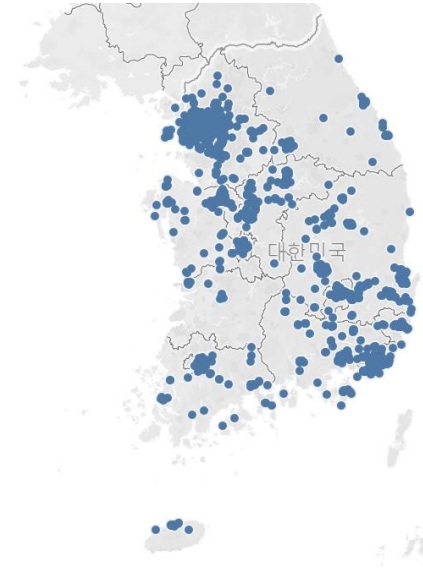
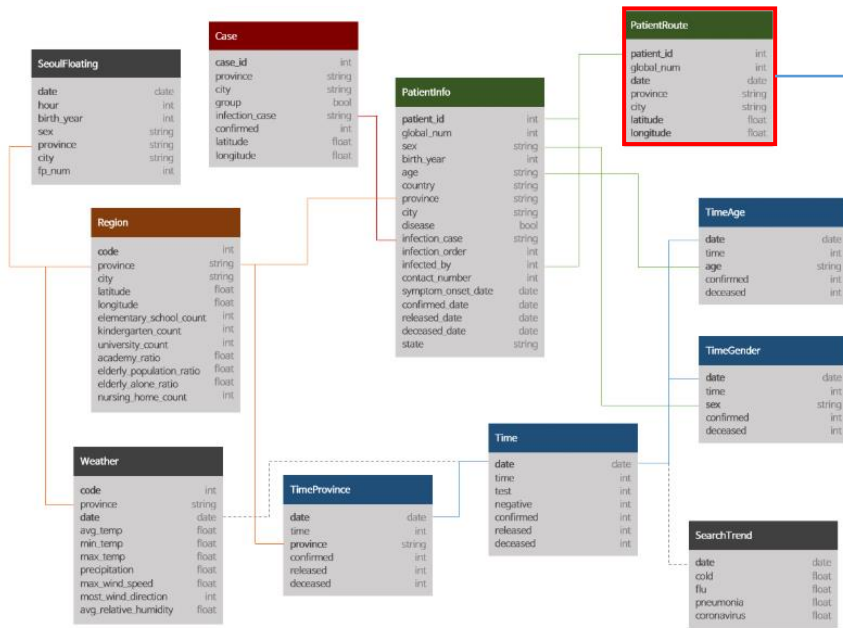


코로나 확산속도 예측

COVID-19 dataset

코로나의 확산속도 예측



	patient_id	global_num	date	province	city	type	latitude	longitude
0	1000000001	2.0	2020-01-22	Gyeonggi-do	Gimpo-si	airport	37.615246	126.715632
1	1000000001	2.0	2020-01-24	Seoul	Jung-gu	hospital	37.567241	127.005659
2	1000000002	5.0	2020-01-25	Seoul	Seongbuk-gu	etc	37.592560	127.017048
3	1000000002	5.0	2020-01-26	Seoul	Seongbuk-gu	store	37.591810	127.016822
4	1000000002	5.0	2020-01-26	Seoul	Seongdong-gu	public_transportation	37.563992	127.029534

