Causal Modelling for Environmental Epidemiology

State-Space Networks from Incomplete Data

Marco Scutari
scutari@bnlearn.com

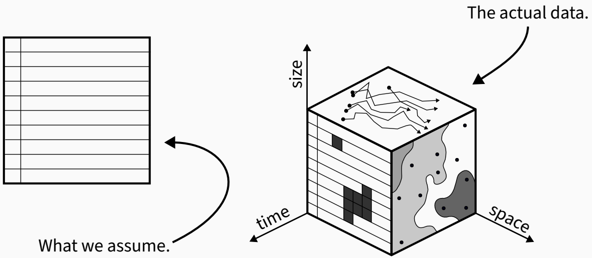
Dalle Molle Institute for
Artificial Intelligence (IDSIA)

February 12, 2025

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 [8, 9]; now we increasingly apply learning algorithms to data [22].



We broadly know how do causal inference [12] once we have $(\mathcal{G}, \Theta)$.

- Combinations of comorbidities are often impossible to study in a classical environmental epidemiology study.
- However, we have massive amounts of Internet-generated data user-contributed health-related content.
- Infodemiology (short for "information epidemiology") draws on this data to replace epidemiological data and improve public health.

We need to assume:

- a non-negligible association between the frequency of online mentions of specific diseases and their incidence;
- a broad coverage of the population.

A motivating example: understanding the effect of pollution and changing weather patterns on mental and 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).

Following up from a previous infodemiology study [14].

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

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.

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

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

Nuisance Parameters, Parameters of Interest

A causal network has two components: the graph $\mathcal{G}$ and the parameters Θ. Causal inference defines queries using $\mathcal{G}$:

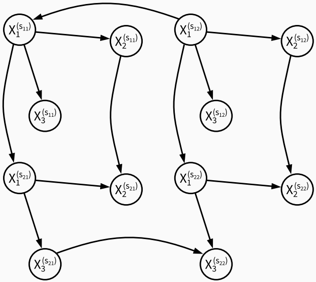
- Conditional independence, via d-separation.
- Intervention, via mutilation.
- Counterfactual, via the twin network.

Our ability to answer scientific questions using the causal network rests on having the right nodes in the network. Without them, we cannot even formulate our question.

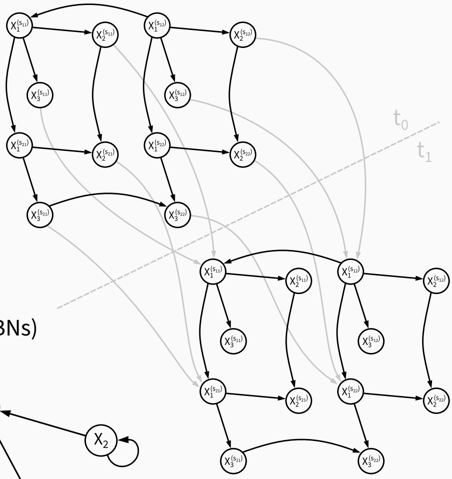
- The dimensions we use in the queries (interest) should be represented as nodes.
- The dimensions we do not (nuisance) should be represented as parameters in the local distributions.

Network Structures: Time vs Space vs State-Space

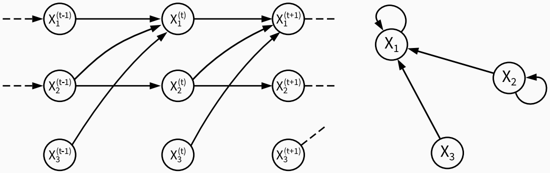
Spatial Structure



State-Space Structure



Temporal Structure (dynamic BNs)



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 [6] (if MCAR) or with inverse-probability weights like HC-aIPW [11] (if MAR or MNAR).

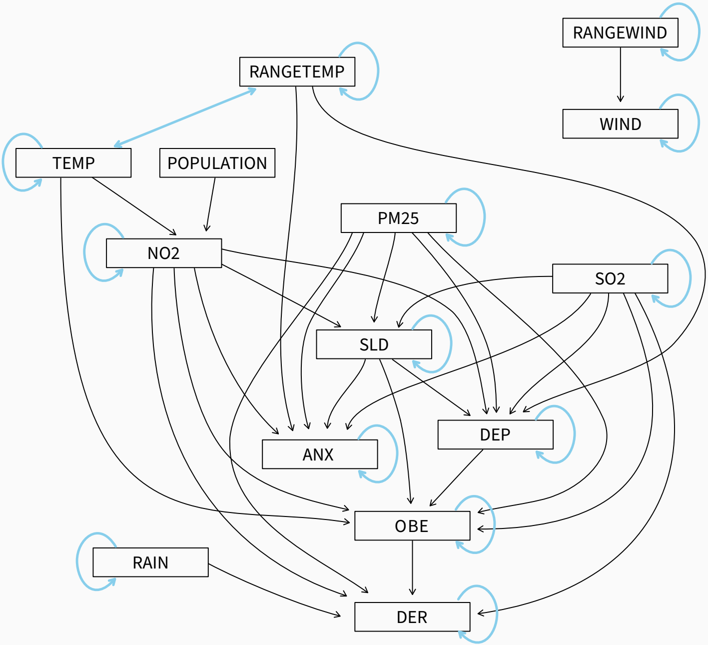
Denoising: bagging and model averaging with data-driven threshold [16].

