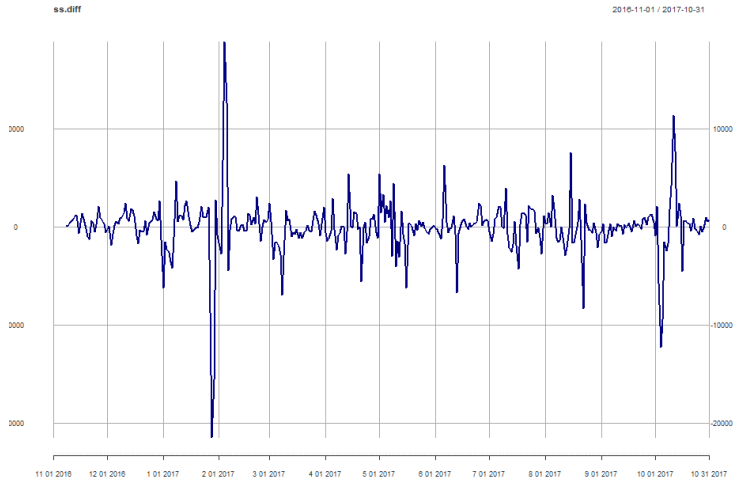


<그림 3-1> 주문량의 ACF 함수와 PACF 함수

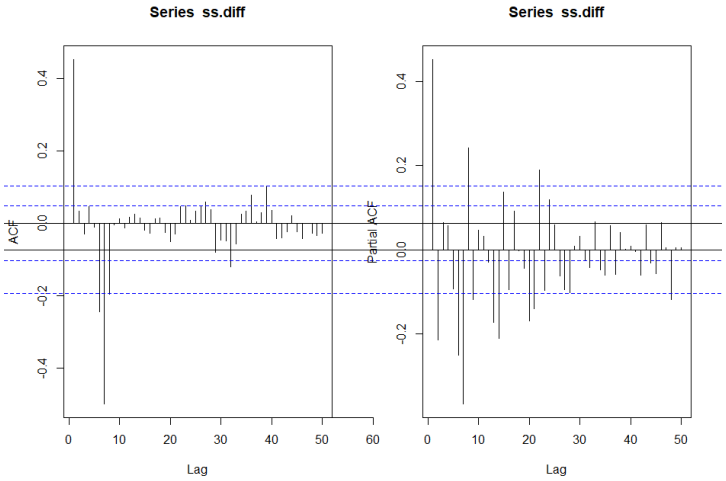


<그림 3-2> 한 달 동안의 주문량 그래프

앞서, 주문량이 계절성이 있는 시계열임을 파악하였기 때문에 계절 차분을 통해 정상시계열로 조정을 해야 한다. **<그림 3-3>**는 원시계열을 계절차분을 한 결과의 타임플랏을 나타낸 것이고 **<그림 3-4>**는 계절차분 후의 ACF, PACF 를 나타낸 것이다.



<그림 3-3> 계절 차분 후 time plot



<그림 3-4> 계절 차분 후 ACF 와 PACF

분산이 급격히 증가하는 부분이 있기 때문에 정상성을 만족하지 않는 것처럼 보일 수도 있지만 단위근 검정을 통해서 계절 차분을 한 주문량의 데이터는 정상성을 만족한다. 결과는 <표 3-3>에서 확인이 가능하다.

<표 3-3. 계절 차분 후 정상성 검정 결과>

Variables	ADF		PP		KPSS
	none	constant	none	constant	constant
call	-8.050687***	-8.051265***	-11.35938***	-11.34392***	0.017215

3-2. ARMAX 모형의 적합과 예측

3.1 절에서 정상성 검정을 통해 주문량 변수에 계절성이 존재하는 것을 확인하였기 때문에 계절성을 고려한 Seasonal ARIMA 모형을 사용하였다. 계절성을 가지고 있는 ARIMA 모형의 표기는 ARIMA(p,d,q)(P,D,Q)$_s$ 로 하며 p 는 자기회귀 (auto-regression)의 차수, d 는 차분

횟수, q 는 이동평균 (moving-average)의 차수, P 는 계절 자기회귀의 차수, D 는 계절차분의 횟수, Q 는 계절 이동평균의 차수 그리고 s 는 계절성의 주기이다.

