KEI
Working
Paper
2025-05

대기오염물질 간의 인과관계 추정을 위한 인공지능 방법론 활용 현황 분석

Analysis of the Utilization of AI Methodologies for Estimating Causal Relationships between Air Pollutants

김도연

KEI 한국환경연구원
Korea Environment Institute

저 자 김도연

연구진

연구책임자 김도연 (한국환경연구원 전문연구원)
참여연구원 진대용 (한국환경연구원 연구위원)
한지현 (한국환경연구원 부연구위원)

연구자문위원 (가나다 순)

강성원 (한국환경연구원 선임연구위원)
석흥일 (고려대학교 인공지능학과 교수)
송완호 (과학기술정보통신부 기초연구진흥과 과장)
최기철 (한국환경연구원 연구위원)

발행인 김 홍 균
발행처 한국환경연구원
(30147) 세종특별자치시 시청대로 370
세종국책연구단지 B동(과학·인프라동)
전화 044-415-7777 팩스 044-415-7799
www.kei.re.kr
발 행 2025년 5월 31일
등 록 제 2015-000009호 (1998년 1월 30일)
ISBN 979-11-5980-422-9 95530

이 보고서를 인용 및 활용 시 아래와 같이 출처를 표시해 주십시오.
김도연(2025), 「대기오염물질 간의 인과관계 추정을 위한 인공지능 방법론 활용 현황 분석」, 한국환경연구원.

요 약

▌연구의 주요 내용

○ 본 연구는 PM2.5, NO_2, SO_2 등 다양한 대기오염물질 간의 인과관계 추정을 위한 인공지능 기반 인과 분석 기법의 적용 가능성을 모색하고 활용 방안을 제시함

- 특히, 대기오염 데이터의 특수성(시간 지연, 비선형 상호작용, 잠재 변수 존재 등)을 고려한 인과 발견 방법론을 중심으로 검토하여 정리함

○ 대기오염 분야 연구에서 분석 목적에 적합한 최적의 인과 발견 방법론과 세부 활용안을 유형별로 표로 정리하여 제시함

- 인과 발견 방법론은 크게 조건부 독립성 기반, 점수 기반 및 딥러닝 기반 방법론 등을 중심으로 이론적 구조, 통계적 가정, 해석 가능성 등을 바탕으로 비교 분석함

○ 서울시 중구 지역의 대기오염 및 기상 데이터를 수집하여, 시간 지연과 잠재 변수 처리가 가능한 LPCMCI(Latent-PCMCI) 기법을 적용하여 사례 분석을 수행함

- 분석 대상은 대기오염물질과 기온, 풍속, 강수량의 기상 데이터로 전처리 후 시계열 데이터에 대한 인과 분석을 수행하고, 분석 결과를 그래프로 시각화하여 표현함

▌정책 제안

○ 본 연구에서 제안한 상황별 방법론 활용(안) 목록 표는 연구자들이 연구 주제에 적합한 분석 도구를 선택하는 데 도움을 줄 수 있으며, 이는 향후 정책 수립 및 평가 과정에서 과학적 근거로 활용될 수 있음

○ 인공지능 기반 인과 분석 기법으로 대기환경의 복잡한 시공간 인과구조를 정밀하게 분석하면, 준실시간으로 시의성 있는 글로벌 정책을 제안할 수 있음

주제어: 대기오염물질, 인과 분석, 인과 발견, 인공지능, PCMCI

차 례

Ⅰ 서 론

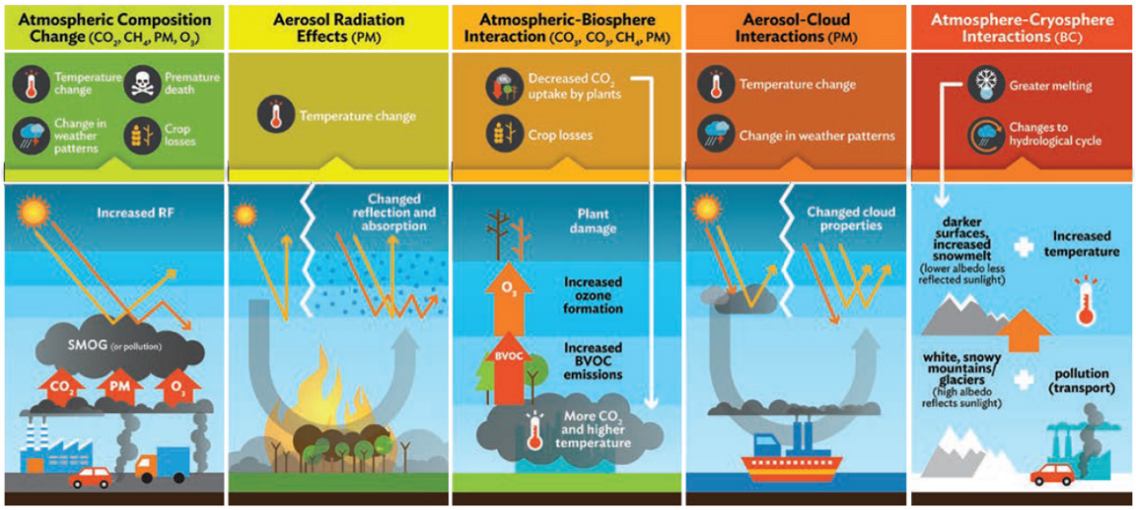
1. 연구 배경 및 필요성

○ 대기오염물질 간의 인과관계를 분석하는 것은 효과적인 대기오염 관리와 정책 수립을 위해 필수적임. 다양한 대기오염물질은 서로 반응하여 새로운 오염물질을 생성하거나 변환되며, 그 영향은 기후변화와도 깊이 연결되어 있음(이승민 외, 2021, pp.86-87; 국립환경과학원, 2015)

- 초미세먼지(PM2.5)는 자동차에서 배출된 질소산화물(NOx) 등과 같은 대기오염물질에서 변환됨(Zhang et al., 2013, pp.7068-7070). 따라서 대기오염 제어의 첫 번째 단계는 이러한 대기오염물질 간의 인과관계를 규명하는 것임(Wang et al., 2016, pp.1031-1032)
- 화학 질량 균형법(CMB: Chemical Mass Balance) 기법이 활용되었으며, 이 방법은 특정 대기오염물질을 수집하고 필터링 후 재구성된 화학 질량과 일치시키는 기법임(Rees et al., 2004, p.3305). 그러나 이러한 방법은 실험 장비의 한계와 높은 비용 문제로 연구 수행에 제약이 있음(Handhayani, 2023, p.1). 이에 따라 비용 및 시간 측면에서 효율적인 대안으로 인공지능(AI: Artificial Intelligence) 기법이 주목받고 있음

○ 대기오염과 기후변화 문제는 서로 밀접하게 연결되어 있으며, 일부의 대기오염물질은 전 지구적 또는 지역적 기후 시스템에 영향을 미침

- 아래 〈그림 1-1〉과 같이 1) 대기 조성 변화, 2) 에어로졸 복사 효과, 3) 대기-생물권 상호작용, 4) 에어로졸-구름 상호작용, 5) 대기-빙권 상호작용 등을 통해 기후변화와 대기오염이 복합적으로 서로 상호작용함



주: BVOC=Biogenic Volatile Organic Compounds, RF=RadioFrequency.
자료: Asian Development Bank(2022.3), p.8, 〈그림 7〉을 발췌하여 저자 작성.

〈그림 1-1〉 대기오염과 기후변화의 상호작용

○ 최근 대기오염과 기후변화는 예측이 어려운 극단적 형태로 변하고 있으며, 이를 해결하기 위해 기존 통계 및 물리 모델을 넘어 인과관계 추정을 통한 정확한 예측의 필요성이 증대되고 있음(Runge et al., 2019a, pp.8-9; Runge et al., 2023, pp.488-489)

- 기존의 대기오염 관리 모델은 주로 오염물질의 물리적·화학적 이동과 반응을 추적하고 예측하는 데 집중되어 있어, 오염물질 확산과 농도 변화는 시뮬레이션할 수 있지만 변수 간의 인과관계를 직접 추정하기에는 한계가 있음(Masson-Delmotte et al., 2021, pp.27-28)
- 특히, 시간 지연(time lag)과 비선형적 변수 간 상호작용을 충분히 반영하지 못해 대기오염의 주요 원인과 결과 간의 복잡한 인과관계를 정밀하게 파악하기 어려움(Runge et al., 2019, p.8)
- 대기오염 문제의 정확한 인과관계를 이해하는 것이 정책 수립 및 대응 전략 마련에 필수적이지만, 기존 모델은 이러한 인과 구조를 포착하기에 부족하며, 이는 복잡한 대기오염 문제를 해결하는 데 제약으로 작용함(Zhang et al., 2018, pp.404-405)

○ 최근 AI를 활용한 인과관계 추정 기법이 활발히 연구되고 있지만, 대기오염 분야 인과 분석 연구에서 실제로 인공지능 방법론이 어떻게 활용될 수 있는지에 대한 내용을 체계적으로 정리한 자료가 미흡함

- AI 기반 인과 추정 기법은 기존 상관분석의 한계(예: 인과 방향성 구분 어려움, 숨은 변수

고려하지 못함 등)를 개선하고 보다 정확한 인과관계를 계산할 수 있으나, 다양한 분석 목적에 맞는 목적별 적합한 방법론 개발이 필요함

○ 대기오염 관리 측면에서 AI 기반 인과 추정은 대기오염의 근본적인 원인을 찾고, 지역 및 시기별 맞춤형 대응을 위한 정책 개발에 활용 가능함

- 대기오염 변수 간 시간적 인과관계를 추정하거나, 지역 간 대기오염 이동을 분석하는 시공간적 인과관계 분석 등 다양한 분석 목적에 따라 최적화된 기법이 요구됨
- 특히, 시간 지연을 식별하는 것은 인과관계 발견의 중요한 요소로, 원인과 결과 사이의 시간 차이를 파악하여 인과관계의 동역학을 이해하고 시계열 예측의 정확도를 높이는 데 필수적임(Runge et al., 2019b, p.2)
- 기존 연구들은 상관분석과 시계열 검증을 통해 변수 간 최적의 시간 지연과 인과관계를 평가해왔음(Spirtes et al., 2000, pp.1-2)

○ AI 기반 인과관계 추정은 기상 및 대기오염 데이터에서 시간 지연과 비선형 상호작용을 반영하여 변수 간의 복잡한 인과관계를 정밀하게 분석할 수 있는 유용한 도구임

- 대표적인 인과관계 방법론 그레인저 인과(granger causality)는 변수 간 선형관계를 가정하고 있으며, 비선형적인 상관관계 계산과 다변량 변수 분석이 어려운 문제가 있음
- 또한 시계열 패턴이 유사한 변수 간 동시성 문제로 불확실한 인과관계 값을 계산할 수 있으나, 직접적 또는 간접적 인과관계에 대한 구분이 어려운 한계가 있음

○ 이러한 한계를 개선하기 위해 정보 이론 기반의 인과관계 추정 방법과 수리 모델 기반의 방법론이 개발됨. 하지만 기존 대부분의 수리 모델들은 복잡한 연산으로 실제 환경 문제에 활용하기에는 어려운 문제가 존재함

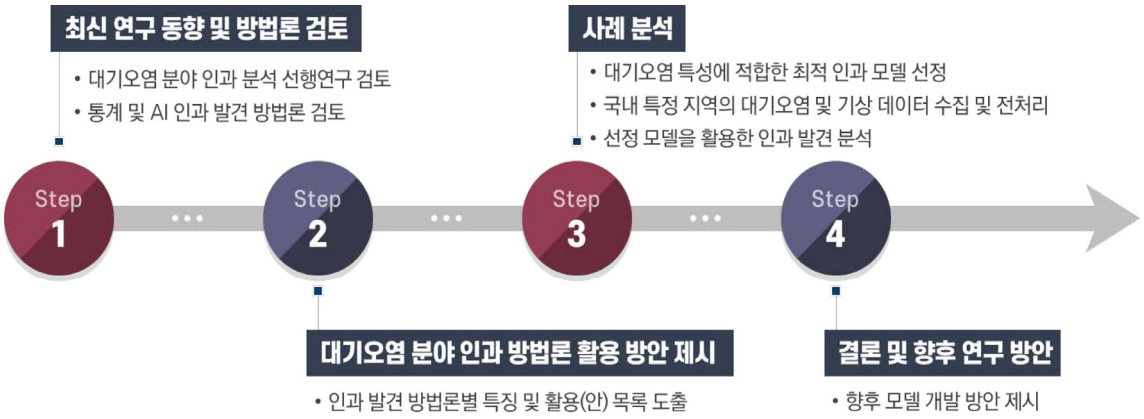
- 대기오염과 같은 복합적인 원인을 기반으로 한 문제는 인과 구조를 정밀하게 이해하는 데 여전히 한계가 존재함

○ 최근에는 시간 지연과 비선형 상호작용을 효율적으로 계산할 수 있는 인공지능 기반 인과관계 추정 기법이 활발히 연구되고 있음. 이는 대기오염 데이터의 복잡한 상호작용 관계를 보다 명확하게 분석할 수 있을 것으로 기대됨

2. 연구 목적

○ 본 연구는 대기오염물질 간의 인과 분석을 위한 통계 및 인공지능 기반 방법론들의 특징 및 활용 방안을 목록화하여 정리하고, 기존 방법론의 한계를 개선하는 방법을 제시하여 환경 정책 분야에서 인공지능 방법론의 활용성을 높이고자 함

○ 선행연구 검토 분석을 통해 대기오염물질 데이터의 특수성(시간 지연, 비선형적 상호작용 등)을 고려한 가장 효과적인 방법론을 선정하여 사례 분석을 수행하고, 활용 가능성 및 한계점 등을 제시함



자료: 저자 작성.

