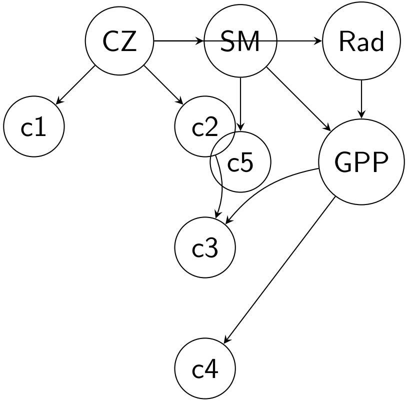
Causal Discovery for Earth System Sciences

Section 1 – What is Causal Discovery?

Emiliano Díaz Salas Porras

Image and Signal Processing @ Universitat de València

Example DAG: Climate and Productivity



- What variables would you use to predict **GPP**?
- What are the possible sources of association between **GPP** and the rest of the variables?
- Is Reichenbach principle right?

Chocolate Consumption vs Nobel Laureates

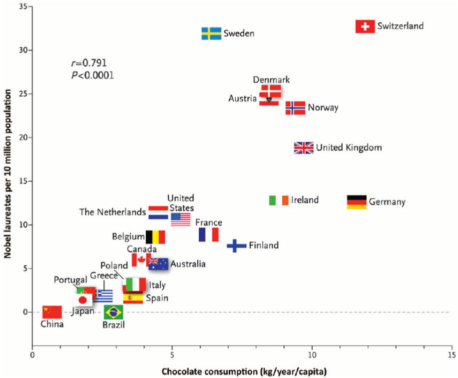


Figure: Source: Messerli (2012)

Correlation is not causation:

- Spurious relationship
- Likely driven by a confounding variable (e.g., national wealth)

Quinn et al. (1999): Myopia and Night Light Exposure

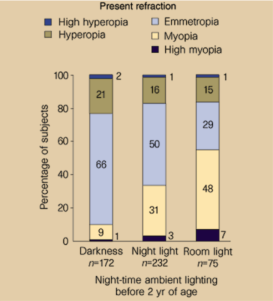
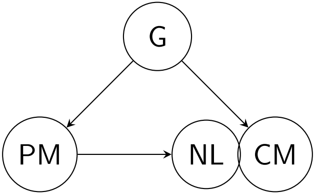


Figure: Prevalence of myopia in children by lighting conditions during sleep in the first two years of life.
Source: Quinn et al., Nature (1999)

Initial Conclusion: Exposure to light during sleep may increase risk of myopia.

Unobserved Confounding: Myopia and Night Light



Legend:

- G: Genetic predisposition
- PM: Parent Myopia
- CM: Child Myopia
- NL: Night Light use

Key Idea: Unobserved common cause (G) induces spurious association between NL and CM. [Zadnik, et al 2000]

Simpson's Paradox: Kidney Stones Example

Study: Patients treated for kidney stones with:

- **Treatment A**
- **Treatment B**

Observed Success Rates:

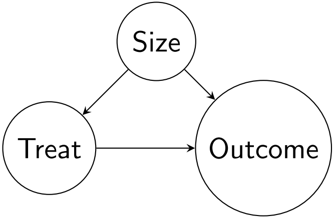
Treatment	Small Stones	Large Stones	All
A	81/87 (93%)	192/263 (73%)	273/350 (78%)
B	234/270 (87%)	55/80 (69%)	289/350 (83%)

Table: Source: Charig. et al 1986

Paradox: Overall B is better, but if we go by size-specific groups then A is better.

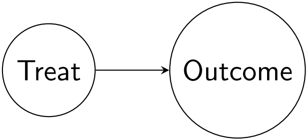
Simpson's Paradox: The Role of the Causal DAG

DAG 1: Size as a Confounder



Stratify by Size to remove confounding.

DAG 2: No Confounding



Overall association reflects causal effect.

Key Insight: Whether we trust the overall result or the subgroup-specific results depends on the true **causal DAG**.

The SEASFIRE Dataset

SEASFIRE: Satellite-based Earth System data for analyzing the drivers of fire occurrence.