먼저, 날씨 변수를 추가하지 않고 주문량의 데이터만을 가지고 SARIMA 모형을 적합해 보았다. ARIMA(1,0,1)(1,1,1)$_7$ 모델을 적용한 결과는 아래 **<그림 3-6>**과 같다. 모형 적합식은 다음과 같이 쓸 수 있다. 모형 적합 결과, AIC 는 6425.39 로 나타났다. 모형 적합 뒤, 잔차가 White Noise 를 따르는지 알아보기 위해서 Q-test 를 수행하였다. **<표 3-4>**에서 보여지는 것처럼 p-value 가 유의수준(α=0.05)보다 크기 때문에 귀무가설을 기각하지 못하므로 White Noise 를 따른다는 결론을 내릴 수 있다.

$$(1-B^7)(1-0.2447B)(1+0.0097B^7)X_t = (1+0.3759B)(1-0.8666B^7)\varepsilon_t,\ \ \varepsilon_t \sim WN(0,\sigma^2)$$

```
Call:
arima(x = call, order = c(1, 0, 1), seasonal = list(order = c(1, 1, 1), period = 7),
    include.mean = T)

Coefficients:
         ar1     ma1     sar1     sma1
      0.2447  0.3759  -0.0097  -0.8666
s.e.  0.0889  0.0808   0.0618   0.0368

sigma^2 estimated as 3449116:  log likelihood = -3207.7,  aic = 6425.39
```

<그림 3-6> ARIMA(1,0,1)(1,1,1)$_7$ 모형 적합 결과

<표 3-4. 모형 적합 후 Q-test 결과>

	Q(7)	Q(14)	Q(21)	Q(28)	Q(35)
p-value	0.7468809	0.8510812	0.9708734	0.9739755	0.8964137

다음으로, 주문량을 예측하는데 날씨 변수가 유의한지 알아보기 위해 날씨 변수를 하나씩 포함시키면서 모형을 적합하였다. 주문량의 데이터가 계절 차분되어 사용되었기 때문에 날씨 변수도 주문량과 같은 주기로 계절 차분한 뒤 모형에 추가하였다. ARIMA 의 차수는 앞서 적합한 모델의 차수와 동일하다. 총 8 개의 날씨 변수가 있었으며, 추가적으로 '요일' 변수도 추가하여 요일에 따른 영향도 보고자 하였다. 모형 적합 순서는 '일 최고기온', '일 최저기온', '일교차', '일 평균기온', '풍속', '강수량', '가시거리', '미세먼지농도' 순으로 적합하였다. 어떤 변수를 추가하였을 때 제일 적합한지 AIC 를 통해 비교하였다. 전체적으로 날씨 변수를 추가하였을 때 AIC 가 추가하지 않았을 때 (AIC = 6425.39)보다 확연히 줄어든 것을 알 수 있다. 그 중에서 '요일' 변수를 추가하였을 때 가장 AIC 가 크게 감소하였고, 그 다음으로 '일 강수량'을 추가하였을 때, '일 최고기온'을 추가하였을 때가 뒤따랐다.

<표 3-5. 각 변수 별 AIC>

변수명	max.temp	min.temp	temp.diff	avg.temp	wind
AIC	**6304.097**	6308.108	6304.602	6305.935	6303.1
변수명	water	visibility	dust	day	
AIC	**6296.260**	6305.166	6307.839	**6261.709**	

이번에는 '일 최고기온', '일 강수량' 변수를 가장 큰 AIC 의 감소를 보인 '요일' 변수와 묶어서 적합하였다.

Model 1	max.temp + day
Model 2	water + day
Model 3	max.temp + water +day

```
Call:
arima(x = call, order = c(1, 0, 1), seasonal = list(order = c(1, 1, 1), period = 7),
    xreg = cbind(dmax.temp, model.matrix(~dat$day)[, -1]), include.mean = T)

Coefficients:
         ar1     ma1     sar1     sma1       ..1  dat.dayTue  dat.dayWed  dat.dayThu  dat.dayFri
      0.5523  0.2763  -0.0055  -0.8673  -46.5920    1574.321    5649.977    9110.140   6762.4539
s.e.  0.0681  0.0868   0.0635   0.0390   18.2319     819.626    1006.694    1422.372    983.3923
      dat.daySat  dat.daySun
       14881.897    7493.538
s.e.    1639.864     664.208

sigma^2 estimated as 2105847:  log likelihood = -3058.63,  aic = 6141.27
```