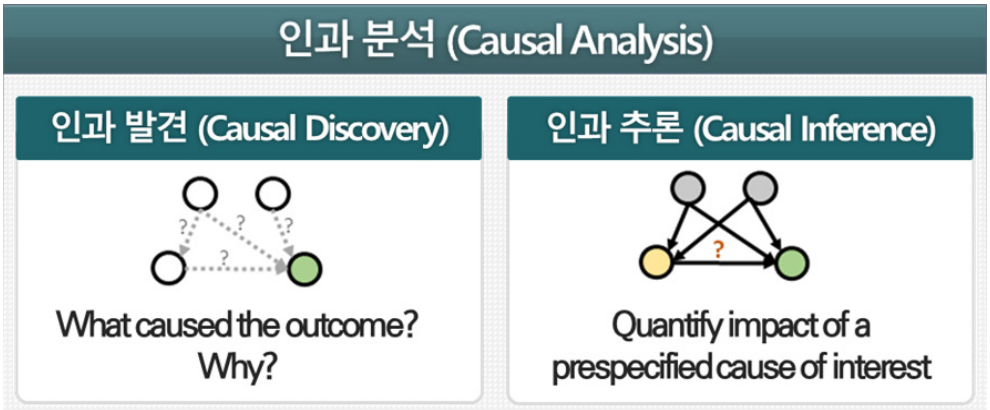
〈그림 1-2〉 연구 범위 및 추진 방법

Ⅱ 인과 분석 방법론 및 연구 동향

1. 인과 분석(causal analysis)

○ 인과 분석은 데이터에서 원인(cause)과 결과(effect)를 발견하는 것을 목표로 하며, 인공지능 방법론의 한계를 해결할 수 있는 중요한 도구로 간주되고 있음(Pearl, 2018, p.7)

- 인과성 연구(causality study)는 다학제적인 역사를 가지고 있어 연구가 분절적으로 이루어져 왔음. 인과성에 대한 연구는 역학(epidemiology), 경제학(economics), 통계학(statistics), 컴퓨터 과학(computer science) 등 다양한 분야에서 이루어져 왔음(Nogueira et al., 2022, p.2)
- 인과 분석은 인과 발견(causal discovery)과 인과 추론(causal inference) 두 가지 주요 과업으로 구분할 수 있음(그림 2-1 참조). 이 두 가지 과업은 서로 반대되는 접근법을 취함
  - 인과 발견은 사전에 변수 간 관계를 가정하지 않으며, 데이터를 기반으로 직접 인과 구조를 학습함
  - 인과 추론은 변수 간 관계가 존재한다고 가정한 후, 실제로 그 관계가 유효한지 검증하고 정량적으로 분석함



자료: Nogueira et al.(2022), pp.6-23을 참조하여 저자 작성.

〈그림 2-1〉 인과 분석 개요

- 일반적으로 인과 모델(causal models)은 특정 시스템 또는 모집단 내의 인과관계를 수학적으로 표현하는 모델을 의미함(Hitchcock, 2020). 인과관계는 변수 간의 확률적 (비)독립성, 개입(intervention)의 효과 또는 반사실적 주장(counterfactual claims)과 같은 개념을 포함함

○ 인과 발견 알고리즘은 구성 방식에 따라 제약 기반(constraint-based) 알고리즘과 점수 기반(score-based) 알고리즘으로 분류될 수 있음. 이러한 분류는 일반적으로 베이지안(Bayesian) 계열 방법에 적용되지만, 유사한 구조를 가진 다른 방법에도 확장하여 적용할 수 있음

- 제약 기반 알고리즘은 독립성 검정(independence tests)을 활용하여 그래프의 에지(edge) 제약 조건을 식별함
  - 예를 들어, G2 검정(Spirtes et al., 2000, pp.1-2)을 사용하여 관찰 데이터에서 변수 간 독립성을 평가함
  - 이후 추가적인 규칙을 적용하여 방향성을 결정하나, 예외적으로 방향성을 지정하지 않고 무방향 그래프(undirected graph)를 생성하는 경우도 있음. 이러한 그래프는 특정 노드의 관계만을 나타내는 국소적(local) 그래프인 경우가 많음
- 점수 기반 알고리즘은 후보 그래프에 대해 베이지안 정보 기준(BIC: Bayesian Information Criterion)과 같은 평가 척도를 사용하여 점수를 할당함
- 그러나 가능한 모든 그래프를 평가해야 하기 때문에 계산 비용이 매우 높음. 이에 따라 탐색 공간을 줄이기 위해 탐욕적 휴리스틱(greedy heuristics) 기법이 적용됨

○ 인과 추론 알고리즘은 특정 변수가 관심 있는 결과에 미치는 인과적 효과(causal effect)를 추정하고 정량화하는 것을 목표로 함. 이는 단순한 상관관계 분석을 넘어, 변수 간의 실제 인과적 관계를 파악하고 특정 개입(intervention)이 결과에 미치는 영향을 정량적으로 평가하는 과정을 포함함

- 분석 맥락에 따라 인과 효과는 서로 다른 평가 지표(metrics)로 정량화될 수 있으며, 각 지표는 분석 수준에 따라 초점을 달리함
  - 예를 들어, 개별 수준(individual-level)에서의 인과 효과를 평가하는 방법과, 전체 모집단(population-level)에서의 평균적인 효과를 측정하는 방법이 존재함
- 대표적인 인과 효과 지표는 평균 처리 효과(ATE: Average Treatment Effect), 조건부 평균 처리 효과(CATE: Conditional Average Treatment Effect), 개별 처리 효과(ITE:

Individual Treatment Effect) 등이 있으며, 데이터 특징 및 연구 목적에 따라 적절한 지표를 선택하여 활용함

- 인과 추론은 관찰 데이터와 실험 데이터 모두에서 수행될 수 있으며, 적용하는 방법론에 따라 비혼란성(unconfoundedness) 가정을 적용하는지 여부가 달라질 수 있음(Nogueira et al., 2022, p.23)

○ 본 연구는 변수 간의 인과관계를 계산하는 인과 발견(causal discovery) 방법론을 중심으로 활용 현황을 조사하여 정리하였음

2. 인과 발견(causal discovery)

가. 조건 기반 접근법(constraint-based approach)

○ 조건 기반 접근법은 모든 데이터가 구조적 인과 모델(SCM: Structural Causal Model)에 의해 생성된다고 가정함

- 이러한 가정하에서, 인과 그래프의 연결 구조는 관찰 데이터의 분포를 형성하며, 변수 간의 주변 및 조건부 독립 또는 종속 관계를 결정함(Geiger et al., 1990, p.507). 이 특성은 인과 마르코프 조건(causal Markov condition)으로 알려져 있음(Spirtes et al., 2000, pp.1-2)
- 조건 기반 인과 발견 방법론은 식(2-1)과 같은 독립성 검정을 순차적으로 수행함

$$H_0 : X \perp\!\!\!\perp Y | Z \quad vs. \quad H_1 : X \not\perp\!\!\!\perp Y | Z \qquad \text{식(2-1)}$$

- 여기서 $\perp\!\!\!\perp$는 독립성을 뜻하고, 모든 변수 쌍(X, Y)과 선택된 변수 집합 Z에 대해 검정을 수행한 후 독립성 검정 결과를 바탕으로 인과 그래프 계산에 활용함
- 이러한 접근법은 인과 마르코프 조건을 활용하며, 이 가정들은 그래프 내 d-분리(d-separation)(Pearl, 1988)와 데이터 분포의 조건부 독립성을 연결함

$$X d\text{-}sep\ Y | Z \ \Leftrightarrow\ H_1 : X \perp\!\!\!\perp Y | Z \qquad \text{식(2-2)}$$

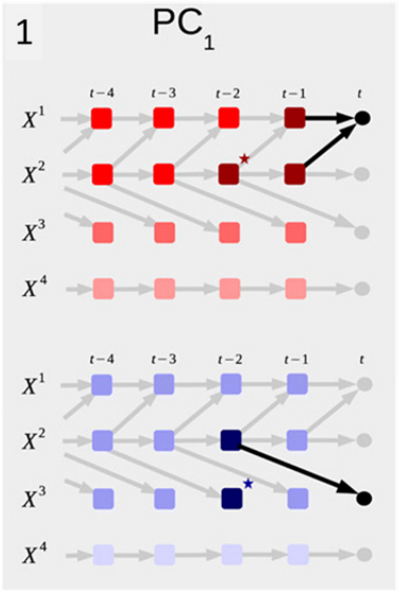
- 두 노드는 조건 집합 Z가 주어졌을 때 모든 경로가 차단되면 d-분리됨(Pearl, 1988). 구체적으로, 조건 집합 Z가 주어졌을 때 두 노드 간 경로는 다음 조건 중 하나를 만족하면 차단됨:
  - 경로에 Z에 속한 비충돌자(non-collider, 예: →•→ or ←•→)가 포함되거나, Z 또는 Z의 조상에 속하지 않는 충돌자(collider, 예: →•←)가 포함됨

○ 계산의 효율성을 높이기 위해 대부분의 방법은 식(2-2)에서 조건 집합 Z를 적응적이고 효율적인 방식으로 선택함

- 대표적인 예로 PC 알고리즘이 있음. PC 알고리즘은 SCM과 인과 충실성 가정 외에도 비순환성과 관찰되지 않은 변수가 없다는 가정, 즉 인과 충분성을 추가로 가정함(Spirtes and Glymour, 1991, pp.62-72)
- 반면, FCI 알고리즘(Zhang, 2008, pp.1873-1896)은 인과 충분성 가정을 완화하여 관찰되지 않은 변수를 허용함. 일부 방법은 비순환성 가정을 요구하지 않음(Bongers et al., 2021, pp.2885-2915)

○ PCMCI(Peter and Clark Momentary Conditional Independence) 검정(Runge et al., 2019a, pp.8-9)과 그 확장 버전인 PCMCI+(Runge, Peters, and Sontag, 2020, pp.1388-1397), Latent-PCMCI(Gerhardus et al., 2020, pp.12615-12625)는 PC와 FCI를 시계열 데이터에 맞게 조정한 방법으로, 시계열 데이터의 특유한 분석 과제를 해결하도록 설계되었음

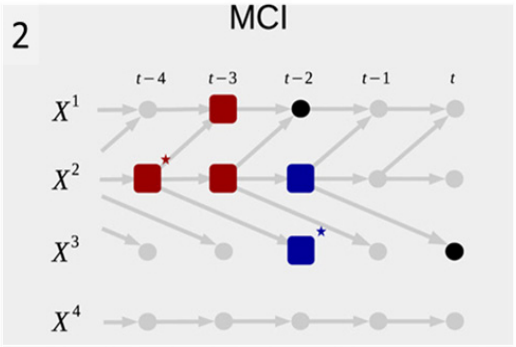
- PCMCI는 조건부 독립을 기반으로 하는 방법으로 시간차(time lag)별로 링크의 존재 유무를 감지하며, 크게 PC(Peter and Clark) 알고리즘과 검정을 수행하는 MCI(Momentary conditional independence) 두 가지 방법으로 구성됨(그림 2-2 참조)
- PC 알고리즘은 n개의 변수 각각에 대해 관련 없는 조건들을 반복적인 독립성 검정으로 제거함(방법: PC- stable 알고리즘에 기반한 Markov 집합 탐색 알고리즘)



Algorithm S1. Pseudo-code for condition-selection algorithm $\mathrm{PC}_{q_{\max}}$ to estimate parents of X_t^j. We use this algorithm as a pre-selection stage in PCMCI with $p_{\max} = N\tau_{\max}$ (i.e., no restriction on the maximum number of parents) and $q_{\max} = 1$. For the standalone PC-stable algorithm, we set $q_{\max}$ to a large value of 10.

Require: Time series dataset $\mathbf{X} = (X^1, X^2, \ldots, X^N)$, selected variable X^j, maximum time lag $\tau_{\max}$, significance threshold α_{PC}, maximum condition dimension $p_{\max}$ (default $p_{\max} = N\tau_{\max}$), maximum number of combinations $q_{\max}$ (default $q_{\max} = 1$), conditional independence test function

function CI(X, Y, $\mathbf{Z}$)
  Test $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ using test statistic measure I
  return p-value, test statistic value I
Initialize preliminary set of parents $\widehat{\mathcal{P}}(X_t^j) = \{X_{t-\tau}^i : i \in \{1, \ldots, N\},\ \tau \in \{1, \ldots, \tau_{\max}\}\}$
Initialize dictionary of test statistic values $I^{\min}(X_{t-\tau}^i \to X_t^j) = \infty\ \ \forall\ X_{t-\tau}^i \in \widehat{\mathcal{P}}(X_t^j)$
for $p = 0, \ldots, p_{\max}$ **do**
  if $|\widehat{\mathcal{P}}(X_t^j)| - 1 < p$ **then**
    Break for-loop
  for all $X_{t-\tau}^i$ in $\widehat{\mathcal{P}}(X_t^j)$ **do**
    $q = -1$
    for all lexicographically chosen subsets $\mathcal{S} \subseteq \widehat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}$ with $|\mathcal{S}| = p$ **do**
      $q = q + 1$
      if $q \geq q_{\max}$ **then**
        Break from inner for-loop
      Run CI test to obtain (p-value, I) $\leftarrow$ CI($X_{t-\tau}^i$, X_t^j, $\mathcal{S}$)
      if $|I| < I^{\min}(X_{t-\tau}^i \to X_t^j)$ **then** ▷ Store min. I of parent among all tests
        $I^{\min}(X_{t-\tau}^i \to X_t^j) = |I|$
      if p-value $> \alpha_{\mathrm{PC}}$ **then** ▷ Removed only after all $X_{t-\tau}^i$ have been tested
        Mark $X_{t-\tau}^i$ for removal from $\widehat{\mathcal{P}}(X_t^j)$
        Break from inner for-loop
  Remove non-significant parents from $\widehat{\mathcal{P}}(X_t^j)$
  Sort parents in $\widehat{\mathcal{P}}(X_t^j)$ by $I^{\min}(X_{t-\tau}^i \to X_t^j)$ from largest to smallest
return $\widehat{\mathcal{P}}(X_t^j)$



Algorithm S2. Pseudo-code for MCI causal discovery stage. Here we state the algorithm for $\tau \geq 0$, then causal links for $\tau = 0$ correspond to contemporaneous links, which are left undirected here.

