장 려 상

인공지능을 활용한 질병확진자수 예측과 방역 정책 평가

| 김근한 · 서민경 · 안지혁 |

본 논문 내용은 저자의 견해이며, 통계청 및 통계개발원 공식 견해와 일치하지 않을 수도 있습니다.

인공지능을 활용한 질병확진자수 예측과 방역 정책 평가

김근한*, 서민경**, 안지혁***

요 약

본 연구는 코로나19에 대응하기 위한 방역 정책의 효과를 분석하고 예측하기 위해 다양한 모델을 활용한다. 전염병 확산의 수학적 모델에 추가로 인공 신경망 모델을 도입하여 방역 정책에 따른 확진자 수 변화를 예측한다. 연구는 순환 신경망 모델과 물리 정보 기반 신경망을 활용하며, 방역 단계 변경 시점과 강도에 따른 전파율, 회복률, 사망률을 예측한다. 또한, 코로나19 변이 바이러스에 대한 예측과 해당 기간의 방역 정책 평가도 수행한다. 연구는 국내의 코로나19 발생 현황 데이터와 인구이동량 등의 다양한 변수를 활용하여 분석한다. 연구는 실제 전파 상황에 대한 예측과 정책 평가를 통해 코로나19 대응 방역 정책 수립에 기여하고자 한다. 본 연구에서 활용한 데이터와 데이터의 전처리, 예측 및 시각화를 위해 사용한 코드는 https://github.com/G-Nan/Prediction-COVID-19에서 확인할 수 있다.

* 계명대학교 기초과학부 수학전공, 학사과정
** 계명대학교 기초과학부 수학전공, 학사과정
*** 계명대학교 기초과학부 수학전공, 학사과정

Ⅰ. 서 론

1. 연구의 배경 및 목적

코로나바이러스 감염증-19(Coronavirus disease 2019, 이하 코로나19)는 SARS-CoV-2가 일으키는 중증 호흡기 증후군이다. 2019년 12월 중국에서 첫 사례가 보고되었고, 2020년 3월 11일 세계보건기구(World Health Organization, WHO)의 세계적 대유행(pandemic) 선언 이후 여러 형태의 바이러스 변이를 거치며 2023년 6월 21일 기준 전 세계적으로 768,187,096명의 누적 확진[1]이 발생하였다. 국내에서는 2020년 1월 20일 첫 번째 사례가 발견된 이후 2023년 6월 26일 기준 국내 32,131,606명의 누적 확진[2]이 보고되었다. 세계보건기구는 2023년 5월 5일 브리핑을 통해 코로나19 국제공중보건위기상황 선포를 해제한다고 발표하였다.

1) 대한민국 정부의 코로나 방역 정책

대한민국 정부는 코로나19 확산에 의한 감염병 위기 경보 단계를 2020년 2월 23일 심각 단계로 격상하였다[3]. 또한 코로나19 대응을 위하여 국무총리를 본부장으로 하는 중앙 재난 안전 대책본부를 가동하여 범정부적으로 방역에 집중하고 있으며, 감염병 대응의 특수성과 전문성을 고려하여 방역 컨트롤타워인 중앙방역대책본부(질병관리청)와 중앙사고수습본부장(보건복지부 장관)이 협력하여 방역업무를 총괄하였다. 감염병 위기 경보 단계는 약 3년 3개월 동안 심각 단계를 유지하였으며, 2023년 5월 11일 코로나19 중앙 재난 안전 대책본부 회의에서 코로나19 감염병 위기 경보를 심각에서 경계로 하향 조정하였다.

코로나19 발생 이후 초기의 정책은 사스 바이러스 국내 유입 당시(2003) 구성된 질병관리본부가 작성한 국가 위기 유형의 전염병을 포함하는 표준 매뉴얼, 실무 매뉴얼과 메르스 바이러스의 국내 유입 당시(2015) 완성한 감염병 위기관리 표준 매뉴얼을 기반으로 진행하였다(질병관리청, 2021)[4]. 초기정책으로써 감염병의

1) https://covid19.who.int/ (2023.06.26.)
2) https://kosis.kr/covid/covid_index.do (2023.06.26.)
3) https://buly.kr/Edp8ZkO (2023.06.27.)
4) https://www.kdca.go.kr/contents.es?mid=a20301020300 (2023.06.20.)

예방 및 관리에 관한 법률[5] 제49조에 의해 다중시설 운영통제를 하였다(국가법령정보센터, 2022). 코로나19 방역 정책은 2020년 3월 22일에 최초로 시행되었다. 그 이후 2020년 4월 20일에 국민의 의견 및 자치회의 회의를 거쳐 거리두기 수위를 조절하는 것으로 결정하였다. 그리고 2020년 5월 6일에 일일 확진자 수 감소로 인해 사회적 거리두기에서 생활 속 거리두기로 변경되었다. 당시 정책은 규정이 정확히 확립되지 않았다. 이에 정부는 2020년 6월 28일부로 코로나19 확산 정도에 따라 1~3단계로 구분하여 시행하기로 하였다. 정책의 형태는 변경되었으나 방역 정책의 강도는 유지되었다.

2020년도 8월 16일까지는 전국적으로 1단계가 유지되고 있었으나 종교단체의 8.15 집회 관련 확진자 발생이 급증하여 수도권의 방역 단계가 2단계로 격상되었다(보건복지부, 2020)[6]. 1주일 뒤 비수도권 모든 지역도 2단계로 격상되었다. 하지만 일일 확진자 수가 줄지 않자 정부는 8월 28일에 강화된 2단계 체제(2.5단계)를 만들면서 8월 30일부로 수도권을 2.5단계로 격상시켰다. 이 조치로 인해 확진자 수가 감소하는 추세를 보여 9월 14일부로 수도권의 방역 단계가 2단계로 격하되었다. 이후 정부는 추석 연휴에 대비하여 추석 특별방역 기간을 지정하여 실시하였고 이는 확진자 수의 감소 추세를 일으켰다. 이에 따라 10월 12일부로 모든 시·도의 방역 단계를 1단계로 격하시켰다. 이렇게 3단계 체제를 이어오면서 집단감염 등 확진자 수 변화에 맞춰 병역단계를 강화하거나 약화하는 선택을 하였다(보건복지부, 2020)[7]. 그러나 단기・장기적인 사회・경제적 비용을 고려하여 이제는 의료체계 여력에 따라, 감당할 수 있는 위험 수준을 설정하고 코로나19 유행을 통제하는 것이 효율적이라는 전문가들의 의견이 제시되었다. 기존 3단계 체제는 단계별로 실행되는 조치의 강도가 큰 차이가 있으며, 의료 대응능력이 향상되었음에도 3단계 체제의 격상 기준이 초기 역량에 맞춰져 있다 보니 낮게 설정되어 있었다. 또한, 코로나19 확산 정도가 지역마다 다르고, 전국적으로 동일한 조치를 실시한 것에 대한 형평성 논란이 발생하였고 3단계 체제에 대한 사회적 수용성이 저하되어 2020년 11월 7일부로 5단계 체제로 전환하였다(보건복지부, 2020)[8].

---

5) https://www.law.go.kr/법령/감염병의예방및관리에관한법률 (2023.06.29.)
6) https://buly.kr/6BslUQJ (2023.05.31.)
7) https://buly.kr/FWOvTgX (2023.05.31.)
8) https://buly.kr/ESuNapN (2023.05.31.)

5단계 체제 변경 이후 처음으로 단계가 변경된 것은 11월 24일이다. 11월 14일에 진행된 민주노총의 대규모 집회로 인해 전국적으로 대유행이 시작됨으로써 수도권 지역은 2단계로 호남 지역은 1.5단계로 격상되었다. 하지만 감염의 확산을 잡지 못하여 12월 8일에는 수도권 2.5단계, 비수도권 2단계로 다시 격상하게 되었다. 이 대유행으로 인해 2021년 1월 17일까지 단계의 격상은 없었지만, 세부적인 조치 강화는 계속되었다. 이후 2월 7일까지 방역 단계가 유지되다 2021년 2월 8일, 2월 15일 이틀에 걸쳐 단계를 격하하게 되었다. 이때부터 4단계 체제로 전환한 2021년 7월 1일까지 단계의 변화 없이 유지하였다(보건복지부, 2021)[9]. 5단계 체제를 이어오면서 단계의 세분화가 다양한 조치의 시행을 가능하게 했지만, 0.5단계 사이의 위험성이 국민의 행동 대응에 대해 불명확한 메시지를 전달한다는 지적이 있었다. 이에 기존 5단계를 4단계로 간소화하고 지자체의 자율권을 강화하였다. 그리고 사적 모임 등 활동에 대한 관리 강화와 다중이용시설의 위험도에 따라 정책을 세분하였다.

2021년 6월부터 시작된 4차 대유행 이후 일일 확진자 수가 유지되거나 감소하여 거리두기를 완화하였다. 그러나 전파력이 강한 델타 변이의 확산이 7월부터 급격히 증가하여 7월 14일부로 수도권의 방역 단계를 4단계로 격상하게 되었다. 격상 이후 일일 확진자 수가 감소 또는 유지하는 추세를 보여 7월 28일과 10월 18일 두 번에 걸쳐 거리두기를 완화하였다. 이후 확산 추세가 안정화되면서 정부는 11월 1일부로 단계적 일상 회복을 선포하였다(보건복지부, 2021)[10]. 이 개편방안은 당시 일일 확진자 수의 감소세가 지속되어 실시되었으나 12월 1일 오미크론의 국내 유입으로 인해 위기를 맞게 되었다. 결국, 정부는 12월 18일부로 전국적으로 4단계 격상[11]을 발표하였고 거리두기를 해제한 2022년 4월 18일까지 소폭 완화를 실행할 뿐 방역 단계를 4단계로 유지하였다. 2022년 4월 18일부로 사회적 거리두기가 해제되었다. 이후 신규 변이(높은 전파력, 높은 치명률, 백신의 중증·사망 방지 효과 저하 등의 특성을 가진 새로운 변이) 등장 또는 겨울철 재유행 등으로 생활 방역만으로 대응이 어려운 경우, 재도입 논의하겠다 하였지만, 현재까지 확산 추세에 따른 생활 방역을 조정하여 시행하고 있을 뿐, 거리두기를 재도입하지는 않았다.[12] 마스크 착용 의무에 대해서도 2020년 10월 다중이용시설

9) https://buly.kr/GOyiNpL (2023.05.31.)
10) https://buly.kr/YaW6Q6 (2023.06.01.)
11) https://url.kr/5bgl37 (2023.06.26.)

중심의 마스크 착용 의무화가 도입된 이후 거리두기가 해제되기까지는 마스크 착용이 의무였지만, 2022년 5월 일부 상황을 제외한 실외 마스크 의무 착용이 완화됨을 시작으로 2023년 1월 30일 일부 시설 외 실내 마스크 착용 권고로 전환, 2023년 3월 20일 대중교통을 포함한 일부 의무시설 내 마스크 착용 의무가 추가 조정되었다[13]. 코로나 위기 단계가 3년 3개월 만에 심각 단계에서 경계 단계로 하향 변경되었다. 이에 따라 격리기간이 7일에서 5일 권고로 전환되었으며 마스크 착용 의무 역기 입소형 감염 취약 시설, 병원급 의료기관을 제외하고 권고로 전환되었다. 검역 또한 입국 후 3일 차 PCR이 권고 사항이었지만, 단계가 하향 조정되면서 종료되었다[14]. 지금까지 방역 정책이 계속 변화하게 된 이유를 살펴보았는데 일일 확진자 수 변동에 따라 성급하게 단계를 변경하거나 코로나19로 인하여 국민의 경제활동에 큰 타격을 주어 자영업의 경우 영업 중단으로 인해 수입이 급감하는 상황을 초래하였다. 또한, 백신접종이 시작된 이후로 접종 유무에 따른 인원 제한 정책을 계속 개편하여 혼란을 주었다.

2) 코로나19 이후 바이러스 대유행 가능성

미국 서부에 위치한 라 호야 면역학연구소(La Jolla Institute for Immunology)의 연구진은 국제 학술지 셀 리포츠(Cell reports)에 기고한 최신 보고서를 통해 변이 지카 바이러스(Zika virus)가 강력한 제2의 코로나 사태 후보라고 보고했다(Regia- Nava et al, 2022). 연구진은 실험을 통해 지카 바이러스가 감염력이 높으며, 변이도 쉽게 일으킨다는 사실을 발견했으며 변이 지카 바이러스가 세상에 나온다면, 세계는 코로나 사태와 비슷한 세계적 대유행을 맞이하게 될 것으로 예측했다. 또한 한 번의 변이로도 지카 바이러스의 확산은 급격히 증가할 것이라고 보고했다.