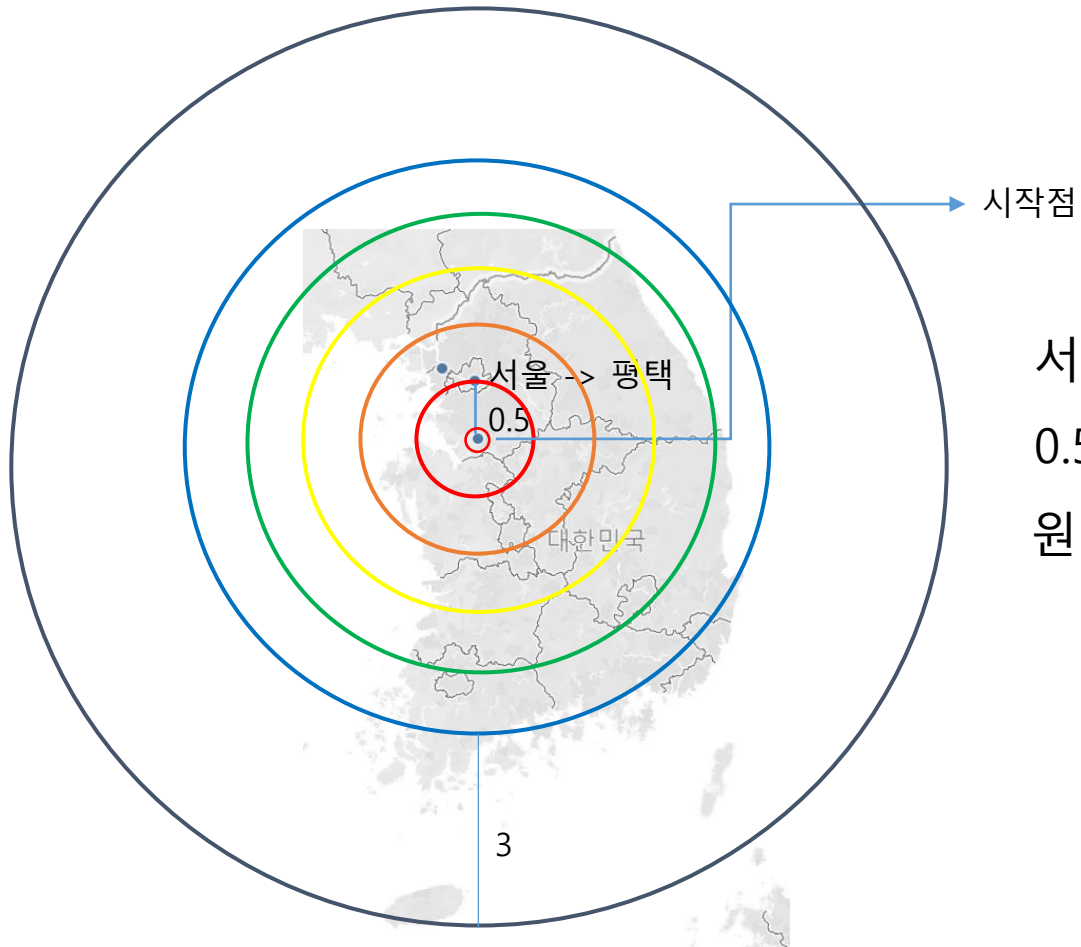


일부 확진자의 경로

가설

확진자의 경로가 시작점에서부터 거리가 멀수록,
코로나 확산에 더 큰 영향을 줬을 것이다.