Require: Time series dataset $\mathbf{X} = (X^1, X^2, \ldots, X^N)$, sorted parents $\widehat{\mathcal{P}}(X_t^j)$ for all variables X^j estimated with Algorithm S1, maximum time lag $\tau_{\max}$, maximum number p_X of parents of variable X^i, and conditional independence test function CI

for all $(X_{t-\tau}^i, X_t^j)$ with $i, j \in \{1, \ldots, N\},\ \tau \in \{0, \ldots, \tau_{\max}\}$, excluding (X_t^j, X_t^j) **do**
  Remove $X_{t-\tau}^i$ from $\widehat{\mathcal{P}}(X_t^j)$ if necessary
  Define $\widehat{\mathcal{P}}_{p_X}(X_{t-\tau}^i)$ as the first p_X parents from $\widehat{\mathcal{P}}(X_t^i)$, shifted by τ
  Run MCI test to obtain (p-value, I) $\leftarrow$ CI($X_{t-\tau}^i$, X_t^j, $\mathbf{Z} = \{\widehat{\mathcal{P}}(X_t^j), \widehat{\mathcal{P}}_{p_X}(X_{t-\tau}^i)\}$)
Optionally adjust p-values of all links by False Discovery Rate-approach (FDR)
return p-values or q-values (for FDR-adjusted tests) and MCI test statistic values

주: PCMCI 모델은 (1) PC Condition-selection 단계와 (2) MCI 검정 단계로 구성.
자료: Runge, J. et al.(2019b), p5, 〈그림 3〉을 수정하여 저자 작성.

〈그림 2-2〉 PCMCI의 구성

- 〈그림 2-2〉의 1. PC 알고리즘에서 변수 X_1과 X_3에 대한 pc condition selection 알고리즘 그림을 보면, 먼저 가장 밝은 색깔에 해당하는 상관관계가 없는 변수가 제거됨. 이후 더 이상 테스트할 조건(condition)이 없을 때까지 변수가 제거되고, 최종적으로 조건부 독립성을 식별하는 데 필요한 parents(〈그림 2-2〉에서 가장 진한 박스)만 남게 됨. 하지만 false positive(거짓을 참으로 잘못 판단)를 포함하는 문제가 있음

- MCI 검정에서 상호의존성이 높은 시계열에 대한 false positive 문제를 해결할 수 있음
- PC에서 추출한 저차원의 condition들은 MCI 조건부 독립성 검정에 사용됨. 〈그림 2-2〉 2.MCI 검정에서 파란색 박스는 조건부 독립성을 확립하기에 충분하며, 빨간색 박스는 자기 상관을 설명하고 MCI의 인과강도의 추정치를 계산함
  - 즉, PC condition-selection 알고리즘은 PCMCI에서 pre-selection 단계이며 MCI causal discovery stage에서는 PC에서 계산된 parents를 조건으로 사용하고, 모든 변수 쌍을 대상으로 테스트함

○ 조건 기반 접근법은 유연한 구조를 가지고 있어, 조건부 독립성을 검정하는 다양한 방법을 상황에 맞게 선택하여 사용할 수 있음

- 예를 들어, 선형 가우시안 데이터에는 부분 상관 기반 검정을, 일반적인 함수 관계에는 조건부 상호 정보 기반 검정을 사용할 수 있음. 이러한 유연성 덕분에 조건 기반 접근법은 비모수적 특성을 가짐(Runge, Storkey, and Perez-Cruz, 2018, pp.938-947)

○ (활용 사례) 조건 기반 접근법은 대기 경로 재구성 및 기후 모델 평가 등 지구과학 분야의 다양한 문제에 적용됨

- 대기 경로 재구성: 대기 변수 간 인과관계를 분석하여 폭풍이나 대기 순환 경로를 예측하는 데 활용(Ebert-Uphoff and Deng, 2012, pp.5648-5665; Runge, Petoukhov, and Kurths, 2014, pp.720-739; Kretschmer et al., 2016, pp.4069-4081)
- 기후 모델 평가: 기후 모델의 예측력을 검증하고 개선하는 데 사용됨(Nowack et al., 2020, pp.1-11)
- 원격 상관(teleconnection) 분석: 북대서양 진동(NAO: North Atlantic Oscillation)과 동아시아 몬순 간의 원격 상관관계를 분석하여 기후 예측을 향상시킴(Di Capua et al., 2020, pp.17-34)

○ Granger 인과관계(Granger, 1969)와 정보이론적 확장(Information-Theoretic Extension)(Schreiber, 2000)은 독립성 관계를 활용하는 별도의 방법론으로, 지구과학에서도 일부 활용됨

- Granger 인과관계: 시계열 데이터에서 한 변수가 다른 변수의 미래 값을 예측하는 데 필요한지 평가하는 방법
  - 해양 온도의 변화가 대기 중 이산화탄소 농도에 미치는 영향을 분석할 수 있음

- 정보이론적 확장: 전이 엔트로피(transfer entropy)와 같은 정보이론 지표를 활용하여 변수 간 정보 흐름을 분석하는 방법
  - 비선형 관계를 다룰 수 있어, 복잡한 기후 현상의 인과관계를 분석하는 데 유용함
  - 이러한 방법들은 해양-대기 상호작용, ENSO와 같은 기후 현상의 인과관계를 분석하는 데 지구과학에서 적용됨(Triacca, 2005, pp.133-135; McGraw et al., 2018, pp.3289-3300)

나. 점수 기반 접근법(score-based approach)

○ 점수 기반 접근법은 점수 함수 S를 정의하여 주어진 데이터셋 D와 가능한 모든 인과 그래프 G에 대해 점수 $S(D, G)$를 계산함

- 점수 함수는 데이터가 주어진 그래프 구조에 얼마나 잘 맞는지를 평가하는 기준으로, 일반적으로 복잡도를 페널티로 부과한 로그 우도(complexity-penalized log-likelihood)를 활용함(Peters, Janzing, and Scholkopf, 2017)
  - 예를 들어, 특정 모수적 통계 모델(예: 가우시안 분포의 선형 모델)을 가정할 경우, 로그 우도는 데이터의 적합도를 측정하고, 복잡도 패널티는 그래프의 복잡성을 제어하여 과적합을 방지함. 이후 가장 높은 점수를 받은 그래프를 선택하여 인과 구조를 추론

○ 점수 기반 인과 발견에서는 방향성 비순환 그래프(DAG: Directed Acyclic Graph)의 수가 변수 개수에 따라 초지수적으로 증가하는 문제가 있음

- 예를 들어, 변수가 5개일 경우 가능한 DAG의 수는 수백 개를 넘고, 변수가 10개로 늘어나면 그 수는 수백만 개로 급증함
- 이러한 계산 복잡성으로 인해 최적의 점수 그래프를 찾는 과정은 대부분 탐욕적(greedy) 탐색 방식으로 진행됨
  - 탐욕적 탐색은 현재 그래프에서 점수를 가장 많이 향상시킬 수 있는 방향으로 그래프를 조금씩 수정해가며 최적의 그래프를 찾아가는 방식임

○ 가정한 통계 모델이 식별 가능성(identifiability)을 보장하지 않는 경우, 즉 동일한 데이터 분포를 생성하는 서로 다른 그래프가 존재할 경우, 탐색 과정을 마르코프 동치 클래스(markov equivalence class) 단위로 재구성함

- 마르코프 동치 클래스는 동일한 독립성 관계를 가지는 그래프들의 집합을 의미
- 이 접근법의 대표적인 알고리즘으로 GES(Greedy Equivalence Search)가 있음

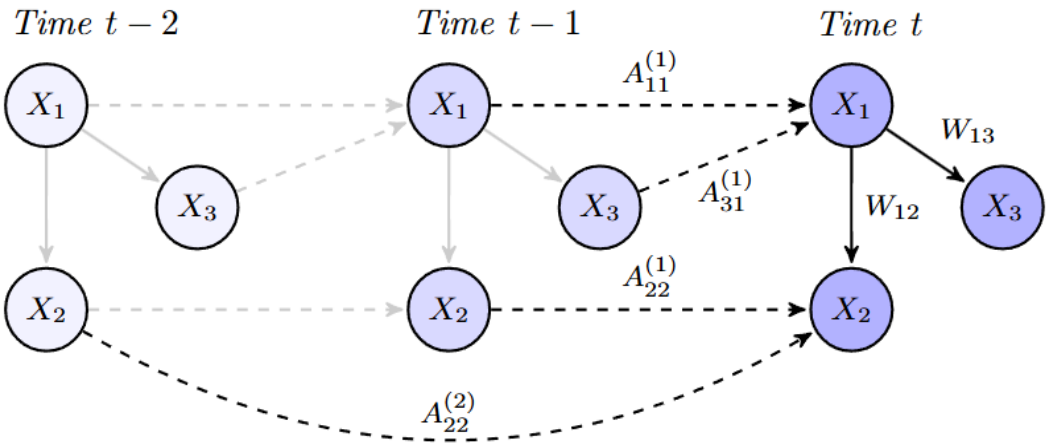
(Chickering, 2002). GES는 다음과 같이 두 단계로 동작함:

- forward phase: 빈 그래프에서 시작하여 점수가 증가하는 방향으로 간선(edge) 추가
- backward phase: 불필요한 간선을 제거하여 점수를 최적화

- 이러한 과정을 통해 GES는 효율적으로 최적의 마르코프 동치 클래스를 찾아냄
- 또한 GES를 기반으로 한 고효율 하이브리드 방법(Tsamardinos, Brown, and Aliferis, 2006)도 개발되어 계산 성능을 더욱 개선하고 있음

○ 최근 점수 기반 인과 발견은 연속 최적화 문제(continuous optimization problem)로 재구성되는 방식이 제안됨(Zheng et al., 2018; Lorch et al., 2021, pp.24111-24123)

- 이 방법은 이산적인 그래프 수에 대한 비효율적인 탐색을 피하기 위해, 그래프 구조를 연속적인 파라미터 공간에서 최적화하는 접근법임
  - 예를 들어, 그래프의 인접 행렬(adjacency matrix)을 연속 변수로 표현하고, 최적화 기법을 사용
- 그러나 이 방식은 개입 데이터(interventional data)의 가용성이나 동일 분산 잡음(equal-variance noise)과 같은 추가적인 가정이 필요함
- 시계열 데이터에 적용되는 방법으로 DYNOTEARS(Pamfil et al., 2020)가 대표적이며, 이는 시계열 데이터의 동적 구조를 고려하여 인과관계를 추론하지만, 아직 초기 단계에 있음
- DYNOTEARS는 M개의 독립적인 정상 시계열 데이터를 분석하며, 각 시계열 데이터는 $\{x_{m,t}\}_{t\in\{0,\ldots,T\}}$로 표현되며, $x_{m,t}$는 d개의 변수로 구성된 벡터로, 이는 d차원의 실수 벡터 공간에 속함
  - 이 모델은 변수들이 동시적 및 시간 지연 방식으로 서로 영향을 준다고 가정함. 이러한 관계를 각각 동일 시간 내 의존성(intra-slice dependency)과 시간 간 의존성(inter-slice dependency)으로 구분됨(그림 2-3 참조)



주: 1) 3개의 노드와 자기회귀 차수 $p=2$를 가정하며, 동일 시간 내 의존성은 실선, 시간 간 의존성은 점선으로 표시.
2) $Time\ t$에서의 변수에 영향을 주지 않는 간선은 가독성을 높이기 위해 연한 색으로 표시.
자료: Pamfil et al. (2020), p.1596, 〈그림 1〉을 발췌하여 저자 작성.