조지타운대학교(Georgetown University)의 월시 외교 대학원(Walsh School of Foreign Service) 내 정책 연구 기관인 안보 신기술센터(Center for Security and Emerging Technology, CSET)는 올해 4월 다음 팬데믹 바이러스를 생성할 가능성이 있는 7가지 바이러스군을 확인했다고 밝혔다. 이 바이러스군은 팬데믹을

12) https://url.kr/7sumxy (2023.06.26.)
13) https://url.kr/cwp8ze (2023.06.26.)
14) https://url.kr/gqkcrw (2023.06.26.)

일으키는 특성을 갖고 있으며, 모든 바이러스 중에 팬데믹 위험성이 가장 높다고 말했다. 현재 도시화와 지구 기후의 변화로 사람들이 동물과 더 가까이 접촉하게 되면서 인수공통전염병의 위험도 더 커지고 있다고 보고했으며, 그 결과 세계화로 인해 바이러스가 국경을 넘어 새로운 지역사회로 전파되면서 새로운 바이러스가 더욱 빠르게 확산이 될 수 있다고 보고했다(Schuerger et al, 2023). 현재 코로나19는 안정세로 접어들었으나 앞으로 코로나19보다 더 큰 팬데믹을 가져올 바이러스가 언제든 나올 수 있다는 것을 의미한다.

3) 연구의 목적

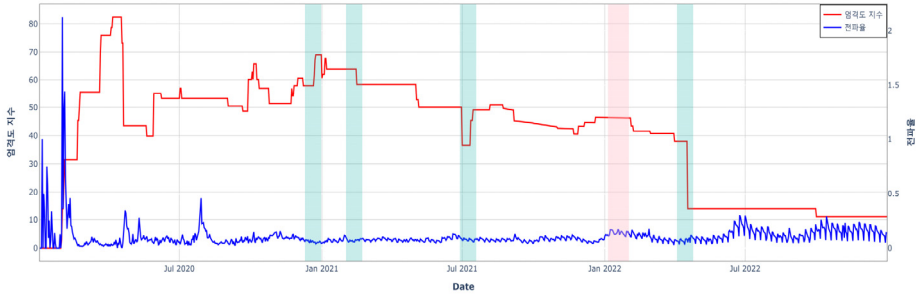
본 연구에서는 코로나19에 대응하기 위해 방역 당국이 시행한 방역 정책에 따른 확진자 수 변화를 예측하고 이를 바탕으로 방역 정책을 분석하고자 한다. 전통적으로 전염병 확산 예측은 통계 모형과 수학적 모형을 통해 진행됐다. 이는 코로나19 분석에도 활용되었으며 다양한 결과가 정부의 정책 수립에 도움을 주고 있다. 그러나 전염병 확산은 매우 복잡하고 많은 요인에 의해 결정되기 때문에 이러한 전통적인 모델은 한계가 있으며 최근에는 좀 더 복잡한 형태의 모델이 연구・개발되고 있다. 이러한 모델은 보통 결정론적 모형과 확률 구조 모형이 함께 사용된다. 확률 구조는 더 현실적이지만 분석하기가 훨씬 더 복잡하다. 한편, 최근에는 영상 분류나 자연어 처리 분야를 비롯한 다양한 분야에서 합성곱 신경망(Convolution Neural Network, CNN), 순환 신경망(Recurrent Neural Network, RNN)과 같은 구조의 인공 신경망 모델이 성공적으로 사용되고 있다. 특히 기간의 경과에 따른 동적 정보를 포함하는 시계열 데이터는 RNN 형태의 모델이 적극적으로 활용되고 있다. 이러한 RNN 기반의 모델들은 예측의 측면에서 통계적 방법보다 우수한 성능을 보이기도 한다. RNN 계열의 모델은 순서, 시간 등을 모델링 할 때 강점을 보이고, 입력과 출력의 길이를 다양하게 적용할 수 있다는 장점이 있다. 본 연구에서는 다양한 형태의 순환 신경망 모델과 함께 물리 정보 기반 신경망(Physics Informed Neural Network, PINN)을 사용하여 확진자 수의 변화를 예측하고 이를 통해 방역 정책에 대한 다양한 평가를 시도하고자 한다. 이를 바탕으로 코로나19 초기에 과거의 전염병 확산에 대한 경험으로 방역 정책 수립에 도움을 받았던 것처럼 코로나19의 경험을 이후 발생할 바이러스 유행에 대처할 수 있는 방역 정책 수립에 도움을 주고자 한다.

2. 연구의 범위 및 방법

Hale et al. (2021)은 코로나19 발생 이후 세계 각국의 방역 정책 강도를 비교하기 위한 데이터 세트인 Oxford COVID-19 Government Response Tracker(OxCGRT)를 소개하고 있다. OxCGRT 데이터는 2020년 1월 1일부터 180개 이상의 국가와 여러 국가의 관할 권내에 대한 폐쇄 및 봉쇄, 보건 및 경제 정책과 관련된 정부 정책을 포함하고 있으며 이를 바탕으로 정책의 엄격도 지표 제시하고 있다. 〈그림 1〉은 대한민국의 방역 정책 엄격도 지표와 본 연구에서 계산된 국내 코로나19 전파율과 연구 대상 기간을 시각화한 그래프이다.

본 연구의 대상은 정부가 코로나19에 대응하기 위해 발표한 단계별 방역 정책이다. 2019년부터 현재까지 정부는 70여 차례 단계별 방역 정책을 발표하였으나 이 중 방역 강도 변화에 따라 4회의 방역 단계 변경 시점을 선택하였으며 이 기간의 정책에 대하여 전염병 확산을 단기(1주) 및 중기(2주)로 예측하고 이를 통해 정책에 대한 평가를 하고자 한다. 4회의 방역 단계 변경 시기, 주요 내용, 지배 바이러스 변이는 〈표 1〉과 같다.

〈그림 1〉 국내 방역 정책의 엄격도 지수와 전파율



주: 엄격도 지수는 휴교, 직장 폐쇄, 여행 금지 등 9가지 대응 지표를 바탕으로 0에서 100까지의 값으로 조정된다(100 = 가장 엄격함). 전파율은 전국을 대상으로 한 확진자 수 데이터와 식 (4)를 통해 계산되었다. 자료: Hale et al. (2021)

〈표 1〉 연구 대상 기간 방역 단계 내용

변경 시기	주요 내용	방역 강도	지배 바이러스 변이
2020.12.24	연말연시 특별 방역 정책 시행	강화	오리지널
2021.02.15	수도권 (2단계 격하, 22시 제한 변경) 비수도권 (1.5단계 격하)	완화	알파 변이
2021.07.14	수도권 4단계 체제 처음으로 4단계 격상 대전, 충북, 충남, 광주, 대구, 부산, 울산, 경남, 강원, 제주 : 2단계 격상	강화	알파 변이
2022.04.18	거리두기 모두 해제	완화	오미크론 변이

자료: 보건복지부 보도자료

또한, 기존의 델타 변이에서 오미크론 변이로 지배 바이러스가 변경된 시기에 대해서 PINN을 통해 신규 변이 바이러스인 오미크론 변이에 대한 전파율, 회복률, 사망률을 예측하여 해당 기간의 정책을 평가한다.

방역 정책의 효과를 분석하기 위해 국내 17개 시·도의 일일 코로나19 발생 현황자료를 사용하였다. 코로나19 발생 현황자료는 공공데이터포털[15]에서 제공하는 Open API를 활용하여 획득한 보건복지부 코로나19 시·도 발생 현황 데이터를 사용하였다. 연구의 시간적 범위는 코로나19가 국내에서 최초 발생한 2019년 1월부터 2023년 6월까지이다. 방역 정책별 코로나19 발생 현황 예측을 위해 보건복지부에서 발표하는 국내 17개 모든 시·도의 발생 현황을 예측 모델 학습 데이터로 활용하였다.

본 연구는 정부에서 발표한 코로나19 방역 정책이 코로나19 확산에 미치는 영향을 분석하고자 한다. 이를 위해 코로나19 발생 현황 데이터 이외에도 통신 모바일 인구이동량 통계 방역 정책 발표일 기준으로 분석하였으며, 또한 확진자 수의 변화와 비교하였다. 마지막으로 방역 정책에 대한 국민 인식은 한국리서치에서 조사한 코로나19 정기 인식조사[16]를 참고하였다.

15) https://www.data.go.kr/ (2023.06.30.)
16) https://hrcopinion.co.kr/covid-19/awareness (2023.06.26.)

Ⅱ. 이론적 배경 및 선행연구 고찰

1. 이론적 고찰

1) 전염병 확산 수리 모델

전염병 확산 연구에 수학을 처음으로 사용한 사람은 천연두 확산을 막기 위해 수학을 활용한 네덜란드 출생의 수학자 다니엘 베르누이(D. Bernoulli)이다. 1766년 베르누이는 확률이론을 활용하여 전염병 때문에 얼마나 많은 사람이 사망하는지를 분석하였으며, 인체에 약한 천연두 균을 투입하여 천연두에 대한 면역력을 형성하면 천연두 확산을 예방할 수 있음을 확인하게 되었다. 전염병 확산에 대한 수리 모델은 베르누이 이후 200여 년이 지난 20세기 초에야 좀 더 체계적인 연구가 시작되었다. 영국의 병리학자 로널드 로스(R. Ross)는 말라리아는 모기가 옮기는 병이라는 사실을 알아내어 말라리아 확산 모델에 대한 연구를 통해 1902년 노벨 생리학·의학상을 받게 된다. 1972년 스코틀랜드 수학자 윌리엄 커맥(W. Kermak)과 예방역학자 앤더슨 맥켄드릭(A. McKendrick)은 전염병 유행의 시작 조건과 확산 형태를 예측하는 SIR모델을 제시했다(Kermak and McKendrik, 1927). SIR 모델은 전염병이 어떻게 확산되는지, 총감염 수, 전염병 지속 기간 등을 예측하고 재생산지수와 같은 역학 변수를 추정하는 데 사용된다. 또한, SIR 모델을 통해 전염병 확산에 있어 방역 정책의 개입이 확산에 어떻게 영향을 주는지 확인할 수 있다. SIR 모델은 전체 인구를 감염 가능성이 있는 사람(Susceptible), 감염된 사람(Infected), 회복되거나 사망하여 감염 위험에서 벗어난 사람(Recovered)으로 구분하여 감염 유행을 예측한다. t시간에서 해당 구분의 크기를 각각 $S(t)$, $I(t)$, $R(t)$라고 할 때 SIR 모델은 식 (1)과 같은 미분 방정식계로 표현된다. 식 (1)에서 N은 전체 인구수를 나타내며, α와 β는 각각 질병 전파율과 질병 회복률이다. 모델에서는 출생과 해당 질병 이외의 사망이 없다고 가정한다. 즉, 모든 시간 t에 대하여 $N = S(t) + I(t) + R(t)$ 이 성립한다.

이러한 SIR 모델은 감염자에게 노출된 사람(Exposed)을 반영한 SEIR 모델, 치료와 격리 등 행동 변화를 고려한 SEIQR 모델, SEIHR 모델 등 다양한 파생 모델이 연구되고 있다.

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{\alpha\, I(t)\, S(t)}{N}, \\ \dfrac{dI(t)}{dt} = \dfrac{\alpha\, I(t)\, S(t)}{N} - \beta\, I(t), \\ \dfrac{dR(t)}{dt} = -\beta\, I(t)\,. \end{cases} \quad (1)$$

전염병의 확산 모델을 통해 확인할 수 있는 중요한 역학변수는 기초재생산지수[17]이다. 기초감염재생산지수(R_0)는 감염이 없는 집단에서 발생한 첫 감염자가 평균적으로 감염시킬 수 있는 환자 수를 나타낸다. R_0값이 1보다 크면 환자 수가 증가하여 확산이 유행할 가능성이 있으며 1보다 작으면 해당 질병은 서서히 소멸하게 된다. 전염병 확산이 코로나19와 같이 장기간 유지될 경우, 기초감염재생산지수보다는 실질감염재생산지수(R_t)를 고려하게 된다. R_t값은 일부 면역이 있는 집단이나 방역이 진행되는 상황에서, 시간별로 발행한 감염자로 인해 신규 감염되는 환자의 평균을 나타낸다. R_t값은 확산 여부를 판단하는 실질적 지표로 사용되며, 방역 당국의 방역 정책효과의 평가에 활용될 수 있다. 즉 방역 당국은 R_t값을 1 이하로 만들기 위한 방역 정책의 수립 하고자 노력하게 된다. 그러나 본 연구에서 활용한 보건복지부 발표 코로나19 감염현황에서는 회복된 사람의 수 $R(t)$를 추정할 수 있는 격리 해제 수 및 격리 환자 수가 2021년 12월 이후 발표되지 않고 있다. 따라서 시간별로 변화는 $\beta(t)$ 값을 변이 별 격리기간의 역수로만 추정할 수 있다. 본 연구에서는 감염재생산지수를 활용한 방역 정책의 효과를 분석하는 대신 방역 정책에 따른 시간별 질병 전파율 $\alpha(t)$와 이에 따른 확진자 수 예측을 통해 방역 정책의 효과를 평가하고자 한다.