<그림 3-6> Model1 모형 적합 결과

```
Call:
arima(x = call, order = c(1, 0, 1), seasonal = list(order = c(1, 1, 1), period = 7),
    xreg = cbind(dwater, model.matrix(~dat$day)[, -1]), include.mean = T)

Coefficients:
         ar1     ma1     sar1     sma1      ..1  dat.dayTue  dat.dayWed  dat.dayThu  dat.dayFri
      0.5682  0.2528  0.0291  -0.8752  10.0307    1834.408   5957.8916    9271.647   6908.1157
s.e.  0.0651  0.0816  0.0634   0.0376   2.9649     825.157    994.3978    1417.296    980.4145
      dat.daySat  dat.daySun
       15063.342   7511.7475
s.e.    1615.008    667.5946

sigma^2 estimated as 2078007:  log likelihood = -3056.28,  aic = 6136.57
```

<그림 3-7> Model2 모형 적합 결과

```
Call:
arima(x = call, order = c(1, 0, 1), seasonal = list(order = c(1, 1, 1), period = 7),
    xreg = cbind(dmax.temp, dwater, model.matrix(~dat$day)[, -1]), include.mean = T)

Coefficients:
         ar1     ma1     sar1     sma1       ..1     ..2  dat.dayTue  dat.dayWed  dat.dayThu
      0.5665  0.2619  0.0240  -0.8729  -37.7283  9.0481   1832.1044   5764.2492    9092.424
s.e.  0.0653  0.0828  0.0639   0.0384   18.0054  2.9753    818.4299    992.8299    1406.391
      dat.dayFri  dat.daySat  dat.daySun
        6760.154   14924.207   7506.8970
s.e.     974.239    1610.795    660.3907

sigma^2 estimated as 2052531:  log likelihood = -3054.1,  aic = 6134.19
```

<그림 3-8> Model3 모형 적합 결과

이 세 모형 모두 잔차의 Q-test 를 기각하지 않아 모형이 잘 적합 되었음을 알 수 있다. Q-test 결과는 아래 **<표 3-6>** ~ **<표 3-8>**에 정리되어 있다.

<표 3-6. Model1 적합 후 Q-test 결과>

	Q(7)	Q(14)	Q(21)	Q(28)	Q(35)
p-value	0.8110399	0.9715369	0.8359121	0.7585047	0.8342204

<표 3-7. Model2 적합 후 Q-test 결과>

	Q(7)	Q(14)	Q(21)	Q(28)	Q(35)
p-value	0.5133029	0.9250990	0.6656400	0.6424509	0.74083

<표 3-8. Model3 적합 후 Q-test 결과>

	Q(7)	Q(14)	Q(21)	Q(28)	Q(35)
p-value	0.5896488	0.9258547	0.7598701	0.6636672	0.7607569

이 세 모형 중 어떤 모형이 주문량 데이터에 제일 잘 적합하였는지 AIC 비교를 통해 알아보았다. '일 최고기온', '일 강수량', '요일' 변수를 모두 넣은 Model 3 이 가장 AIC 가 작았으며 이 결과는 앞에서 날씨 변수를 추가하지 않았을 때 보다 300 가까이 줄어들었다. 이는 날씨 변수를 추가하지 않은 모형보다 날씨 변수와 요일 변수를 추가하였을 때 모형의 설명력이 증가하는 것을 의미한다.

<표 3-9. Model1, Model2, Model3 AIC 비교>

Model	AIC
Model 1	6141.270
Model 2	6136.566
Model 3	**6134.190**

본 연구는 주문량의 예측에 의의를 두고 있으므로 Out-of-Sample Forecasting 을 통해 각 모델의 예측력을 계산해 보고자 한다.

앞서 적합한 Seasonal ARIMA 모형을 사용해 예측 값을 구하여 모델 성능을 시험해 보고 적절한 모형을 선정할 것이다. 예측 기간 15 일 즉, 2017 년 10 월 17 일부터 2017 년 10 월 31 일까지를 예측 기간으로 두고, Rolling Window 방법을 사용하여 RMSE 를 통해 비교해 보았다.

