

파이썬 프로그래밍



국내 입국 관광객 분석 프로젝트

조: 4 조

조원: 김태훈, 남경혜, 이상윤, 정재영, 방수영



서울대학교

주요 관광 5 개국 시계열 데이터 분석

- SARIMA 모형을 중심으로

1. 서론

2020년 1월 이후, 전 세계적인 코로나 19의 전파로 인해 미증유의 언택트 시대가 도래했다. 이로 인해 각국 경제 전반의 산업들이 타격을 입고 있으며, 그 중에서도 단연코 관광업이 가장 큰 타격을 입은 산업으로 꼽하고 있다. 하지만 미국, 영국 등 주요 국가로부터의 코로나 19 백신이 개발 및 보급이 되기 시작했고, 추후 코로나 국면이 종식된 이후의 미래의 관광 수요에 대비한 계획이 요구된다 할 수 있다.

따라서 이러한 목적 의식 하에서 우리는 2019년 기준 관광객 입국 TOP5 개국인 중국, 일본, 미국, 대만, 홍콩의 2011~2019년에 걸친 관광 목적 입국자 수 데이터로 시계열 분석을 진행하기로 결정하였다. 코로나 19로 입국 제한이 걸린 특수 상황으로 인해 관광객 수가 급감한 2020년의 데이터는 제외하였으며, '관광'을 목적으로 하는 입국은 강한 계절성을 가지는 데이터라고 판단하여, SARIMA(Seasonal Autoregressive Integrated Moving Average)모형을 본 시계열 분석의 모델링에 활용하기로 하였다.

2. 본론

2.1. 전처리

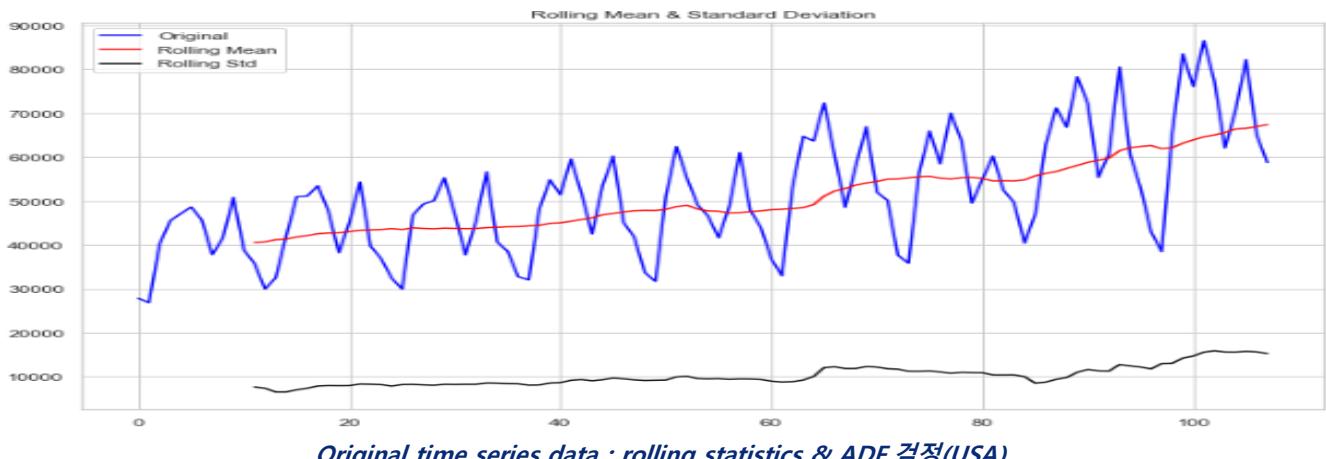
우선 분석을 위해 <https://know.tour.go.kr/>의 '월별 입국 관광 통계' 자료를 활용하였다. 2020년 초 이래로 지속되고 있는 코로나 국면 동안에는 관광을 목적으로 한 입국이 정부 정책 등으로 거의 증발하다시피 했기 때문에, 입국자 수 동향의 일반적인 동향을 파악하고, 나아가 코로나 국면 이후의 data를 모형에 fit 하려는 우리의 목적을 감안하면, 코로나 기간 동안의 데이터를 제외하는 것이 적절하다고 생각하여 2011년 1월 1일부터 2019년 12월 31일에 이르는 기간 동안의 데이터를 추출하여 분석을 진행하였다.

	china	japan	usa	taiwan	hk
2011년01월	55070.0	2011년01월 189601.0	2011년01월 27866.0	2011년01월 24653.0	2011년01월 14083.0
2011년02월	53863.0	2011년02월 213970.0	2011년02월 26837.0	2011년02월 23463.0	2011년02월 21075.0
2011년03월	72003.0	2011년03월 262003.0	2011년03월 40121.0	2011년03월 24907.0	2011년03월 17752.0
2011년04월	86397.0	2011년04월 213645.0	2011년04월 45489.0	2011년04월 31169.0	2011년04월 25070.0
2011년05월	85668.0	2011년05월 229017.0	2011년05월 47110.0	2011년05월 27611.0	2011년05월 18603.0
...
2019년09월	432018.0	2019년09월 242475.0	2019년09월 70634.0	2019년09월 100888.0	2019년09월 44708.0
2019년10월	476460.0	2019년10월 241484.0	2019년10월 82230.0	2019년10월 126421.0	2019년10월 64439.0
2019년11월	426849.0	2019년11월 251663.0	2019년11월 65116.0	2019년11월 100595.0	2019년11월 57442.0
2019년12월	433577.0	2019년12월 248793.0	2019년12월 58743.0	2019년12월 88881.0	2019년12월 70638.0
계	35081798.0	계 24066060.0	계 5659970.0	계 6825888.0	계 4786867.0