Formulation



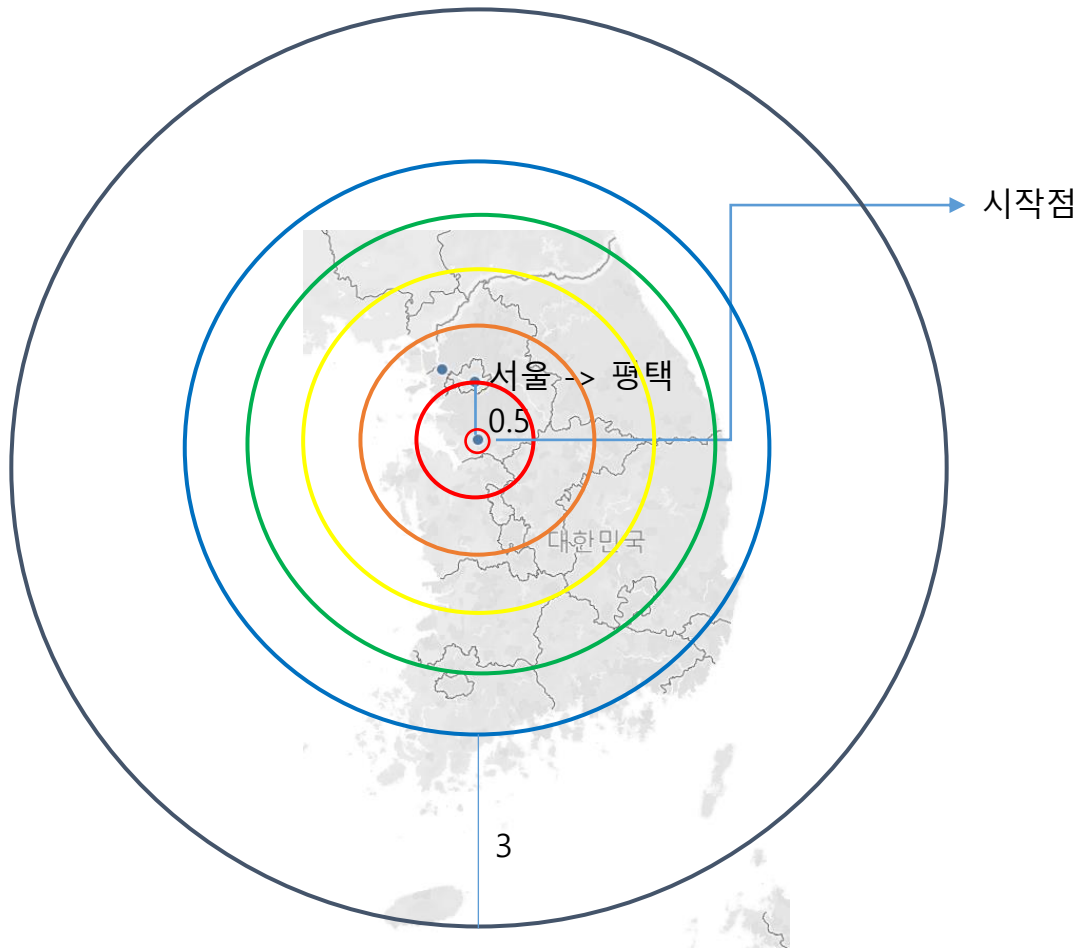
서울에서 경기도(평택)까지의 차이: 위도: 0.5

0.5를 반지름으로 해서 원을 하나씩 만들고 추가시킴

원안에 확진자의 특정 경로가 있으면 원의 weight를 경로에 줌

원은 위도와 경도를 기준으로 설정한 크기다.

Formulation



원은 위도와 경도를 기준으로 설정한 크기다.

$$\text{Diffusion rate} = \sum_1^{\text{days}} (w_{\text{CN}} + x_i)$$

(확산속도)

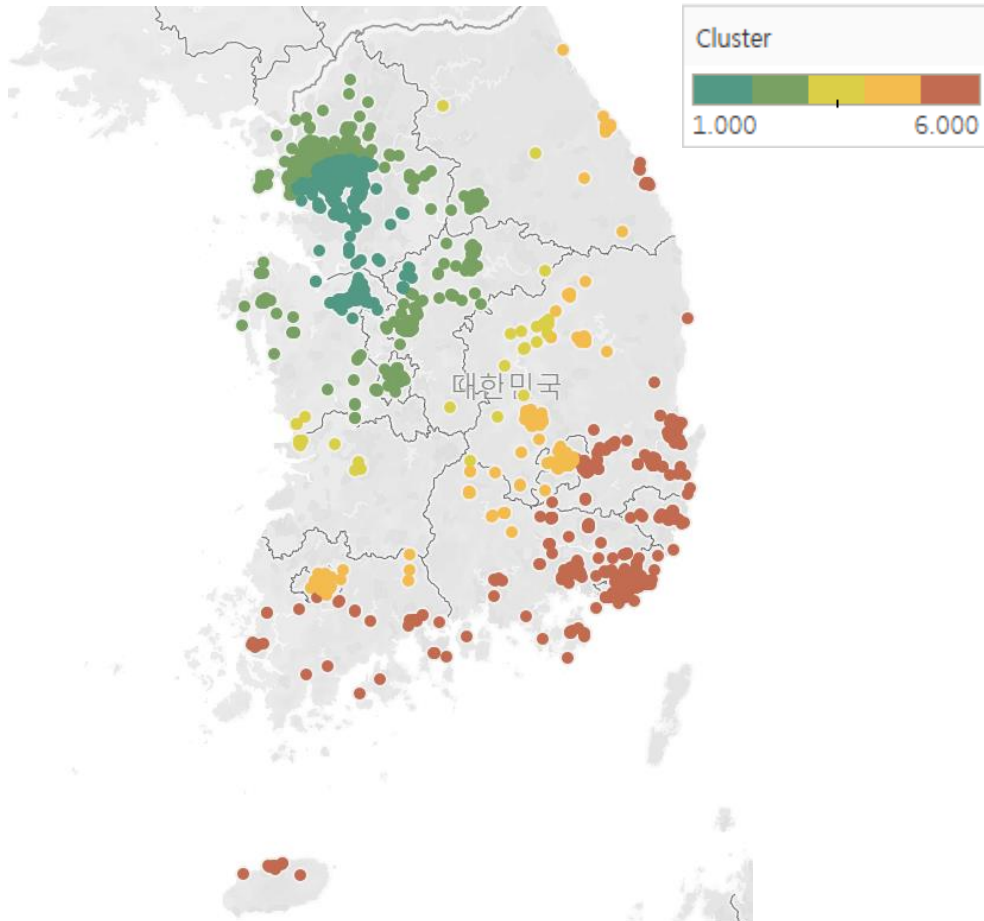
w_{CN} : 특정 원에 해당하는 weight값 \longrightarrow 하이퍼파라미터

