공간회귀모형을 활용한 지역별 실업률과 당뇨병 유병률의 관계 분석

배지원 경제학과 2021150323

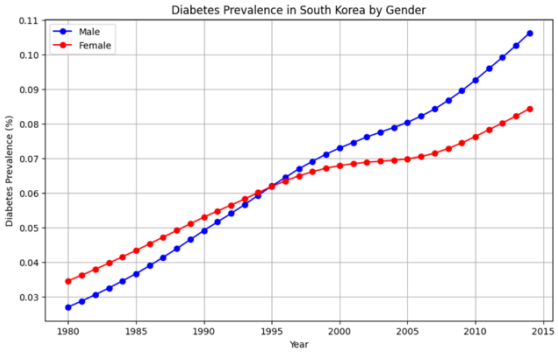
DATA301 데이터 실험 설계 및 분석

2024.06.26

0. Abstract

본 연구는 전 세계적으로 증가하는 당뇨병 유병률에 영향을 미치는 사회경제적 요인들을 분석하는 것을 목적으로 하고 있습니다. 'The Lancet' 의학 저널과 세계은행, Our World in Data에서 제공하는 데이터셋을 활용하여, 다변량 분석을 위해 Lasso, Random Forest, Gradient Boosting과 같은 다변량 분석 모델들로 각 요인의 영향력을 평가하였습니다. 초기 분석 결과, 실업률이 당뇨병과의 유의미한 상관관계가 나타나 우리나라 시·군별 작은 단위에서는 지역적으로 어떤 관계가 있는지 살펴보았습니다. 시·군별 실업률, 당뇨병 진단 경험률, 그리고 위치 데이터를 수집하여 공간자기회귀모형 SAR(Spatial Autoregression Model)을 활용한 공간회귀분석을 진행하였습니다. 이를 통해 실업률과 당뇨병 유병률 간의 관계를 더욱 세부적으로 이해하고, 지역적 특성을 고려한 공중보건 정책 개발에 기여할 수 있는 통찰을 제공하고자 합니다.

1. Introduction



△ 그림 1 성별 한국 당뇨병 유병률 추이

당뇨병은 세계적으로 환자 수가 꾸준히 증가하고 있는 주요 만성 질환 중 하나입니다. 특히 우리나라 당뇨병 유병률은 1970년대의 약 2%를 시작으로 점차적인 증가현상을 보이면서 90년대 초부터 10%에 육박하는 결과를 보이고 있습니다.[1] 이러한 증가 추세는 다양한 사회경제적, 환경적, 유전적 요인의 복합적인 작용에서 기인할 수 있습니다. 특히 최근 연구에서는 생활습관의 변화와 도시화가 당뇨병 유병률 증가와 밀접한 관련이 있다고 보고되고 있습니다.[2] 따라서 이러한 경향을 분석하고 이해하는 것은 공중보건 정책과 개인의 건강관리 전략 개선에 중요한 역할을 할 수 있습니다.

본 프로젝트의 목적은 각국의 당뇨병 발생률

[1] 김응진, 김명환, 김상희, 김동렬, 환정운, 이근식, 전영균, 김영건, & 이정섭. (1970). 한국인 당뇨병의 역학적 연구. *서울 의대 잡지, 11*, 25-30

[2] 조남한. (2005). 우리나라 당뇨병의 유병률과 관리 상태. *대한내과학회지, 68*(1), 1-3.

에 미치는 다양한 사회경제적 요인이 미치는 영향을 종합적으로 분석하여 새로운 요인을 파악하는 것입니다. 이를 위해 각국의 데이터를 활용하여 다변량 분석을 통해 당뇨병 유병률과 당뇨병에 영향을 미치는 잠재적 요인의 상관관계를 조사하려 합니다.

또한, 우리나라에서 해당 사회경제적 요인과 당뇨병 유병률의 관계에 대해 탐구하고자 합니다. 특히, 우리나라 내에서 시·군 별 지역들의 지리적 위치에 따라 두 변수들이 어떻게 작용하는지를 분석하고자 합니다. 이를 통해 실업률과 당뇨병 유병률 간의 관계를 더욱 세부적으로 이해하고, 지역적 특성을 고려하는 것이 가능할 것입니다.