<표 3-10. 각 모델 예측 값의 RMSE 결과>

Model	Model 0	Model 1	Model 2	Model 3
RMSE	886.6657	**505.1159**	621.1023	562.5688

Model 0 는 어떤 변수도 추가하지 않은 Seasonal ARIMA 모형이다. <표 3-9>의 결과에서 알 수 있듯이, Model1 의 RMSE 가 다른 모델들보다 현저히 낮았다. 두 모델의 Performance 를 비교하는 Diebold-Mariano test 를 사용하여 검증해보았다.

<표 3-11. Diebold-Mariano Test 의 p-value>

P-value	Model0	Model1	Model2	Model3
Model 0	------			
Model 1	**0.04522***	------		
Model 2	0.11	0.2289	------	
Model 3	0.08768	0.5751	0.4713	------

결과표를 보면 알 수 있듯이, Model0 과 Model1 에서 P-value 가 0.045 정도로 유의수준 0.05 에서 H_0 : "두 모형은 예측력에 차이가 없다."를 기각한다. 그러므로 Model 1 이 Model 0 보다 예측력이 더 좋다는 대립가설을 채택하게 된다. 결국 '요일'과 '일 최고기온'을 같이 고려한 Model 1 이 다른 모델들보다 RMSE 가 현저히 낮으면서 Model 0 보다 더 좋은 예측력을 보여준다는 사실을 알 수 있다.

3-3. VAR 모형의 적합과 예측

이 절에서는 여러 변수들을 한번에 고려할 수 있는 VAR 모형을 이용하여 주문량 데이터를 적합하고 예측해보고자 한다. 먼저, Granger causality 검정을 통해 주문량을 예측하는 데 어떤 변수가 유의미한 영향을 주는지 알아보았다. 7 일의 주기를 고려하여 lag 는 7 로 설정하였다. 이 검정의 귀무가설은 다음과 같다. 주문량을 예측할 때 날씨 변수들의 영향력을 Granger Causality 로 알아보고 **<표 3-12>**에 그 p-value 를 명시했다. 주문량을 예측하는 것이 주요한 관심이기 때문에 주문량이 날씨 변수를 Granger Cause 하는지는 생략하였다.

H_0 : x 가 y 를 Granger Cause 하지 않는다. (예측력에 도움을 주지 않는다.)

<표 3-12. 각 변수의 Granger Causality 결과>

변수명	max.temp	min.temp	temp.diff	avg.temp
p-value	0.1505	0.1711	0.3957	**0.0327**
변수명	water	visibility	dust	wind
p-value	0.1274	0.2633	**0.0073**	0.4145