CN: Cluster Number

x_i : 확진자가 지나간 경로

days : 일 단위 \longrightarrow 하이퍼파라미터

Clustering



ED: Euclidean Distance

Cluster1: $ED \leq 0.5$

Cluster2: $0.5 < ED \leq 1.0$

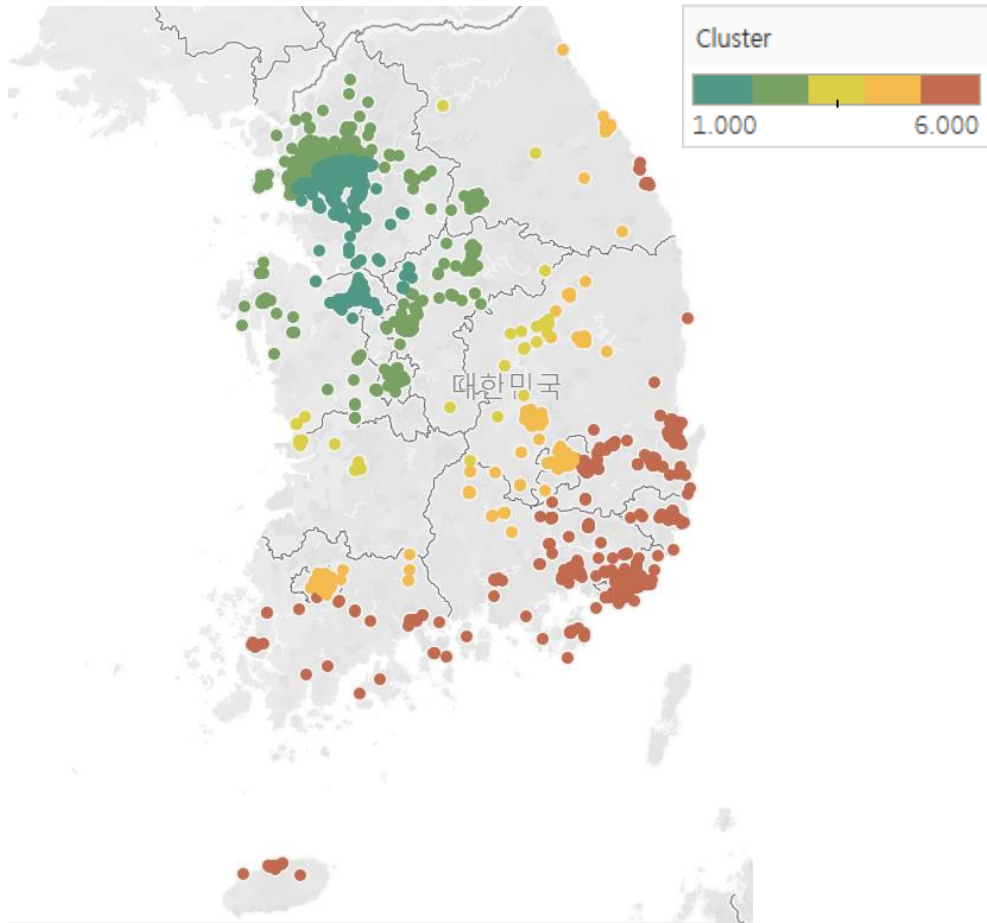
Cluster3: $1.0 < ED \leq 1.5$

Cluster4: $1.5 < ED \leq 2.0$

Cluster5: $2.0 < ED \leq 2.5$

Cluster6: $ED > 2.5$

Weight per cluster



하이퍼파라미터