2. Global Analysis of Socioeconomic Factors and Diabetes Prevalence

1) 데이터 수집 및 출처

크게 당뇨병 유병률 데이터와 각국 나라별 사회경제적 요인 데이터를 이용하였습니다.

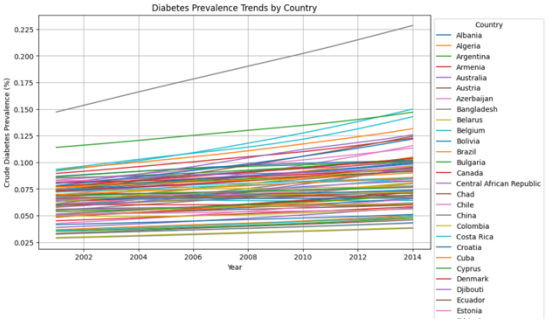
먼저, 당뇨병 유병률 데이터셋은 'The Lancet' 의학 저널에 2016년 게시된 것으로, 전 세계 199개 국가의 18세 이상 성인 인구 중 당뇨병 유병률 정보를 담고 있습니다. 당뇨병 유병률은 1980년부터 2014년까지의 인구 기반 연구와 국가 건강 조사를 통해 수집된 데이터에 기반하며, 연령 및 성별 분포를 조정한 원시 추정치와 95% 불확실성 구간(UI)을 포함합니다.

각국 나라별 사회경제적 요인 데이터셋은 세계은행(World Bank)과 Our World in Data에서 제공하는 정보를 바탕으로, 2000년부터 2019년까지 174개 국가의 데이터를 포함하고 있습니다. 주요 변수로는 각국의 지역, 소득 수준, 연도별 기대 수명, 영양 부족 인구 비율, 이산화탄소 배출량(킬로톤), GDP 대비 현재 건강 지출 비율, 교육 지출 비율, 실업률, 공공부문의 투명성과 부패 정도를 평가하는 CPIA 평점, 안전한 위생 시설 이용 비율, 사고로 인한 장애보정생명연수(DALYs), 전염병 및 비전염성 질병으로 인한 장애보정생명연수 등이 포함됩니다.

2) 데이터 전처리 및 분석(EDA)

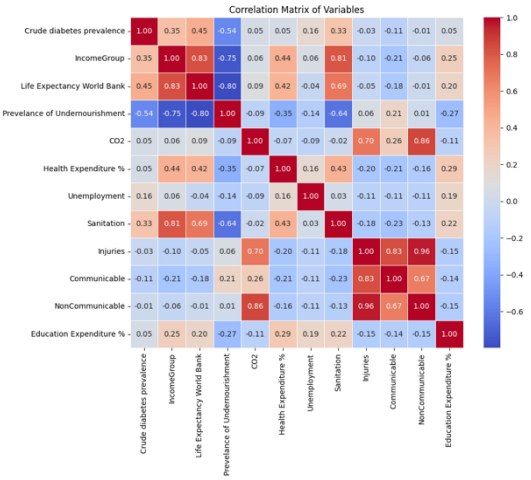
당뇨병 데이터셋은 전 세계 199개 국가를 포함하고 있으며, 사회경제적 요인 데이터셋은 174개 국가에 대한 정보를 제공합니다. 이 프로젝트에서는 두 데이터셋에서 공통적으로 포함된 국가들과 2001년부터 2014년까지의 공통 연도 데이터만을 추출하여 분석하였습니다. 당뇨병 유병률은 남녀의 평균으로 구하였습니다. 데이터 정제 과정에서는 사회경제적 요인 데이터셋의 부패 정도 변수가 전체 행의 절반이 넘는 결측치를 포함하고 있어 이를 제거하였습니다. 그 외 변수들 중에서도 결측치가 많은 경우, 해당 국가를 데이터셋에서 제외시키되, 결측치 비율이 상대적으로 낮고(33% 미만) 중요한 영향을 미칠 가능성이 있는 변수들은 interpolate 보강법과 앞과 뒤 관측치로 보강하는 방법을 통해 결측치를 처리하였습니다. 이러한 과정을 통해 최종적으로 83개 국가의 데이터로 구성된 데이터셋을 확보하였습니다.