2) 시계열 데이터 예측

시계열 데이터 예측은 시간의 흐름에 따라 순차적으로 기록된 데이터를 분석하여 미래를 예측하는 것을 말한다. 이러한 예측에는 몇 가지 도전이 있다. 첫째, 예측값은 완벽하지 않고 불확실성을 내포하므로 적절한 확률 분포 모델을 고려해야 한다. 둘째, 데이터에는 숨겨진 패턴이 존재하며 이를 찾기 어렵게 만드는 요인들이

17) $R_0(t) = \dfrac{\alpha(t)}{\beta(t)}$

있다. 셋째, 다변량 시계열 데이터의 경우 변수의 수가 많아 처리와 학습이 어려울 수 있다. 넷째, 데이터에는 잡음(Noise)이나 결측값이 존재하므로 적절한 전처리가 필요하다. 인공지능 기술의 발전으로 다양한 신경망이 시계열 데이터 예측에 사용되고 있습니다. 본 연구에서는 다음의 인공 신경망을 사용하여 코로나19 관련 예측을 진행하고자 한다. 각 신경망의 특징은 다음과 같다.

- RNN은 순환신경망으로, 순차적인 데이터 처리에 주로 사용된다. 하지만 장기 의존성 문제로 인해 긴 시퀀스 처리에 제약이 있다.
- LSTM은 RNN의 변형으로, 장기 의존성 문제를 해결하기 위해 고안되었다. 메모리 셀을 도입하여 이전 상태의 정보를 더 오래 기억할 수 있게 한다.
- GRU는 LSTM과 유사한 기능을 제공하는 RNN의 변형이다. 더 간단한 구조를 가지고 있으며, 계산 비용이 적고 데이터가 적을 때에도 좋은 성능을 보인다.
- Bidirectional RNN은 입력 시퀀스를 앞뒤로 두 개의 RNN으로 처리하는 방식입니다. 이를 통해 현재 상태의 예측에 이전과 이후의 정보를 동시에 활용할 수 있다.
- Seq2Seq는 시퀀스 입력을 다른 시퀀스 출력으로 변환하는 모델 아키텍처이다. 인코더와 디코더로 구성되어 있으며, 입력 시퀀스와 출력 시퀀스의 길이가 다를 수 있어 유용하다.
- Many-to-One 기법은 시계열 예측에서 사용되는 방법 중 하나이다. 이 방법은 여러 개의 입력 데이터로부터 하나의 출력 값을 예측한다. 입력 시퀀스의 많은 요소들이 마지막 출력에만 영향을 준다.
- Many-to-Many 기법은 입력 시퀀스의 각각의 요소에 대해 여러 개의 출력 값을 예측하는 방식이다. Many-to-One 기법과 달리 여러 개의 출력을 한 번에 예측할 수 있다. 대체로 Seq2Seq 모델과 같은 아키텍처를 사용하여 구현되며, LSTM, GRU와 같은 순환신경망 또는 Transformer와 어텐션을 사용한 신경망 구조로 구현된다.
- 물리 정보 기반 신경망(PINN)은 인공 신경망과 물리학 원리를 결합한 머신러닝 프레임워크이다. 복잡한 물리적 또는 과학적 문제를 해결하는 데 유용합니다. PINN은 신경망을 사용하여 데이터에서 직접 기본 물리법칙을 학습하는 대안적인 접근 방식을 제공한다. PINN은 신경망의 훈련 중에 물리 방정식을 추가 제약 조건으로 사용하여 데이터가 함축하는 기본 물리 방정식을 만족시킨다. PINN은

희소하거나 노이즈가 있는 데이터로부터 학습할 수 있으며, 물리 기반 제약 조건을 결합하여 정확한 예측을 할 수 있다. PINN은 물리학의 기본 법칙을 이해하고 해석 가능성을 제공하여 문제를 해결하는 데 도움을 주며 코로나19 확진자 수와 같은 시계열 데이터 예측을 포함한 광범위한 응용 분야에서 그 가능성을 보여주고 있다.

2. 선행연구 검토

국내 코로나19 발생 상황에 대한 예측연구는 국가수리과학연구소 코로나19 수리모델링 TF 레포트[18]를 통해 수리모델링으로 분석한 코로나19 유행 예측을 약 2주 간격으로 제공하고 있다. 보고서는 국내 여러 대학 및 연구기관의 전염병 역학 관련 연구자들의 코로나19 관련 예측을 포함하고 있다. 또한 보고서에는 역학조사 데이터, 이동 데이터, 정책 데이터, 백신접종 데이터 등을 기반으로 하여 수리 모델 및 인공지능 모델을 사용하여 예측이 진행되고 있다. 특히 재생산지수 변화에 따른 예측과 분석이 포함되어있다.

정재웅·권혁무·홍성훈·이민구(2020)는 2020년 1월 20일부터 7월 31일까지의 기간을 5일 간격으로 나눠 코로나19 R_t값을 추정하였다. 그 결과, 신천지를 통한 집단 감염이 확산한 시기(2020년 2월)와 이태원 클럽발 확산이 발생했던 시기(2020년 5월)의 R_t값이 2 이상으로 나타난 것을 확인했다. Shim et al.(2021)은 코로나19 확진자 발생 사례가 가장 많았던 서울·경기 및 대구·경북 지역에 대해 R_t값과 더블링 타임[19]을 분석했다. 연구 결과, 3월 초와 6월 초의 R_t값이 3~4인 것으로 분석하였다. Choi et al(2020)은 2020년 2월 1일부터 6월 15일까지의 서울·경기 지역의 실제 역학 데이터를 이용해 연령에 따른 사회적 거리두기의 효과를 분석하여 확산 추세 예측뿐만 아니라 정부 방역 정책 평가에 대한 연구가 진행되었다. 연구 결과에 따르면 2020년 4월 24일까지 사회적 거리두기 강도가 높아져서 R_t값은 2.197에서 1 이하의 값까지 감소하였다. 강도 높은 사회적 거리두기가 감염률 감소에 영향을 미친다는 것을 확인하였다. 연구팀은 사회적 거리두기가 완화되고 2차 유행이 시작될 무렵인 4월 24일부터 사회적 거리두기

18) https://www.nims.re.kr/research/post/covid19_2 (2023.06.27.)
19) 코로나19 확진자 수가 두 배 느는 데 걸리는 시간

강도를 실제 시행되었던 것과 비교하여 다양한 시나리오에 대해서 연령별 감염 환자의 변화를 살펴보았다. 연구 결과 사회적 거리두기의 영향력이 연령대별로 다르고, 젊은 인구보다는 고령 인구에게 더 영향력이 크다는 것을 확인하였다. 또한, 연구팀은 7개 기간 모두에서 강력한 사회적 거리두기가 시행됐다면 총감염자 수는 44.6% 감소했을 것이라는 예상도 내놨다. 그러나 모든 기간 약한 사회적 거리두기를 유지했다면, 감염자 수는 지금보다 29.2% 증가했을 것으로 분석했다. 유명수·김연주·백수진·권동혁(2021)은 감염재생산지수가 코로나19 방역 정책 효과에 따른 감염병 전파양상을 추정함에 있어 효과적이나 코로나19 완화 세에도 지속해서 환자가 발생하고 있어 감염재생산지수 산출에 기반을 두어 방역 정책 효과를 검토할 필요가 있다고 분석하였다.

해외에서도 코로나19 방역 정책효과를 분석하기 위한 많은 연구가 진행됐다. Siraj et al(2020)은 인구의 규모, 도시의 크기와 같은 사회 지리적 요인과 함께 사회적 거리두기 강도, 접촉자 추적 여부, 마스크 효과 등과 같은 다양한 초기 방역 정책이 코로나19에 미치는 영향을 연구하였다. Tariq et al(2021)은 새로운 형태의 코로나19 확산 형태가 발생하고 있어, 지속적인 사회적 거리두기 조정과 적극적인 격리 정책이 필요하다고 하였다. 이 밖에도 통신 이동량 빅데이터를 이용한 인구이동과 바이러스 전염 관계 추정(Xiong et al, 2020) 등 다양한 수리 모델이 개발되고 있다.

인공지능 모델을 활용한 코로나19 확산의 경우, Chimmula and Zhang(2020)는 캐나다에서 COVID-19 전염의 시계열 예측을 위해 LSTM 신경망을 사용하여 매우 근접한 예측을 하였다. Chakraborty and Ghosh(2020)는 캐나다, 프랑스, 인도, 한국 및 영국의 일일 확진 사례를 단기 예측을 위해 하이브리드 ARIMA 및 웨이브릿 기반 예측 모델을 사용했다. 노윤아·정승원·문재욱·황인준(2022)은 코로나19 예측을 위하여 확진자 수 데이터 이외에도 “코로나” 키워드 검색량, 백신 접종자 수, 지하철 승차 인원 등의 데이터를 변수로 활용하여 LSTM 기반의 예측 기법을 제안하였다. 홍태경·김은서·이희상(2023)에서는 코로나19 확산 예측에 적합한 딥러닝 모델이 선택하고, 바이러스 확산 패턴의 변화에 따라 모델의 입력 변수의 조정을 제안하였다. 또한, 많은 연구진이 SIR 모델과 인공지능을 통해 여러 사회적 변수를 고려하여 코로나19 확산을 예측하고있다(Ajbar, Alqahtani, and M. Boumaza, 2021; Kim, Kim, Lee, and Oh, 2020; Kim, Seo, and Yeom, 2020; Perakis, Singhvi, Lami and Thayaparan, 2022; 김진오 외, 2020).

3. 선행연구와의 차별성

앞서 살펴본 것처럼 코로나19 확산과 관련하여 상당히 많은 연구가 이루어졌다. 선행연구와 비교하여 본 연구의 가장 두드러진 차이점은 코로나19 확산과 관련한 역학변수의 선택에 있다. 선행연구에서는 주요 역학변수로 일일 확진자 수, 재생산지수 등을 고려하였으나, 본 연구에서는 SIR 모델의 이산 모형을 통해 전파율을 계산하고 전파율의 차분값과 2차 차분값을 역학변수로 활용하여 예측하였다. 또한 새로운 변이 바이러스가 출현했을 경우의 예측을 위해 감염된 사람의 수를 기존 변이 바이러스에 의해 감염된 사람과 신규 변이 바이러스에 의해 감염된 사람의 수로 나누어 고려하는 SIsRD모형을 고려하였다. 이를 통해 우리는 방역 단계 변경의 시기와 강도의 선택이 실제 코로나19 확산에 어떤 영향을 미쳤는지 예측고자 한다. 더불어 지역별 인구, 면적, 인구 밀도, 인구이동량 통계를 활용하여 방역 당국이 발표한 방역 정책이 지역별로 실제 국민 행동에 영향을 주었는지도 함께 분석하여 앞서 예측한 방역 단계의 강도와 시기에 따른 정책효과에 대하여 비교 분석하였다.

Ⅲ. 연구 설계

1. 활용 데이터 및 변수 설정

1) 활용 데이터

본 연구에서 활용한 시계열 데이터는 보건복지부에서 공공데이터포털을 통해 제공하는 코로나19 감염증 시·도 발생 현황 데이터와 행정안전부에서 제공되는 시도별 인구[20], 통계청 국가통계포털(KOSIS)[21]와 통계데이터센터(SDC)[22]를 통해 제공되는 시도별 면적과 SK텔레콤의 통신 모바일 인구이동량 통계이다.

---

20) https://jumin.mois.go.kr/ (2023.06.24.)
21) https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1ZGA17 (2023.06.24.)
22) http://bigdata.kostat.go.kr/foresight/mobMoventSido.do (2023.06.24.)