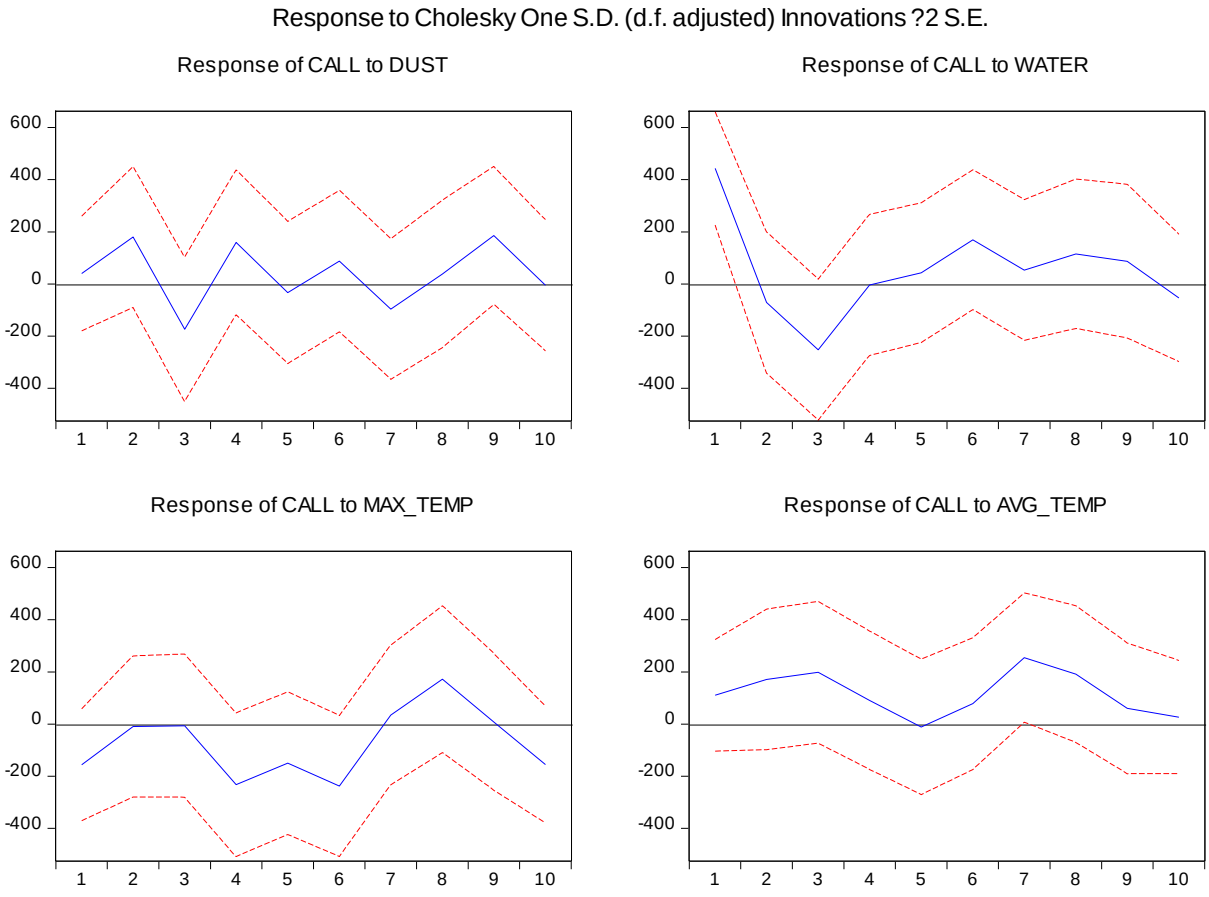
Granger Causality 검정 결과, '일 평균기온'과 '미세먼지농도'의 p-value 가 유의수준보다 작아 중국음식의 배달 주문량을 예측하는데 도움을 준다는 결론을 지을 수 있다. 앞서 Seasonal ARIMA 모형에서 선택한 변수들과 Granger-Causality Test 에서 도출된 변수들을 이용하여 VAR 모형을 적합하였다. Cholesky order 는 '미세먼지농도' – '강수량' – '일 최고기온' – '일 평균기온' – '주문량' 순으로 설정하였고 AR order 은 default 인 2 로 설정하였다. 적합식은 다음과 같다. (y_t)는 주문량을 의미한다. 수정된 R-squared 는 0.37 이었다.

$$y_t = 14546.85 + 10.39dust_{t-1} - 7.29dust_{t-2} - 23.12water_{t-1} + 2.67water_{t-2} - 94.40max.temp_{t-1} - 209.83max.temp_{t-2} + 144.38avg.temp_{t-1} + 103.33avg.temp_{t-2}$$

적절한 order 을 찾기 위해서 Eviews 에서 제공하는 'Lag Length Criteria'를 활용하였다. 분석 결과, Lag 8 이 적절하다는 결론이 도출되었다. Lag 8 으로 다시 한 번 적합하였다. 수정된 R-squared 가 0.54 로 증가하였다.

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-8422.040	NA	2.19e+14	47.21031	47.26462	47.23191
1	-7680.153	1458.837	3.95e+12	43.19414	43.52000*	43.32374
2	-7629.993	97.22907	3.43e+12	43.05318	43.65059	43.29080*
3	-7600.508	56.32720	3.35e+12	43.02806	43.89702	43.37368
4	-7581.017	36.68934	3.45e+12	43.05892	44.19943	43.51255
5	-7549.667	58.13372	3.33e+12	43.02334	44.43540	43.58498
6	-7503.449	84.40971	2.96e+12	42.90447	44.58808	43.57412
7	-7457.559	82.52405	2.64e+12	42.78745	44.74261	43.56510
8	-7430.162	48.50195*	2.61e+12*	42.77401*	45.00073	43.65967

<그림 3-9> VAR 모형의 order selection



<그림 3-10> 충격반응에 대한 plot

<그림 3-10>은 VAR 모형 적합 후 여러 날씨 변수에 대한 주문량의 반응을 plot 으로 그린 것이다. '일 강수량'에 충격이 가해지면 바로 다음 날과 그 다음 날 배달 주문량이 유의하게 증가하는 것을 알 수 있다. 비가 오거나 눈이 오는 날에 사람들이 중국 음식을 포함한

배달음식을 많이 시켜먹는 것과 같은 맥락이다. 또한 '일 평균기온'에 충격이 가해지면 7 일 후에 주문량에 유의한 증가가 발생한다. 이러한 VAR 의 충격 반응 함수는 중국음식 자영업자들은 미리 재료와 인력의 수급 같은 단기적인 전략을 수립하는 데 도움을 줄 것이다.