2011~2019년 중, 일, 미, 대, 홍 5개국 관광 입국자 수 데이터

2.2. 국가별 분석

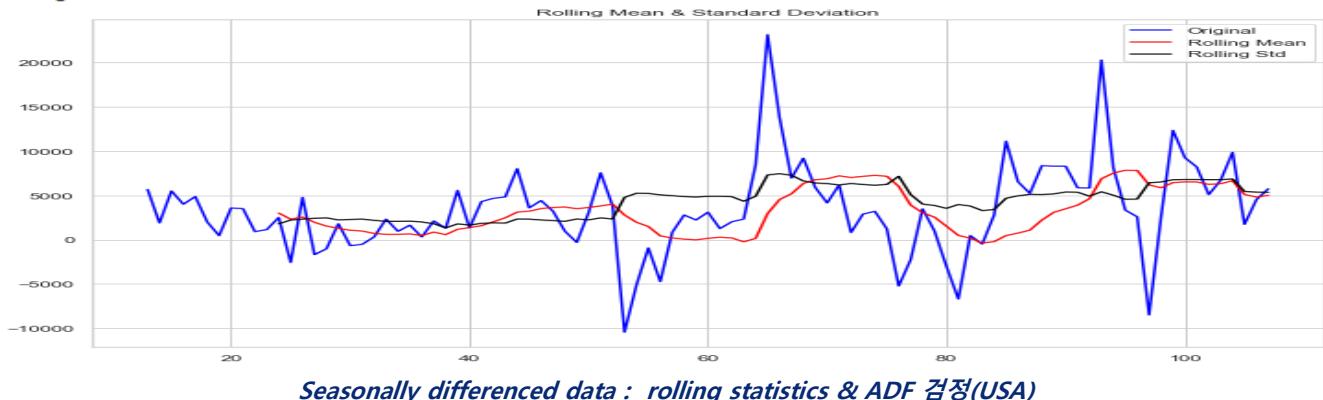
2.2.1 미국



관광 목적의 미국 입국자 수 데이터를 살펴보자. 추가적으로 다른 국가들의 데이터를 살펴보면 보다 분명해지겠지만, 관광의 계절성이 가장 분명하게 관찰된다. 지리적 기후적 유사성을 가진 같은 아시아 국가들보다는 다른 대륙 국가의 관광객들에게 관광의 계절적 동기가 더 강하게 작용하는 것으로 판단된다. 또한 한미 간 관광에 영향을 줄 별다른 이슈가 없었기 때문에 연 단위의 계절적 패턴이 장기적 상승 추세 상에서 일정하게 반복됨을 확인할 수 있다. Rolling statistics 를 살펴보면, 평균과 분산은 점차 커지는 것으로 보여 비정상성이 의심된다. 아래의 ADF 검정 결과를 보면, P 값이 0.99로 비정상임을 확인할 수 있다.

Results of Dickey-Fuller Test:

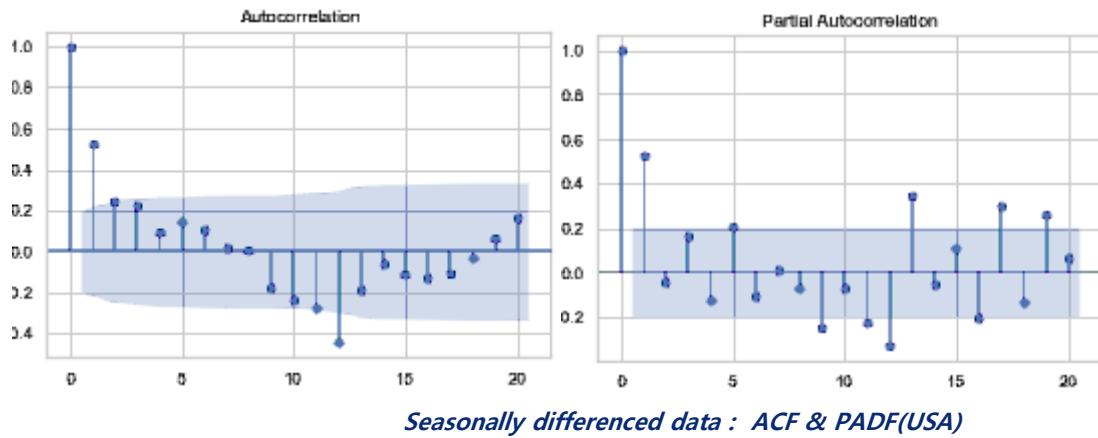
Test Statistic	1.034483
p-value	0.994617



12 개월 전에 대해 계절 차분한 시계열과 rolling statistics 를 살펴보면, 계절성이 눈에 띄게 사라지고 rolling statistics 가 많이 안정화되었음을 확인할 수 있다. 추가적으로 ADF 검정 결과를 보면, P 값은 5% 유의수준에서는 boundary significant 하고 10% 유의수준에서는 insignificant 하여 정상시계열로 판단해도 크게 무리가 없을 것 같다.

Results of Dickey-Fuller Test:

Test Statistic	-2.791419
p-value	0.059510



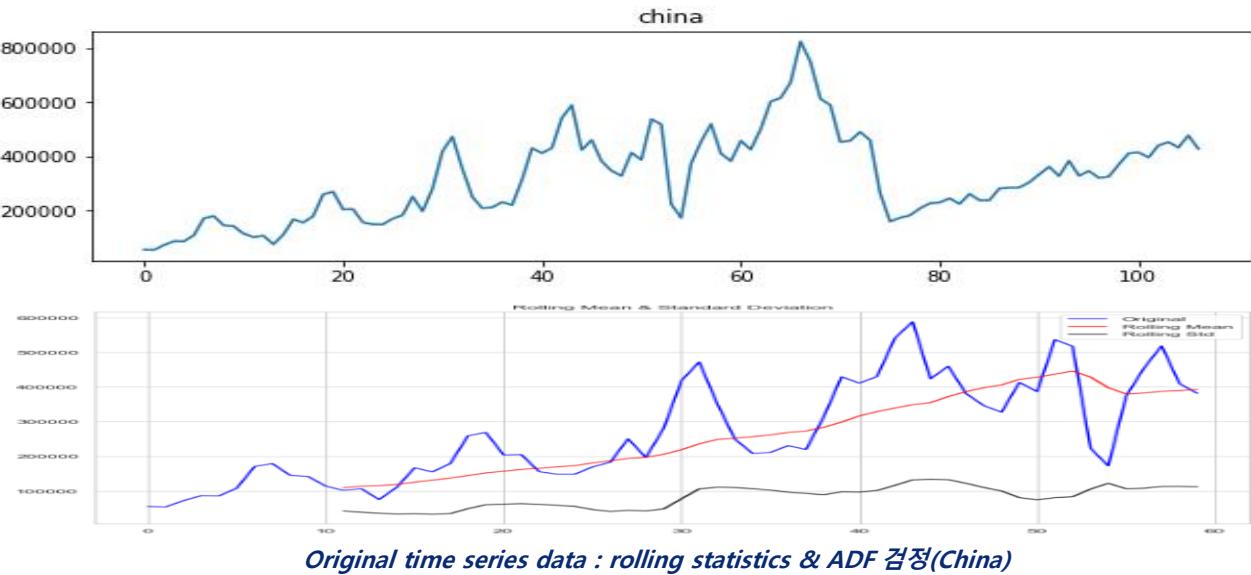
ACF 를 살펴보면 첫 번째, 두 번째, 열 두번째 값만이 significant 함을 관찰할 수 있다. PACF 를 살펴보면 첫 번째, 두 번째, 열두 번째 값이 significant 하며, 이 외에도 significant 한 값들이 관찰되지만, boundary significant 수준과 별 차이가 없는데다, 계절성을 강하게 띠는 관광 데이터 특성을 감안하여 나머지 significant 한 값들은 무시해도 될 것 같다고 판단했다. 요약하자면, 미국 관광 목적 입국자 수 시계열은 SARIMA(2,0,2)(1,1,1,12) 모형을 따르는 것으로 예상된다.