(1) 코로나19 지역별 전파율

코로나19 감염증 시·도 발생 현황 데이터에서는 지역별 일일 확진자 수 데이터를 제공하고 있다. 이를 통해 지역별 전파율(transmission rate)을 다음과 같이 계산하였다. 식 (1)은 연속형 SIR 모델이다. 우리는 식 (1)에 사망자 수를 포함한 SIRD 이산 모델을 고려한다. 여기서 α, β, γ는 각각 전파율, 회복률, 사망률을 의미한다.

$$\begin{cases} S(t+1)-S(t) &= -\dfrac{\alpha(t)}{N}S(t)I(t) \\ I(t+1)-I(t) &= \dfrac{\alpha(t)}{N}S(t)I(t)-\beta I(t)-\gamma I(t) \\ R(t+1)-R(t) &= \beta I(t) \\ D(t+1)-D(t) &= \gamma I(t) \end{cases} \tag{2}$$

위 식을 통해 시간에 따른 전파율 $\alpha(t)$를 다음과 같이 계산하였다.

$$\alpha(t)=\frac{N(S(t)-S(t+1))}{S(t)I(t)} \tag{3}$$

신규 바이러스 변이가 출현한 경우의 추정을 위해서는 식 (4)와 같은 SIsRD 이산 모형을 사용하였다.

$$\begin{cases} S(t+1)-S(t) &= -\dfrac{\alpha_1}{N}S(t)I_1(t)-\dfrac{\alpha_2}{N}S(t)I_2(t) \\ I_i(t+1)-I_i(t) &= \dfrac{\alpha_i}{N}S(t)I_i(t)-\beta_i I_i(t)-\gamma_i I_i(t) \\ R_i(t+1)-R_i(t) &= \beta_i I_i(t) \\ D_i(t+1)-D_i(t) &= \gamma_i I_i(t) \end{cases} \quad (i=1,\ 2) \tag{4}$$

여기서 I_1, I_2는 각각 기존 변이 바이러스에 의해 감염된 사람의 수와 신규 변이 바이러스에 의해 감염된 사람의 수를 나타내며 이 값은 해당 기간 확진자 수에 각 변이의 분포 비율을 곱하여 계산할 수 있다. 변이의 분포 비율은 국제인플루엔자정보공유기구(GISAID)[23]에서 제공되는 데이터를 활용하였다. 또한, α_i, β_i, γ_i

($i = 1, 2$)는 각 변이 바이러스에 대한 전파율, 회복률, 사망률이다. 우리는 기존 변이 바이러스에 대한 충분한 정보를 확보하고 있다고 가정하고 신규 변이 바이러스에 대한 전파율, 회복률, 사망률을 예측한다.

(2) 통신 모바일 인구이동량 통계

통신 모바일 인구이동량 통계는 전국에 거주하는 모든 사람을 대상으로 주차별 일평균 이동건수를 관내, 관외로 구분하여 작성되었으며, SK텔레콤에서 통계청으로 이관 후 통계데이터센터 홈페이지를 통해 제공되고 있다. 통신 모바일 인구이동량 통계에서 관내 이동[24] 데이터를 사용하였다.

(3) 시도별 인구, 면적, 인구 밀도

시도별 인구는 주민등록 인구 및 세대현황을 기반으로 월별 작성되었으며, 행정안전부에서 제공되고 있다. 시도별 인구를 국내 코로나 발병 시점부터 현재까지 3년간의 시도별 인구수를 평균 내어 사용하였다. 면적은 한반도 면적을 시도별로 나누어 km^2 단위로 작성된 데이터를 사용하였다. 인구 밀도는 시도별 인구를 시도별 면적으로 나누어 계산하였다.

2) 데이터 전처리

본 연구의 주요 데이터인 일일 코로나19 확진자 현황자료는 공공데이터포털에서 제공하는 Open API 서비스를 통해 보건복지부에서 발표하는 자료를 수집하였다. 수집한 데이터를 시도별로 나누어 정리하였으며, 코로나 첫 발생인 2020년 1월 20일부터 2023년 5월 31일까지의 데이터를 수집하였다. 수집한 데이터에 날짜가 동일하거나 누적 데이터가 줄어드는 등의 이상치를 보건복지부와 질병관리청 발표를 참고하여 직접 수정하였다. 이후 바이러스 변이 점유율 데이터를 바탕으로 변이 바이러스별 확진자 현황을 달리 집계하였으며, 수집한 각 변이 바이러스별 확진자 수 데이터를 식 (3)을 활용하여 시도별 $S(t)$, $I(t)$, $R(t)$, $D(t)$, $\alpha(t)$, $\beta(t)$, $\gamma(t)$ 데이터를 생성하였다.

---

23) https://gisaid.org/ (2023.05.31.)

24) 본인이 실거주하는 시군구 내 타 행정동을 방문하여 30분 이상 체류한경우를 이동량(건수)로 집계

(1) 단기 예측(7일)

방역 정책 전환 시기를 기준으로 200일 전부터 이후 7일까지의 데이터를 학습에 사용하였으며, 이후 7일을 예측하였다. 이는 잠복기와 정책 적응 기간을 고려하였기 때문이다.

(2) 중기 예측(14일)

방역 정책 전환 시기를 기준으로 300일 전부터 이후 7일까지의 데이터를 학습에 사용하였으며, 이후 14일을 예측하였다.

(3) 장기 예측(21일)

장기 예측은 방역 정책 전환과는 별도로 신규 변이 바이러스 출현 시 신규 변이 바이러스의 전파율, 회복률, 사망률의 예측이 목적이므로 지배 바이러스가 델타 변이에서 오미크론 변이로 전환되는 시점을 기준으로 이전 168일(24주)간의 데이터를 학습에 사용하였고, 이후 21일(3주)간을 예측하였다. 또한, 기존 변이 바이러스의 전파율, 회복률[25], 사망률[26]을 함께 학습에 사용하였다.

2. 연구 방법론

1) 예측 모델 작성

예측 모델에 사용된 신경망은 Intel(R) Core(TM) i3-10100 CPU@3.60GHz을 탑재한 PC를 사용하여 학습하였으며, RTX 2070 SUPER D6 8GB GPU 가속을 사용하였다. 단일 모델 훈련 시간은 최대 30분이 소요되었으며, 훈련에 사용된 가상환경은 Python 3.8.16과 Pytorch 1.9.0를 기반으로 구축하였다.

단기와 중기 예측을 위해 앞서 언급한 인공 신경망 중에서 PINN을 제외한 12개의 인공 신경망을 학습하였으며, RMSE와 MAPE를 사용한 정확도 평가를 통해 6개의 모델을 선정하여 단기와 중기 예측에 활용하였다. 장기 예측에는 PINN을 사용하였다.

---

25) 0.0714 = 1/14 = 1/(격리기간)
26) https://m.dongascience.com/news.php?idx=51900 (2023.06.03.)

2) 예측 모델 정확도 평가

예측에 활용한 모델의 정확도를 평가하기 위하여 코로나19 실제 현황과 예측한 결과의 평균 제곱근 오차(Root Mean Square Error, RMSE)와 평균 절대 백분율 오차(Mean Absolute Percentage Error, MAPE)를 사용하였다. A_t를 실제 시계열 데이터, P_t를 예측 시계열 데이터라고 할 때 두 오차에 대한 정의는 다음과 같다.

$$\mathrm{SE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_t - P_t)^2} \tag{8}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_t - P_t}{A_t}\right| \tag{9}$$

Ⅳ. 분 석

1. 모델 분석

1) 모델 성능

(1) 단기 예측 모델

17개 시도별 단기 예측에 대하여 정확도를 RMSE와 MAPE로 측정하여 가장 우수한 모델 6개를 선정하였다. 단기 예측의 경우 선정된 6개의 모델은 seq2seq_RNN, seq2seq_GRU, seq2seq_LSTM, LSTM, GRU, BiLSTM 순이며 이중 seq2seq 기법을 사용한 네트워크가 가장 좋은 성능을 나타내는 것을 확인하였다.

(2) 중기 예측 모델

단기 예측과 같은 방법으로 가장 우수한 모델 6개를 선정한 결과 BiRNN, seq2seq_RNN, BiLSTM, BiGRU, LSTM, RNN 순으로 우수한 결과를 보였다. 중기 예측의 경우 한 번에 14일을 예측하는 seq2seq 기법 보다는 하루씩 예측하는

Many-to-One 기법이 전반적으로 좋은 성능을 나타내는 것을 확인하였다. 기본 RNN보다 성능이 좋다고 알려진 LSTM, GRU에 비해 BiRNN이 더 좋은 성능을 보여주었다. 단기 및 중기 예측에서 공통적으로 seq2seq에 양방향을 결합한 모델은 정확도에서 최하위권을 기록하였는데 해당 모델은 12개 중에 가장 복잡한 모델이며, 7일 이상의 비교적 장기간을 예측할 때 낮은 정확도를 기록했다.

(3) 장기 예측 모델

장기 예측은 지배 변이 바이러스가 변하는 시기에서 신규 변이 바이러스의 전파율 예측을 주목적으로 하였다. 이를 위해 PINN 모델을 사용하였다. 계층 수와 노드 수, 활성화함수에 대한 정확도 평가를 통하여 계층 수는 8층, 노드 수는 64개, 활성화함수는 하이퍼볼릭 탄젠트함수(Hyperbolic Tangent)와 정류된 선형함수(ReLU)를 사용하였다.

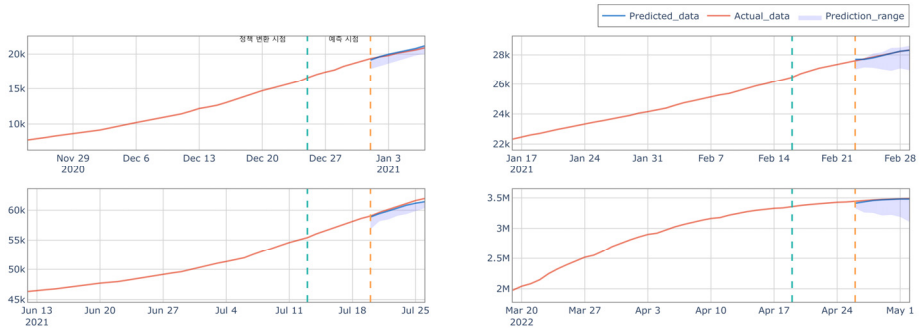
2) 하이퍼 파라미터

학습 네트워크의 하이퍼 파라미터는 격자 탐색 기법을 통하여 최적의 파라미터를 결정하였다. 최종적으로 선택된 하이퍼 파라미터는 학습률(learning rate): 10^{-3}, 인내 기간(patience): 50, 계층 수(num layers): 1, 은닉 상태 크기(hidden size): 16, 배치 크기(batch size): 64, 손실 함수(criterion): MSELoss를 사용하여 학습하였다.

3) 예측 결과

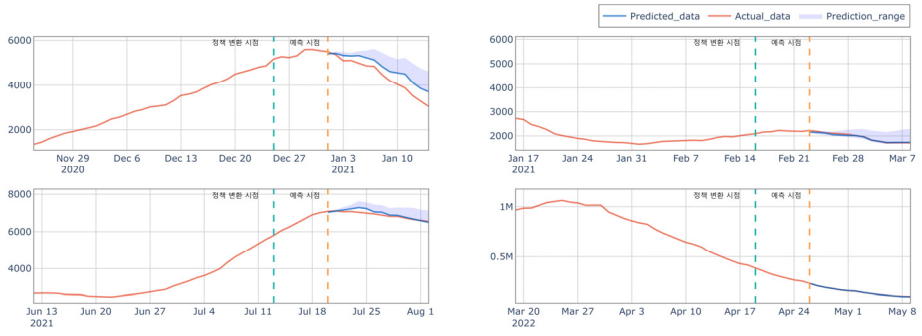
〈그림 2〉, 〈그림 3〉, 〈그림 4〉는 각각 서울특별시의 단기·중기·장기 예측의 시각화이다. 이외 시도별 단기·중기 예측 결과의 시각화는 각각 〈부록 1〉, 〈부록 2〉에 제시되어있다. 단기·중기 예측의 시각화를 살펴보면 실제값이 대부분 예측 범위 내에 있으며, 예측값의 차이가 크지 않음을 확인할 수 있다. 시도별 신규 변이 바이러스 전파율의 장기 예측 결과는 〈표 2〉에서 확인할 수 있다. 〈표 2〉를 통해 기존 바이러스 변이인 델타 변이의 전파율보다 신규 바이러스 변이인 오미크론 변이의 전파율이 비교적 높은 것을 확인할 수 있다. 그러나 제주와 충남의 경우는 오미크론 변이의 전파율이 낮게 나타난다. 이는 제주와 충남 지역의 오미크론 변이 급격한 확산이 다른 지역에 비해 다소 늦게 시작하였기 때문이다.

〈그림 2〉 서울특별시의 단기 예측 시각화



주: 세로축은 확진자 수이며 가로축은 날짜이다.