이를 바탕으로 각 국가 및 연도별로 요인들의 차이와 추이를 분석해보았습니다. 2001년부터 2014년까지 각 나라별 당뇨병 유병률을 선그래프로 나타내어 보았을 때, 거의 모든 나라에서 비율이 증가하는 추세를 볼 수 있었습니다.

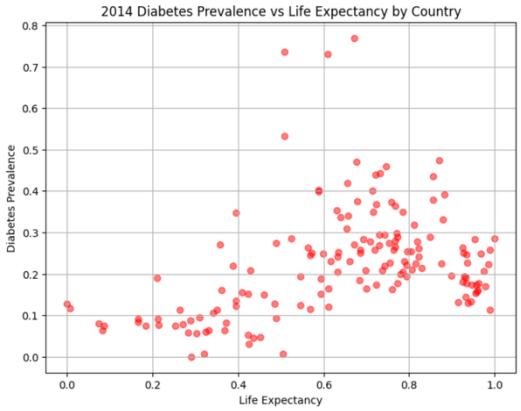


△ 그림 2 각 나라별 당뇨병 유병률 추이

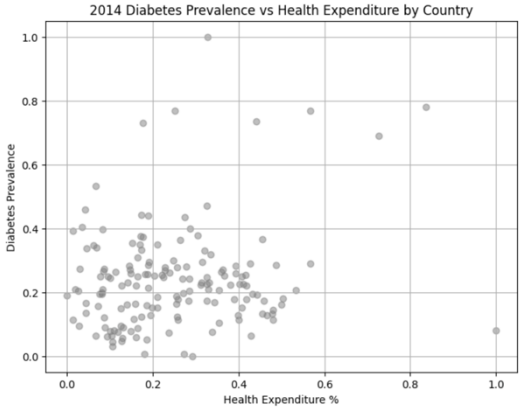
그리고, 가장 최근인 2014년의 데이터에서 변수들끼리의 상관관계 히트맵을 그려보았을 때, 이에 영양부족 인구비율은 강한 음의 관계(-0.54), 수입(0.35)과 기대수명(0.45), 위생시설 접근성(0.33)은 강한 양의 관계를 보인다는 것을 알 수 있었습니다. 하지만 그만큼 수입과 기대수명(0.83), 그리고 장애보정생명연수와 탄소배출량(0.86) 등 타겟 변수가 아닌 요인들끼리도 높은 상관 관계를 보였습니다. 따라서, 이들 간의 패턴이나 비선형 관계들을 포착하기 위해 모델링이 필요함이 강조되었습니다. 스캐터 플롯을 통한 시각적 분석도 시도했습니다. 모든 변수들에 대해서는 스케일링을 적용해주었습니다. 히트맵에서 보였던 상관계수에 따라 얼마나 관계가 있는지를 확인할 수 있었습니다.