SARIMAX Results						
Dep. Variable:	usa	No. Observations:	108			
Model:	SARIMAX(2, 0, 2)x(1, 1, [1], 12)	Log Likelihood	-937.612			
Date:	Sun, 31 Jan 2021	AIC	1889.225			
Time:	15:15:35	BIC	1907.175			
Sample:	0 - 108	HQIC	1896.480			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	1.2275	0.228	5.378	0.000	0.780	1.675
ar.L2	-0.2288	0.225	-1.015	0.310	-0.671	0.213
ma.L1	-0.7898	0.256	-3.088	0.002	-1.291	-0.288
ma.L2	-0.1987	0.242	-0.821	0.411	-0.673	0.275
ar.S.L12	0.4185	0.225	1.862	0.063	-0.022	0.859
ma.S.L12	-0.6175	0.267	-2.310	0.021	-1.141	-0.094
sigma2	1.832e+07	2.83e-08	6.47e+14	0.000	1.83e+07	1.83e+07
Ljung-Box (L1) (Q):	1.50	Jarque-Bera (JB):	28.77			
Prob(Q):	0.22	Prob(JB):	0.00			
Heteroskedasticity (H):	1.83	Skew:	0.25			
Prob(H) (two-sided):	0.09	Kurtosis:	5.63			

SARIMA(2,0,2)(1,1,1,12) 모형에 fit 한 결과 summary(USA)

모형 추정 결과 ar.L2, ma.L2 변수가 10% 유의수준에서도 insignificant 함이 확인되어, 최종적으로는 SARIMA(1,0,1)(1,1,1,12)모형이 가장 적합한 것으로 판단하겠다.

2.2.2 중국

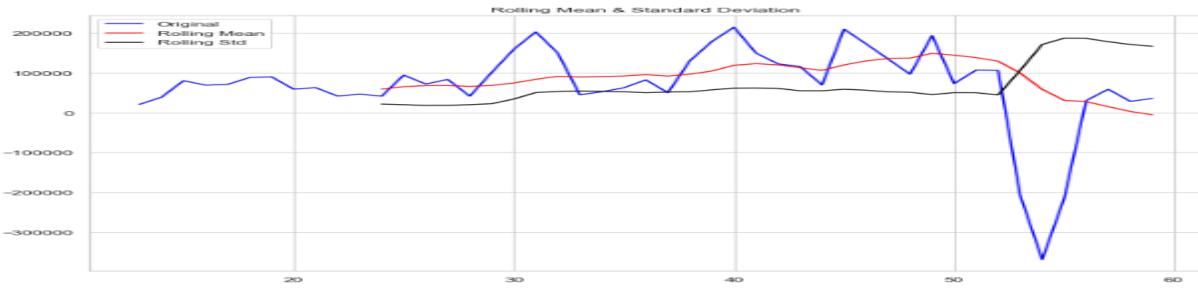


관광 목적의 중국 입국자 수 데이터를 살펴보자. 첫 번째 데이터는 2011~2019 동안의 데이터인데, 가장 눈에 띄는 점은 2016년 사드 배치 확정 이후 이에 대한 보복으로 적용된 '한한령'으로 관광객이 급감하였다는 것이다. 2020년 들어 중국의 한한령 해제 움직임이 시작됐으나 코로나 19 여파로 다시금 중단되었다. 따라서 중국의 경우 다른 나라들과는 달리 한한령 이전인 2011~2015년의 데이터(두 번째 그래프에 해당)를 이용한 분석을 진행하였다.

미국보다는 약한 계절성을 보이며 평균과 분산이 점차 증가하는 비정상 시계열로 예측된다. 아래의 ADF 검정 결과를 보면 P 값이 0.82로 비정상 시계열임을 확인할 수 있다.

Results of Dickey-Fuller Test:

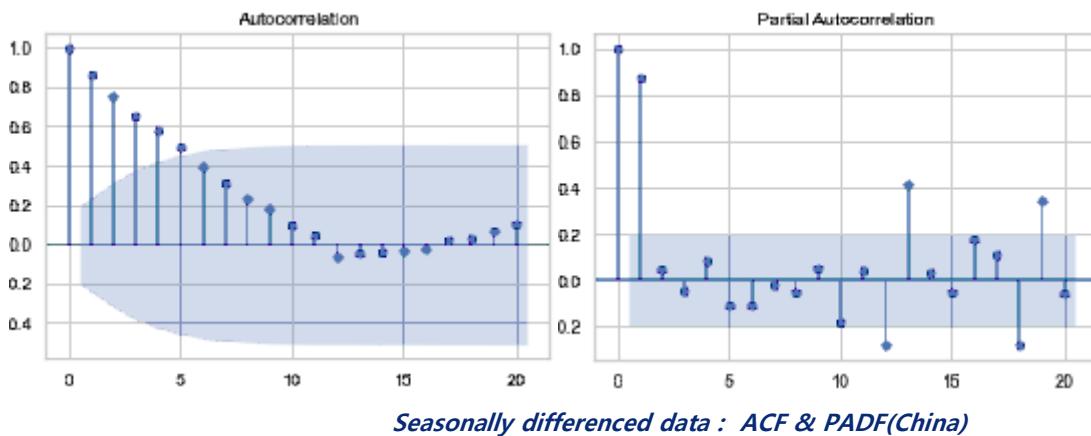
Test Statistic	-0.780110
p-value	0.824895



한한령 이전 데이터에 대해 전년 동월값을 차분한 데이터를 살펴보면, rolling statistics 가 훨씬 안정적으로 바뀌었고, ADF 검정 결과로 얻은 P 값도 0.003으로 정상시계열이 되었음을 알 수 있다.