〈그림 2-3〉 동일 시간 내 의존성과 시간 간 의존성

- DYNOTEARS는 표준 SVAR(Structural Vector Autoregressive) 모델을 기반으로 데이터를 모델링함(Demiralp and Hoover, 2003, pp.745-767). SVAR 모델은 다음과 같이 정의됨:

$$x_{m,t}^{\top} = x_{m,t}^{\top} \quad W + x_{m,t}^{\top} \qquad \text{식(2-3)}$$

  • 여기서 W는 동시적 및 시간 지연 관계를 나타내는 가중치 행렬이며, DYNOTEARS는 연속 최적화 기법을 활용하여 시계열 데이터의 동적 인과 구조를 효과적으로 추론함

○ (활용 사례) 점수 기반 접근법은 지구과학 분야에서도 다양한 연구에 적용됨

- 기후 및 대기 변수 간의 인과관계 분석: 온실가스 농도, 해양 온도, 대기 순환과 같은 변수들 간의 인과 구조를 추론하여 기후 시스템의 동역학을 이해하는 데 기여(Liu and Niyogi, 2020, pp.1-7)
- 지구 시스템 모델 구조 평가: 모델이 실제 데이터와 얼마나 일치하는지를 점수 기반으로 검증함(Mäkelä et al., 2022, pp.2095-2099)
- 이러한 연구들은 기후 및 대기변화의 장기적 영향을 예측하거나, 모델의 신뢰도를 높이는 데 중요한 역할을 함

다. 딥러닝 기반 접근법(deep learning-based approach)

○ 딥러닝 기반 인과 발견 기법은 입력 변수와 출력 변수 간 비선형적이고 복잡한 관계를 효과적으로 학습할 수 있음

- 특히, 비선형 Granger 인과성(Nonlinear Granger Causality)을 심층신경망을 활용하여 분석하는 방법론들이 최근 활발히 연구되고 있음(Gong et al., 2024, p.17). 이러한 방법론들은 복잡한 시계열 데이터에서 유용하게 적용됨
- 시계열 길이가 T인 시계열 데이터 $X=\left(x_1^{1:T},...,x_d^{1:T}\right)$일 때, 비선형 함수 g_j에 대해 $x_j^{t+1}=g_j\left(x_1^{1:t},...,x_d^{1:t}\right)+u_j^{t+1}$로 계산되고, 여기서 u_j^{t+1}는 독립적인 노이즈를 의미함
  만약 모든 $\left(x_1^{1:t},...,x_d^{1:t}\right)$에 대해,

$$g_j\left(x_1^{1:t},...,x_i^{1:t},...,x_d^{1:t}\right)=g_j\left(x_1^{1:t},...,x'^{1:t}_i,...,x_d^{1:t}\right) \quad \text{식(2-4)}$$

  식(2-4)가 성립되고 모든 $x_i^{1:t}\neq x'^{1:t}_i$가 만족되면, 시계열 x_i는 시계열 x_j에 대해 Granger 비인과적(non-causal) 임

○ Neural Granger Causality(NGC)는 비선형 Granger 인과성 추정을 위해 구성요소별 MLP(cMLP)와 구성요소별 LSTM(cLSTM)을 활용한 방법으로, 입력층 가중치의 희소성을 통해 변수 간 Granger 인과관계를 자동으로 학습함(Tank et al., 2021, pp.4267-4279)

- 개별 MLP를 사용하여 각 변수의 영향을 구별하며, 첫 번째 층의 계산식은 다음과 같음:

$$h_1^t=\sigma\left(\sum_{k=1}^{K}W_1^k x^{t-k}+b_1\right) \quad \text{식(2-5)}$$

- x^{t-k}는 시간 $t-k$에서 입력값, b_1는 bias, $\sigma(\bullet)$은 활성화 함수를 의미함. 가중치 행렬 W_1^k의 특정 칼럼이 모든 시점 t에 대해 0이면, 해당 변수는 Granger 비인과적임
- 학습 과정은 데이터 피팅 항(data-fitting term)과 희소성 유도 항(sparse-inducing term)으로 구성되며, 최적화 목적 함수는 다음과 같이 정의됨:

$$\min W\sum_{t=\tau}^{T}(x_j^t - g_j(x_{(t-1):(t-\tau)})) + \lambda\sum_{i=1}^{d}R((W_1)_{:i}) \quad \text{식(2-6)}$$

- 첫 번째 항은 예측값과 실젯값의 차이를 최소화하는 데이터 피팅 항을 의미하며, 두 번째 항은 희소성을 유도하기 위한 희소성 유도 항을 의미함. $R(\bullet)$는 그룹 라쏘 정규화를 사용하여 특정 지연 값을 명시적으로 지정하지 않고도 인과관계를 추출할 수 있도록 함
- cLSTM은 시간 지연 선택 문제를 회피하며, LSTM의 입력 가중치를 통해 Granger 인과관계 정보를 학습으로 기존 LSTM 구조를 유지하면서 변수 간의 인과적 영향력을 학습하는 데 중점을 둠

○ Marcinkevics and Vogt(2021)는 Generalized Vector Autoregression(GVAR) 모델을 제안하여 Self-Explaining Neural Network(SENN) 개념을 Granger 인과 분석에 접목함

- GVAR는 기존 벡터 자기회귀(VAR) 모델을 신경망으로 일반화한 것으로, 시계열 데이터를 입력으로 받아 인과관계 계수 행렬을 산출하는 함수를 도입함으로써 동적이고 해석 가능한 인과 구조를 학습함. GVAR의 모형 식은 다음과 같음:

$$x_t = \sum_{l=1}^{\tau}\Psi_{\theta_l}(x^{t-1})x^{t-l} + u^t \quad \text{식(2-7)}$$

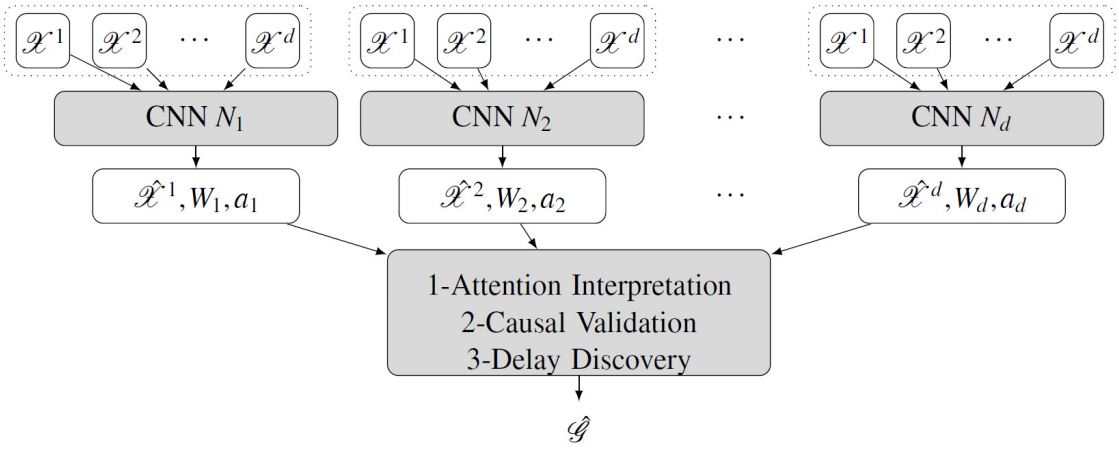
- 여기서 $\Psi_{\theta_l} : \mathbb{R}^d \rightarrow \mathbb{R}^{d\times d}$는 입력 벡터 x^{t-l}에 따라 가중치 행렬을 출력하는 신경망으로 이 행렬의 각 성분은 변수 간 시계열적 인과 영향력을 나타냄
- 학습은 예측 오차(MSE), 인과 행렬의 희소성 유도 항, 인과 행렬의 시간적 변화에 대한 smooth penalty를 포함한 손실 함수를 최소화하는 방식으로 수행됨

○ Xu, Huang, and Yoo(2019)는 고차원 시계열 데이터에서 Granger 인과 구조를 학습하기 위해 Scalable Causal Graph Learning(SCGL) 기법을 제안함. SCGL은 인과 행렬 $A \in \mathbb{R}^{d\times d}$를 Low-lank 행렬로 근사함으로써 계산 효율성을 확보함

- 데이터의 비선형성을 단변량 수준(univariate-level)과 다변량 수준(multivariate-level)으로 분해하여 각각 독립적으로 모델링함으로써 인과 구조의 해석 가능성과 정확도를 동시에 높임

○ Nauta, Bucur, and Seifert(2019)는 Attention-based Dilated CNN 기법을 활용해 Granger 인과관계를 계산하는 Temporal Causal Discovery Framework (TCDF) 모델을 개발함

- TCDF 모델은 〈그림 2-4〉와 같이 각각의 목표 변수에 d개의 독립적인 Attention-based Dilated CNN을 계산하며, 각 목표 변수에 대해 예측값, Attention Scores, Kernel Weights를 계산하는 모델임
- 변수를 예측할 때 활용된 변수에 대한 높은 Attention 점수는 변수 간 인과관계를 정량적으로 수치화하여 계산할 수 있게 함
- TCDF는 순열로 변수 중요도를 계산하여 인과관계를 연결하여, 자기 인과(self-causation)와 원인-결과 간 시간 지연까지 계산이 가능함
- 또한 숨겨진 교란 변수(hidden confounders)의 존재도 계산할 수 있음



주: 1) d개의 $\{N_q\}_{1 \le q \le d}$로 구성된 독립적인 CNN은 전체 시계열 데이터 $X_1, X_2, \ldots, X_d$를 입력으로 사용함.
2) 각 네트워크 N_q는 대상 변수 X_q+를 예측하여 $\hat{X}_q$를 생성하며 커널 가중치 $W_{q,p,k}$ 및 어텐션 점수 $a_{q,p}$를 출력함.

자료: Assaad, Devijver, and Gaussier(2022b), p.778, 〈그림 7〉을 발췌하여 저자 작성.

〈그림 2-4〉 TCDF(Temporal Causal Discovery Framework)에 사용된 신경망 구조

○ Löwe et al.(2022)는 다양한 샘플 간의 공통된 동적 특성을 활용하여 Granger 인과 구조를 학습할 수 있도록 Amortized Causal Discovery(ACD) 프레임워크를 제안함

- ACD는 encoder-decoder 구조로 구성되며, 인코더는 입력 시계열로부터 Granger 인과관계를 학습하고, 디코더는 학습된 인과 구조를 바탕으로 다음 시점의 값을 예측함

- 이 방법은 새로운 데이터에 대해서 추가적인 학습 과정이 필요 없음. 이에 따라 ACD 방법은 다양한 데이터에서 일반화가 가능한 장점이 있고 다양한 인과 구조를 가진 데이터에 대해서도 분석이 가능함

○ Li, Yu, and Principe(2023)은 Granger 인과 분석을 시계열 분석에 결합한 형태인 Causal Recurrent Variational Autoencoder(CR-VAE) 모델을 개발함

- CR-VAE는 디코딩 단계에서만 분석이 가능하게 개발되어 기존 Granger 인과 분석의 개념을 바탕으로 하고 있음
- 각각의 변수들은 독립된 개별 디코더 헤드에서 계산됨. 하지만 인과관계가 없는 변수에 대해서는 디코딩 과정에서 계산되지 않음
- error-compensation 모듈은 시차가 0인 즉각적인 영향을 계산해 예측 성능을 향상시킬 수 있도록 설계함

○ (활용 사례) 딥러닝 기반 인과 발견 접근법은 기후 및 대기과학 분야에서 복잡한 시계열 구조와 고차원 데이터를 효과적으로 분석하는 데 활용되고 있음

- Granger 인과성 기반 심층신경망은 기온, 강수량 등 다양한 기후 변수 간의 비선형적·동적 상호작용을 포착하는 데 사용되며, 이를 통해 기후 시스템의 내재된 인과 메커니즘을 해석하고 예측 정확도를 높이는 데 기여함(Tank et al., 2021, pp.4267-4279; Gong et al., 2024, p.7)
- SCGL은 고차원 위성 자료나 재분석 데이터를 기반으로 지역 간 또는 대기층 간 변수들의 인과 구조를 저랭크 행렬 형태로 추론하며, 대기 이동 경로나 환경 영향 경로 분석 등에 활용됨(Xu, Huang, and Yoo, 2019, pp.1853-1862)
- ACD와 CR-VAE는 다양한 시공간적 샘플에 걸쳐 일반화된 인과관계를 학습하거나, 인과 구조를 반영한 시계열 데이터를 생성함으로써 극한 현상 예측, 정책 개입 시뮬레이션, 기후 시나리오 평가 등 응용 가능성을 넓히고 있음(Löwe et al., 2022, pp.509-525; Li, Yu, and Principe, 2023, pp.8562-8570)

3. 대기오염 분야 인과 분석 방법론 활용 방안

○ 대기오염 분야에서 연구 목적에 적합한 인과 발견(causal discovery) 분석이 가능하도록 〈표 2-1〉과 같이 다양한 방법론들을 분류하고 체계적으로 정리함

- 인과 분석 방법론 분류표 〈표 2-1〉은 실무진 또는 연구자들이 방법론들을 한눈에 비교 분석하여 상황에 맞게 활용 가능하도록 구체적인 활용 가이드라인을 제공하는 데 목적이 있음

○ 〈표 2-1〉은 인과 발견 방법론을 1) 추론 대상, 2) 접근법, 3) 프로세스 가정, 4) 데이터 가정, 5) 동시적 관계, 6) 순환 구조, 7) 숨겨진 변수 7개로 분류하여 정리함