$$w_{CN} = 0.1 + CN$$

CN: Cluster Number

Cluster1: 0.1의 weight

Cluster2: 0.12의 weight

Cluster3: 0.14의 weight

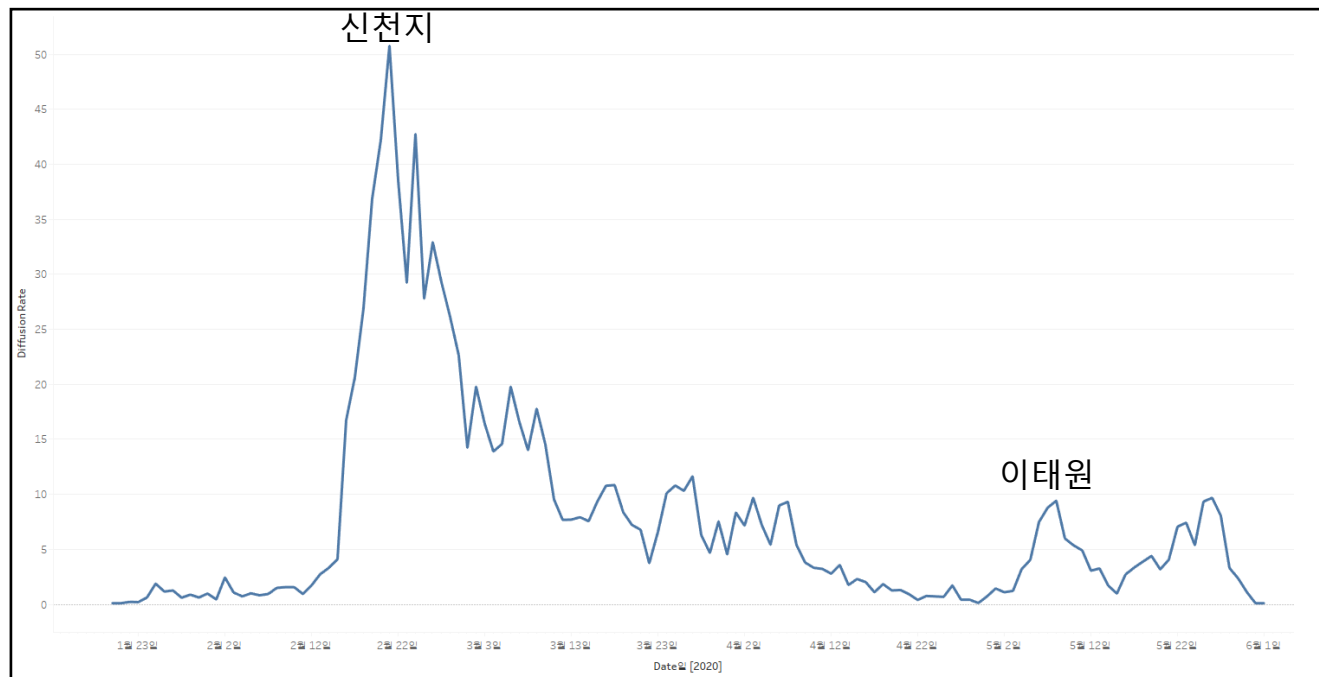
Cluster4: 0.16의 weight

Cluster5: 0.18의 weight

Cluster6: 0.2의 weight

Diffusion rate

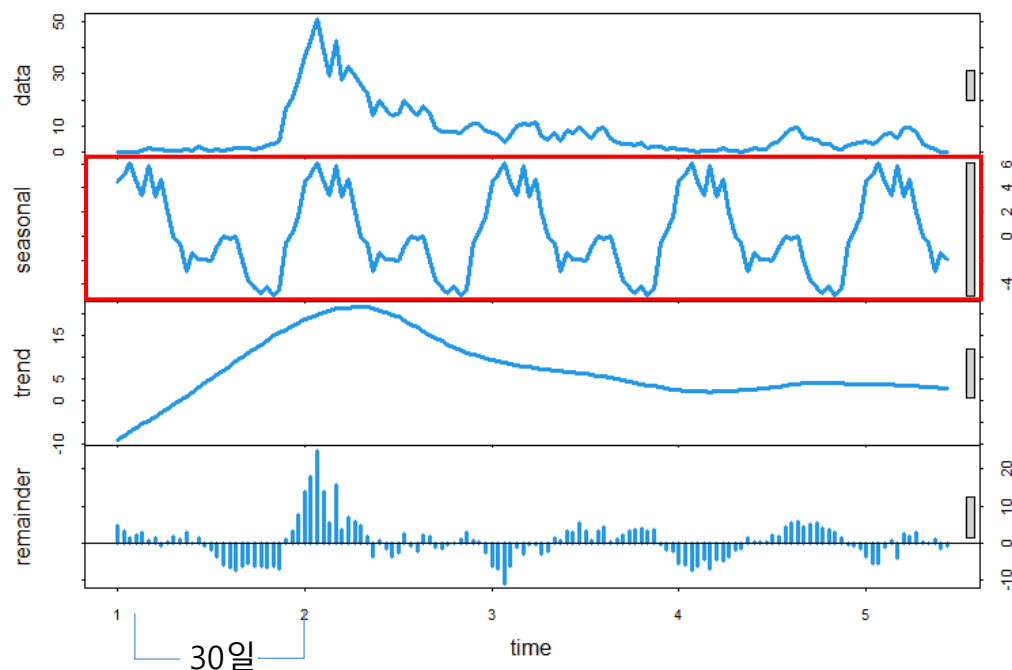
코로나 확산속도



처음에 잠잠했다가 신천지 땀에 한번 확산속도가 크게 커지고
다시 잠잠해졌다가 이태원, 콜센터 등으로 인해 확산속도가 오르락 내리락 하는걸 볼 수 있음