〈그림 3〉 서울특별시의 중기 예측 시각화



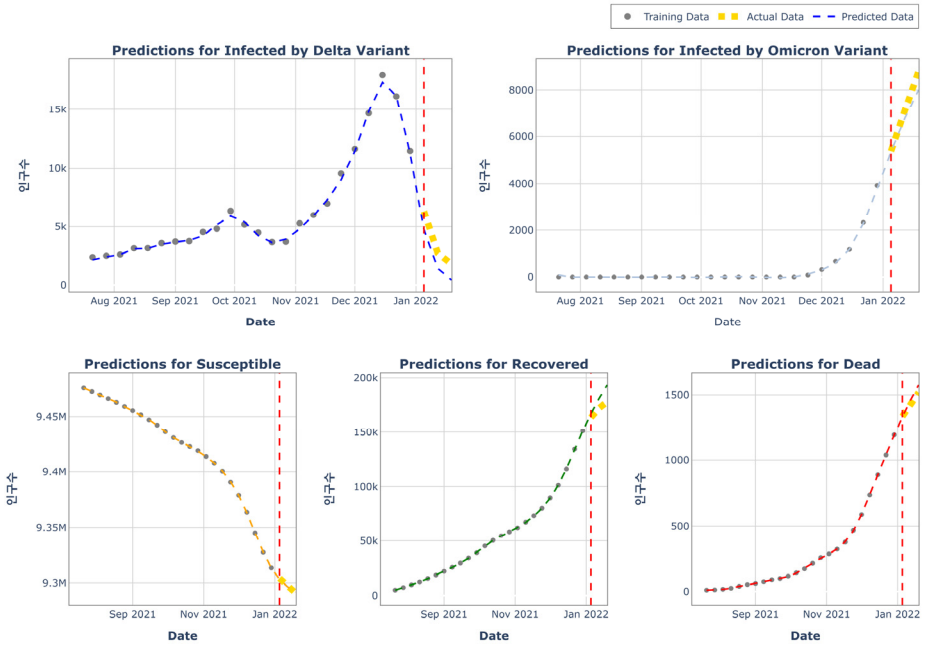
주: 세로축은 확진자 수이며 가로축은 날짜이다.

〈표 2〉 시도별 신규 변이 바이러스 전파율 장기 예측 결과

시·도	강원	경기	경남	경북	광주	대구	대전	부산	서울
전파율1	0.2169	0.2180	0.2031	0.2125	0.2175	0.2144	0.1934	0.2518	0.2699
전파율2	0.2580	0.6031	0.2374	0.4454	0.6448	0.5328	0.3414	0.5174	0.3119
시도	세종	울산	인천	전남	전북	제주	충남	충북	
전파율1	0.1816	0.2492	0.2716	0.1938	0.2124	0.2325	0.2130	0.2752	
전파율2	0.3261	0.2460	0.4296	0.3868	0.4178	0.0675	0.1527	0.6066	

주: 해당 기간에는 델타 변이에서 오미크론 변이로 지배 바이러스가 전환되는 시기였다. 전파율1은 기존 지배 바이러스였던 델타 변이의 전파율이며, 전파율2는 신규 변이 바이러스인 오미크론 변이의 전파율이다. 각각의 전파율은 7일 단위로 계산되었다.

〈그림 4〉 서울특별시의 장기 예측 시각화



장기 예측의 경우 신규 변이 바이러스에 대한 전파율 예측이 주목적이나, 이외의 모든 시계열 데이터에 대해서 충분히 좋은 예측을 보임을 확인할 수 있다. 또한, 신규 변이 바이러스의 학습데이터는 상당수가 0으로 학습할 수 있는 정보가 부족하나, 장기 예측에 사용한 PINN은 전체 바이러스 확산에 대한 지배 방정식 식 (5)를 함께 학습에 사용하기 때문에 기존 변이 바이러스의 확진자 수와 같은 다른 시계열 데이터에서 정보를 받아서 신규 변이 바이러스 학습에 사용할 수 있다.

2. 통계 분석

본 연구에서는 지역별 특성과 정책 변화에 따른 영향을 분석하기 위해 지역별 특성을 나타내는 데이터로써 인구수, 면적, 인구 밀도, 노인 비율, 이동량 데이터를 사용하였다. 우리는 실제값과 예측값의 차이를 방역 정책 변화의 영향력과 비례한다고 가정하고 이를 정책효과(실제값-예측값)라고 하겠다. 즉, 실제값은 방역 정책 전환에 의해 영향을 받은 확진자의 수이고, 예측값은 기존 방역 정책에 영향을 받은 확진자의 수라고 가정하였다.

1) 대응 표본 t-검정

다음은 대응표본 t-검정을 통해 방역 대책 전환이 실제 코로나 확진자 수에 유의미한 영향을 미쳤는지 시도별로 검정해보았다. 독립표본 t-검정이 아닌 대응표본 t-검정을 사용한 이유는 코로나 확진자 수는 시계열 데이터의 특성상 외부요인의 영향을 많이 받게 된다. 그러므로 독립표본이 아닌 대응표본 t-검정을 사용하였다.

해당 분석은 단기 예측에서는 7개의 예측값, 중기 예측에서는 14개의 예측값 데이터를 사용했고, 정규성을 만족하면 T-검정, 만족하지 않으면 윌콕슨(Wilcoxon) 부호순위 검정하였으며, 〈표 3〉은 대응되는 정책효과의 평균값이다. 분석 결과에 따르면 방역 정책 강화1의 경우 단기 예측에서는 모든 도시에서 정책 변화가 확진자 수에 미치는 영향력을 확인할 수 없었지만, 중기 예측에서는 몇몇 도시에서 영향력을 확인할 수 있었다. 방역 정책 강화2의 경우에는 단기, 중기 모두 모든 도시에서 정책 변화가 확진자 수에 미치는 영향력을 확인할 수 없었다. 방역 정책 완화1은 몇몇 도시에서 영향력이 존재했지만 대부분 존재하지 않았다. 방역 정책 완화2의 경우 대부분 도시에서 영향력이 있었음을 확인하였다.

이처럼 방역 정책 변환에 따른 효과는 방역 정책을 얼마나 변환시켰냐에 따라서도 차이가 나지만 도시마다 차이가 나는 것을 확인할 수 있다. 이를 통해 방역 정책을 도시마다 다르게 적용해야 한다는 결론을 도출할 수 있다.

〈표 3〉 시도 및 구간별 정책효과 분석

시·도	단기 예측				중기 예측			
	강화1	강화2	완화1	완화2	강화1	강화2	완화1	완화2
강원	44.62	52.90	15.41*	4678.60***	-29.75***	-6.99	9.08***	233.85
경기	318.31	50.39	259.19***	12780.29*	122.36	107.96	1.63	-612.07
경남	57.41	95.41	-6.05	3863.73**	-13.77**	35.97	-4.72	628.00**
경북	0.12	11.96	-18.88	8682.17***	11.33	28.18	-8.03	-685.02
광주	48.24	5.53	-12.58	540.63	1.47	-14.58	-2.82	1193.93
대구	25.49	39.72	-2.72	9252.85***	19.74	46.45	2.92	226.06
대전	-0.62	81.00	2.76	8545.27***	-14.68**	-11.23	0.16	326.32***

시·도	단기 예측				중기 예측			
	강화1	강화2	완화1	완화2	강화1	강화2	완화1	완화2
부산	67.56	121.93	12.26*	3812.85***	-208.88***	100.38	-18.20	-269.08
서울	-108.26	308.59	-3.18	12679.11*	-534.49***	-43.25	-10.58	485.63
세종	-0.20	10.60	0.21	1207.08**	-1.47*	39.21	2.22*	-54.01
울산	63.77	4.57	7.65**	2877.22***	20.15	7.39	8.16***	303.86*
인천	45.95	145.39	-3.39	5663.99***	-4.92	-6.10	-4.54	-840.02
전남	17.78	7.59	-0.94	9850.64***	5.97	9.21	-4.07	1095.55**
전북	24.49	-9.18	22.77***	3408.66***	-44.88***	69.30	-13.18	395.80
제주	-3.51	58.67	-4.49	2144.54***	7.90	30.81	0.23	777.81***
충남	-1.29	48.04	77.24**	6788.07**	-13.07**	-3.26	-88.64	-671.94
충북	37.61	67.13	-8.12	4144.10**	-40.72***	57.34	8.49*	40.61
합계	337.14	1100.25	337.14	100919.79	-717.73	446.81	-121.91	2575.27
평균	19.83	64.72	19.83	5936.46	-42.22	26.28	-7.17	151.49

* p<0.05 ** p<0.01 *** p<0.001

2) 상관 분석

다음은 상관 분석을 통해 방역 정책효과가 어떤 변수와 상관이 있는지를 확인하였다. 우선 데이터 수가 17개밖에 되지 않으므로 샤피로-윌크(shapiro-wilk) 검정을 통해 정규성 검정을 하였다. 정규성 검정의 결과는 〈부록 3〉에 제시되어 있다. 검정 결과 노령 인구 비율과 방역 정책 완화2의 정책효과만이 정규분포를 따르고 있다. 정규성이 확보된 변수끼리는 피어슨(Pearson) 상관 분석을 했으며, 나머지 변수들은 스피어만(Spearman) 상관 분석과 켄달(Kendall)의 타우(tau) 상관 분석을 하였다.

〈표 4〉은 각 변수에 대한 상관 분석 결과이다. 상관 분석 결과에 따르면 방역 정책 강화1의 정책효과와 인구수, 방역 정책 완화2의 정책효과와 인구수 및 인구이동량이 정(+)적인 상관을 갖는 것으로 나타났다. 이는 인구수 및 인구이동량이

많을수록 정책 변화에 따른 실제 확진자 수가 더 많다는 것을 의미하며, 이를 통해 인구수 및 인구이동량이 많은 곳일수록 방역 정책을 더 높게 유지해야 한다는 것을 확인할 수 있다.

〈표 4〉 상관 분석 결과표

구분	변수1	변수2	Pearson	Spearman	Kendall-tau
2020.12.24. (방역정책 강화1)	정책효과	인구수	-	0.260	0.177
		면적	-	0.115	0.030
		인구 밀도	-	0.152	0.118
		노령인구 비율	-	-0.171	-0.163
		인구이동량	-	0.304	0.235
2021.07.14. (방역정책 강화2)		인구수	-	0.500*	0.353
		면적	-	-0.096	-0.088
		인구 밀도	-	0.375	0.265
		노령인구 비율	-	-0.056	-0.059
		인구 이동량	-	0.441	0.294
2021.02.15. (방역정책 완화1)		인구수	-	0.093	0.074
		면적	-	0.069	0.044
		인구 밀도	-	-0.022	0.015
		노령인구 비율	-	-0.014	-0.015
		인구이동량	-	0.110	0.074
2022.04.18. (방역정책 완화2)		인구수	-	0.674**	0.471**
		면적	-	0.373	0.265
		인구 밀도	-	-0.017	-0.059
		노령인구 비율	0.1983	-	-
		인구이동량	-	0.608**	0.441*

* p<0.05 ** p<0.01 *** p<0.001

V. 연구의 결과

1. 분석 결과 요약

지역별 방역 정책 전환에 따른 영향을 분석하기 위해 지역별 인구수, 면적, 인구 밀도, 노령 인구 비율, 인구이동량 데이터와 방역 정책 전환 이후의 실제 코로나 확진자 수와 예측된 확진자 수의 차이를 분석하였다. 대응 표본 t-검정을 통해 중기 예측에서 방역 정책 전환에 따른 예측값과 실제값의 차이를 분석하였다. 이를 통해 방역 정책 전환이 확진자 수 변화에 미치는 영향은 시도마다 차이가 나는 것을 확인하였다. 또한, 상관 분석을 통해 인구수 및 인구이동량이 많은 지역일수록 더욱 엄격한 방역 정책을 시행해야 한다는 것을 확인하였다.

2. 연구의 의의 및 정책적 함의

본 연구는 인공 신경망을 통한 코로나19 확진자의 예측을 통해 코로나19 방역 정책을 평가하고자 하였다. 코로나19 방역 정책은 사회적, 경제적, 문화적 측면에 모두 영향을 미치므로 단순히 확진자 수로만 평가할 수는 없다. 그러나 확진자 수는 명확히 정책의 결과에 의한 정량화된 데이터이다. 때문에, 본 연구는 방역 정책의 영향력을 평가하고 지역별 맞춤형 방역 정책의 적용을 제안하는데 의의가 있다.

지역 간 편차를 고려한 유연성 있는 방역 정책의 적용은 새로운 전염병 확산 시 필수적이다. 이를 위해 다양한 모델을 활용한 시뮬레이션으로 정확한 예측을 수행하고, 이를 지역별 확진자 예측에 적극적으로 반영해야 한다. 또한, 새로운 전염병 발생 시 상황에 맞는 모델의 개발과 적용을 추진하여 예측력을 높여야 한다. 예를 들어 본 연구에서처럼 여러 변이 바이러스가 동시에 출현하는 상황에도 대응할 수 있는 모델의 연구가 필요하다. 더불어, 감염병 관련 데이터와 분석 방법을 체계화하고 데이터베이스를 구축하여 효과적인 분석을 지원해야 한다. 이를 위해 제도적인 체계를 마련하고 관련 법규에 따라 감염병 대응과 관리에 필요한 데이터의 원활한 분석과 공유가 가능하도록 한다. 연구와 개발을 활성화하여 감염병 예방과 대응에 필요한 지식을 확보하고 이를 위한 감염병 관련 데이터의 데이터베이스화가 필요하다.