Code: The R Implementation

```
# provide an initial estimate.
model = nlme::gls(as.formula(f), data = full, method = "ML",
            cor = nlme::corExp(value = args$spatial[, node],
                    form = ~ LAT + LON | WEEK, nugget = TRUE, fixed = TRUE))
old.logl = as.numeric(nlme:::logLik.gls(model), REML = FALSE)

# iteratively reweighted least squares.
for (iter in 1:(args$irls.max.iter)) {

  # compute the per-state variances...
  weights = sapply(levels(full[, "STATE"]), function(s) var(resid(model)[full[, "STATE"] == s]) )
  for (i in seq(nrow(full)))
    full[i, "w"] = weights[names(weights) == full[i, "STATE"]]
  # ... and re-estimate the model.
  model = nlme::gls(as.formula(f), data = full, method = "ML",
              cor = nlme::corExp(value = args$spatial[, node],
                      form = ~ LAT + LON | WEEK, nugget = TRUE, fixed = TRUE),
              weights = nlme::varFixed(~ w))
  new.logl = as.numeric(nlme:::logLik.gls(model, REML = FALSE))

  # check convergence.
  if (isTRUE(all.equal(old.logl, new.logl)))
    break
  else
    old.logl = new.logl

}#FOR
```

- The causal network is completely identifiable because:
  - Arc directions across time points are fixed.
  - Heteroscedastic residuals + Gaussian noise [10, 18, 19].
  - Even if all $\mathbf{w}_{it} = 1$, the actual residuals $\mathbf{\Sigma}_i(\mathbf{L}; \xi_i)^{-1/2}\varepsilon_{it}$ are heteroscedastic unless $\mathbf{\Sigma}_i(\mathbf{L}; \xi_i) \propto \mathbf{I}_n$.
- The causal network can be statistically validated using:
  - Autocorrelation tests at different lags in each location.
  - Moran's I [5] at each time point, and fit variograms to explore the proportion of variance attributable to spatial structure [13].
  - Bartlett's heterogeneity test [3] on $\mathbf{\Sigma}_i^{-1/2}\varepsilon_{it}$.
- Causal inference over time and space via σ-calculus [7].
- $\mathbf{\Sigma}_i(\mathbf{L}; \xi_i)$ can accommodate irregularly spaced locations.

Incomplete Data + Time (Looks Very Wrong)

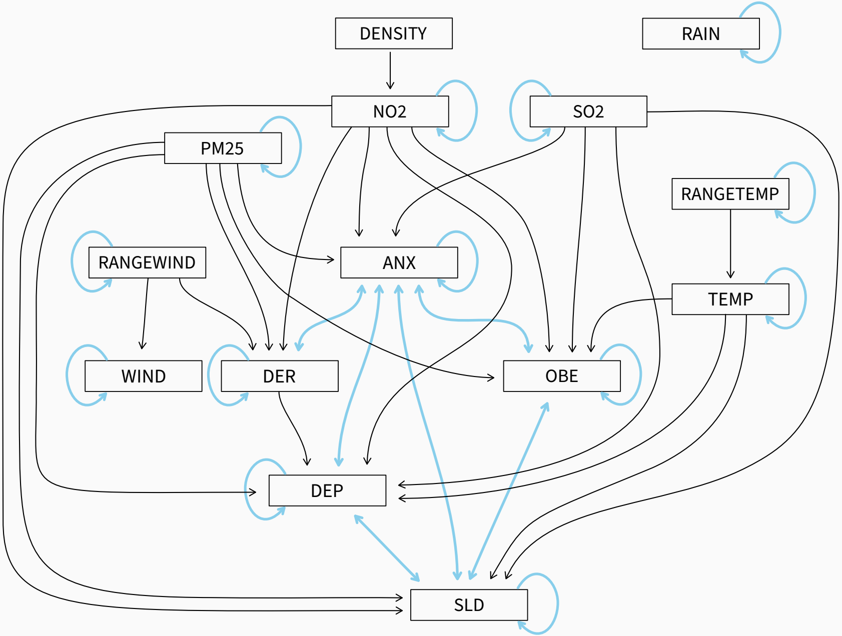


Residuals are largely free from autocorrelation! ✓

	lag 1	lag 2	lag 3	lag 4
ANX	0.008	0.000	0.000	0.008
DEP	0.000	0.000	0.000	0.000
DER	0.032	0.000	0.000	0.000
OBE	0.000	0.000	0.000	0.000
SLD	0.078	0.007	0.007	0.000