△ 그림 3 사회경제적 요인들 간 상관관계 히트맵



△ 그림 4 기대수명과 당뇨병 유병률의 관계 산점도



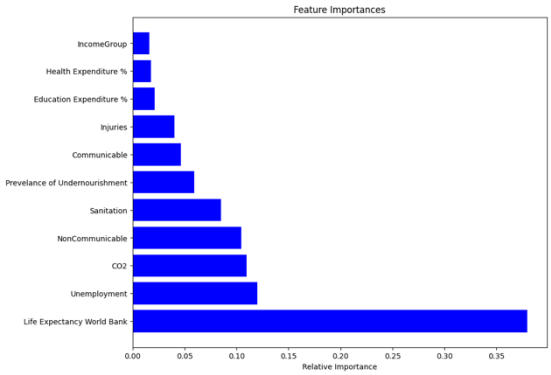
△ 그림 5 건강지출비율과 당뇨병 유병률의 관계 산점도

3) 다변량 분석 모델 및 결과

전처리된 당뇨병 유병률 데이터와 사회경제적 요인 데이터셋을 병합하여 다변량 분석을 수행하고자 하였습니다. Ridge, Lasso, Random Forest, Gradient Boosting 모델들의 하이퍼파라미터들을 각각 튜닝하여 가장 높은 정확도를 나타내는 모델과 그 파라미터들을 최종적으로 채택하였습니다. 이 모델들은 다변량 분석에 적합하며, 각 변수의 영향력을 명확하게 파악할 수 있게 해주기 때문에 선정되었습니다.

MSE(Mean Squared Error) 비교를 통해 랜덤 포레스트와 그라디언트 부스팅 방법이 현저히 낮은 오차 값을 나타내며 뛰어난 성능을 보여준다는 결과를 도출할 수 있었습니다. 거의 완벽에 가까운 성능을 보인 랜덤 포레스트는 n_estimators를 50으로 설정했을 때 R-제곱 값이 0.999로 최적의 정확도를 달성하였고, 그

라디언트 부스팅은 n_estimatores를 200, 학습률을 0.5로 조절했을 때 R-제곱 값이 0.998로 매우 높게 나왔습니다. 이는 두 모델 모두 오버피팅이 의심되는 결과이지만, 각 변수들의 영향력을 비교하기 위해 모델의 결과를 이용하기로 결정하였습니다.



△ 그림 6 랜덤포레스트 결과: 각 변수의 영향력

분석 결과, 가장 영향력이 높은 변수로 기대수명이 확인되었습니다. 그러나, 기대수명과 당뇨병의 관계에 대한 연구는 이미 풍부하기 때문에[3], 새로운 통찰을 얻기 위해 두 번째로 영향력이 높은 변수인 실업률에 주목하게 되었습니다. 이에 따라, 실업률과 당뇨병 유병률 간의 관계를 더 깊이 탐구하고자 합니다.

3. Data and Methods for Regional Analysis in Korea

1) 데이터 수집 및 출처

우리나라의 시·군별 실업률과 당뇨병 진단 경험률 데이터, 그리고 대한민국 행정구역에 대한 시도, 시·군·구에 대한 공간 데이터를 이용하였습니다.

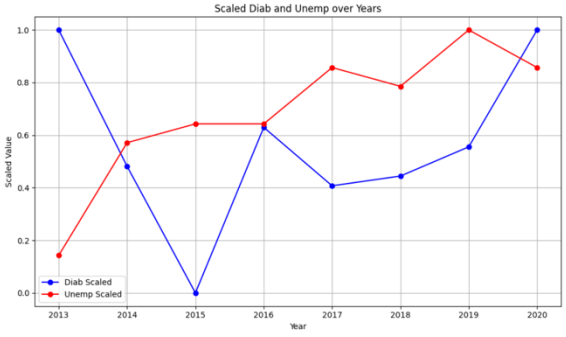
먼저, 우리나라 시·군별 실업률 정보는 통계청의 지역별 고용조사, 경제활동인구 총괄 데이터[4]를 참고했습니다. 이 데이터 셋은 지역고용정책 수립을 위해 시·군·구 단위의 고용현황을 파악하는 기본통계로, 전국 만 15세 이상 가구원을 대상으로 하였으며, 여러 노동 분야 칼럼 중 실업률을 선정하였습니다.

당뇨병 진단 경험률 데이터[5]는 시·군·구 단위 주민 건강수준 및 관련 요인 현황을 파악하기 위해 질병관리청에서 진행한 지역사회건강조사를 통해 수집할 수 있었습니다.

그리고 지역의 위치적 정보를 고려하기 위해 지오서비스웹(GEO-SERVICE-WEB)의 행정구역(SHP) 데이터[6]를 수집하였습니다. 이 데이터는 시·군·구 단위의 행정경계를 포함합니다.