3. 연구의 한계

본 연구의 주요 데이터인 보건복지부 코로나19 시·도 발생 현황 데이터는 2022년 3월 29일부터 회복자에 대한 데이터를 얻을 수 없었으며, 일부 데이터에 대하여 오기 및 누락을 포함하고 있었다. 때문에, 학습용 데이터에 많은 노이즈와 결측치를 포함하였고 이는 예측 성능을 낮추는 요인이 되었다. 특히 회복자 데이터에 대한 누락은 연구 결과에 상당한 영향을 미쳤다고 할 수 있다. 회복자에 정확한 데이터를 확보하고 있었다면 좀 더 정확하고 유의미한 결론에 도달했을 것으로 예상한다. 연구에서 사용한 바이러스 변이의 분포 데이터는 2주간 간격으로 작성되어 있었다. 좀 더 세밀한 간격으로 변이의 분포를 알 수 있었다면, 좀 더 정확한 데이터를 생성하여 연구에 활용했을 것으로 예상한다. 또한, 코로나19 확진자 수에 대한 데이터는 보건복지부 외에도 각 지방자치단체에서 지역별로 발표하고 있으며, 이를 보건복지부가 취합하여 시도 발생 현황으로 발표하고 데이터베이스화하여 제공한다. 이때, 시도별뿐만 아니라 모든 시군에 대한 데이터를 제공했으면 더 정확한 연구를 할 수 있었을 것이다. 예측을 위해 연구에서 사용한 인공 신경망은 주로 순환신경망과 물리 정보 기반 신경망을 사용하였으나, 최근 인공지능 기술의 발전으로 시계열 데이터에 대한 높은 정확도를 보이는 다양한 모델들이 소개되고 있다. 특히 현재 최고 수준의 결과를 보여주는 SOTA(State-of-the-art)[27]를 통해 소개된 N-beats[28], Temporal Fusion Transformer[29] 등의 시계열 예측 모델을 코로나19 확산 예측에 적용한다면 본 연구의 예측보다 높은 수준의 예측이 가능할 것이다.

김남종(2020)은 코로나19 방역 정책은 경제활동의 편익과 감염병 확산으로 인한 비용 간 트레이드오프(tradeoff)가 명백히 존재함으로 코로나19 통제를 위한 방역 정책 시행에는 어려움을 분석하였다. 그러나 본 연구에서는 단순히 감염병의 확산에 집중하여 정책을 평가하고 있다. 이후의 연구에서는 경제활동의 정량적 지표를 활용하여 필요가 있다.

---

27) https://paperswithcode.com/sota (2020.07.04.)
28) https://paperswithcode.com/paper/n-beats-neural-basis-expansion-analysis-for (2020.07.04.)
29) https://paperswithcode.com/paper/temporal-fusion-transformers-for (2020. 07.04.)

참고문헌

김남종(2020), “감염재생산 지수를 고려한 코로나19 대응의 경제학적 논의와 시사점”, 『주간금융브리프』, 29(24) : 3~11 (한국금융연구원).

김진오·김지수·김희원·박효상·김지혜·옥혜원·문혜원·고석주(2020), “SIR 모형과 LSTM 기반의 COVID-19 확진자 예측 비교 및 분석”, 『한국지능정보시스템학회 학술대회논문집』, 59~64.

노윤아·정승원·문재욱·황인준(2022), “사회적 변수를 고려한 LSTM 기반 코로나19 일별 확진수 예측 기법”, 『정보과학회 컴퓨팅의 실제 논문지』 28(2): 116-121한국정보과학회.

유명수·김연주·백수진·권동혁(2021), “감염재생산지수 개념 및 방역 정책에 따른 변화”,『주간 건강과 질병』 14(6) : 282~287.

정재웅·권혁무·홍성훈·이민구(2020), “국내 코로나바이러스감염증-19의 감염재생산지수 추정” 『품질경형학회지』 48(3) 493-510.

홍태경·김은서·이희상(2023), “코로나19 확진자 수 예측을 위한 딥러닝 모델: 확산 패러다임 변화를 반영한 입력변수 조정” 『대한산업공학회지』 49(2) 120-132.

Ajbar, A., R. T. Alqahtani, and M. Boumaza(2021), “Dynamics of an SIR- Based COVID-19 Model With Linear Incidence Rate, Nonlinear Removal Rate, and Public Awareness”, Frontiers in Physics, 9: 1~13.

Schuerger, C., Batalis, S., Quinn, S., Adalja, A., and Puglisi, A.(2023) "Viral Families and Disease X: A Framework for U.S. Pandemic Preparedness Policy" (Center for Security and Emerging Technology).

Chakraborty, T. and I. Ghosh(2020), “Real-time forecasts and risk assessment of novel coronalvirus(COVID-19) cases: A data-driven analysis”, Chaos, Solition & Franctals, 135:108850, Elsevier.

Chimmula, V. and L. Zhang(2020), “Time series forecasting of Covid-19 transmission in Canada using LSTM networks”, Chaos, Soliton & Fractals, 135:109864, Elsevier.

Choi, Y., J. Kim, H. Choi, H. Lee, and C. Lee(2020) “Assessment of Social Distancing for Controlling COVID-19 in Korea: An Age-Structured Modeling Approach” International Journal of Environmental Research and Public Health, 17(20): 7474, MDPI.

Hale, T., N. Angrist, R. Goldszmidt, et al(2021), “A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)”. Nat Hum Behav, 5:529–538.

Kermak, W. O. and A. G. McKendrik(1927), “A contribution to the mathematical theory of epidemics”, Proceedings of the Royal Society A, 115:700-721.

Kim, B., E. Kim, S. Lee, and C. Oh(2020), “Mathematical Model of COVID-19 Transmission Dynamics in South Korea: The Impacts of Travel Restrictions, Social Distancing, and Early Detection”, Processes, 8(10): 1~18, MDPI.

Kim, Y., M. Seo, and H. Yeom,(2020), “Estimating a breakpoint in the pattern of spread of COVID-19 in South Korea”, International Journal of Infectious Diseases, 97: 360~364, Elsevier.

Perakis, G., D. Singhvi, O. S. Lami and L. Thayaparan(2022), “COVID-19: A multiwave SIR-based model for learning waves”, Production and Operations management, Sepcial issue: 1~19, Wiley.

Regia-Nava, J. A., Y. Wang, C. Fontes-Garfias, Y. Liu, T. Syed, M. Susantono, A. Gonzalez, K. Viramontes, S, Verma, K. Kim, S. Landeras-Bueno, C. Huang(2022), “A Zika virus mutation enhances transmissionpotential and confers escape fromprotective dengue virus immunity”, Cell Reports, 39: 110655, CellPress.

Shim, E., A. Tariq, and G. Chowell(2021), “Spatial variability in reproduction number and doubling time across two waves of the COVID-19 pandemic in South Korea, February to July, 2020”, International Journal of Infectious Diseases, 102: 1-9, Elsevier.

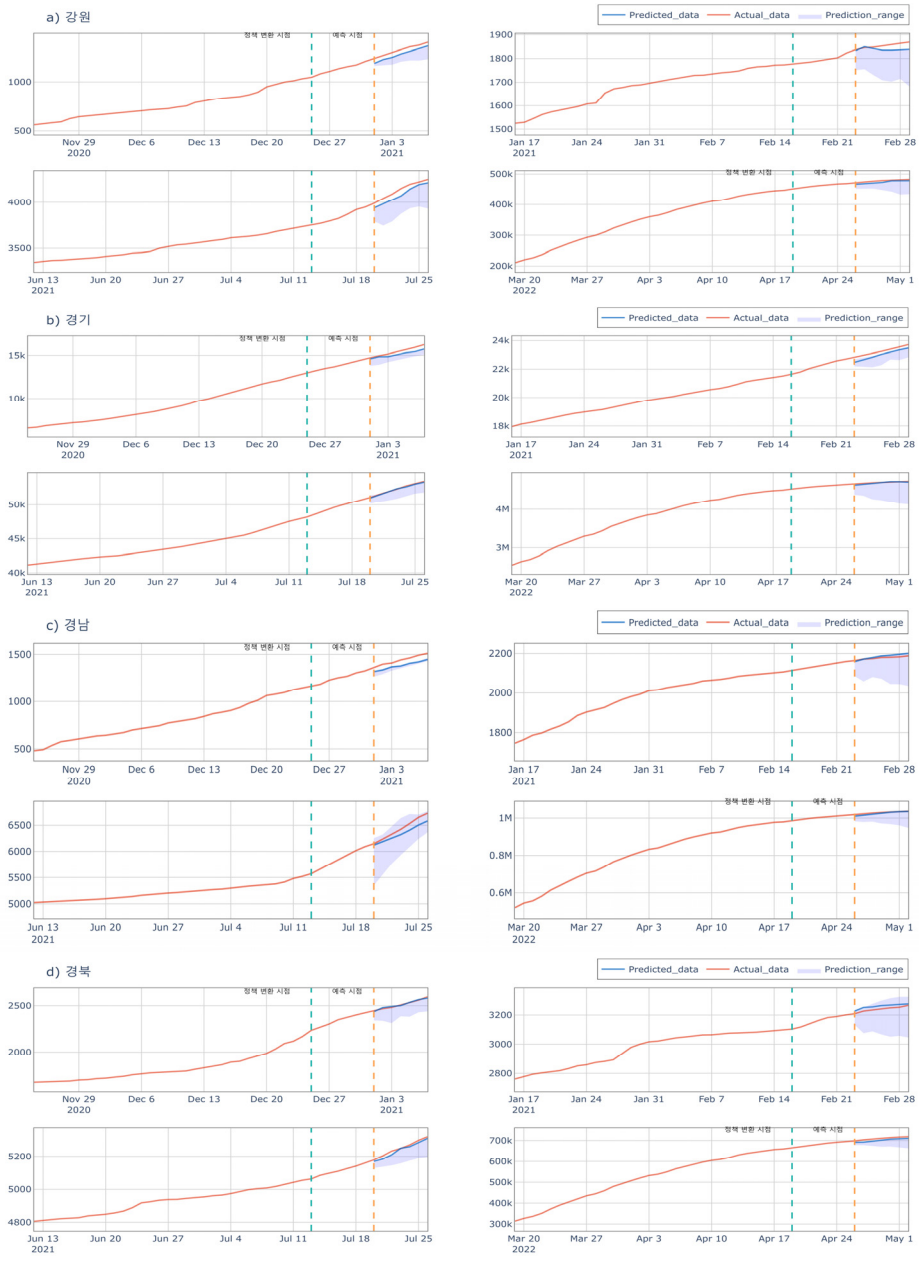
Siraj, A., A. Worku, K. Berhane, M. Aregawi, M. Eshetu M, A. Mirkuzie, Y. Berhane, and D. Siraj(2020), "Early estimates of COVID-19 infections in small, medium and large population clusters", BMJ Glob Health, 5(9): e003055.

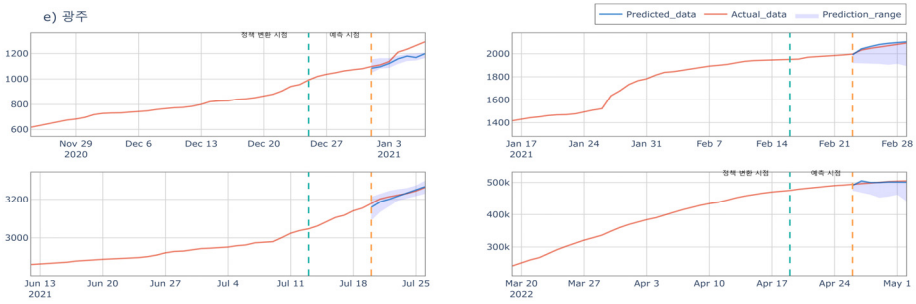
Tariq, A., E. Undurrag, C. C. Laborde CC, K. Vogt-Geisse, R. Luo, R. Rothenberg, and G. Chowell(2021), "Transmission dynamics and control of COVID-19 in Chile, March-October, 2020", PLOS Negl Trop Dis, 15(1): e0009070.

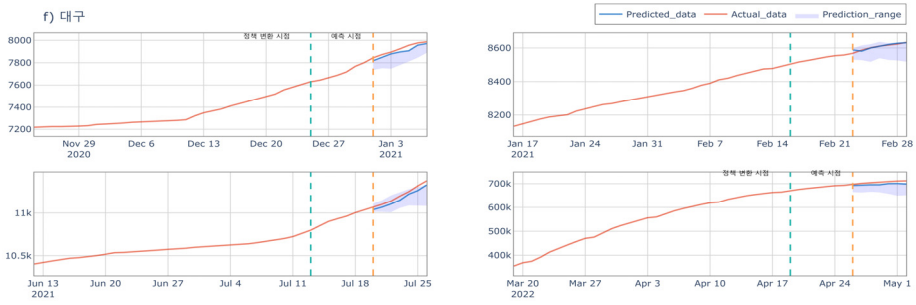
Xiong, C., S. Hu, M. Yang, and L. Zhang,(2020). "Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections", Proceedings of the National Academy of Sciences, 117(44): 27087-27089.