Period	S.E.	DUST	WATER	MAX_TEMP	AVG_TEMP	CALL
1	15.98079	0.036936	4.558427	0.562909	0.283322	94.55841
...						
29	20.53519	1.695189	3.942039	6.975119	2.902094	84.48556
30	20.55896	1.692293	3.924481	6.934702	2.884539	84.56398
Cholesky Ordering: DUST WATER MAX_TEMP AVG_TEMP CALL						

<그림 3-11> Variance Decomposition

<그림 3-11>은 예측오차의 분산분해(Variance Decomposition)를 나타낸 것으로 특정 변수의 움직임에 대한 설명이다. 결과표에서 날씨 변수들과 주문량에 대하여 자체 충격 및 상대방 변수에 대한 분산분해를 보여주고 있는데, 총 30 일을 예측하였다. 장기적으로 보았을 때, '일 최고기온'은 약 7%의 크기로 주문량에 영향을 준다고 할 수 있다.

마지막으로, Out-of-Sample Forecasting 을 통해 VAR 모형의 예측력을 확인하였다. VAR 모형의 RMSE 는 1330.122 로 Seasonal ARIMA 모형보다 좋지 않은 예측력을 보였다.

4. 결론

이번 분석에서 배달음식, 특히 중국음식점의 주문 수요를 예측하는 다양한 모델을 수립하고 비교해 보았다.

먼저, 사용된 데이터로는 SKT 빅데이터 허브에서 가져온 '배달 업종 이용 현황'이었고, 그 중에서 2016년 10월부터 2017년 11월까지 총 1년의 중국음식점 주문 전화 수를 사용했다. 다른 업종의 주문 전화 수를 사용할 수도 있었지만, 다른 업종의 주문 전화 수는 결측값이 많이 존재하기도 했고 주문 전화 수 자체가 적은 업종이었기 때문에 중국음식점의 주문 전화 수를 사용했다. 또한, 주문 수요와 관련된 독립 변수로는 기상청에서 가져온 날씨 데이터, 강수량, 일 최고기온, 일 최저기온, 일교차, 일 평균기온, 풍속, 강수량, 가시거리, 미세먼지 농도, 요일을 사용했다. 데이터 탐색과정에서 중국음식의 주문량은 설과 추석에 급감했음을 확인할 수 있었고, 평일에 비해 주말에 주문량이 급증하는 경향을 볼 수 있었다. 또한 다수의 변수에서 계절성을 갖고 있기 때문에 정상성을 만족하지 않았음을 알 수 있었고, ACF와 PACF함수를 통해 7의 주기를 갖는 것을 확인했으며 이러한 계절성을 고려하기 위해 모델로 Seasonal ARIMA와 VAR을 사용했다.

Seasonal ARIMA에서는 각 날씨변수를 하나씩 포함시키면서 AIC를 비교해보았고, 그 중 요일 변수가 가장 낮은 AIC를 보여주면서, 요일 변수와 다른 변수들을 묶어보면서 AIC를 비교해 보았다. '일 최고기온', '일 강수량', '요일' 변수를 모두 넣은 Model3이 AIC가 가장 낮았음을 확인했

다. 또한, VAR에서는 Granger-Causality검정 결과, '일 평균기온', '미세먼지 농도' 변수가 중국음식 주문 수요를 예측하는데 도움을 준다는 결론을 얻었다. Lag 8로 모델을 적합한 결과, Adjusted R-Squared가 0.54임을 알 수 있었고, '일 강수량'에 충격이 가해지면, 다음 날과 그 다음날의 주문량이 유의하게 증가함을 확인할 수 있었고, '일 평균기온'에 충격이 가해지면, 7일 후 주문량이 유의하게 증가함을 파악 할 수 있었다.