Coverage:

- **Spatial resolution:** 0.25° x 0.25° grid (approx. 25 km)
- **Temporal resolution:** 8-day
- **Time span:** 2001–2020
- **Global extent**

Data Types:

- Fire occurrence
- Vegetation & productivity
- Climate variables (temperature, radiation, etc.)
- Hydrology & soil moisture
- Anthropogenic pressure
- Drought: added from EDIT- copernicus

Selected Variables:

Variable	Description
Fire_CCI	Burned area fraction per grid cell
GPP	Gross Primary Productivity
LAI	Leaf Area Index
SoilMoist	Root-zone soil moisture
T2M	2-meter air temperature
Radiation	Surface shortwave radiation
Precip	Precipitation (monthly)
PopDensity	Population density
LandUse	Dominant land cover class

Causal Discovery Questions with SEASFIRE + Droughts

Potential Causal Discovery Questions:

- **What are the direct causes of fire occurrence?**
  - Role of drought events vs. gradual dryness (e.g., Drought Code)
  - Climatic vs. anthropogenic drivers
- **Do different types of drought (meteorological, hydrological) have distinct causal effects on fire?**
  - Can we disentangle their roles across ecosystems?
- **How does vegetation productivity (e.g., GPP) mediate the impact of drought on fire?**
- **Is the effect of drought on fire modulated by land cover or population density?**
- **Are fire regimes in different regions driven by the same causal structure?**

Two SEASFIRE Datasets for Causal Discovery

To explore different causal discovery settings, we define two datasets:

1. IID Dataset (Static Causal Discovery)

- Each data point: A spatial grid cell at a specific month
- Features: Climate, vegetation, fire, anthropogenic variables at that timestep
- Goal: Discover static causal relationships assuming i.i.d. samples
- Example: Does GPP cause fire occurrence across space?

2. Time Series Dataset (Temporal Causal Discovery)

- Each data point: A sequence of monthly observations per grid cell
- Features: Lagged variables, temporal dependencies
- Goal: Discover causal relations with temporal ordering and memory
- Example: Does drought event at $t-2$ affect fire occurrence at t?

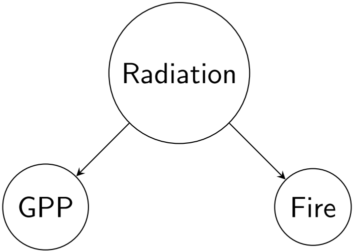
Both datasets use the same spatial grid and variables, but differ in assumptions and methods.

Challenges in Causal Discovery with SEASFIRE

1. **Spurious Associations**
   - *Example:* Fire occurrence and vegetation productivity (GPP) may be correlated due to shared seasonality with solar radiation.
   - **Solution:** Need to condition on confounders (e.g., radiation) or remove seasonal trends.
2. **Simpson's Paradox**
   - *Example:* Population density appears negatively associated with fire at global scale.
   - But stratifying by land cover (e.g., forest, cropland) reveals the opposite.
   - **Lesson:** Aggregated patterns can reverse when stratified.
3. **Unobserved Confounding**
   - *Example:* Land management practices or local policies may affect both vegetation and fire, but are not observed.
   - **Consequence:** Causal estimates may be biased.

Spurious Association via Common Cause (Seasonality)

Example: GPP and fire may appear correlated across months.



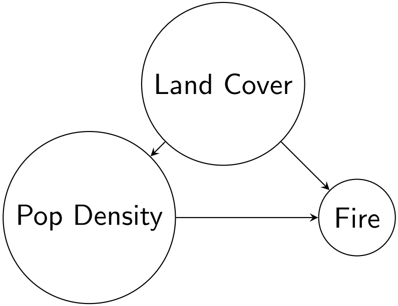
Interpretation:

- Solar radiation varies seasonally and affects both productivity and fire risk.
- GPP and Fire are spuriously correlated due to a shared common cause.

Fix: Condition on seasonality or radiation when estimating causal links.

Simpson's Paradox: Land Cover as a Confounder

Example: Population density and fire appear negatively correlated overall.



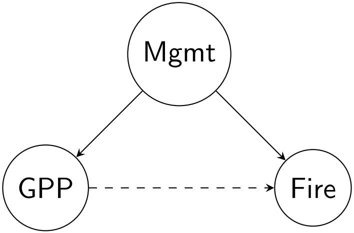
Interpretation:

- Urban areas: high population, low fire
- Forest/cropland: lower population, higher fire
- Aggregated view hides these group-specific trends

Lesson: Conditioning on land cover reveals the true direction of association.

Unobserved Confounding: Policy and Land Management

Example: Fire and vegetation condition may be influenced by unseen management decisions.



Interpretation:

- Land management (e.g., fire suppression, deforestation) is often unobserved.
- It may affect both vegetation productivity and fire occurrence.
- GPP–Fire association could be spurious without accounting for policy.

Implication: Causal sufficiency may not hold — need methods that handle latent confounders (e.g., FCI, LPCMCI).