- (추론 대상) 이변량 분석은 단일 인과관계를 분석하며, 다변량 분석은 PM2.5, NO_2, SO_2 등 다양한 오염물질 간의 상호작용을 종합적으로 파악 가능함
- (접근법) 요약 그래프는 정적인 분석에 활용되고, 시계열 그래프는 동적 분석에 사용됨. 확장 요약 그래프는 세부 분석이 가능하며, 대기오염 데이터 특성에 따라 접근법이 달라짐
- (프로세스 가정) 간단한 인과 분석이 가능한 독립성 기반 분석, 점수 기반 방법, 인과 방향성 계산이 가능한 비대칭 기반으로 분류함
  - 독립성 기반은 대기오염 변수 간 단순 상관성을 신속하게 계산하며, 점수 기반 모델은 대기오염 모델의 최적 구조를 평가할 수 있고, 비대칭 기반 모델은 변수 간 인과 방향성을 계산하는 데 사용 가능함
- (데이터 가정) 선형/비선형, 확률적/결정적 가정으로 대기오염 동역학 분석에 활용 가능
  - 선형 가정은 대기오염 배출과 농도 간 단순 관계를 분석하고, 비선형 가정은 복잡한 대기 화학 반응을 모델링하며, 확률적 가정은 기상 불확실성을, 결정적 가정은 특정 조건에서의 오염 패턴을 분석하는 데 유리함
- (동시적 관계) 허용 여부는 실시간 상호작용 분석에 유용함
  - 동시적 관계 허용 시, 대기오염과 기온 간 즉각적인 상호작용을 모니터링하거나 급격한 오염 사건(예: 산불)의 실시간 영향을 분석하는 데 활용될 수 있음
- (순환 구조) 허용 여부는 피드백 루프나 장기적 인과관계 분석에 기여함
  - 순환 구조 허용은 오염 배출과 기상 조건 간 피드백 루프(예: 오염이 기후를 변화시켜 더 많은 오염을 유발)를 모델링하며, 시간 지연만 고려 시 장기적인 오염 누적 효과를 분석하는 데 적합함
- (숨겨진 변수) 허용 여부는 미관측 요인을 포함한 포괄적 분석을 가능하게 함
  - 숨겨진 변수 허용은 관측되지 않은 오염원(예: 비공식 배출지)이나 기상 요인(예: 고층 대기 흐름)을 포함하여 대기오염의 전체적인 인과 구조를 탐구하는 데 기여함

〈표 2-1〉 인과 발견 방법론 분류 목록

<table>
<tr><th>방법론</th><th>추론 대상</th><th>접근법</th><th>프로세스 가정</th><th>데이터 가정</th><th>동시적 관계</th><th>순환 구조</th><th>숨겨진 변수</th></tr>
<tr><td>GC(Granger, 1969)</td><td>이변량</td><td rowspan="7">요약 그래프</td><td rowspan="12">독립성 기반</td><td>선형, 확률적</td><td rowspan="3">✗</td><td rowspan="3">시간 지연만</td><td rowspan="3">✗</td></tr>
<tr><td>Multi-GC(Geweke, 1982)</td><td rowspan="3">다변량</td></tr>
<tr><td>Multi-nonlin-GC
(Bueso, Piles, and Camps-Valls, 2020)</td><td rowspan="2">비선형, 확률적</td></tr>
<tr><td>Multi-TE
(Barnett, Barrett, and Seth, 2009)</td></tr>
<tr><td>TE(Schreiber, 2000)</td><td rowspan="3">이변량</td></tr>
<tr><td>CCM(Sugihara et al., 2012)</td><td>비선형, 결정적</td><td>✓ (완화 가능)</td><td>미확인</td><td>부분적으로</td></tr>
<tr><td>Ext.-CCM(Diaz et al., 2022)</td></tr>
<tr><td>PCMCI(Runge et al., 2015)</td><td rowspan="12">다변량</td><td rowspan="3">시계열 그래프</td><td rowspan="9">선형/비선형,
확률적</td><td>✗</td><td>시간 지연만</td><td rowspan="3">✗</td></tr>
<tr><td>tsPC(Runge, 2020)</td><td rowspan="8">✓ (완화 가능)</td><td rowspan="8">✓(완화 가능)</td></tr>
<tr><td>PCMCI+(Runge, 2020)</td></tr>
<tr><td>PCGCE(Assaad, Devijver, and Gaussier, 2022a)</td><td rowspan="3">확장 요약
그래프</td></tr>
<tr><td>FCIGCE(Assaad, Devijver, and Gaussier, 2022a)</td><td rowspan="6">✓</td></tr>
<tr><td>tsFCI(Entner and Hoyer, 2010)</td><td rowspan="5">시계열 그래프</td></tr>
<tr><td>SVAR-FCI
(Assaad, Devijver, and Gaussier, 2022a)</td></tr>
<tr><td>SVAR-GFCI
(Mallinsky and Spirtes, 2018)</td><td>점수 및 독립성
기반</td></tr>
<tr><td>LPCMCI
(Gerhardus and Runge, 2020)</td><td>독립성 기반</td></tr>
</table>

〈표 2-1〉의 계속

방법론	추론 대상	접근법	프로세스 가정	데이터 가정	동시적 관계	순환 구조	숨겨진 변수
(F)GES (Meek, 1997; Chickering 2002b; Ramsey, 2017)	다변량	요약 그래프	점수 기반	선형, 확률적	✓(완화 가능)	시간 지연만	✗
DYNOTEARS(Pamfil et al., 2020)		시계열 그래프					
IDYNO(Gao et al., 2022)				선형/비선형, 확률적			
NTS-NOTEARS(Sun, Schulte, and Liu, 2023)							
TiMiNo(Peters, Janzing, and Scholkopf, 2013)			독립성 기반				
RHINO(Gong et al., 2022)							
VARLiNGAM(Shimizu et al., 2006)			비대칭 기반	선형, 확률적			
NGC(Tank et al., 2021)			점수 기반	비선형, 확률적	✗	✗	✗
GVAR(Marcinkevics and Vogt, 2021)		확장 요약 그래프			✓(완화 가능)	✓(완화 가능)	✓
SCGL(Xu, Huang, and Yoo, 2019)						✓	
TCDF(Nauta et al., 2019)		시계열 그래프			✓		
ACD(Löwe et al., 2022)		확장 요약 그래프					
CR-VAE(Li, Yu, and Principe, 2023)							

주: 1) 동시적 관계 및 순환 구조 허용 여부(✓: 허용, ✗: 비허용), “완화 가능”은 기본 설정과 다르게 조정할 수 있음을 의미함.
예: “동시적 관계: ✓ (완화 가능)”은 기본적으로 동시적 관계를 허용하지만, 필요시 이를 허용하지 않는 버전으로 사용할 수 있음을 의미.
2) 요약 그래프(Summary Graph): 시간적 의존성을 하나의 정적 그래프로 요약하여 변수 간 인과관계를 분석(시간 축을 단순화한 형태).
시계열 그래프(TSG: Time Series Graph): 시간에 따른 동적 관계를 명시적으로 표현하며, 시간 지연과 동시적 관계를 모두 포함.
확장 요약 그래프(Extended Summary Graph): 요약 그래프를 확장하여 추가적인 시간적 또는 구조적 정보를 포함.
자료: Camps-Valls et al.(2023), p.9, 〈표 2〉를 바탕으로 저자 작성.

○ 〈표 2-1〉의 인과 발견 방법론 내용을 1) 회귀, 2) 정보 이론, 3) 비선형 동역학, 4) 조건부 독립성, 5) 점수 기반, 6) 혼합형, 7) 딥러닝 기반 방법론으로 분류하여 각각의 방법론별 특징 및 장점을 정리하고 대기오염 정책 연구에서의 활용(안)을 〈표 2-2〉에 정리하였음

○ 인과 발견 방법론을 특징별로 분류하고, 각각의 방법론의 장점과 특징을 고려한 대기오염의 원인 분석, 전파 경로 탐지, 정책 평가, 지역적 대응 전략 수립 등 다양한 연구 목적에 맞춘 맞춤형 활용(안)을 목록화하여 〈표 2-2〉에 제시함

- PCMCI와 같은 조건부 독립성 기반의 방법론들은 조건부 독립성 검정을 통해 인과관계를 분석함. 대기오염의 전파 경로 또는 정책 효과를 정량적으로 계산하는 데 활용 가능함
  - 특히, 숨겨진 변수 허용 방법(FCIGCE, tsFCI, LPCMCI 등)은 관측되지 않은 요인(예: 비공식 배출지, 고층 대기 흐름)을 포함하여 분석이 가능하므로 미세먼지의 잠재적 위험 요인을 파악하는 데 중요한 역할을 함
- 점수 기반 방법은 점수 함수를 활용해 인과 그래프를 최적화 계산하는 데 사용되므로 대기오염의 전체적인 인과 구조를 계산할 때 활용 가능함
  - (F)GES, DYNOTEARS, IDYNO 등의 방법론들은 효율적인 계산이 가능하며, 장기 예측에도 활용이 가능함. 구체적으로 대기오염 정책 효과 평가 또는 예측한 정책 시나리오의 결과를 시뮬레이션하는 방법에 활용될 수 있음
- 딥러닝 기반의 방법론은 비선형적인 시계열 데이터의 복잡한 비선형적인 인과 구조를 학습할 수 있음
  - Granger 인과관계를 딥러닝과 결합한 다양한 방법론들(NGC, GVAR, SCGL 등)은 복잡한 비선형 관계를 학습할 수 있어, 대기오염 농도 변화의 원인 분석 및 지역 간 또는 측정소 간의 오염 이동 경로 계산 분석에 효과적임
  - 예를 들어, SCGL은 고차원 대기 변수 간 인과 그래프를 저차원 공간에서 근사하여 계산 효율성을 확보하면서도, 도심과 교외 간 오염 확산 경로를 구조적으로 모델링하는 데 활용될 수 있음(Xu, Huang, and Yoo, 2019, pp.1853-1862)
  - GVAR는 해석 가능한 인과 행렬을 출력하여 변수 간 영향력의 방향성과 부호까지 파악할 수 있어, 기상 요소(예: 풍속, 온도)가 특정 오염물질에 미치는 시간적 영향 패턴을 정량적으로 분석하는 데 유리함(Marcinkevics and Vogt, 2021, pp.1-9)
  - TCDF는 어텐션 기반의 CNN 구조를 통해 변수 간 인과관계를 자동으로 탐지하고, 숨겨진 교란 변수나 시간 지연 효과까지도 반영할 수 있어, 고해상도 대기 시계열 데이터에 대한 실시간 모니터링 및 원인 규명 도구로 유용함(Nauta, Bucur, and Seifert, 2019, p.19)

- ACD 및 CR-VAE는 다양한 지역·기간·상황에 걸친 데이터를 학습하여 일반화된 인과 구조를 도출하거나, 학습된 인과관계를 바탕으로 새로운 시계열 데이터를 생성함으로써 극단 기상현상 시뮬레이션, 정책 개입 효과 예측, 시나리오 기반 대기질 평가 등 실질적 정책 활용 가능성을 확장시킴(Löwe et al., 2022, pp.509-525; Li, Yu, and Principe, 2023, pp.8562-8570)

○ 이처럼 딥러닝 기반 인과 발견 방법은 전통적 통계 기반 접근법에 비해 데이터의 복잡성, 비선형성, 고차원성을 보다 유연하게 수용하며, 대기오염의 원인 분석, 전파 경로 탐지, 정책 시뮬레이션 및 대응 전략 수립에 있어 강력한 보조 도구로 활용될 수 있음

○ 본 연구는 다양한 인과 발견 방법론 중 비선형 LPCMCI를 시범 분석 모델로 선정함

- LPCMCI가 관측되지 않은 대기오염 요인의 시간적 영향력을 정밀하게 추정할 수 있는 숨겨진 변수 처리 능력과 시계열 데이터의 시간적 특성을 효과적으로 분석하는 데 강점을 가지고 있기 때문임
- 또한 LPCMCI는 숨겨진 변수와 관측 변수 간의 상호작용을 분리하여 정책 입안자들이 숨겨진 요인을 고려한 대응 전략을 수립하는 데 기여할 수 있음

〈표 2-2〉 인과 발견 방법론 특징 및 활용(안) 목록

유형	방법론	특징	장점	활용(안)	세부 활용(안)
회귀 기반 방법	GC (Granger, 1969)	· Granger 인과관계 · 시간 선행성을 기반으로 예측력 평가	· 선형관계 분석에 효과적 · 계산 효율 높음	· 선형 및 비선형 인과관계를 분석하여 대기오염물질 간 영향 평가	· 대기오염 규제 정책의 단기 효과를 신속히 평가
	Multi-GC (Geweke, 1982)	· 다변량 Granger 인과관계 · GC를 다변량으로 확장	· 다변량 선형관계 분석 가능 · 시간 지연 관계에 강력		· 다중 오염원의 장기적 대기오염 영향 평가
	Multi-nonlin-GC (Bueso, Piles, and Camps-Valls, 2020)	· 비선형 다변량 Granger 인과관계 · 비선형 관계 허용	· 비선형 관계 포착 가능 · 다변량 분석에 유연		· 계절별 대기오염의 비선형 패턴 변화 예측
정보 이론 기반 방법	Multi-TE (Barnett, Barrett, and Seth, 2009)	· 다변량 전이 엔트로피 · TE를 다변량으로 확장	· 다변량 비선형 정보 흐름 분석 가능	· 정보 흐름을 통해 대기오염의 비선형 전파 경로와 복잡한 상호작용 분석	· 단일 오염원의 다변량 전파 경로를 비선형 정보 분석으로 학습
	TE (Schreiber, 2000)	· 전이 엔트로피 · 정보 흐름으로 인과관계 추론	· 비선형 정보 흐름 분석에 적합		· 복합 오염원의 지역 간 대기오염 전파 경로를 정보 흐름 분석으로 추적
비선형 동역학 기반 방법	CCM (Sugihara et al., 2012)	· 수렴 교차 매핑 · 비선형 동적 시스템에서 인과관계 추론	· 결정적 비선형 시스템 분석에 강력	· 비선형 동적 시스템에서 대기오염의 복잡한 패턴과 기상 조건의 영향을 분석	· 계절별 대기오염의 비선형 동적 패턴을 탐지
	Ext.-CCM (Diaz et al., 2022)	· 확장 수렴 교차 매핑 · CCM의 확장 버전	· CCM보다 복잡한 비선형 관계 포착 가능		· 극단 기상 조건이 대기오염에 미치는 비선형 영향을 평가