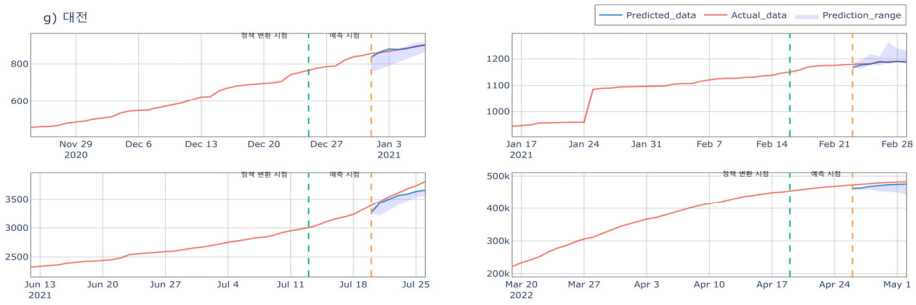
부 록

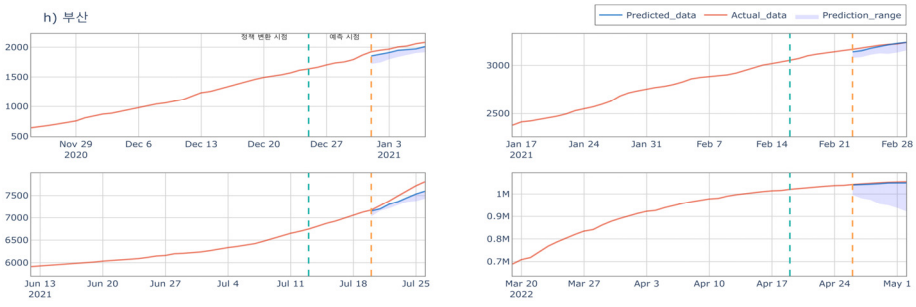
〈부록 1〉 시도별 단기 예측 결과 시각화

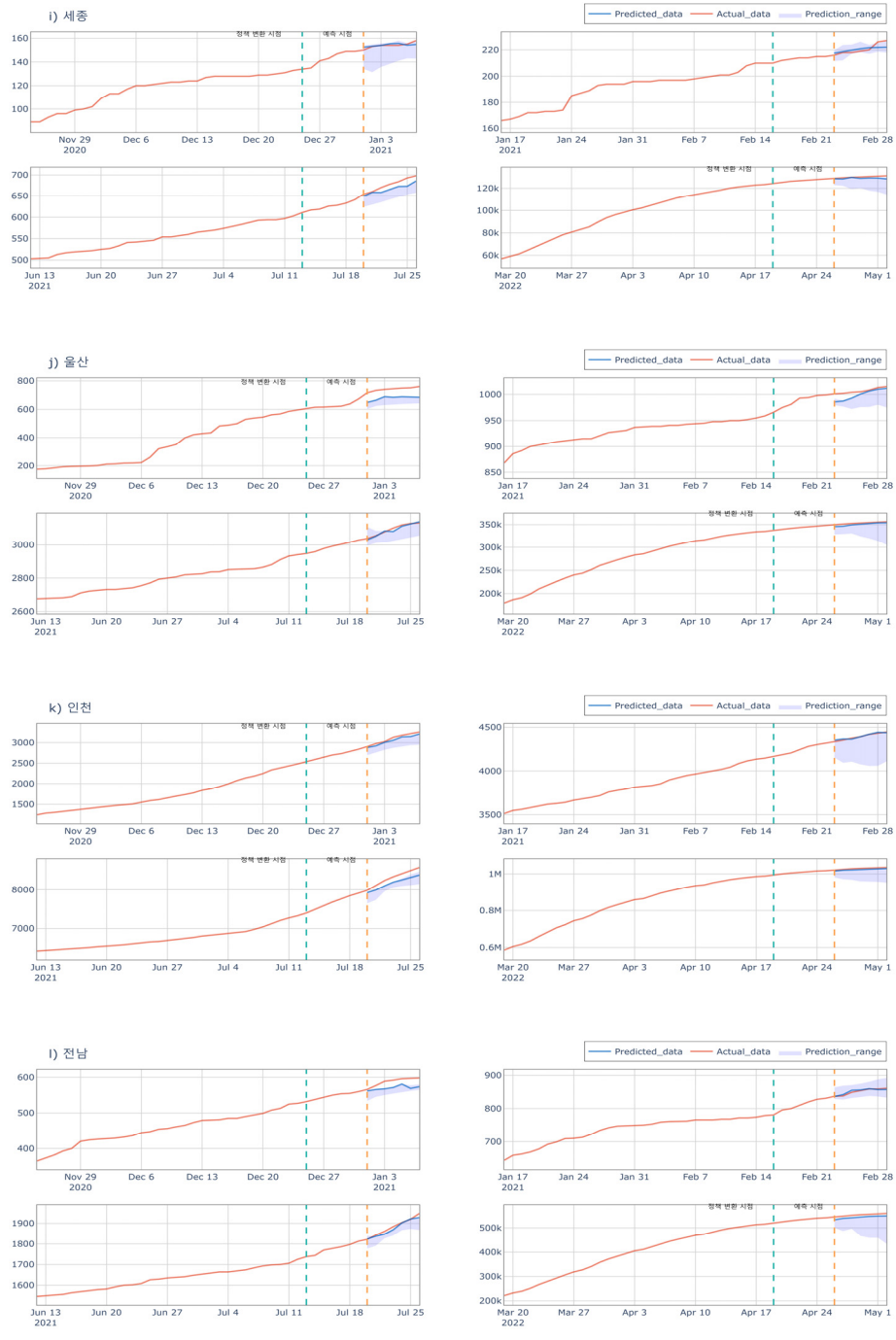


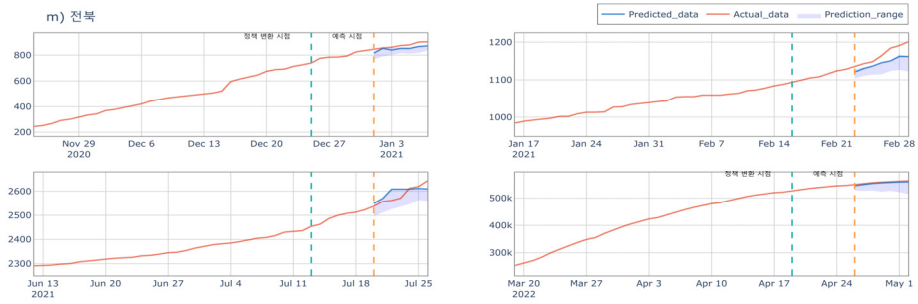

e) 광주
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

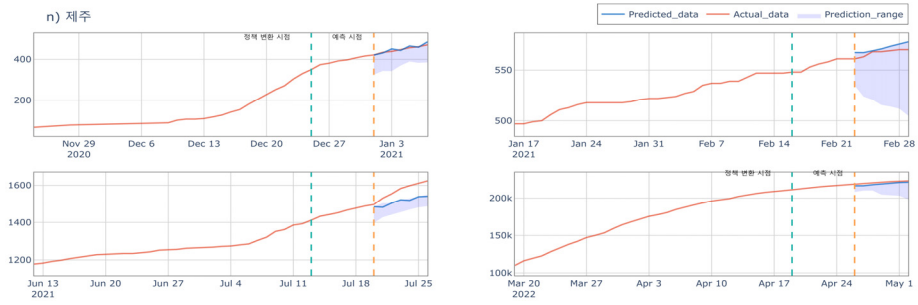

f) 대구
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

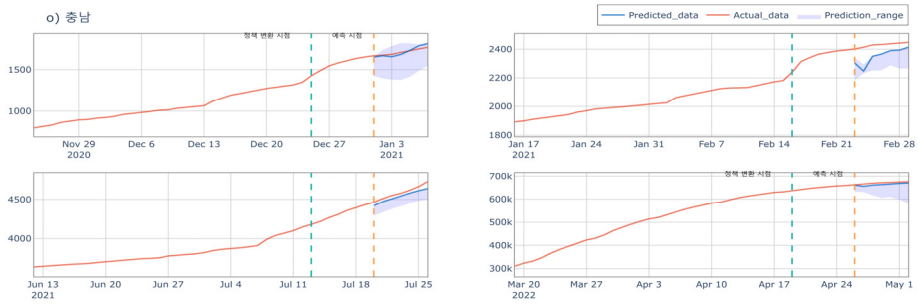

g) 대전
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

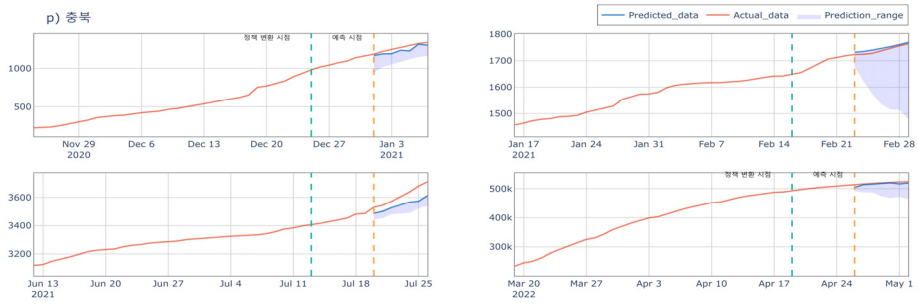

h) 부산
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range


i) 세종
j) 울산
k) 인천
l) 전남
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점

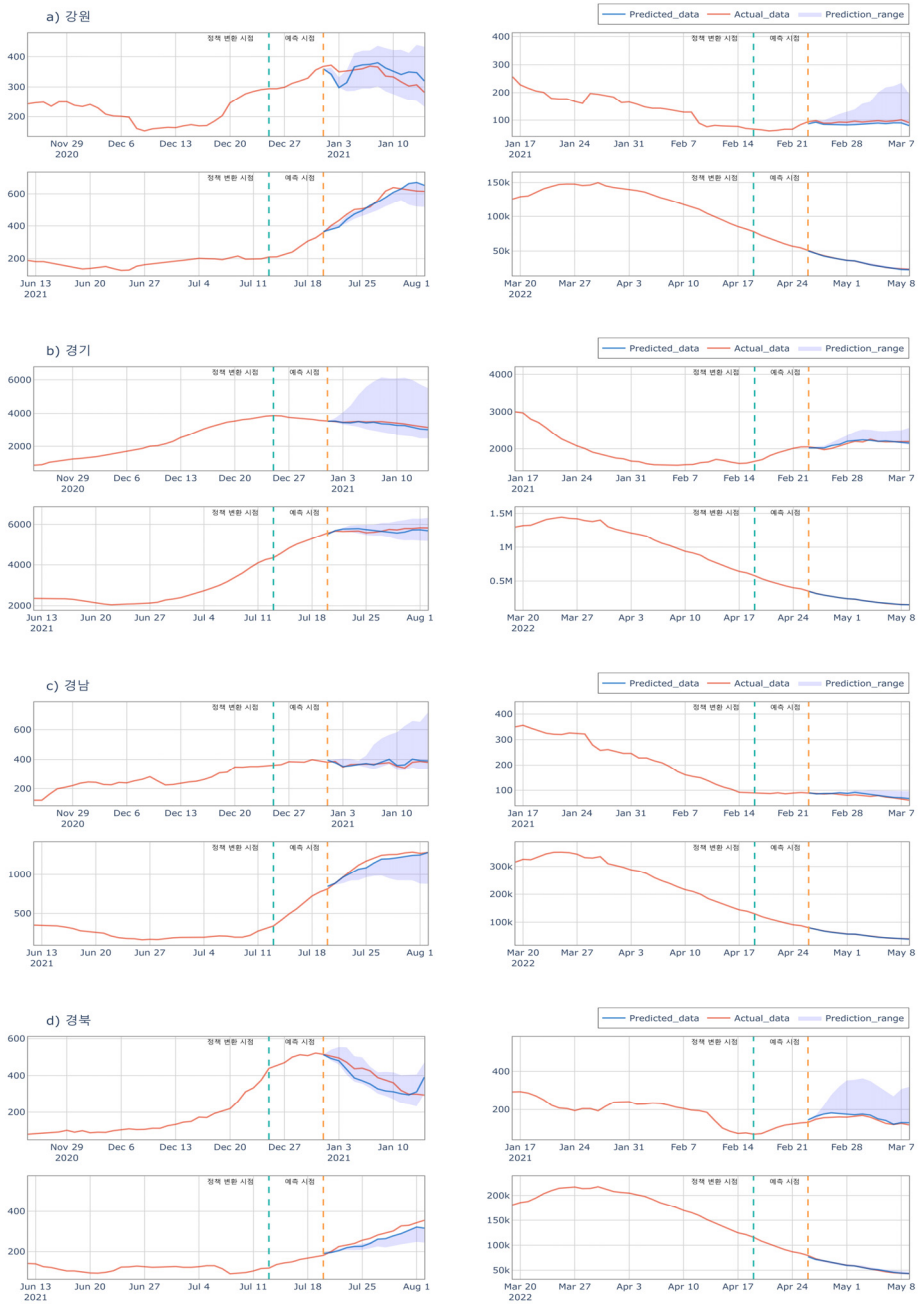

m) 전북
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

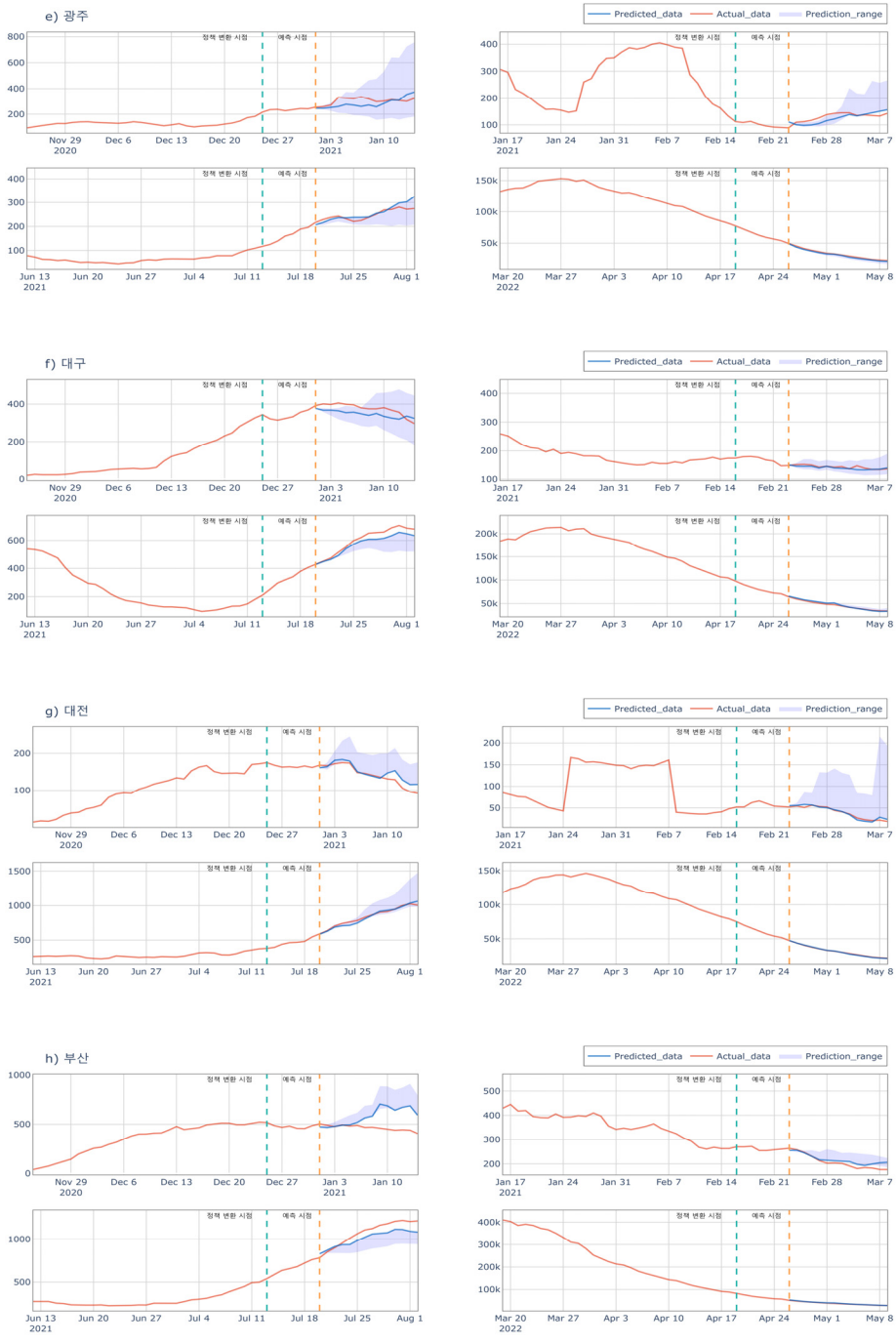

n) 제주
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

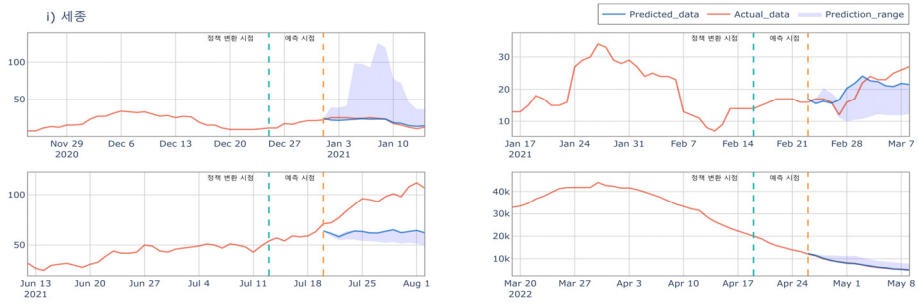

o) 충남
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

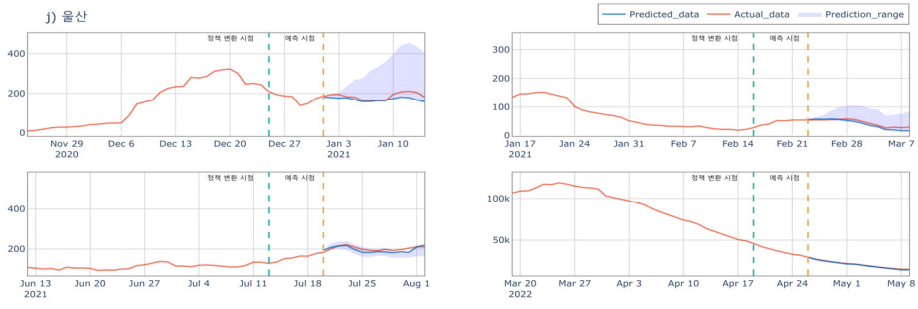