ARIMA 분석(시계열자료 특성 분석)

시계열 요소분해 시각화



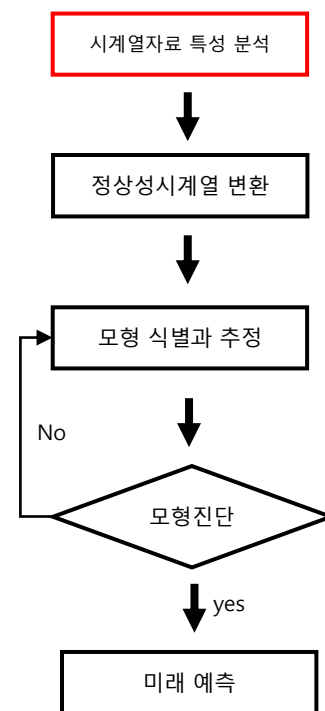
Augmented Dickey-Fuller Test

```
data: tsdata[, 2]  
Dickey-Fuller = -3.0558, Lag order = 5, p-value = 0.1374  
alternative hypothesis: stationary
```

데이터가 계절성을 띤다 → 계절 차분을 적용

귀무가설(H_0): 시계열 데이터가 비정상성이다. ✓
대립가설(H_1): 시계열 데이터가 정상성을 만족한다.

→ 현 데이터는 비정상성 데이터다.

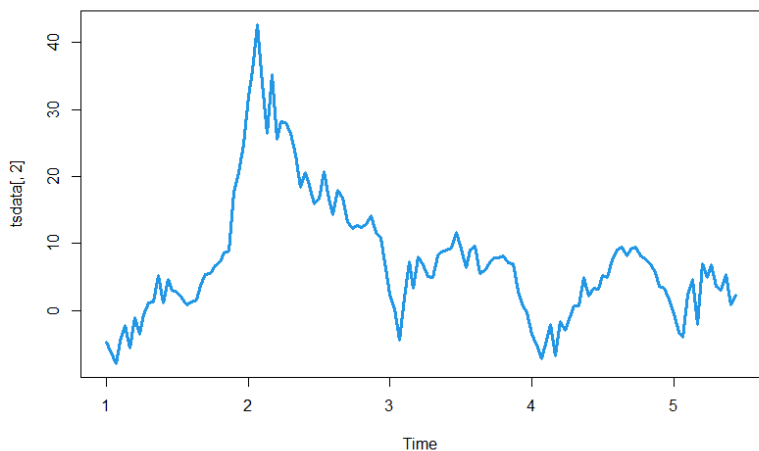


ARIMA 분석(정상성시계열 변환)

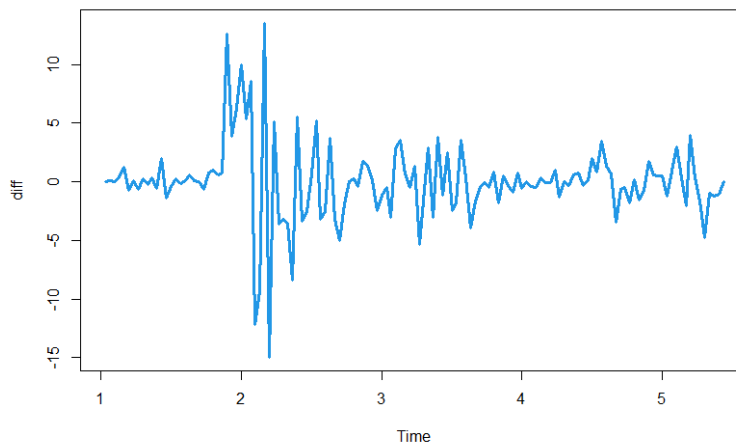
특정 패턴이 보이는 계절성 제거

차분을 통해 비정상성시계열 자료를 정상성시계열로 변환

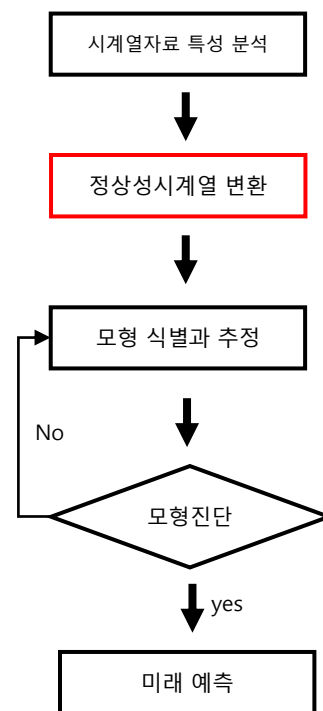
차분: 현재 시점에서 이전 시점의 자료를 빼는 연산



계절성 제거



차분



ARIMA 분석(모형 식별과 추정)