〈표 2-2〉의 계속

유형	방법론	특징	장점	활용(안)	세부 활용(안)
조건부 독립성 기반 방법 (PCMCI 계열)	PCMCI (Runge et al., 2015)	· 시계열 조건 기반 방법 · PC를 시계열에 최적화	· 시계열 데이터에서 시간 지연 관계 분석에 효율적	· 시계열 데이터에서 동시적 및 시간 지연 인과관계를 분석하여 대기오염 전파 경로를 학습	· 실시간 대기오염 전파 분석을 통해 동시적 및 시간적 인과관계를 파악하고 정책 효과를 평가
	tsPC (Runge, 2020)	· 시계열 PC 알고리즘 · PC를 시계열에 적용	· 시계열 데이터에서 동시적 관계 분석 가능		· 교통 기반 대기오염의 시간 지연 영향을 빠르게 분석하여 정책 시나리오를 평가
	PCMCI+ (Runge, 2020)	· PCMCI 확장 · 동시적 관계와 순환 구조 허용	· PCMCI보다 복잡한 관계(동시적, 순환) 분석 가능		· 대기오염 상호작용 분석으로 동시적 및 순환적 인과관계를 파악하여 정책의 간접 효과를 평가
	LPCMCI (Gerhardus and Runge, 2020)	· 지역적 PCMCI · 시계열 데이터에 최적화	· 지역적 분석으로 복잡한 시계열 데이터 처리 가능		· 지역별 대기오염 시간적 인과 분석으로 지역별 저감 정책 효과를 비교 평가

〈표 2-2〉의 계속

유형	방법론	특징	장점	활용(안)	세부 활용(안)
조건부 독립성 기반 방법 (기타)	PCGCE (Assaad, Devijver, and Gaussier, 2022a)	· PC 기반 확장 요약 그래프 · 요약 그래프를 확장	· 확장된 요약 그래프를 통해 세부 관계 분석 가능	· 조건부 독립성 검정 기반 대기오염의 세부 인과 구조와 숨겨진 요인을 분석	· 지역별 대기오염의 세부 인과관계를 분석하여 잠재적 지역 오염 구조를 학습
	FCIGCE (Assaad, Devijver, and Gaussier, 2022a)	· FCI 기반 확장 요약 그래프 · 숨겨진 변수 허용	· 숨겨진 변수를 고려한 요약 그래프 분석 가능		· 숨겨진 변수를 포함한 대기오염 인과 분석으로 미관측 요인의 영향을 탐색
	tsFCI (Entner and Hoyer, 2010)	· 시계열 FCI · FCI를 시계열에 적용	· 숨겨진 변수를 포함한 시계열 데이터 분석 가능		· 미관측 기상 요인의 대기오염 영향을 시간적 인과 분석으로 규명
	TiMiNo (Peters, Janzing, and Scholkopf, 2013)	· 시계열 데이터에 특화된 독립성 기반 방법	· 시계열 데이터에서 독립성 기반 분석 효율적		· 도시 환경 대기오염 영향을 독립성 기반 시간적 분석으로 효율적으로 평가
	RHINO (Gong et al., 2022)	· 시계열 데이터에 특화된 독립성 기반 방법	· 복잡한 시계열 데이터 분석에 유연		· 복잡한 시간적 대기오염 인과 분석으로 지역별 잠재 구조적 요인을 학습

〈표 2-2〉의 계속

유형	방법론	특징	장점	활용(안)	세부 활용(안)
점수 기반 방법	(F)GES (Meek, 1997; Chickering 2002b; Ramsey, 2017)	· GES와 F-GES · 점수 기반 탐욕적 탐색	· 탐욕적 탐색으로 계산 효율성 높음	· 점수 기반으로 인과 구조를 학습하여 대기오염의 장단기 영향을 예측	· 대기오염 규제 정책의 단기 효과를 선형 인과 분석으로 신속히 평가
	DYNOTEARS (Pamfil et al., 2020)	· 시계열 데이터에 특화된 점수 기반 방법	· 시계열 데이터에서 동적 구조 분석에 효과적		· 대기오염의 장기적 동적 변화를 시간적 선형 구조 분석으로 평가
	IDYNO (Gao et al., 2022)	· 비선형 시계열 데이터에 특화된 점수 기반 방법	· 비선형 시계열 관계 분석 가능		· 대기오염의 비선형적 장기 영향을 시간적 비선형 인과 학습으로 분석
	NTS-NOTEARS (Sun, Schulte, and Liu, 2023)	· 비선형 시계열 데이터에 특화된 NOTEARS 확장	· 비선형 관계와 시계열 구조 동시 분석		· 비선형 상호작용과 시계열 구조를 분석하여 대기오염의 잠재 구조를 학습
혼합형 방법	SVAR-FCI (Assaad, Devijver, and Gaussier, 2022a)	· SVAR 기반 FCI · 구조적 VAR 모델 사용	· SVAR 모델 기반으로 구조적 분석 가능	· 여러 접근법을 결합하여 대기오염의 복잡한 인과 구조와 정책 효과를 분석	· 구조적 대기오염 정책의 효과를 시간적 인과 분석으로 시뮬레이션
	SVAR-GFCI (Mallinsky and Spirtes, 2018)	· SVAR 기반 GFCI · 점수와 독립성 결합	· 점수 기반과 독립성 기반의 장점 결합		· 미관측 요인을 포함한 대기오염 인과 분석으로 원인을 규명
	VARLiNGAM (Shimizu et al., 2006)	· VAR 기반 LiNGAM · 비대칭성 활용	· 비대칭성을 활용한 선형관계 분석에 효과적		· 대기오염의 비대칭적 영향을 분석하고 정책 효과를 평가

〈표 2-2〉의 계속

유형	방법론	특징	장점	활용(안)	세부 활용(안)
딥러닝 기반 방법	NGC (Tank et al., 2021)	· 비선형 Granger 인과관계를 MLP/LSTM 구조로 추정	· 희소성 반영한 비선형 다변량 인과 추정 가능	· 비선형·고차원 시계열에서 인과 구조 학습 및 해석을 통해 대기오염 원인 분석, 전파 경로 탐지 및 정책 효과 평가	· 교통량-대기질 간 비선형 효과 분석
	GVAR (Marcinkevics and Vogt, 2021)	· 설명 가능한 신경망 기반의 시간에 따라 변하는 인과 행렬 도출	· 해석 가능성 높음 · 시계열 내 영향력 방향성 파악		· 기상 변화가 특정 오염물질에 미치는 시차별 영향 해석
	SCGL (Xu, Huang, and Yoo, 2019)	· 저랭크 근사 기반 고차원 인과 그래프 학습	· 고차원 데이터에서 효율적 · 노이즈 억제 우수		· 도심 및 교외 간 오염 확산 구조의 인과 그래프 모델링
	TCDF (Nauta, Bucur, and Seifert, 2019)	· 어텐션 기반 dilated CNN으로 인과 구조 학습	· 시간 지연, 숨겨진 변수, 동시성 등 복잡한 구조 반영		· 산불 등 급격한 사건 발생 시 인과 탐지 및 대응 정보 제공
	ACD (Löwe et al., 2022)	· Encoder-Decoder 구조 · 샘플 간 일반화 가능한 인과 구조 학습	· 다양한 지역 및 시간대에도 일반화 가능 · 추가 학습 불필요		· 서로 다른 도시에서 공통된 오염 인과 패턴 자동 학습
	CR-VAE (Li, Yu, and Principe, 2023)	· VAE 기반 인과 제약 구조를 포함한 시계열 생성 모델	· 미래 생성 및 정책 시나리오 시뮬레이션에 강력		· 탄소 중립 정책 시나리오별 대기질 예측 시뮬레이션

주: 1) 〈표 2-1〉을 바탕으로 각 방법론의 특징과 장점을 요약하고, 대기오염 정책 연구 분야에서 활용 가능한 방안을 제시함.
2) "활용(안)"은 대기 분야 연구자들이 대기오염의 원인, 전파 경로, 정책 효과 등을 분석할 수 있도록 각 방법론의 강점을 반영하여 작성됨.
3) LPCMCI(Gerhardus and Runge, 2020) 방법론은 3장 사례 분석에서 대기오염물질 간 인과 발견 분석에 사용됨.
자료: 저자 작성.

Ⅲ 사례 분석

1. 분석 모델 선정

○ 본 연구는 LPCMCI(Latent-PCMCI) 기법을 활용하여 대기오염물질 간의 인과관계에 대한 사례 분석을 시범적으로 수행함으로써, 방법론의 실질적 적용 가능성을 검증함

- 전통적인 상관관계 분석이나 회귀 분석은 변수 간의 동시적 관계만을 파악하는 한계가 있으며, 변수 간의 시간 차(time-lag)에 따른 영향을 고려하기 어려움
- 이에 따라, 본 연구에서는 시간 지연을 고려한 인과 발견 기법인 LPCMCI를 적용하여 대기오염물질 및 기상 요소 간의 동적 인과관계를 규명하고자 함

○ PCMCI에는 여러 변형 기법이 존재하지만, 본 연구에서 LPCMCI를 선택한 이유는 잠재 변수 고려 가능, 노이즈 및 다중공선성 문제 완화 및 시간 지연 분석 강화 특징 때문임

- 잠재 변수 고려: 직접 관측되지 않는 배출원, 기상 변화, 지역적 요인 등 숨겨진 환경 요소의 영향을 반영할 수 있어, 보다 신뢰성 높은 인과관계 도출 가능
- 노이즈 및 다중공선성 문제 완화: 대기오염 데이터는 측정 오차와 변수 간 높은 상관성을 포함하는 경우가 많아, LPCMCI가 보다 안정적인 인과관계를 추출하는 데 유리함
- 시간 지연 분석: LPCMCI는 변수 간 유효한 시간 지연을 효율적으로 계산하여 단기 및 장기 영향 분석이 가능함

○ 본 연구에서 선정한 LPCMCI 기법은 대기오염물질 변수 간의 시간에 따른 변화와 인과관계를 계산하고 시각화하여 표현할 수 있음

- LPCMCI 분석 결과는 대기오염 정책 수립 및 평가 등에 활용할 수 있을 것으로 기대됨

2. 데이터 수집 및 전처리

○ 대기오염물질 간의 인과관계 사례 분석을 위해 서울시 중구 대기오염 및 기상 측정소의 시간 단위(hourly) 관측 데이터(2022년)를 수집하고 분석에 활용하였음

- 대기오염 데이터: 에어코리아(Air Korea)[1]에서 제공하는 이산화황(SO_2), 일산화탄소(CO), 오존(O_3), 이산화질소(NO_2), 미세먼지(PM10), 초미세먼지(PM2.5) 농도 값 데이터 수집
- 기상 데이터: 기상청 API허브[2]에서 제공하는 방재기상관측 자료 기온(°C), 풍속(m/s), 강수량(mm) 데이터 수집

○ 대기오염 및 기상 데이터는 대부분 멱함수(power-law) 형태의 긴꼬리 분포를 보이며, 변수 간 단위 및 스케일 차이가 크고 이상치의 영향이 존재하여 정규화 과정이 필수적임

- 모든 변수는 평균 0, 표준편차 1로 표준화하여 스케일 차이를 제거하고 변수 간 비교 가능성 확보를 위해 스케일 정규화(z-score standardization)를 수행함
- 데이터 내 일부 시간대에 결측값(NaN)이 발견되었으며, 이는 먼저 선형 보간법(linear interpolation)을 통해 보완하였음. 단, 특정 변수에서 연속적으로 결측값이 길게 발생한 경우에는 해당 시간 구간을 전체 분석에서 제외함

3. 분석 결과

○ LPCMCI 모델을 활용해 변수 간 인과관계를 분석하고, 시간을 최대 6시간까지만 고려하여 개별 변수 간 시차(time-lag) 영향을 정량적으로 계산함

○ ParCorr(Partial Correlation) 검정을 바탕으로 변수 간 독립성을 계산하였고, p-value 값은 0.1을 기준으로 0.1 미만의 유의미한 관계만을 분석 대상으로 하여 연결함