Results of Dickey-Fuller Test:

Test Statistic	-3.760913
p-value	0.003334



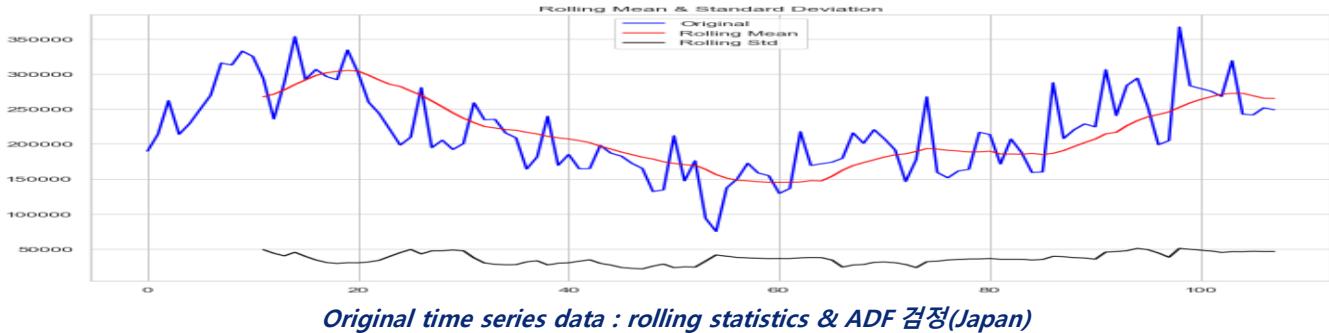
ACF 를 살펴보면 1~6 번째 값만 significant 함을 관찰할 수 있다. PACF 를 살펴보면 1,2,12,13,18,19 번째 값이 significant 하다. 13, 18, 19 번째 값은 significant 하지만 그 값이 크지 않고, 계절성을 강하게 띠는 관광 데이터 특성을 감안하여 첫 번째와 두 번째 값 이외에 12 번째 값만 반영하기로 결정했다. 중국 관광 목적 입국자 수 시계열은 SARIMA(5,0,2)(0,1,1,12) 모형을 따르는 것으로 예상된다.

SARIMAX Results						
Dep. Variable:	china	No. Observations:	60			
Model:	SARIMAX(5, 0, 2)x(0, 1, [1], 12)	Log Likelihood	-600.073			
Date:	Sun, 31 Jan 2021	AIC	1218.145			
Time:	16:55:01	BIC	1234.986			
Sample:	0 - 60	HQIC	1224.509			
Covariance Type:	opg					
coef	std err	z	P> z	[0.025	0.975]	
ar.L1	-0.2097	0.262	-0.801	0.423	-0.722	0.303
ar.L2	0.1440	0.348	0.414	0.679	-0.538	0.826
ar.L3	0.2507	0.209	1.198	0.231	-0.159	0.661
ar.L4	0.0855	0.349	0.245	0.807	-0.599	0.770
ar.L5	0.0847	0.338	0.250	0.802	-0.578	0.748
ma.L1	1.4572	0.187	7.807	0.000	1.091	1.823
ma.L2	0.9564	0.176	5.434	0.000	0.611	1.301
ma.S.L12	0.0040	0.389	0.010	0.992	-0.759	0.767
sigma2	5.387e+09	1.1e-10	4.9e+19	0.000	5.39e+09	5.39e+09
Ljung-Box (L1) (Q):	0.08	Jarque-Bera (JB):	139.88			
Prob(Q):	0.78	Prob(JB):	0.00			
Heteroskedasticity (H):	14.70	Skew:	-1.44			
Prob(H) (two-sided):	0.00	Kurtosis:	10.85			

SARIMA(5,0,2)(0,1,1,12) 모형에 fit 한 결과 summary(China)

ma.L1, ma.L2 를 제외하고는 대부분의 변수들이 유의하지 않게 나타났다. 이 결과에 따르면 중국 관광 목적 입국자 수 데이터는 SARIMA(0,0,2)(0,1,0,12) 모형을 따르는 것으로 판단된다. 그러나 중국의 경우 고려 기간이 짧고, 또한 고려 기간이 짧음으로 인해 메르스의 영향이 모형 적합에 큰 영향을 끼칠 것이므로 더 긴 시계열에서 확장된 모형을 통해 다시 적합해 볼 필요성이 있을 것으로 판단된다.

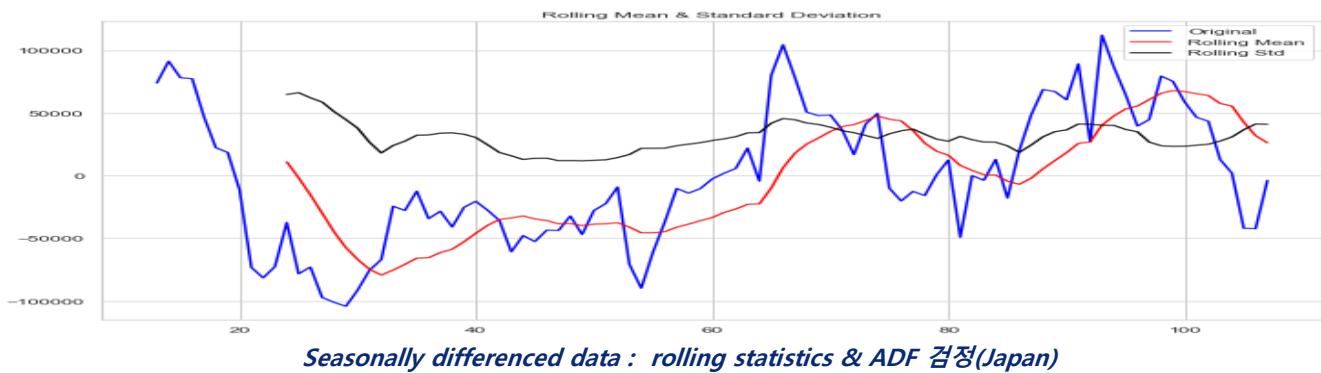
2.2.3 일본



관광 목적의 일본 입국자 수 데이터를 살펴보자. 인접 국가 인만큼 미국 보다는 약한 계절성을 보인다. 2015년 이전에는 서서히 감소하는 추세를 보이다가 메르스 사태때 저점을 찍은 이후 다시 서서히 증가하는 형태를 보인다. 랜덤워크 시계열처럼 보이지만 분명하지 않아 추가적인 확인이 필요해 보인다. ADF 검정 결과를 보면, P 값은 0.11로 10% 유의수준에서도 insignificant 하다. 따라서 비정상 시계열이라고 판단할 수 있다.