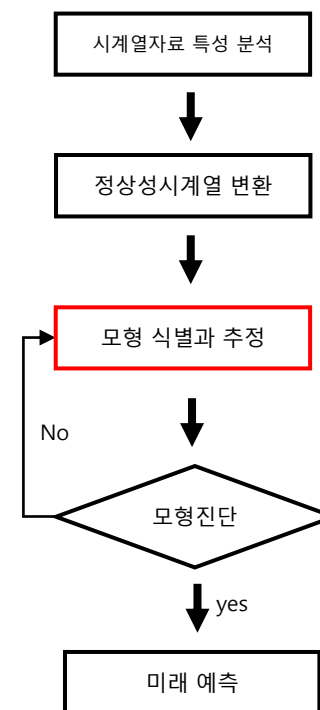
Series: diff[, 2]
ARIMA(4,0,1)(0,0,1)[30] with zero mean

AR모형의 차수: 4
차분 차수: 0
MA모형의 차수: 1

계절성을 갖는 MA모형 차수: 1
계절의 차수: 30일

Coefficients:

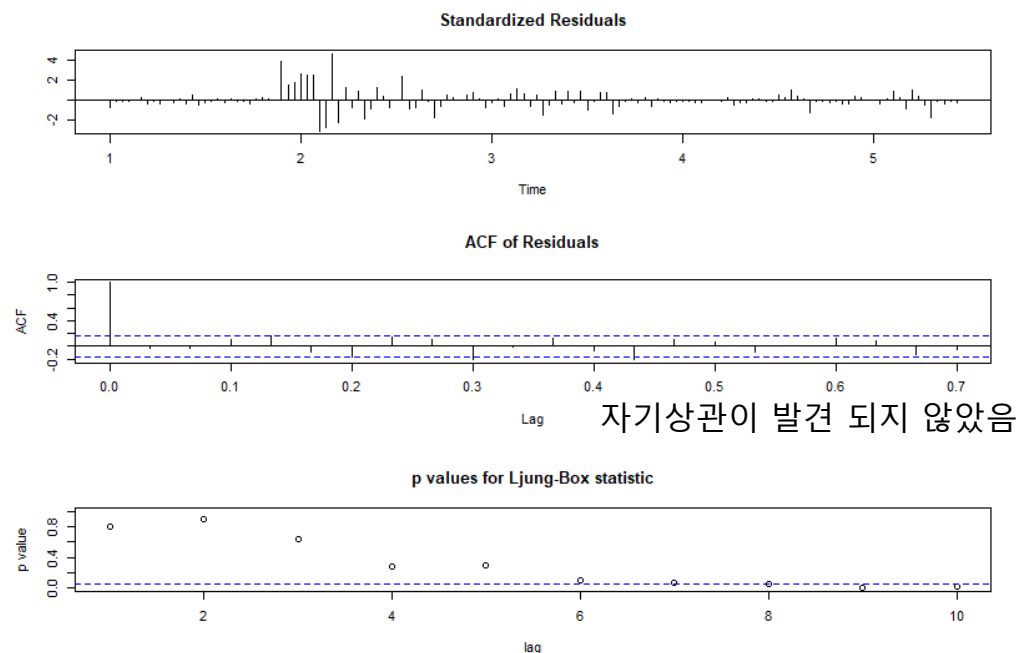
	ar1	ar2	ar3	ar4	ma1	sma1
	-1.0548	-0.0857	0.3082	0.0929	0.8525	0.1484
s.e.	0.1664	0.1310	0.1274	0.1162	0.1355	0.0785



차수: 시간 지연 수
AR(자기상관성)모형
MA(이동평균)모형

ARIMA 분석(모형 진단)

