Causal Networks of Infodemiological Data

Modelling Dermatitis

Marco Scutari [1]
scutari@bnlearn.com

Samir Salah

Delphine Kerob

Jean Krutmann

[1] Dalle Molle Institute for Artificial Intelligence (IDSIA)

June 24, 2025

The Problem: Dermatitis, Pollution and Climate Change

An epidemiological problem: understanding the effect of pollution and changing weather patterns on mental and (especially) dermatological conditions.

- Main variables: 3 pollutants (NO_2, SO_2, PM2.5), 3 mental conditions (anxiety, depression, sleep disorders), obesity, dermatitis, weather patterns (temperatures, wind speed, precipitations; both mean and spread).
- Possible confounders: education level, unemployment, income, household size and population density.
- Size: $\approx$53k observations over $\approx$500 US counties and 134 weeks.
- Missing values: between 0% (the conditions) and 55% (pollutants).

Data Sources: Google Trends, NOAA, EPA, US Census

Google

Google COVID-19 Open Data: 400 health conditions, 4 countries (county-level in the US), weekly search frequencies for 2020-2023 normalised by NLP.

Weather stations in 1652 counties with and satellite images.

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

Monitoring stations in 1470 counties with hourly measurements of NOx, SOx, O3, PMx.

Socio-economic data at the population level to avoid confounding.

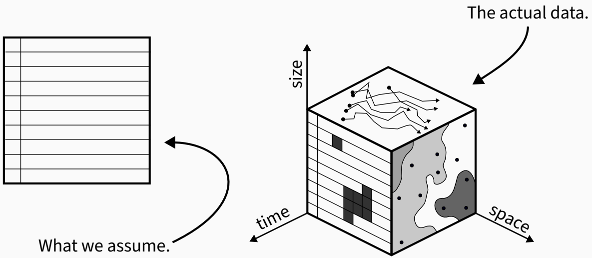
U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

Causal Discovery Meets These Data

Causal discovery means learning a network $\mathcal{G}$ and parameters Θ:

$$\underbrace{\mathrm{P}(\mathcal{G}, \Theta \mid \mathcal{D})}_{\text{learning}} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{\text{structure learning}} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{\text{parameter learning}} .$$

We used to rely on domain experts [5, 6]; now we increasingly apply learning algorithms to data [14].



We broadly know how do causal inference [9] once we have $(\mathcal{G}, \Theta)$.

GIRLS Networks: GLS + IRLS

I propose to learn a dynamic network that encodes a first-order vector auto-regressive process (VAR):

$$X_{it} = f_i(\Pi_{X_{it}}\boldsymbol{\beta}_{it}) + \varepsilon_{it}; \quad \mathrm{E}(\varepsilon_{it}) = 0, \mathrm{COV}(\varepsilon_{it}) = \mathbf{w}_{it}^{\mathrm{T}}\boldsymbol{\Sigma}_i(\mathbf{L}; \xi_i)\mathbf{w}_{it}.$$

where:

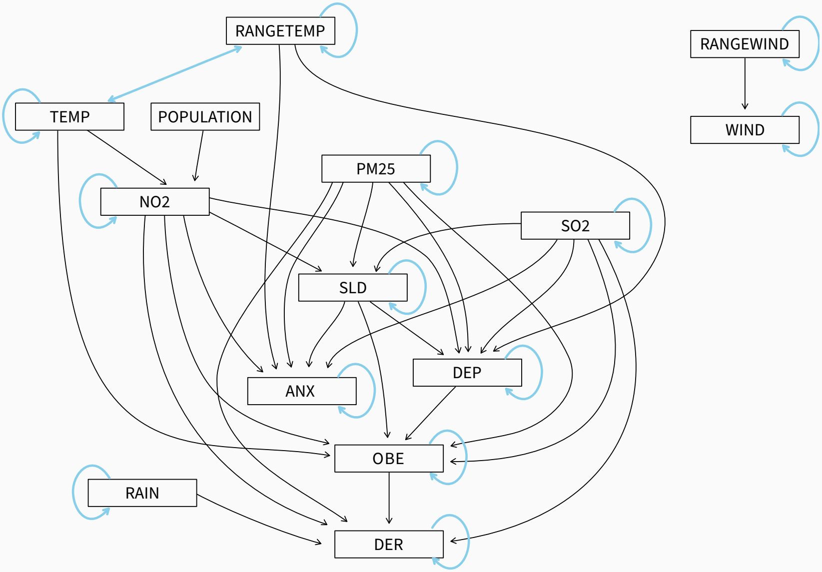
- $\boldsymbol{\Sigma}_i(\mathbf{L}; \xi_i)$ models spatial correlation from location coordinates $\mathbf{L}$ via generalised least squares (GLS); ξ_i model correlation decay.
- The $\mathbf{w}_{it}$ handle
  - heteroscedasticity, via iteratively reweighted least squares (IRLS);
  - missing values, either with 0-1 weights like the PNAL score [3] (if MCAR) or with inverse-probability weights like HC-aIPW [8] (if MAR or MNAR).

Denoising: bagging and model averaging with data-driven threshold [11].

Theoretical Considerations

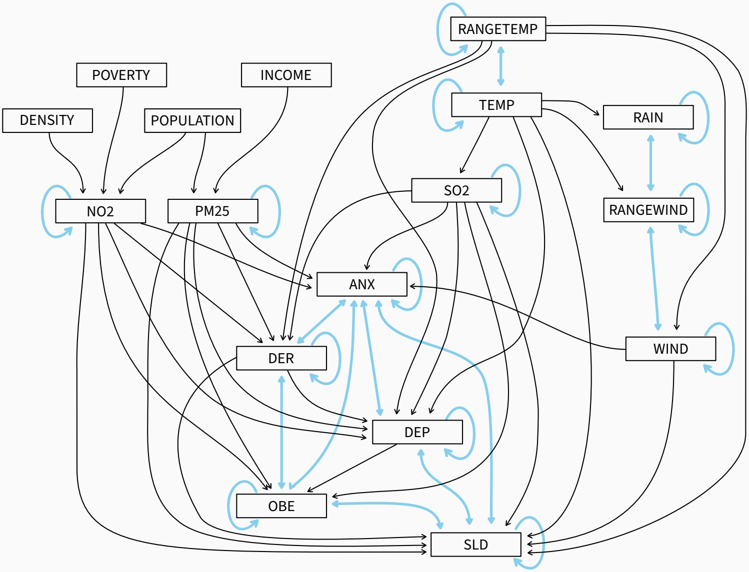
- The causal network is completely identifiable because:
  - Arc directions across time points are fixed.
  - Heteroscedastic residuals + Gaussian noise [7, 12, 13].
  - Even if all $\mathbf{w}_{it} = 1$, the actual residuals $\boldsymbol{\Sigma}_i(\mathbf{L}; \xi_i)^{-1/2}\varepsilon_{it}$ are heteroscedastic unless $\boldsymbol{\Sigma}_i(\mathbf{L}; \xi_i) \propto \mathbf{I}_n$.
- The causal network can be statistically validated using:
  - Autocorrelation tests at different lags in each location.
  - Moran's I [2] at each time point, and fit variograms to explore the proportion of variance attributable to spatial structure [10].
  - Bartlett's heterogeneity test [1] on $\boldsymbol{\Sigma}_i^{-1/2}\varepsilon_{it}$.
- Causal inference over time and space via σ-calculus [4].
- $\boldsymbol{\Sigma}_i(\mathbf{L}; \xi_i)$ can accommodate irregularly spaced locations.

My Baseline Model (Looks Very Wrong)



Does not pass any of the tests, predictive accuracy $R^2 \approx 0.70 - 0.75$.