2) 데이터 전처리 및 분석(EDA)

우선, 우리나라 전체의 실업률과 당뇨병 유병률 간의 상관관계를 알아보았습니다. 두 데이터셋이 모두 포함하고 있는 기간인 2013년에서 2022년까지의 연간 값들을 MinMaxScaler로 스케일링한 후 비교해 보았더니 상관관계 -0.29의 결과를 얻을 수 있었습니다.



△ 그림 7 2013~2022년 한국의 실업률과 당뇨병 유병률 추이

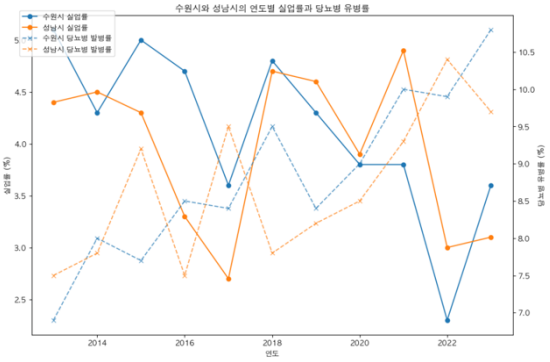
이 자체로는 결과의 유의미성을 알기가 쉽지 않았습니다. 대한민국의 모든 지역들에서 두

[3] Baik, S. H., & Choi, K. M. (2003). Diabetes Mellitus in Elderly Korean. *Korean Diabetes Journal*, *27*(4), 299-303.
[4] 통계청, 「지역별고용조사」, 2023 2/2, 2024.06.26, 시·군·구 경제활동인구 총괄
[5] 질병관리청, 「지역사회건강조사」, 2023, 2024.06.26, 시·군·구별 당뇨병 진단 경험률(30세 이상)
[6] 지오서비스웹. 행정구역(SHP) 데이터. Retrieved from https://www.geo-service-web.com

변수가 서로 같은 관계를 가지고 있다고 판단하기 어려웠기에 더 세부적으로 지역을 나누어 분석을 진행해보았습니다.

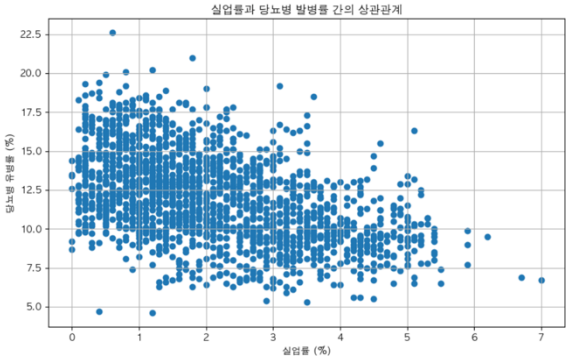
두 변수의 시·군별 데이터를 병합하며 중복되는 지역 총 161 구역들을 조인 키로 이용하였습니다. 변수 값은 모두 표준화된 그 지역 전체 인구 중 해당 인구의 비율이었기에 추가적인 스케일링을 필요로 하지 않았습니다. 결측치도 존재하지 않았습니다. 기간은 2013년부터 2023년을 적용하였습니다.

여러 구역들 중 서울과 가까운 수원시와 성남시를 골라 두 변수의 추이를 비교해보았습니다. 전반적으로 수원시와 성남시 모두 최근 10년간 실업률은 감소하는 추세를, 당뇨병 유병률은 증가하는 추세를 보이고 있었습니다. 그리고 두 지역에서 모두 실업률이 증가하면, 당뇨병 유병률이 감소하고, 그 반대 방향도 보이는 것을 확인할 수 있었습니다.



△ 그림 8 수원시와 성남시의
연도별 실업률과 당뇨병 유병률 추이

두 변수 간의 관계가 전반적인 우리나라 지역들에서 어떻게 나타나는지 확인하기 위해 산점도도 그려보았습니다. 그래프를 통해 음의 상관관계를 유추할 수 있었습니다.