Results of Dickey-Fuller Test:

Test Statistic	-2.493196
p-value	0.117139



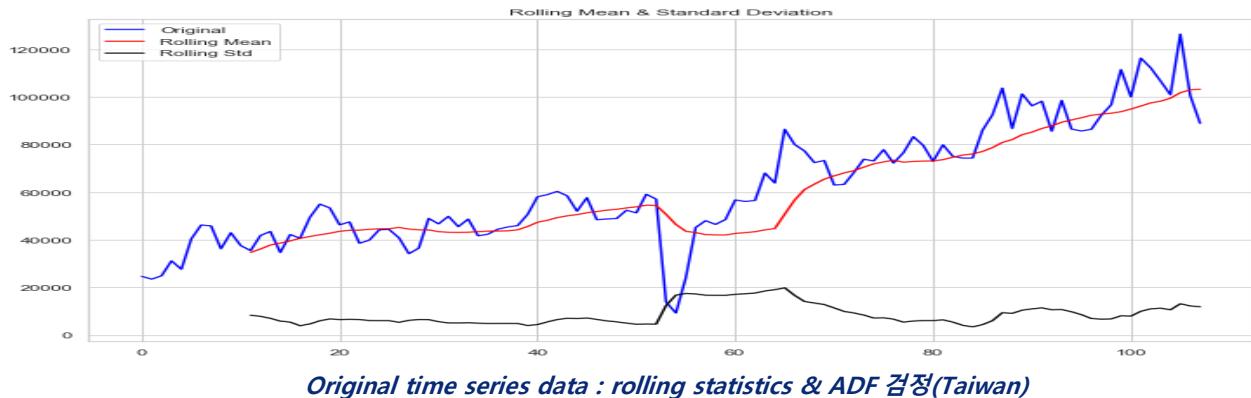
12개월 전에 대해 계절 차분한 시계열과 rolling statistics를 살펴보면, 계절차분을 수행했음에도 안정화되는 양상을 찾기는 힘들어 보인다. 오히려 원시계열의 경우보다 시계열 자체와 rolling statistics가 더욱 불안정한 양상을 보여 여전히 불안정 시계열인 것으로 판단된다. 추가적으로 ADF 검정 결과를 보면, P 값은 0.31로 비정상성이 더 크게 나타남을 알 수 있다.

Results of Dickey-Fuller Test:

Test Statistic	-1.931447
p-value	0.317415

이러한 결과는 한일 간 정치경제적 갈등과 무관하지 않아 보인다. 한일간에는 위안부 문제와 아베 정권의 반한 정책으로 인한 불매운동 등으로 갈등이 지속되어 왔고, 이러한 문제가 시계열이 불안정해지는 형태로 반영된 것으로 보인다. 정치경제적 문제를 정량화된 변수로 모형에 반영하기는 힘들기 때문에 추가적으로 ACF와 PACF를 확인하여 SARIMA 모형에 fit 하는 작업은 수행하지 않았다.

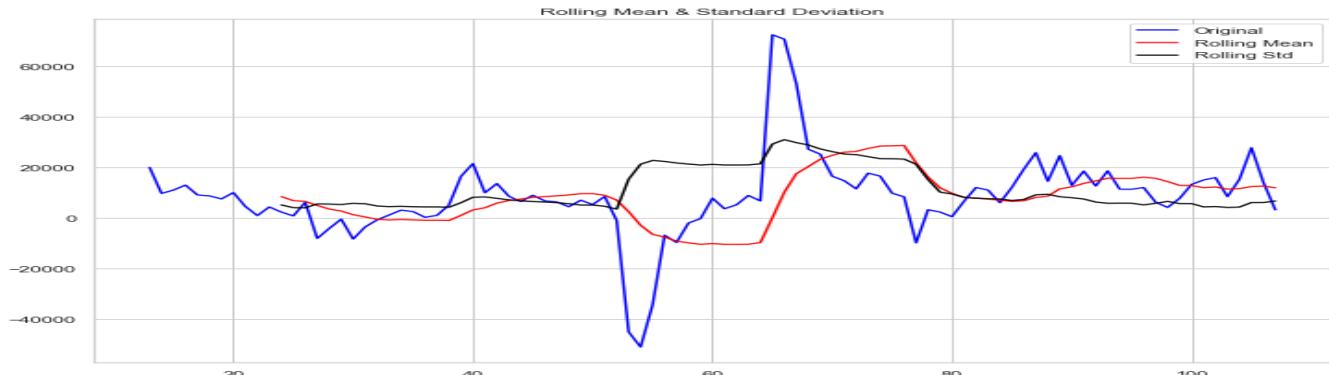
2.2.4 대만



관광 목적의 대만 입국자 수 데이터를 살펴보자. 같은 아시아 국가 인만큼 미국 보다는 다소 완화된 계절성을 보인다. 가장 눈에 띄는 점은 2015년 메르스 사태로 인해 6~8월 경 관광객이 급감한 것인데, 이때를 제외하고는 완만한 상승 추세를 보여준다. Time trend 만 제거하면 안정화되는 시계열인지, 아니면 Time trend 상의 비정상 시계열, 즉 drift on trend의 경우인지 그래프만으로는 판단하기 어렵다. 아래의 ADF 검정 결과를 보면, P 값이 0.84로 비정상이 확인된다.

Results of Dickey-Fuller Test:

Test Statistic	-0.715402
p-value	0.842720

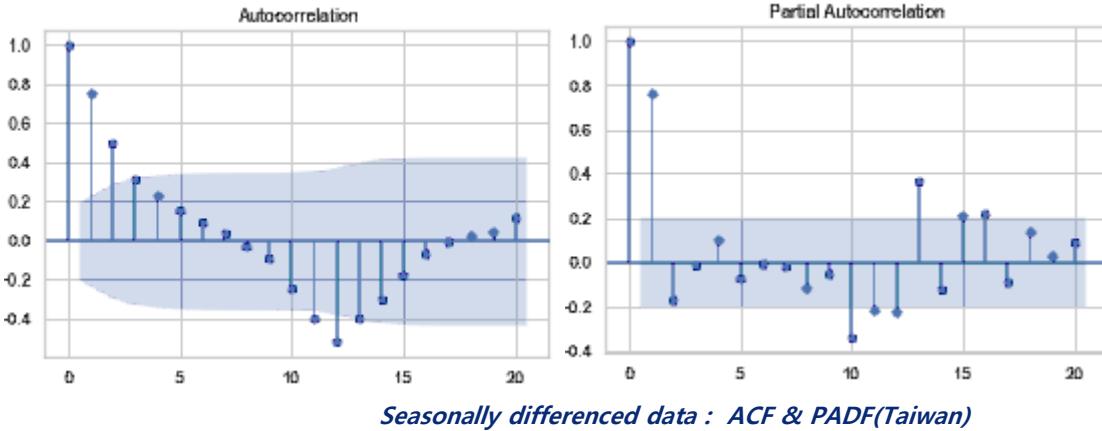


Seasonally differenced data : rolling statistics & ADF 검정(Taiwan)