My Final Model (Looks the Best So Far)



$\text{BF} \geqslant 3400$, passes all the tests, predictive accuracy is still $R^2 \approx 0.70 - 0.75$.

Causal Inference: What Conclusions Can We Draw?

- What is the relative impact of the direct risk factors?

  ANX (0.574), NO_2 (0.339), OBE (0.077), PM2.5, RANGETEMP, SO_2 (0.01).

- What proportion of pollution effects is mediated?

  PM2.5, NO_2 and SO_2 change by 0.54x, 0.93x and 0.56x.

- What proportion of weather effects is mediated?

  TEMP/RANGETEMP, WIND/RANGEWIND, RAIN change by 0.29x, 0.38x, 0.02x

- What would be the impact of tightening environmental regulations?

  PM2.5 $12 \to 9\mu g/m^3$ for 1 year: -18% DER. $12 \to 8\mu g/m^3$: -21% DER.

- How long must a cold spell last before dermatitis increases?

  DER +5% after 4 weeks.

Conclusions

- Causal discovery makes simplifying assumptions that are too strong for infodemiological (and epidemiological) data.
- Classical statistics gives us flexible and scalable tools to model complex structures in the data.
- State-space data, mixed variable types, missing values, population structure, non-stationarity: we can deal with them!
- GIRLS produces causal networks we can trust (because we validate their assumptions) and that generalise well (in time and space).

That's all!

Happy to discuss in more detail.

References I

- M. S. Bartlett.
  Properties of Sufficiency and Statistical Tests.
  Proceedings of the Royal Society of London, Series A, 160(901):268–282, 1937.
- R. S. Bivand and D. W. S. Wong.
  Comparing Implementations of Global and Local Indicators of Spatial Association.
  Test, 27(3):716–748, 2018.
- T. Bodewes and M. Scutari.
  Learning Bayesian Networks From Incomplete Data with the Node-Averaged Likelihood.
  International Journal of Approximate Reasoning, 138:145–160, 2021.
- J. Correa and E. Bareinboim.
  A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments.
  Proceedings of the AAAI Conference on Artificial Intelligence, 34(06):10093–10100, 2020.
- M. J. Druzdel and L. C. van der Gaag.
  Elicitation of Probabilities for Belief Networks: Combining Qualitative and Quantitative Information.
  In *UAI*, pages 141–148, 1995.

- M. J. Druzdel and L. C. van der Gaag.
  Building Probabilistic Networks: "Where Do the Numbers Come From?".
  IEEE Transactions on Knowledge and Data Engineering, 12(4):481–486, 200.
- B. Duong and T. Nguyen.
  Heteroscedastic Causal Structure Learning.
  In *European Conference on Artificial Intelligence*, pages 598–605, 2023.
- Y. Liu and A. C. Constantinou.
  Greedy Structure Learning From Data That Contain Systematic Missing Values.
  Machine Learning, 111(10):3867–3896, 2022.
- J. Pearl and D. Mackenzie.
  The Book of Why: The New Science of Cause and Effect.
  Basic Books, 2018.
- J. Pinheiro and D. Bates.
  Mixed-Effects Models in S and S-Plus.
  Springer, 2000.

References III

- M. Scutari and R. Nagarajan.
  On Identifying Significant Edges in Graphical Models of Molecular Networks.
  Artificial Intelligence in Medicine, 57(3):207–217, 2013.
- S. Xu, O. A. Mian, A. Marx, and J. Vreeken.
  Inferring Cause and Effect in the Presence of Heteroscedastic Noise.
  Icml, 162:24615–24630, 2022.
- N. Yin, T. Gao, Y. Yu, and Q. Ji.
  Effective Causal Discovery Under Identifiable Heteroscedastic Noise Model.
  In *AAAI Conference on Artificial Intelligence*, pages 16486–16494, 2024.
- A. Zanga, E. Ozkirimli, and F. Stella.
  A Survey on Causal Discovery: Theory and Practice.
  Int. J. Approx. Reason., 151:101–129, 2022.