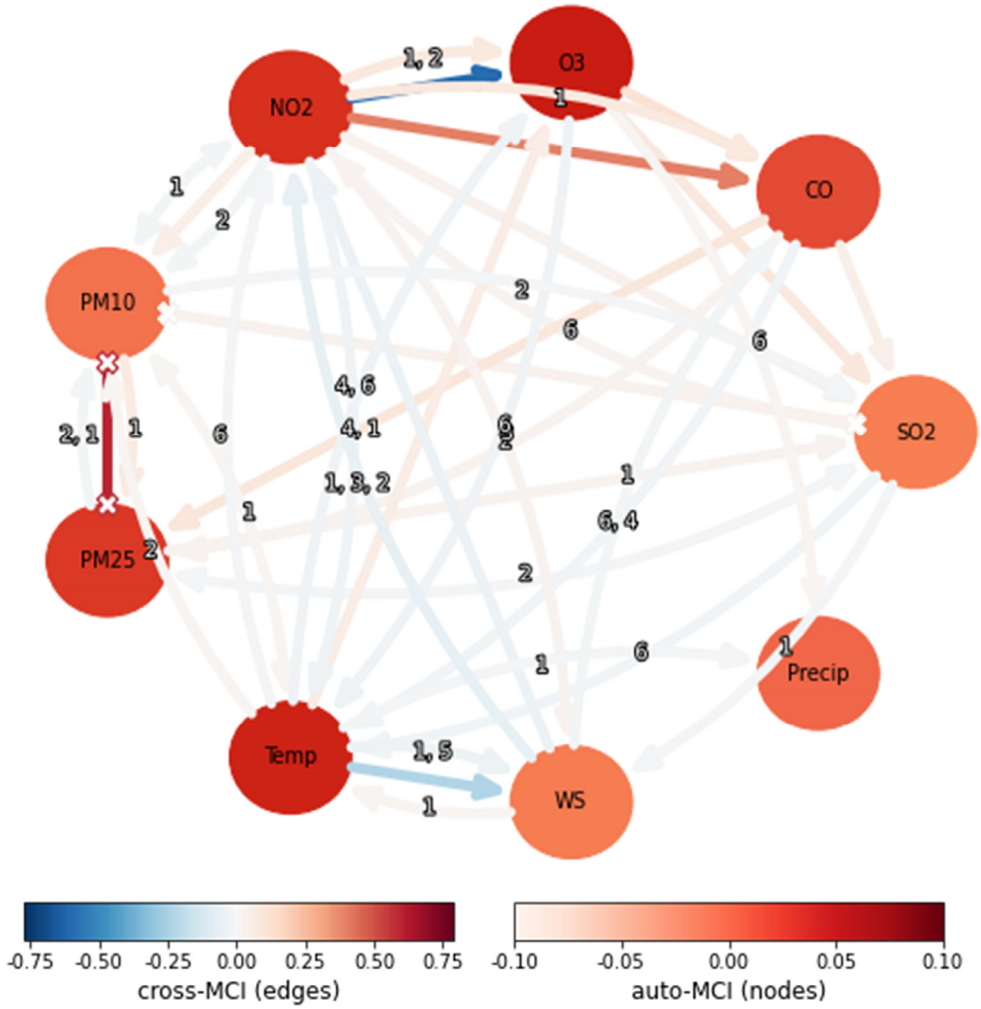
○ 분석 결과는 Directed Partially Aggregated Graph(DPAG) 형태로 시각화하여 표현했음. 본 시각화 결과는 〈그림 3-1〉과 같이 변수 간의 인과관계를 직관적으로 이해할 수 있도록 구성된 방향성이 있는 네트워크 그래프임

- 노드(node): 각 변수(대기오염물질 및 기상 요소)를 나타냄
- 에지(edge): 두 노드 간 통계적으로 유의미한 인과관계가 존재할 경우 연결되며, 이는

1) 에어코리아, “최종확정 측정자료”, 검색일: 2025.3.25.
2) 기상청 API허브, “방재기상관측(AWS)”, 검색일: 2025.3.25.

원인 → 결과 변수의 흐름을 나타냄

- 노드 색상: 해당 변수의 자기회귀(auto-MCI) 강도를 반영하며, 색이 짙을수록 해당 변수가 자신의 과거 값에서 영향을 강하게 받는 경향이 있음을 의미함
- 에지 색상: 색상(빨강: 양의 영향, 파랑: 음의 영향)이 진할수록 관계(cross-MCI) 강도가 크다는 것을 나타냄
- 에지 라벨: 에지 위 표시된 숫자는 해당 인과관계가 발생한 시간 지연(lag τ)을 의미함
- 엑스(×) 기호가 표시된 에지: 시차가 0인 경우(lag=0), 즉 동일 시간대에 발생한 동시적 관계(contemporary link)를 의미하며, 이러한 관계는 방향성이 없음



주: 서울시 중구 지역의 대기오염 및 기상 변수 간 인과관계 네트워크를 시각화하여 표현한 것임.
자료: 저자 작성.

〈그림 3-1〉 LPCMCI 기반 대기오염물질 및 기상 변수 간 인과 네트워크 분석 시각화 결과

○ 전체 인과 네트워크 분석 시각화 결과 〈그림 3-1〉을 살펴보면, 총 9개의 변수 간 p-value 값 기준 유의미한 인과관계만 연결되고, 다음과 같은 주요 특징을 보였음

- 상대적으로 많은 에지가 연결된 PM2.5, NO_2, 기온(Temp) 변수는 중심 노드를 의미하는 변수임을 확인함
- NO_2 → O_3는 강한 음의 인과관계(파란색 에지)를 나타내며, 이는 실세계에서 광화학 반응에 의한 영향이 반영되어 나타난 결과로 보임
- 기온 상승은 대기 정체와 오염물질의 화학 반응에 영향을 미치는 변수로 나타남. 또한 풍속(WS)은 NO_2, CO 등에 음의 방향으로 파란색 에지가 연결되었으며, 이를 통해 풍속의 증가는 대기 중 오염물질이 희석되고 확산되는 물리적 영향이 반영된 결과라는 점을 알 수 있음
- 시간 지연(lag) 효과: 각각의 개별 에지는 1~2시간 정도의 짧은 시간 지연과 4~6시간 지연이 많이 나타남. 단기 시차(1~2시간)에서는 오염물질 간 즉각적인 반응 또는 같은 배출원에서의 영향에 의해 나타남. 중기 시차(4~6시간)는 대기 중 화학적인 변화 또는 기상 조건의 간접 영향 등을 반영함. 이는 오염물질 간 상호작용이 실시간이 아니라 일정 시간이 지난 후 작용한다는 것을 의미함
  - 본 사례분석은 서울시 중구 단일 대기 및 기상 측정소를 대상으로 하여 최대 6시간까지만 lag를 고려하여 분석을 수행함. 향후 광역 시공간 모델로 확장하고 오염물질의 장거리 이동 및 누적 효과 등을 분석하기 위해 lag 구간을 12시간, 24시간 등 장기 시차까지 확대 적용하여 분석 가능함
- 자기회귀적 구조(auto-dependence): 노드 색상은 자기 회귀(auto-MCI) 강도를 의미함. 기온, O_3, PM2.5, NO_2는 모두 자기회귀성이 강하게 나타난 변수로, 이들은 자신의 과거 값만으로도 일정 수준의 미래 예측이 가능함. 이는 해당 변수들이 명확한 시간적 패턴 또는 누적 효과를 가지고 있음을 시사함.

○ 주요 변수 간 인과 분석 결과에 대한 설명은 다음과 같음

- PM2.5 ↔ PM10 (lag=1~2): PM2.5와 PM10은 1~2시간 시차에서 양방향의 유의미한 인과관계를 보였으며, 이는 두 입자상 물질이 물리적으로 밀접한 연관성을 갖고 있다는 의미로 해석됨
- PM2.5 ━ PM10 (lag=0): 시차가 0인 동시간대에서 인과 방향이 없는 강한 양의 상관관계(양끝이 ×로 표시된 빨간색 에지)로 나타남. 이는 두 물질이 동일한 환경 조건에서

동시에 발생하거나 유사한 패턴으로 변동했음을 의미함

- SO_2 ━ PM10 (lag=0): SO_2와 PM10은 산업 배출이나 석탄 연소와 같은 공통의 1차 배출원에서 동시 배출되는 경우가 많아, 동시간대에서의 무방향 양의 상관관계(양끝이 ×로 표시된 빨간색 에지)로 나타난 것으로 해석됨
- NO_2 → O_3 (lag=1~2): NO_2가 1~2시간 후 O_3 농도에 음의 영향(파란색 에지)을 미치는 인과관계가 나타났음. 이는 오염지역에서 NO_2와 함께 배출되는 NO가 O_3를 제거하는 반응($NO + O_3 \rightarrow NO_2 + O_2$)을 유도해 O_3 농도가 낮아지는 현상으로 해석될 수 있음. 오존 단기 예측 및 선제적 대응 시 이러한 시간 지연 특성을 고려할 필요가 있음
- 기온(Temp) → O_3: 기온 상승 시 대기오염물질 중 O_3가 가장 강한 양의 인과관계(빨간색 에지)를 갖는 변수로 나타남. 이는 고온 조건에서 광화학 반응이 활발해져 오존 생성이 증가하기 때문임. 실제로 여름철 폭염 기간 동안 오존 농도 급증 현상이 자주 나타남
- 기온(Temp) → NO_2 (lag=6): 기온 상승은 약 6시간 후 NO_2 농도에 음의 영향(파란색 에지)을 미치는 것으로 나타남. 이는 기온 상승 시 햇빛과 자외선량이 증가하면서 NO_2가 광분해되어 O_3 생성에 활용되기 때문에, 일정 시간 경과 후 NO_2 농도가 감소하는 것으로 해석할 수 있음
- 풍속(WS) → NO_2, CO (lag=1~3): 풍속 증가 시 NO_2 및 CO 농도가 1~3시간 후 감소하는 것으로 나타남. 이는 풍속에 따른 확산 및 희석 효과로 설명되며, 풍속 감소 시 대기오염물질이 정체될 수 있으므로 예보와 연계한 대응 체계가 필요함

○ 본 사례 분석은 최신 인과 발견 모델을 실제 대기오염 및 기상 자료에 적용하여 모델의 적용 가능성과 실효성을 확인함. 다만, 서울시 중구 지역 하나의 측정소만을 대상으로 인과 분석을 수행하여 대기오염의 이동 및 확산, 그리고 지역 간 영향을 계산할 수 없음

- 대기오염물질은 지리적인 형태, 배출원 위치 등 다양한 공간적인 특징에 따라 미치는 영향이 다르므로 하나의 측정소를 대상으로 한 분석 결과는 한계가 있음
- 향후, 국내 또는 전 세계의 측정소 데이터를 통합한 시공간 모델링 개발 연구가 필요하며, 이를 통해 지역 및 측정소 간의 오염 영향 이동경로를 계산하는 인과 분석이 가능할 것으로 기대됨

Ⅳ 결론 및 향후 연구 방안

1. 결론

○ 본 연구는 대기오염물질 간의 인과관계 분석을 위해 통계 및 인공지능 기반 방법론의 활용 현황을 〈표 2-1〉에 정리하고, 이를 바탕으로 방법론별 대기오염 연구 분야에서의 활용 가능성을 분석하고 활용(안)을 제시함

- 특히, 변수 간 시간 지연(lag) 효과 및 다변량 상호작용을 반영할 수 있는 기법을 중심으로 검토함(표 2-2 참조)

○ 현재 대기오염 분야에서는 Granger 인과성 방법론 뿐만 아니라, 시계열 확장형 인과 발견 기법인 PCMCI 및 LPCMCI와 같은 방법론이 다변량 시계열 인과 분석에서 활발히 활용되고 있음

- 특히, LPCMCI 기법은 시간 지연을 고려한 조건부 독립성 기반 인과 발견이 가능한 알고리즘으로 비정상적 특성을 가진 시계열이나 관측되지 않은 잠재 변수(latent confounders)가 존재하는 문제에서도 활용 가능함

○ 실제 적용 가능성 평가를 위해, 특정 측정소(서울시 중구)의 대기오염물질 및 기상 데이터를 수집하여 LPCMCI 인과 분석을 수행하였음

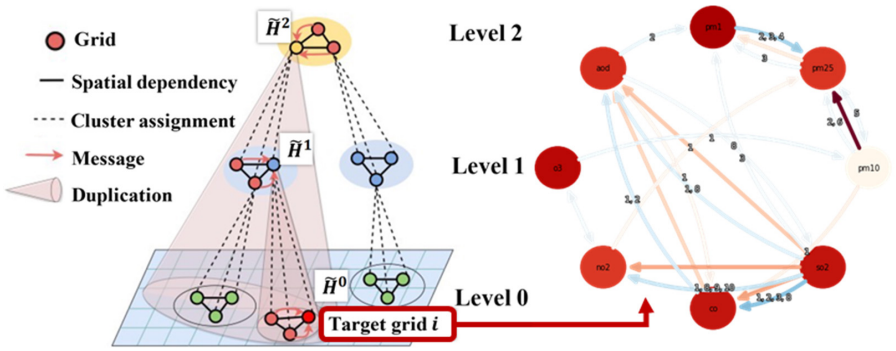
- 분석 결과, PM2.5, NO_2, 기온 변수는 중심 노드로 나타남. 풍속, 기온 기상 변수는 오염물질 확산 및 농도 변화에 영향을 미치는 변수로 나타남. 또한 $NO_2 \rightarrow O_3$ 간 음의 인과관계 등은 광화학 반응 메커니즘 결과로 해석할 수 있으며, 대부분의 변수 간 시간 지연이 존재하였음

- LPCMCI 기법은 변수 수와 시계열 길이가 증가할수록 연산량이 대폭 증가하는 특성이 있음. 이는 고해상도 공간 데이터나 다지역 데이터를 분석할 때 한계가 있을 것으로 판단됨

2. 향후 연구 방안

○ 기존 방법론의 가장 큰 한계점인 계산 복잡도 문제를 해결하기 위해 계층적 메시지 전달(Hierarchical Message Passing) 기법을 활용한 방법론 개발이 필요함

- 기존 시계열 인과 모델은 변수 수 또는 공간 범위가 커질수록 계산량이 급격히 증가하는 문제가 있음. 특히 대기오염 데이터처럼 고해상도의 데이터를 분석할 경우, 효율적인 계산이 매우 중요한 이슈임
- 이에 따라, 최근 제안된 Fast Hierarchical Message Passing 구조(그림 4-1 참조)는 다층 구조에서의 공간 의존성을 단계적으로 축약·확산시킴으로써 계산 효율을 높이고, 넓은 지역 간 상호작용까지 고려할 수 있는 해법으로 주목받고 있음(Wu et al., 2024, pp.2433-2441)
- 〈그림 4-1〉의 모델 구조는 각 공간 격자(grid)를 클러스터 단위로 그룹화한 뒤, 클러스터 간 상위 레벨에서 메시지를 요약하고 전달하는 방식으로 구성되어 있음. 이러한 계층적 설계는 지역 수준(level 0)에서의 세부 정보와 광역 수준(level 2)에서의 상호작용을 동시에 반영할 수 있도록 함



자료: Wu et al.(2024), p.2437, 〈그림 4〉를 수정하여 저자 작성.