우리 분석의 목적은 예측인 만큼, 각 모델에서 예측력이 얼마나 정확한지가 기준이었기 때문에 RMSE를 활용하여 모델 선정을 진행했다. 먼저 SARIMA모형에서 가장 낮았던 RMSE가 505.1159이었고, VAR에서는 1330.122로 SARIMA보다 좋은 예측력을 보여주지 못했다. 따라서, 최종 모델은 '요일'과 '일 최고기온'을 같이 고려한 ARIMA(1,0,1) $(1,1,1)_7$이 채택되었다. 이 모델은 어떠한 변수도 고려하지 않은 SARIMA모델보다 유의수준 0.04522에서 더 좋은 Performance를 보여준다는 결론을 얻었다.

5. 의의 및 한계

이번 분석은 최근 1인가구의 증가와 함께 배달 어플리케이션의 폭풍 같은 성장에 힘입은 외식업에 대해, 특히 연구가 많이 부족했던 배달 외식업에 날씨와 배달 주문량을 연관시키며 기존의 연구와 같이 날씨가 배달음식의 주문량에 영향을 준다는 점을 다시 한 번 확인하였고, 이를 SARIMA와 VAR모델을 통해 어떤 변수가 유의하게 영향을 주는지 그 영향력을 정량화 했다는 점에서 의의를 가진다. 또한, 이는 기상청에서의 날씨 예보 데이터를 활용한다면, 언제든지 미래 주문량의 수요를 예측할 수 있음을 의미한다. 따라서, 우리가 세운 모델을 자영업자나 기업 등에서 잘 활용한다면 음식의 재료나 배달 인력의 수급과 같은 단기적인 전략의 수립에 큰 도움을 줄 수 있을 것으로 기대한다. 또한, 추가적인 분석을 통해 연령대와 지역을 같이 고려한다면, 마케팅전략과 최적의 배달 경로루트 수립에도 도움을 줄 수 있을 것이다. 게다가, 주문량 데이터가 더 확장된다면, 배달음식 시장에 진입하고자 하는 예비창업자들에게도 사전에 주문 수요를 예측함으로써 경영전략을 수립하는데도 도움이 될 수 있다고 기대한다. 특히, 서울시에서는 '우리 마을 가게 상권분석 서비스[1]'라는 창업에 도움을 주는 서비스를 운영 중인데, 경쟁점포, 가구수, 직장인구 수, 유동인구 수와 같은 데이터를 사용하여 창업위험도를 설정해 주고 있다. 이 곳에 추가적으로 배달주문 수요도 함께 고려해 준다면, 예비 창업자들의 폭발적인 호응을 얻을 것이라 기대한다.

다만, 우리 분석은 1년의 데이터만을 사용하였다는 것에 아쉬운 점이 있었다. 시간의 한계상, 수 개년의 데이터를 사용하였다면, 정량화된 영향력을 제대로 파악했을 것이고, 예측 값도 더 정확할 수 있었다. 게다가, 통신사 1곳 SKT에서만의 데이터를 사용하였기 때문에 국내 전체로의 확장은 힘들 것으로 보인다. 하지만, 기존 연구와 같이 아직 연구가 많이 존재하지 않는 배달 외식업시장에 대해 연구의 활용도 측면에서 정부기관과의 협력 가능성을 제시했고, 추가적인 데이터의 분석 활용 가능성을 제시했다는 측면에서 또 다른 의의를 갖는다고 할 수 있겠다.

---

[1] 우리마을 가게 상권분석 서비스, http://golmok.seoul.go.kr/sgmc/main.do.

참고자료

우아한형제들, RPG코리아 (2016). "2016 배달 음식점 보고서".

박상민 (2014). "배달의 시대". 신한금융투자.

통계청 (2017). "장래가구추계".

농림축산식품부 (2017). "외식 트렌드 보고서".

박길수 (2017). "홀+배달 복합매장 창업으로 불경기 이겨낸 걸작떡볶이". <디트뉴스>. 2017.01.12 (http://www.dtnews24.com/news/article.html?no=410960).

김다영, 김대룡, 변수지 (2016). "날씨에 따른 배달음식 주문 건수 예측". 한국기상학회 학술대회 논문집, Vol.2016 No.10.

윤현준 (2016) "계절 ARIMA 모형을 이용한 배달음식 주문 수요예측 연구". 연세대학교 공학대학원 석사학위논문.

정수미 (2017). "날씨가 배달음식 매출에 미치는 영향: 체감 기상변수와 계절 간 차이를 중심으로". 이화여자대학교 빅데이터분석학협동과정 석사학위논문.