But they are full of spatial correlation! ✗

	proportion
ANX	0.468
DEP	0.397
DER	0.738
OBE	0.579
SLD	0.381


DENSITY
RAIN
NO2
SO2
PM25
RANGETEMP
RANGEWIND
ANX
TEMP
WIND
DER
OBE
DEP
SLD

The causal network fits the data much better!

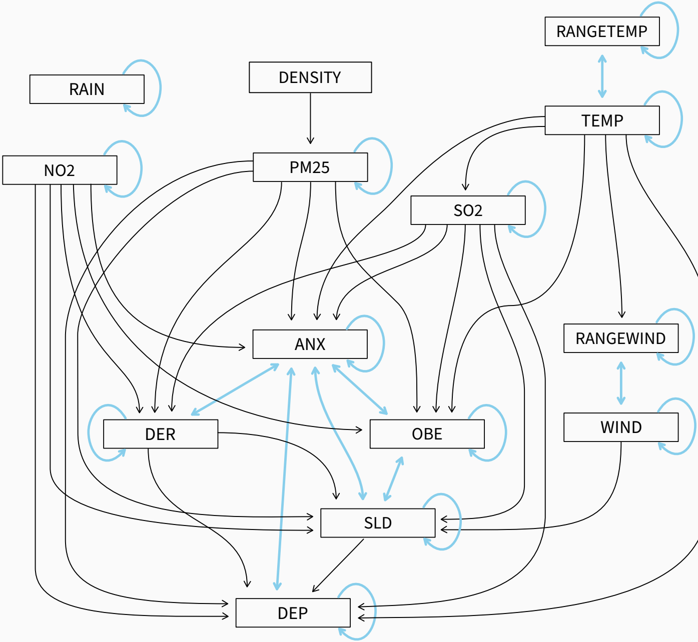
$$\log \mathrm{BF} = (-39.77) - (-44.33) = 4.56 \qquad \Longrightarrow \qquad \mathrm{BF} = 95.92.$$

But the residuals are markedly heteroscedastic!

	p-value
ANX	8×10^{-182}
DEP	9×10^{-217}
DER	0
OBE	8×10^{-100}
SLD	1×10^{-147}

One more time...

Incomplete Data + Space + Time + Heteroscedasticity (Looks OK)



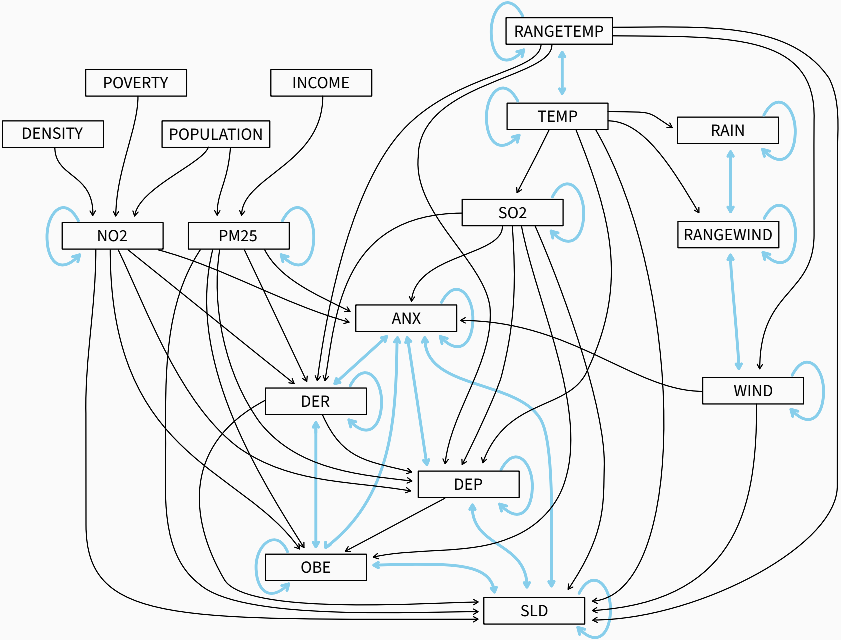
The causal network fits the data much better! ✅

$$\log \mathrm{BF} = (-36.55) - (-39.77) = 3.22 \qquad \Longrightarrow \qquad \mathrm{BF} = 25.$$

The weighted residuals are completely homoscedastic! ✅

	p-value
ANX	1
DEP	1
DER	1
OBE	1
SLD	1

Some arcs are obviously missing, reduce sparsity a bit...