12개월 전에 대해 계절 차분한 시계열과 rolling statistics를 살펴보면, 우선 계절성이 눈에 띄게 사라졌음을 확인할 수 있고, 메르스 사태 당시의 데이터를 제외하면 차분된 시계열과 rolling statistics 모두 훨씬 안정화된 모습을 확인할 수 있다. 추가적으로 ADF 검정 결과를 보면, P 값은 0.08로 5% 유의수준에서는 insignificant하고 10% 유의수준에서는 significant하여 10% 유의수준에서는 정상시계열로 판단된다.

Results of Dickey-Fuller Test:

Test Statistic	-2.662317
p-value	0.080789



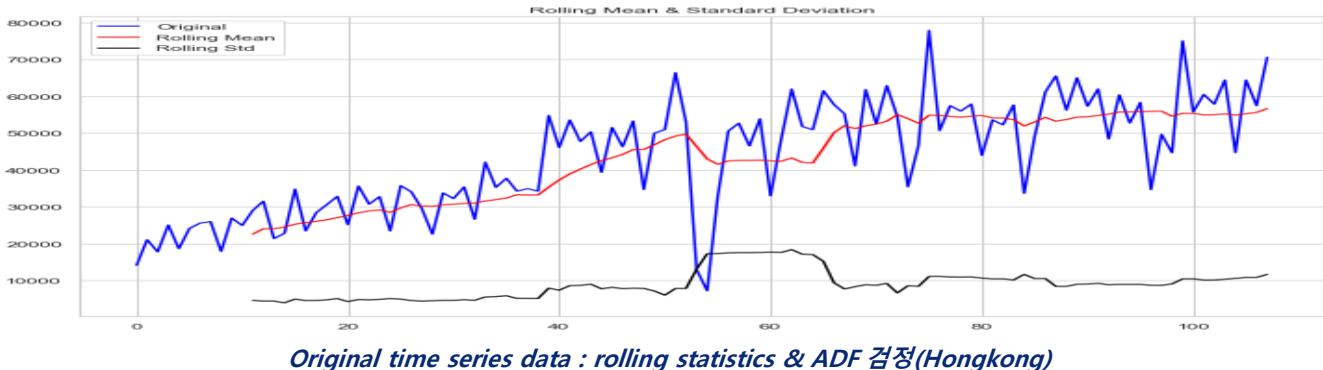
ACF의 경우 첫 번째, 두 번째, 세 번째, 열두 번째 값이 significant 하며, insignificant 한 것으로 간주해도 될 것 같다. PACF의 경우 첫 번째, 두 번째, 열세 번째 값이 significant 하고, 나머지는 boundary 안에 위치하거나 boundary significant 한 것으로 보여, 역시 insignificant 한 것으로 간주해도 될 것 같다. 다만, 13 개월을 주기로 움직인다고 생각하기는 힘들어서, 13 개월 대신 12 개월로 대체하여, 대만 관광 목적 입국자 수 시계열은 SARIMA(3,0,2)(1,1,1,12) 모형을 따를 것으로 예측했다.

SARIMAX Results						
Dep. Variable:	taiwan	No. Observations:	108			
Model:	SARIMAX(3, 0, 2)x(1, 1, [1], 12)	Log Likelihood	-1010.912			
Date:	Sun, 31 Jan 2021	AIC	2037.824			
Time:	15:17:38	BIC	2058.339			
Sample:	0 - 108	HQIC	2046.116			
Covariance Type:	opg					
coef	std err	z	P> z	[0.025	0.975]	
ar.L1	1.7551	4.062	0.432	0.666	-6.206	9.716
ar.L2	-0.8072	7.053	-0.114	0.909	-14.631	13.016
ar.L3	0.0514	2.996	0.017	0.986	-5.821	5.924
ma.L1	-0.9249	4.071	-0.227	0.820	-8.904	7.055
ma.L2	-0.0419	4.002	-0.010	0.992	-7.886	7.802
ar.S.L12	-0.0135	0.345	-0.039	0.969	-0.689	0.663
ma.S.L12	-0.6719	0.340	-1.973	0.048	-1.339	-0.005
sigma2	1.146e+08	1.03e-06	1.11e+14	0.000	1.15e+08	1.15e+08
Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	721.57			
Prob(Q):	0.97	Prob(JB):	0.00			
Heteroskedasticity (H):	2.81	Skew:	-1.43			
Prob(H) (two-sided):	0.00	Kurtosis:	16.12			

SARIMA(3,0,2)(1,1,1,12) 모형에 fit 한 결과 summary(Taiwan)

대부분의 변수들이 유의하지 않게 나타났다. 대만의 경우 유독 메르스로 인한 관광객 급감의 효과가 크게 나타났는데, 이런 변수들이 모형에 반영되지 않음으로 인해 모형 fitting 이 잘 되지 않은 결과가 아닌가 생각된다. 대만의 데이터가 어떤 모형을 따르는지는 판단을 유보하고 좀 더 긴 시계열 데이터에 추가적인 설명변수를 도입한 확장된 모형으로 다시 적합해 볼 필요가 있어 보인다.