모형의 적합성 검정

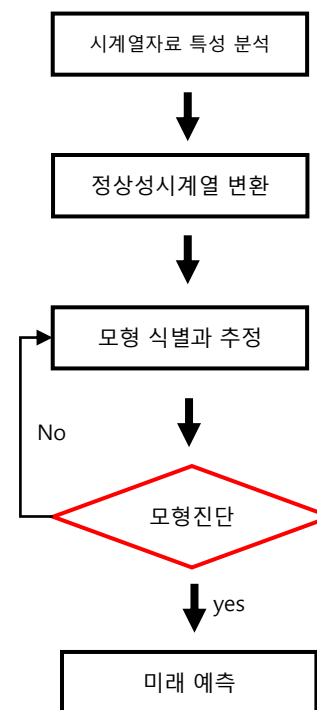


Box-Ljung test

```
data: arima_md$residuals
x-squared = 0.063959, df = 1, p-value = 0.8003
```

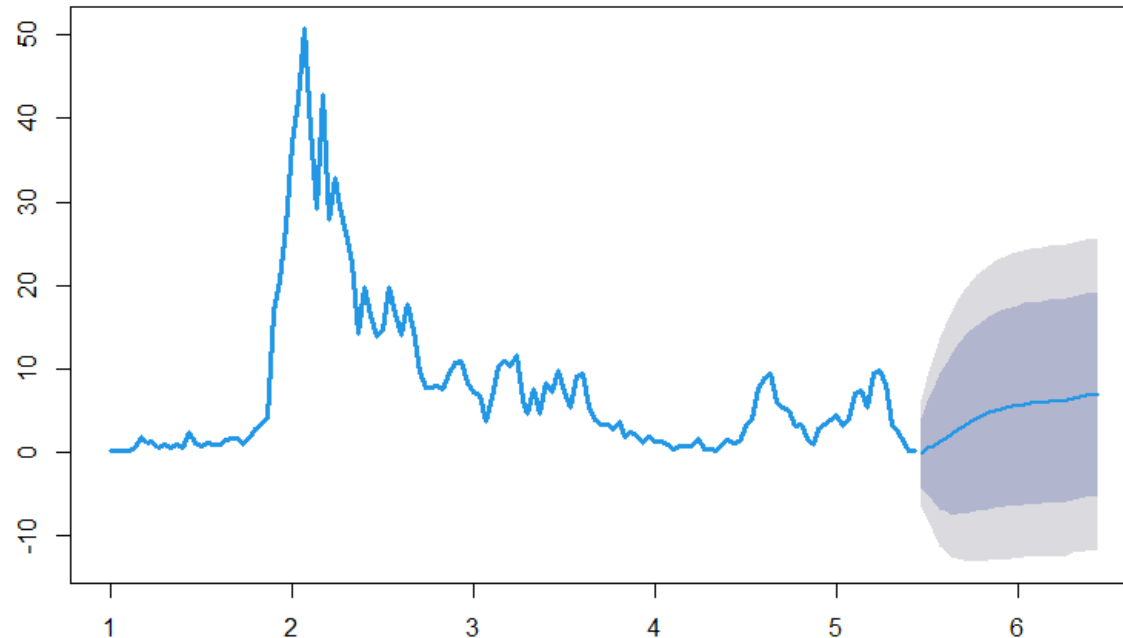
모형의 잔차가 불규칙하고 독립적으로 분포되어 있다.

통계적으로 적절한 모형이다

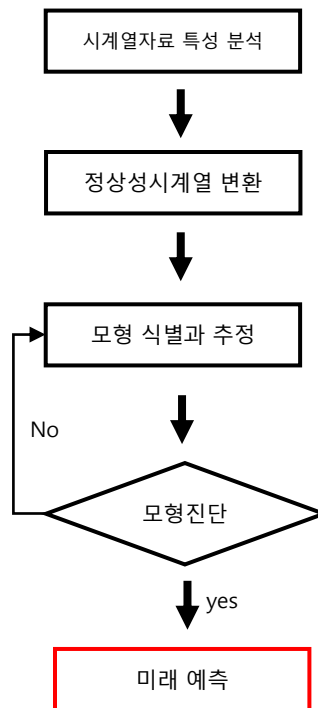


ARIMA 분석(미래 예측)

30일 예측



코로나의 확산속도는 다시 증가하는 추세를 보인다.



결론 및 향후 방향

결론

확산속도가 다시 증가하고 있는 추세다.

향후방향

Clustering과 하이퍼파라미터를 이용해서 weight를 할당했는데, 더 최적화된 값이 필요

인구밀도와 지역 별 거리가 반비례 관계를 가져서 formulation한 식에 weight를 반영하지 않았는데, 인구밀도를 이용해서 식을 update하는 방안 고려