△ 그림 9 지역 실업률과 당뇨병 유병률 상관관계 산점도

시·군 경계 데이터를 정제하는 것에 있어서는, 위치 정보를 나타내는 geometry 칼럼의 데이터들을 Polygon, 혹은 Multipolygon 형식으로 바꾸어 주어야 했습니다. 총 161개의 지역들의 위치 정보들을 수집하여 실업률, 당뇨병 유병률, 그리고 위치 데이터를 하나의 데이터 프레임으로 병합시켰습니다.

4. Methodology for Spatial Regression Analysis

1) 공간회귀분석

한국 내에서는 사회경제적, 그리고 의료적인 차원에서 지역별로 차이가 크다고 할 수 있습니다. 수원시와 성남시는 지리적으로도 산업적으로도 다른 지역들에 비해 상대적으로 비슷함에도 불구하고 위의 그래프에서 보이다시피 비율 값에 차이가 존재했습니다. 그렇기에 실업률과 당뇨병 유병률이 지역에 따라 어떻게 다르며, 인접 지역과는 어떤 영향을 주고 받는지에 대한 심층적인 연구가 필요합니다.

공간 자료는 인접 지역과 서로 영향을 주고 받는 상호의존적 특성인 공간자기상관성을 보이는데, 이러한 경우 공간오차모형 SEM(Spatial Error Model), 공간자기회귀모형 SAR(Spatial Autoregressive Model), 그리고 일반공간모형 SAC(Spatial Autoregressive

Combined model)을 구축하여 분석합니다.[7] 그 중 SAR 모형은 공간적 상관성이 분석 단위의 측정값(종속변수), 즉 당뇨병 유병률에서 발생한다는 것을 가정하기 때문에, 본 프로젝트의 취지에 가장 부합한다고 판단하였습니다.[8] 그리고 공간 인접성을 기준으로 행렬을 구축하였고, 인접행렬을 산출하는 방법으로 Queen's contiguity 방식을 적용하였습니다. 이는 두 지역이 변 또는 모서리를 공유하는 경우로 두 구역이 동-서, 남-북, 양 대각선 방향의 어느 방향으로든지 인접되어 있는 형태를 말합니다.[9] 먼저, 각 지역 위치 데이터를 이용하여 가중행렬을 만들고, 그 가중행렬을 이용하여 공간적 자기상관분석을 실시하는 분석을 진행하였습니다.

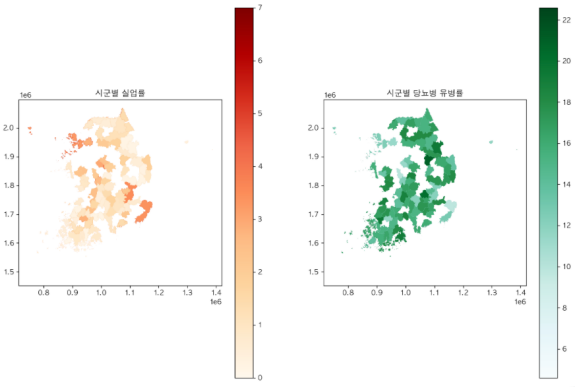
2) 결과 해석 및 논의

```
REGRESSION RESULTS
------------------

SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :Health Data Analysis
Weights matrix      :Queen contiguity weights
Dependent Variable  :     당뇨병 유병률                Number of Observations:        1839
Mean dependent var  :     11.9896              Number of Variables   :           3
S.D. dependent var  :      2.7881              Degrees of Freedom    :        1836
Pseudo R-squared    :      0.2356
Spatial Pseudo R-squared:  0.2307

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       13.58841         0.16992        79.97092         0.00000
              실업률         -0.90281         0.04784       -18.86945         0.00000
      W_당뇨병 유병률          0.00106         0.00032         3.27261         0.00107
------------------------------------------------------------------------------------
Instrumented: W_당뇨병 유병률
Instruments: W_실업률
================================ END OF REPORT =====================================
```