p) 충북
정책 변환 시점
예측 시점
Predicted_data
Actual_data
Prediction_range

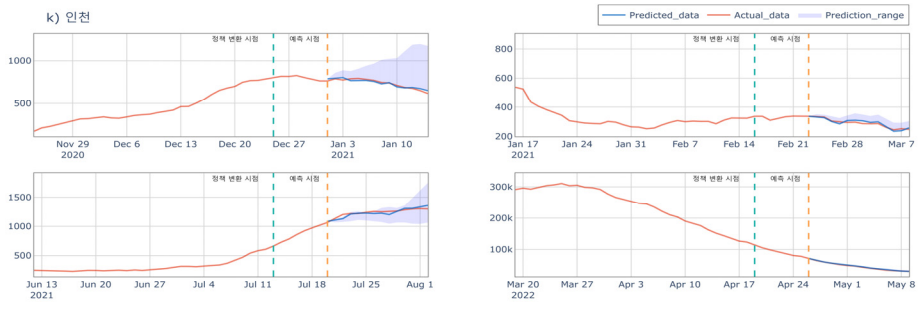
〈부록 2〉 시도별 증기 예측 결과 시각화

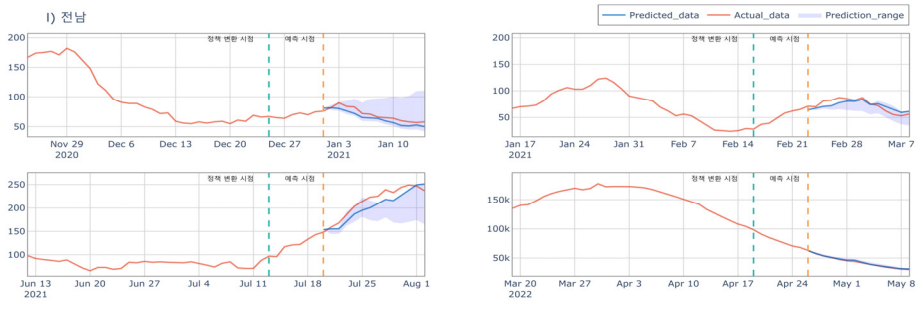



e) 광주
f) 대구
g) 대전
h) 부산
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점

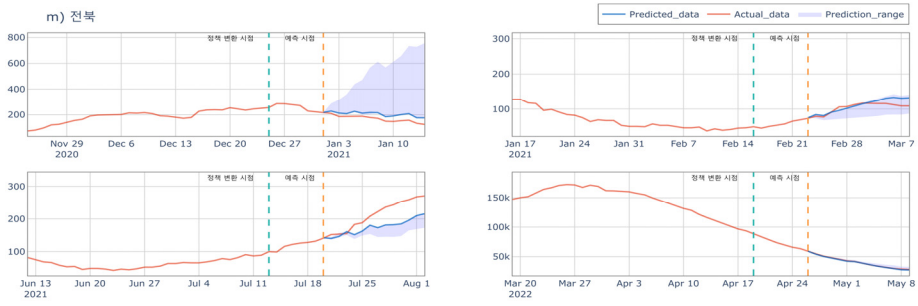

i) 세종
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점


j) 울산
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점

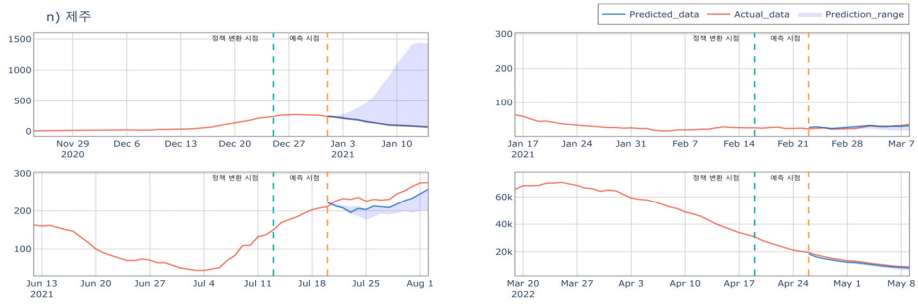

k) 인천
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점


l) 전남
Predicted_data
Actual_data
Prediction_range
정책 변환 시점
예측 시점

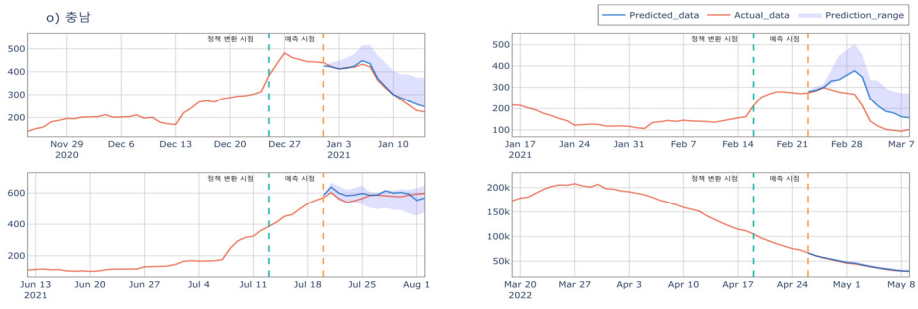
m) 전북



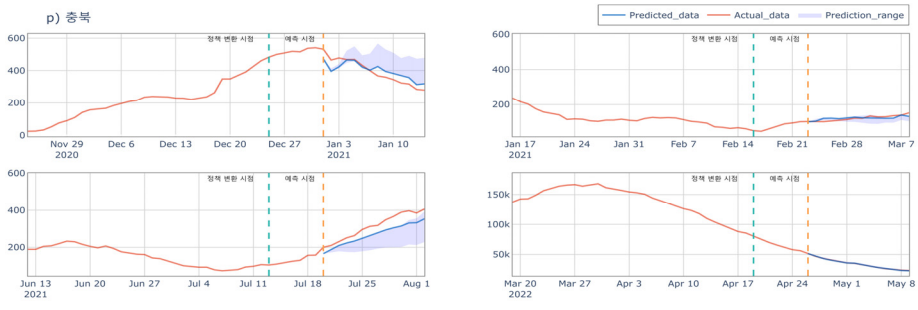
n) 제주



o) 충남



p) 충북



〈부록 3〉 정규성 검정

구분		변수	검정 통계량
시도		인구수	0.638***
		면적	0.824**
		인구 밀도	0.548***
		노령인구 비율	0.976
		이동량	0.671***
단기 예측	강화1	실제값 - 예측값	0.700***
	강화2		0.778**
	완화1		0.497***
	완화2		0.938
중기 예측	강화1		0.573***
	강화2		0.897
	완화1		0.673***
	완화2		0.959

* p<0.05 ** p<0.01 *** p<0.001