RANGETEMP
POVERTY
INCOME
TEMP
RAIN
DENSITY
POPULATION
SO2
RANGEWIND
NO2
PM25
ANX
WIND
DER
DEP
OBE
SLD

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

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

  DER +5% after 4 weeks.

- Using GLMs is straightforward because we can estimate them with IRLS, which we already use, and allows for discrete variables.
- Bringing change point detection from the literature on VARs [1, 2].
- A more robust handling of missing values, proving that PNAL works under MAR or leveraging my students' work on causal discovery under MNAR [4, 20, 21].
- Incorporating random effects to separate global and local effects (in time/space/sub-populations) from my previous work [15, 17].

- Causal discovery makes simplifying assumptions that are too strong.
- Classical statistics gives us flexible and scalable tools to model complex structures in the data.
- Pose the research question first: model the data dimensions you need graphically and hide the rest in the local distributions.
- State-space data, mixed variable types, missing values, population structure, non-stationarity: we can deal with them!

Acknowledgements

Alice Bernasconi
Alessio Zanga
Fabio Stella
Università degli Studi di Milano-Bicocca

Samir Salah
Delphine Kerob
L'Oréal, La Roche-Posay

Jean Krutmann
Leibniz Research Institute for Environmental Medicine
Medical Faculty, Heinrich Heine University

My former student, Tjebbe Bodewes (University of Oxford).

That's all!

Happy to discuss in more detail.

References I

- P. Bai, A. Safikhani, and G. Michailidis.
  Multiple Change Points Detection in Low Rank and Sparse High Dimensional Vector Autoregressive Models.
  IEEE Transactions on Signal Processing, 68:3074–3089, 2020.
- P. Bai, A. Safikhani, and G. Michailidis.
  Multiple Change Point Detection in Reduced Rank High Dimensional Vector Autoregressive Models.
  Journal of the American Statistical Association, 118(544):2776–2792, 2023.
- M. S. Bartlett.
  Properties of Sufficiency and Statistical Tests.
  Proceedings of the Royal Society of London, Series A, 160(901):268–282, 1937.
- A. Bernasconi, A. Zanga, P. J. F. Lucas, M. Scutari, and F. Stella.
  Towards a Transportable Causal Network Model Based on Observational Healthcare Data.
  In *AIxIA*, pages 67–82, 2023.
- R. S. Bivand and D. W. S. Wong.
  Comparing Implementations of Global and Local Indicators of Spatial Association.
  TEST, 27(3):716–748, 2018.

References II

- 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, 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.
- M. Scutari, D. Kerob, and S. Salah.
  Inferring Skin-Brain-Skin Connections from Infodemiology Data Using Dynamic Bayesian Networks.
  Scientific Reports, 14:10266, 2024.
- M. Scutari, C. Marquis, and L. Azzimonti.
  Using Mixed-Effect Models to Learn Bayesian Networks from Related Data Sets.
  Proceedings of Machine Learning Research (PGM 2022), 186:73–84, 2022.

- M. Scutari and R. Nagarajan.
  On Identifying Significant Edges in Graphical Models of Molecular Networks.
  Artificial Intelligence in Medicine, 57(3):207–217, 2013.
- L. Valleggi, M. Scutari, and F. M. Stefanini.
  Learning Bayesian Networks with Heterogeneous Agronomic Datasets via Mixed-Effect Models and Hierarchical Clustering.
  Engineering Applications of Artificial Intelligence, 131:107867, 2024.
- 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*, volume 38, pages 16486–16494, 2024.
- A. Zanga, A. Bernasconi, P. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, and F. Stella.
  Risk Assessment of Lymph Node Metastases in Endometrial Cancer Patients: A Causal Approach.
  In *AIxIA*, pages 1–15, 2022.

- A. Zanga, A. Bernasconi, P. J. F. Lucas, H. Pijnenborg, C. Rejinen, M. Scutari, and F. Stella.
  Causal Discovery with Missing Data in a Multicentric Clinical Study.
  In *AIME*, pages 40–44, 2023.
- A. Zanga, E. Ozkirimli, and F. Stella.
  A Survey on Causal Discovery: Theory and Practice.
  Int. J. Approx. Reason., 151:101–129, 2022.