△ 그림 10 SAR 모형 공간회귀분석 결과

독립 변수인 실업률의 계수는 -0.90281로, 이는 실업률이 1% 증가할 때 당뇨병 유병률이 약 0.903% 감소함을 의미합니다. 유의확률이 0에 가까운 숫자로 결과는 매우 유의미하다고 할 수 있습니다. 공간적 종속 변수의 계수는 인접 지역의 당뇨병 유병률이 높을수록 해당 지역의 당뇨병 유병률에 미치는 영향을 나타냅니다. 결과값은 0.00106으로 인접 지역과 해당 지역의 유병률이 양(+)적인 방향으로 상관관계가 있음을 나타내며 유의확률 0.001로 매우 유의합니다.

전체적인 모델 해석을 위해서 Pseudo R-squared 값을 살펴보았을 때, 0.2356으로 모델이 종속 변수의 변동성을 약 23.56% 설명하고 있음을 의미합니다. 또한 공간적 Pseudo R-squared 값은 0.2307로, 공간적 요인을 포함한 모델도 비슷한 수준의 설명력을 가지고 있음을 알려줍니다. 이는 사회과학 데이터 분석에서 비교적 괜찮은 설명력이라고 할 수 있습니다.[10]



△ 그림 11 지역별 실업률과 당뇨병 유병률 지도 시각화

시·군별 지역 위치 정보를 이용하여, 두 변수들의 10년간 평균 값을 지도로 시각화 해보았습니다. 확연히 실업률이 낮은 지역들은(왼쪽 지도의 옅은 부분) 유병률이 높은(오른쪽 지도의 진한 부분) 경향을 보이는 것 또한 확인할 수 있었습니다.

5. Conclusion

1) 연구 요약 및 주요 발견

본 프로젝트는 전 세계 199개 국가의 사회경제적 요인과 당뇨병 유병률 데이터를 분석하

[7] 구, 본 진, & 장, 덕 현. (2021). 공간회귀분석을 이용한 부산지역 공공도서관 접근성 영향 요인 분석. *한국문헌정보학회지*, 55(4), 67-87.
[8] 최유진. (2018). 공간회귀모형을 활용한 사회적경제 규모의 결정요인 분석. *지방정부연구*, 22(2), 455-476.
[9] 정경숙. (2017). 행정구역 통합과 도시공간분포 변화분석: (통합)창원시 사례. *지방정부연구*, 20(4), 345-364.
[10] McFadden, D. (1974) "Conditional logit analysis of qualitative choice behavior." Pp. 105-142 in P. Zarembka (ed.), Frontiers in Econometrics.

여 실업률이 당뇨병 유병률에 미치는 영향을 밝혀냈습니다. 다변량 분석 결과, 기대 수명이 가장 큰 영향을 미치는 요인으로 나타났지만, 실업률도 중요한 영향을 미치는 변수로 확인되었습니다. 이를 바탕으로 한국의 시·군별 데이터를 분석하여 실업률과 당뇨병 유병률 간의 관계를 공간회귀분석을 통해 탐구하였습니다. 분석 결과, 한국의 지역들에서는 실업률이 높을수록 당뇨병 유병률이 낮아지는 경향(계수: -0.90281, 유의확률: 0.000)이 있으며, 인접 지역 간의 당뇨병 유병률도 양의 상관관계(계수: 0.00106, 유의확률: 0.001)를 보였습니다.

2) 정책적 시사점

실업률이 높아질수록 당뇨병 유병률이 감소하는 경향을 통해 실업 상태에서 사람들이 당뇨병 관리에 더 집중할 수 있는 시간적 여유가 있을 수 있음을 시사할 수 있었습니다. 또한 공간적 상관성 분석을 통해, 인접 지역의 당뇨병 유병률이 해당 지역에 영향을 미치는 것이 나타났습니다. 이는 보건 정책 수립 시 인접 지역과의 협력 및 공동 대응이 필요함을 시사합니다.