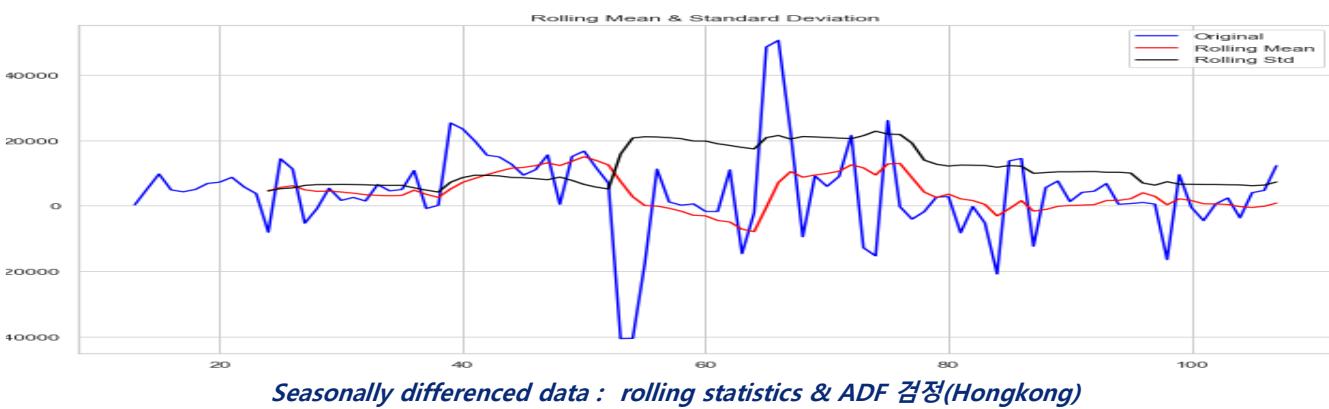
2.2.5 홍콩



관광 목적의 홍콩 입국자 수 데이터를 살펴보자. 같은 아시아 국가인만큼 미국 보다는 다소 완화된 계절성을 보인다. 가장 눈에 띄는 점은 2015년 메르스 사태로 인해 6~8월 경 관광객이 급감한 것인데, 이때를 제외하고는 완만한 상승 추세를 보여준다. 평균과 분산이 점차 증가하는 비정상 시계열의 형태를 보여준다. 아래의 ADF 검정 결과를 보면, P 값이 0.60으로 비정상성이 확인된다.

Results of Dickey-Fuller Test:

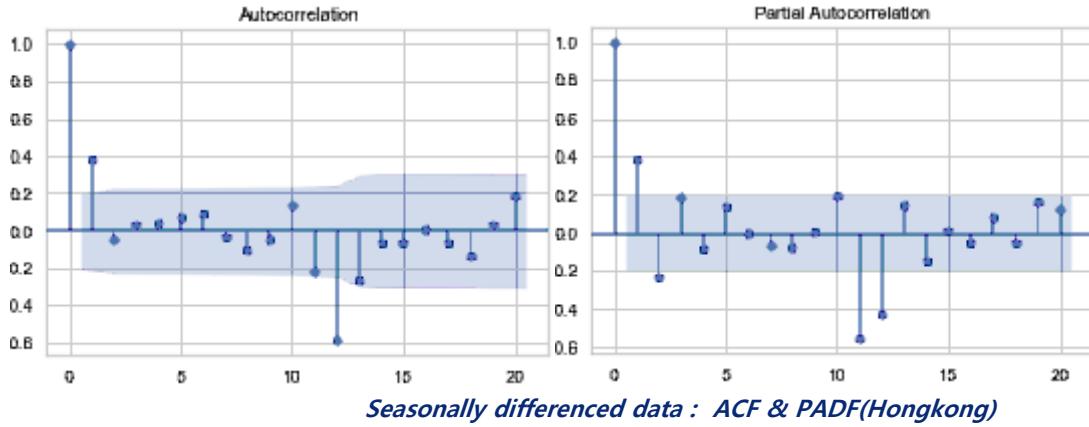
Test Statistic	-1.356676
p-value	0.602861



12개월 전에 대해 계절 차분한 시계열과 rolling statistics를 살펴보면, 우선 계절성이 눈에 띄게 사라졌음을 확인할 수 있고, 메르스 사태 당시의 데이터를 제외하면 차분된 시계열과 rolling statistics 모두 훨씬 안정화된 모습을 확인할 수 있다. 추가적으로 ADF 검정 결과를 보면, P 값은 0.00003로 아주 낮은 유의수준에서도 significant하여 정상시계열이 되었음을 알 수 있다.

Results of Dickey-Fuller Test:

Test Statistic	-4.933451
p-value	0.000030



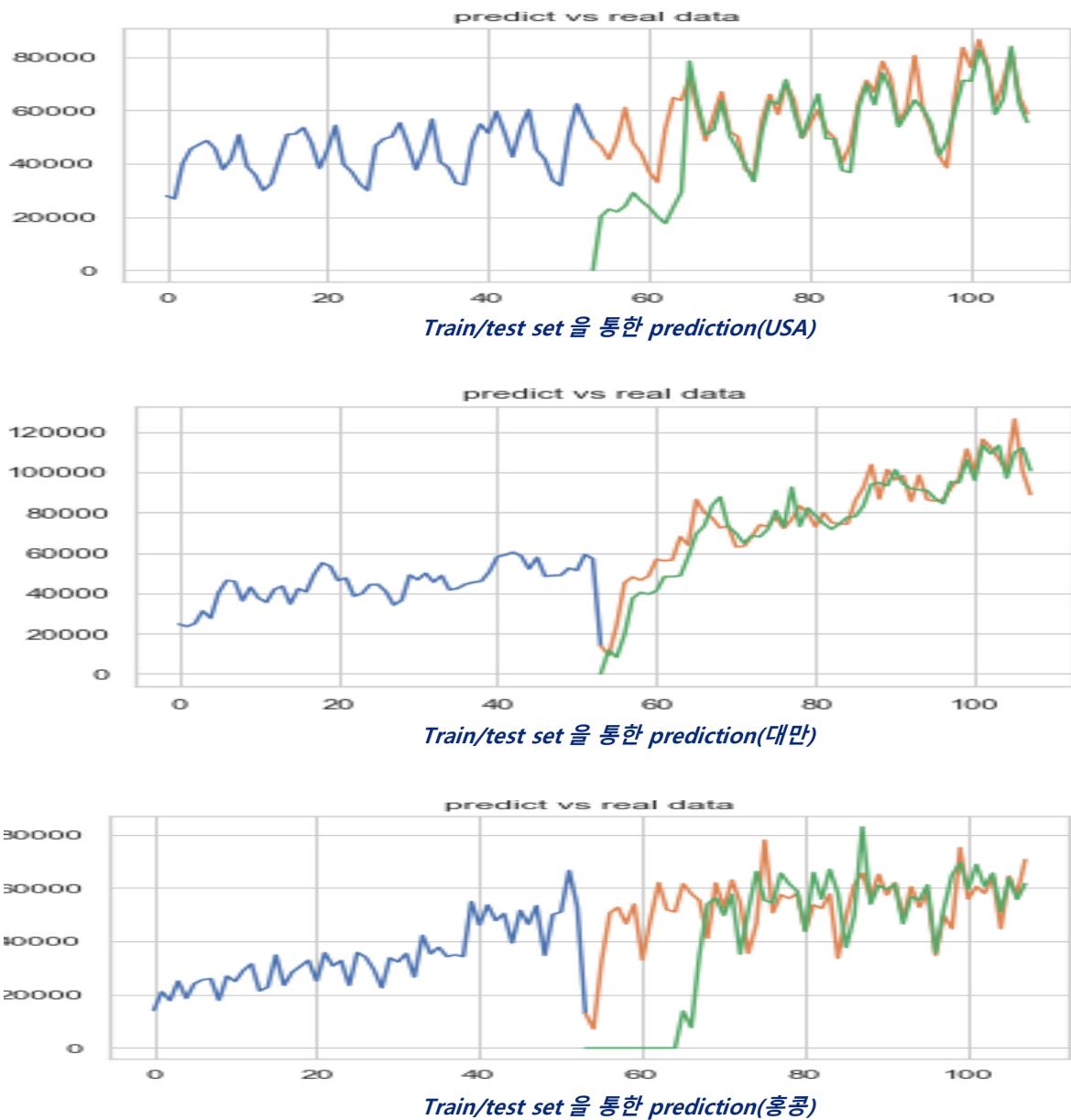
ACF의 경우 첫 째, 두 째, 열두 째 값까지는 통계적으로 유의한 것이 관찰되며, PACF의 경우 첫 번째, 두 번째, 열한 번째, 열두 번째 값이 통계적으로 유의함이 관찰된다. 따라서 대만 관광 목적 입국자 수 시계열은 SARIMA(2,0,2)(1,1,1,12) 모형을 따르는 것으로 볼 수 있을 것 같다.