〈그림 4-1〉 고속 계층적 메시지 전달(Fast Hierarchical Message Passing) 구조

○ 이러한 모델(그림4-1 참조)을 인과 분석 모델에 통합하면, 기존 분석의 한계점이었던 계산량 증가 문제를 해결할 수 있고, 광역 대기환경의 원인을 (준)실시간으로 분석할 수 있을 것으로 기대됨

- 향후 광범위한 대기환경의 복잡한 공간 및 시간 인과관계를 분석하기 위해서는 이러한 계층 기반 메시지 전달 기법을 포함한 구조적 고도화가 필수적이라고 사료됨

참고문헌

[국내문헌]

이승민 외(2021),「기후변화에 따른 미세먼지 대기질 변화 추정 및 관련 정책 지원 연구」, 한국환경연구원. pp.86-87.

국립환경과학원(2015),「GAINS-Korea를 이용한 기후·대기환경 통합 정책 연구(Ⅱ)」.

[국외문헌]

Asian Development Bank(2022.3), *Air Quality in Asia: Why Is It Important, and What Can We Do?*, p.8.

Assaad, C. K., E. Devijver, and E. Gaussier(2022a), "Discovery of Extended Summary Graphs in Time Series", *Proceedings of The Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, Vol.180, pp.96-106.

Assaad, C. K., E. Devijver, and E. Gaussier(2022b), "Survey and Evaluation of Causal Discovery Methods for Time Series", *Journal of Artificial Intelligence Research*, Vol.73, p.778.

Barnett, L., A. B. Barrett, and A. K. Seth(2009), "Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables", *Physical Review Letters*, 103(23): 238701, pp.1-10.

Bongers, S. et al.(2021), "Foundations of Structural Causal Models with Cycles and Latent Variables", *The Annals of Statistics*, 49(5), pp.2885-2915.

Bueso, D., M. Piles, and G. Camps-Valls(2020), "Explicit Granger Causality in Kernel Hilbert Spaces", *Physical Review E*, 102(6): 062201, pp.1-8.

Camps-Valls, G. et al.(2023), "Discovering Causal Relations and Equations from Data", *Physics Reports*, Vol.1044, pp.1-68.

Chickering, D. M.(2002a), "Optimal Structure Identification with Greedy Search", *Journal of Machine Learning Research*, 3(Nov), pp.507-554.

Chickering, D. M.(2002b), "Learning Equivalence Classes of Bayesian-Network Structures",

Journal of Machine Learning Research, 2, pp.445-498.

Demiralp, S. and K. D. Hoover(2003), "Searching for The Causal Structure of a Vector Autoregression", *Oxford Bulletin of Economics and Statistics*, Vol.65, pp.745-767.

Diaz, E. et al.(2022), "Inferring Causal Relations from Observational Long-Term Carb on and Water Fluxes Records", *Scientific Reports*, 12(1), 1610, pp.1-10.

Di Capua, G. et al.(2020), "Tropical and Mid-Latitude Teleconnections Interacting With the Indian Summer Monsoon Rainfall: A Theory-Guided Causal Effect Network Approach", *Earth System Dynamics*, 11(1), pp.17-34.

Ebert-Uphoff, I. and Y. Deng(2012), "Causal Discovery for Climate Research Using Graphical Models", *Journal of Climate*, 25(17), pp.5648-5665.

Entner, D. and P. O. Hoyer(2010), "On Causal Discovery from Time Series Data Using FCI", *Proceedings of the European Workshop on Probabilistic Graphical Models*, pp.121-128.

Gao, T. et al.(2022), "IDYNO: Learning Nonparametric DAGs from Interventional Dynamic Data", *Proceedings of Machine Learning Research(PMLR, 2022)*, Vol. 162, pp.6988-7001.

Geiger, D. et al.(1990), "Identifying Independence in Bayesian Networks", *Networks*, 20(5), p.507.

Gerhardus, A. and J. Runge(2020), "High-Recall Causal Discovery for Autocorrelated Time Series with Latent Confounders", *Advances in Neural Information Processing Systems*, 33, pp.12615-12625.

Geweke, J.(1982), "Measurement of Linear Dependence and Feedback Between Multiple Time Series", *Journal of The American Statistical Association*, 77(378), pp.304-313.

Gong, C. et al.(2024), "Causal Discovery from Temporal Data: An Overview and New Perspectives", *ACM Computing Surveys*, 57(4), p.17.

Gong, W. et al.(2022), "Rhino: Deep Causal Temporal Relationship Learning with Hi story-Dependent Noise", arXiv preprint arXiv:2210.14706, pp.1-28.

Granger, C. W.(1969), "Investigating Causal Relations by Econometric Models and

Cross-Spectral Methods", *Econometrica: Journal of the Econometric Society*, pp.424-438.

Handhayani, T.(2023), "An Integrated Analysis of Air Pollution and Meteorological Conditions in Jakarta", *Scientific Reports*, 13(1), 5798, p.1.

Hitchcock, C.(2020), "Causal Models", In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, pp.1-11.

Kretschmer, M. et al.(2016), "Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation", *Journal of Climate*, 29(11), pp.4069-4081.

Li, H., S. Yu, and J. Principe(2023). "Causal Recurrent Variational Autoencoder for Medical Time Series Generation", *In Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7), pp.8562-8570.

Liu, J. and D.Niyogi(2020), "Identification of Linkages Between Urban Heat Island Magnitude and Urban Rainfall Modification by Use of Causal Discovery Algorithms", *Urban Climate*, 33, pp.1-7.

Lorch, L. et al.(2021), "Dibs: Differentiable bayesian structure learning", *Advances in Neural Information Processing Systems*, Vol.34, pp.24111-24123.

Löwe, S. et al.(2022). "Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data", *In Conference on Causal Learning and Reasoning,* pp. 509-525.

Malinsky, D. and P. Spirtes(2018, August), "Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding", *In Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, pp. 23-47.

Marcinkevičs, R. and J. E. Vogt(2021). "Interpretable Models for Granger Causality Using Self-Explaining Neural Networks", *In Proceedings of the International Conference on Learning Representations(ICLR, 2021)*, pp.1-23.

Masson-Delmotte, V. et al.(2021), *Summary for Policymakers*, IPCC, pp.27-28.

McGraw, M. C. and E. A. Barnes(2018), "Memory Matters: A case for Granger Causality in Climate Variability Studies", *Journal of Climate*, 31(8), pp.3289-3300.

Meek, C.(1997), "Graphical Models: Selecting Causal and Statistical Models", Ph.D. thesis, Carnegie Mellon University, pp.19-36.

Mäkelä, J. et al.(2022), "Incorporating Expert Domain Knowledge into Causal Structure Discovery Workflows", *Biogeosciences*, 19(8), pp.2095-2099.

Nauta, M., D. Bucur, and C. Seifert(2019). "Causal Discovery with Attention-Based Convolutional Neural Networks", *Machine Learning and Knowledge Extraction*, 1(1), p.19.

Nogueira, A. R. et al.(2022), "Methods and Tools for Causal Discovery and Causal Inference", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(2): e1449. p.2, pp.2-23.

Nowack, P. et al.(2020), "Causal Networks for Climate Model Evaluation and Constrained Projections", *Nature Communications*, 11(1), pp.1-11.

Pamfil, R. et al.(2020), "DYNOTEARS: Structure Learning from Timeseries Data", *Proceedings of Machine Learning Research(PMLR, 2020)*, Vol.108, pp.1595-1605.

Pearl, J.(2014), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Elsevier, pp.118-138.

Pearl, J.(2018), "Theoretical Impediments to Machine Learning with Seven Sparks From the Causal Revolution", arXiv preprint arXiv:1801.04016. p.7.

Peters, J., D. Janzing, and B. Schölkopf(2013), "Causal Inference on Time Series using Restricted Structural Equation Models", *Advances in Neural Information Processing Systems*, Vol.26, pp.1-8.

Peters, J., D. Janzing, and B. Schölkopf(2017), *Elements of Causal Inference: Foundations and Learning Algorithms(The MIT Press, 2017)*, pp.50-56.

Ramsey, J.(2017), "A Million Variables and More: The Fast Greedy Equivalence Search Algorithm for Learning High-Dimensional Graphical Causal Models, with an Application to Functional Magnetic Resonance Images", *International Journal of Data Science and Analytics*, 3(2), pp.121-129.

Rees, S. L. et al.(2004), "Mass Balance Closure and The Federal Reference Method for PM2.5 in Pittsburgh, Pennsylvania", *Atmospheric Environment*, 38(20), p.3305.

Runge, J., V. Petoukhov, and J. Kurths(2014), "Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models", *Journal of Climate*, 27(2), pp.720-739.

Runge, J. et al.(2015), "Identifying Causal Gateways and Mediators in Complex Spatio-Temporal Systems", *Nature Communications*, 6(1), pp.1-10.

Runge, J., A. Storkey, and F. Perez-Cruz(2018), "Conditional Independence Testing Based on a Nearest-Neighbor Estimator of Conditional Mutual Information", *Proceedings of Machine Learning Research (PMLR, 2018)*, Vol. 84, pp.938-947.

Runge, J. et al.(2019a), "Inferring Causation From Time Series in Earth System Sciences", *Nature Communications*, 10(1), pp.8-9.

Runge, J. et al.(2019b), "Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets", *Science Advances*, 5(11), eaau4996, p.2.

Runge, J.(2020), "Discovering Contemporaneous and Lagged Causal Relations in Auto correlated Nonlinear Time Series Datasets", *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence(UAI)*, Vol.124, pp.1388-1397.

Runge, J., J. Peters, and D. Sontag(2020), "Discovering Contemporaneous and Lagged Causal Relations in Autocorrelated Nonlinear Time Series Datasets", *Proceedings of Machine Learning Research(PMLR, 2020)*, Vol. 124, pp.1388-1397.

Runge, J. et al.(2023), "Causal Inference for Time Series", *Nature Reviews Earth & Environment*, 4(7), pp.488-489.

Schreiber, T.(2000), "Measuring Information Transfer", *Physical Review Letters*, 85(2), p.461.

Shimizu, S. et al.(2006), "Linear Non-Gaussian Acyclic Model for Causal Discovery", *Journal of Machine Learning Research*, 7(10), pp.2003-2030.

Spirtes, P. and C. Glymour(1991), "An Algorithm for Fast Recovery of Sparse Causal Graphs", *Social Science Computer Review*, 9(1), pp.62-72.

Spirtes, P. et al.(2000), "Constructing Bayesian Network Models of Gene Expression Networks from Microarray Data", pp.1-2.

Sugihara, G. et al.(2012), "Detecting Causality in Complex Ecosystems", *Science*, 338

(6106), pp.496-500.

Sun, X., O. Schulte, and G. Liu, P. Poupart(2023), "NTS-NOTEARS: Learning Nonparametric DBNs with Prior Knowledge", *International Conference on Artificial Intelligence and Statistics(AISTATS),* pp.1-23.

Tank, A. et al.(2021), "Neural Granger Causality", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8), pp.4267-4279.

Triacca, U.(2005), "Is Granger Causality Analysis Appropriate to Investigate the Relationship Between Atmospheric Concentration of Carbon Dioxide and Global Surface Air Temperature?", *Theoretical and Applied Climatology*, 81(3), pp.133-135.

Tsamardinos, I., L.E. Brown, and C. F. Aliferis(2006), "The Max-Min Hillclimbing Bayesian Network Structure Learning Algorithm", *Machine learning*, 65(1), pp.31-78.

Wang, Y. et al.(2016), "Chemical Characterization and Source Apportionment of PM 2.5 in a Semi-Arid and Petrochemical-Industrialized City, Northwest China", *Science of the Total Environment*, 573, pp.1031-1032.

Wu, B. et al.(2024). "WeatherGNN: Exploiting Meteo-and Spatial-Dependencies for Local Numerical Weather Prediction Bias-Correction", *InProceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2433-2441.

Xu, C., H. Huang, and S. Yoo(2019). "Scalable Causal Graph Learning Through a Deep Neural Network", *InProceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp.1853-1862.

Zhang, J.(2008), "On The Completeness of Orientation Rules for Causal Discovery in The Presence of Latent Confounders and Selection Bias", *Artificial Intelligence*, 172(16), pp.1873-1896 .

Zhang, R. et al.(2013), "Chemical Characterization and Source Apportionment of PM 2.5 in Beijing: Seasonal Perspective", *Atmospheric Chemistry and Physics*, 13(14), pp.7068-7070.

Zheng, X. et al.(2018), "DAGs with NO TEARS: Continuous Optimization for Structure Learning", *Advances in Neural Information Processing Systems*, Vol.31(Curran

Associates, Inc., 2018), pp.1-9.

Zhang, Y. et al.(2018), “Causal Direction Inference for Air Pollutants Data”, *Computers & Electrical Engineering*, 68, pp.404-405.

[온라인 자료]

에어코리아, “최종확정 측정자료”, https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123, 검색일: 2025.3.25.

기상청 API허브, “방재기상관측(AWS)”, https://apihub.kma.go.kr, 검색일: 2025.3.25.

※ 본 책자는 환경표지 인증을 받은 용지로 인쇄되었습니다.

대기오염물질 간의 인과관계 추정을 위한 인공지능 방법론 활용 현황 분석

KEI 한국환경연구원
Korea Environment Institute

(30147) 세종특별자치시 시청대로 370 세종국책연구단지 B동
전화 044-415-7777 팩스 044-415-7799 www.kei.re.kr

95530
9 791159 804229
ISBN 979-11-5980-422-9