이러한 결과를 바탕으로 지역 보건 정책 및 경제 정책을 연계하여 종합적인 대책을 마련할 수 있을 것 입니다. 예를 들어, 실업률이 높은 지역에 대해 당뇨병 예방 및 관리 프로그램을 강화하거나, 인접 지역 간의 협력 프로그램을 마련하는 등의 방안을 고려할 수 있습니다.

3) 연구의 한계 및 향후 연구 방향

실업률과 당뇨병 유병률 간의 관계가 보통 양적 상관관계를 나타내는 연구[11]가 있음에도 불구하고, 본 프로젝트에서 완전히 반대의 결과가 나온 것을 비판적으로 바라볼 필요가 있습니다.

먼저, 혼란 변수들이 있을 가능성이 있습니다. 예를 들어, 소득수준, 의료 접근성, 식습관 등은 실업률과 당뇨병 유병률 모두와 상관 관계가 있을 수 있으므로 종속 변수와 독립 변수의 관계가 잘못 해석되었을 수도 있습니다. 그리고, 경제 위기, 지역 보건 정책 등 외부 요인이 특정 기간 동안 변수들에 영향을 미쳤을 가능성도 있기에 분석 결과가 왜곡될 수 있습니다.

향후 연구에서는 외부 요인 및 다른 변수들의 영향을 배재하면서, 실업률과 당뇨병 유병률 간의 관계를 더 정확하게 나타내고자 합니다. 시간에 따른 외부 요인의 영향을 통제하는 패널 데이터 분석을 진행하거나, 혼란 변수들을 추가하여 공간회귀모델을 활용하면 실업률의 순수한 영향을 파악할 수 있을 거라 기대합니다.

[11] Kim, S. (2016). 실업이 건강에 미치는 영향: 고혈압 및 당뇨의 질병 진단을 중심으로. 서울대학교 대학원.

7. References

1) 김응진, 김명환, 김상희, 김동렬, 환정운, 이근식, 전영균, 김영건, & 이정섭. (1970). 한국인 당뇨병의 역학적 연구. *서울 의대 잡지, 11,* 25-30

2) 조남한. (2005). 우리나라 당뇨병의 유병률과 관리 상태. *대한내과학회지, 68*(1), 1-3.

3) Baik, S. H., & Choi, K. M. (2003). Diabetes Mellitus in Elderly Korean. *Korean Diabetes Journal, 27*(4), 299-303.

4) World Bank. 2019. Life Expectancy and Socio-Economic Data. Kaggle. (https://www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank)

5) Meeratif. 2016. Diabetes Specific Data - All Countries. Kaggle. (https://www.kaggle.com/datasets/meeratif/diabetes-specific-data-all-countries)

6) 통계청, 「지역별고용조사」, 2023 2/2, 2024.06.26, 시·군·구 경제활동인구 총괄

7) 질병관리청, 「지역사회건강조사」, 2023, 2024.06.26, 시·군·구별 당뇨병 진단 경험률(30세 이상)

8) 지오서비스웹(GEO-SERVICE-WEB) (https://www.geoservice.co.kr)

9) 구, 본 진, & 장, 덕 현. (2021). 공간회귀분석을 이용한 부산지역 공공도서관 접근성 영향 요인 분석. *한국문헌정보학회지*, 55(4), 67-87.

10) 최유진. (2018). 공간회귀모형을 활용한 사회적경제 규모의 결정요인 분석. *지방정부연구*, 22(2), 455-476.

11) 정경숙. (2017). 행정구역 통합과 도시공간분포 변화분석: (통합)창원시 사례. *지방정부연구*, 20(4), 345-364.

12) McFadden, D. (1974) “Conditional logit analysis of qualitative choice behavior.” Pp. 105-142 in P. Zarembka (ed.), Frontiers in Econometrics.

13) Kim, S. (2016). 실업이 건강에 미치는 영향: 고혈압 및 당뇨의 질병진단을 중심으로. 서울대학교 대학원.