SARIMAX Results						
Dep. Variable:	hk	No. Observations:	108			
Model:	SARIMAX(2, 0, 2)x(1, 1, [1], 12)	Log Likelihood	-1017.285			
Date:	Sun, 31 Jan 2021	AIC	2048.571			
Time:	15:21:44	BIC	2066.521			
Sample:	0 - 108	HQIC	2055.826			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.9866	0.497	1.985	0.047	0.012	1.961
ar.L2	0.0053	0.491	0.011	0.991	-0.958	0.968
ma.L1	-0.5065	0.514	-0.986	0.324	-1.513	0.500
ma.L2	-0.3658	0.437	-0.837	0.403	-1.223	0.491
ar.S.L12	-0.2852	0.229	-1.247	0.212	-0.733	0.163
ma.S.L12	-0.5314	0.252	-2.109	0.035	-1.025	-0.038
sigma2	1.439e+08	1.41e-08	1.02e+16	0.000	1.44e+08	1.44e+08
Ljung-Box (L1) (Q):	0.01	Jarque-Bera (JB):	98.21			
Prob(Q):	0.92	Prob(JB):	0.00			
Heteroskedasticity (H):	1.60	Skew:	-1.41			
Prob(H) (two-sided):	0.19	Kurtosis:	7.08			

SARIMA(2,0,2)(1,1,1,12) 모형에 fit 한 결과 summary(Hongkong)

모형 추정 결과 ar.L12, ma.L12를 제외한 나머지 변수들은 유의하지 않은 것으로 관찰되었다. 이 결과에 따르면 홍콩 관광 입국자 수 시계열은 SARIMA(0,0,0)(1,1,1,12)를 따른다고 해석할 수 있다. 그러나 홍콩 역시 대만과 마찬가지로 메르스의 영향이 아주 크게 나타났는데, 이것이 모형에 반영되지 않았으므로 선불리 판단을 내리기보다는 좀 더 긴 시계열 데이터에 추가적인 설명변수를 도입한 확장된 모형으로 다시 적합해 볼 필요가 있다고 판단된다.

3. Prediction



모형 fitting 이 가장 잘 되지 않는 일본과 데이터가 가장 적은 중국을 제외한 3 개국의 데이터를 train / test 으로 분할하여 test 기간에 대한 prediction 을 해본 결과가 위의 그래프이다.

분할 시점이 메르스 사태 시점과 일치하여, 분할 시점 직후의 예측은 교란됨을 보이나 나머지 기간에 대해서는 우수한 예측력을 보임이 확인된다. 코로나 국면 이후의 예측에도 이러한 모형 활용이 가능할 것으로 보이나, 코로나 이후 관광의 대기수요가 폭발할지 혹은 여행을 꺼리는 세태가 지속될 것인지에 따라 모형이 제시하는 추세보다 더 높거나 낮은 관광객 수가 관찰될 것이므로, 코로나 사태에 대한 지속적인 monitoring 과 연구가 필요할 것이다.

4. 결론

지금까지 SARIMA 모형을 중심으로 주요 5 개국의 관광 입국자 수 데이터를 분석해 보았다. 분석 결과 관광 입국자 수 시계열의 특징은 다음으로 요약할 수 있을 것 같다.

첫째, 강한 계절성을 보이며, 지리적 기후적 유사성이 낮아질수록 계절성이 더 크게 나타나는 것으로 보인다. 이는 다른 아시아 국가들보다 미국의 경우 계절적 패턴이 더 강하게 나타나는 데서 유추할 수 있다.

둘째, 한한령, 위안부 배상 문제, 불매 운동 등 정치경제적 요인과 메르스, 코로나와 같은 전염병 요인이 아주 큰 영향을 미침이 관찰된다. 특히 이런 변수들은 정량화하기가 어렵고 정확히 같은 요인이 반복되지 않는 일회적 성격을 지녀 그 영향의 크기를 파악하기가 어렵기 때문에, 모형 적합에 있어 큰 방해요인이 된다. 정치경제적 갈등이 거의 없었고, 메르스로 인한 관광 감소의 영향도 작았던 미국의 데이터가 SARIMA 모형에 가장 잘 적합 되는 이유가 여기에 있을 것이다.

만약 정부가 해외 관광객을 대상으로 한 관광업을 주요 산업으로 육성하고자 한다면, 주요국들과 정치경제적 갈등을 최소화하여 불확실성을 제어하고, 전염병 확산에 대한 보다 신속하고 확실한 대응체계를 갖추는 것이 중요해 보인다.

셋째, 대체로 관광 입국 수요가 꾸준히 증가해 왔다는 점이다. 정치경제적 갈등과 전염병을 제외하고 보면 각국의 관광 수요는 꾸준히 우상향 추세를 보여 왔다. 코로나 국면이 끝나면, 코로나 19로 인하여 제한 되어 왔던 관광 대기 수요가 전 세계적으로 폭발할 것이므로, 이처럼 폭증하는 여행수요를 잘 끌어들일 수 있도록 선제적인 계획 수립이 필요하다고 생각된다.