Learning causal drivers of PyroCb

Emiliano Diaz Salas Porras, Gherardo Varando, Fernando Iglesias-Sua
Gustau Camps-Valls, Kenza Tazi, Kara D. Lamb, Duncan Watson-Parris, As
Braude, Daniel Okoh, Paula Harder, Nis Meinert

Machine Learning for Science

http://isp.uv.es @isp_uv_es
emiliano.diaz@uv.es

Image Signal Processing - IS

Motivation:

Causal discovery in Earth System science: **no experiments possible** on glol
scale, but different regimes act as “natural” interventions to create **experimer
like data**.

Goal:

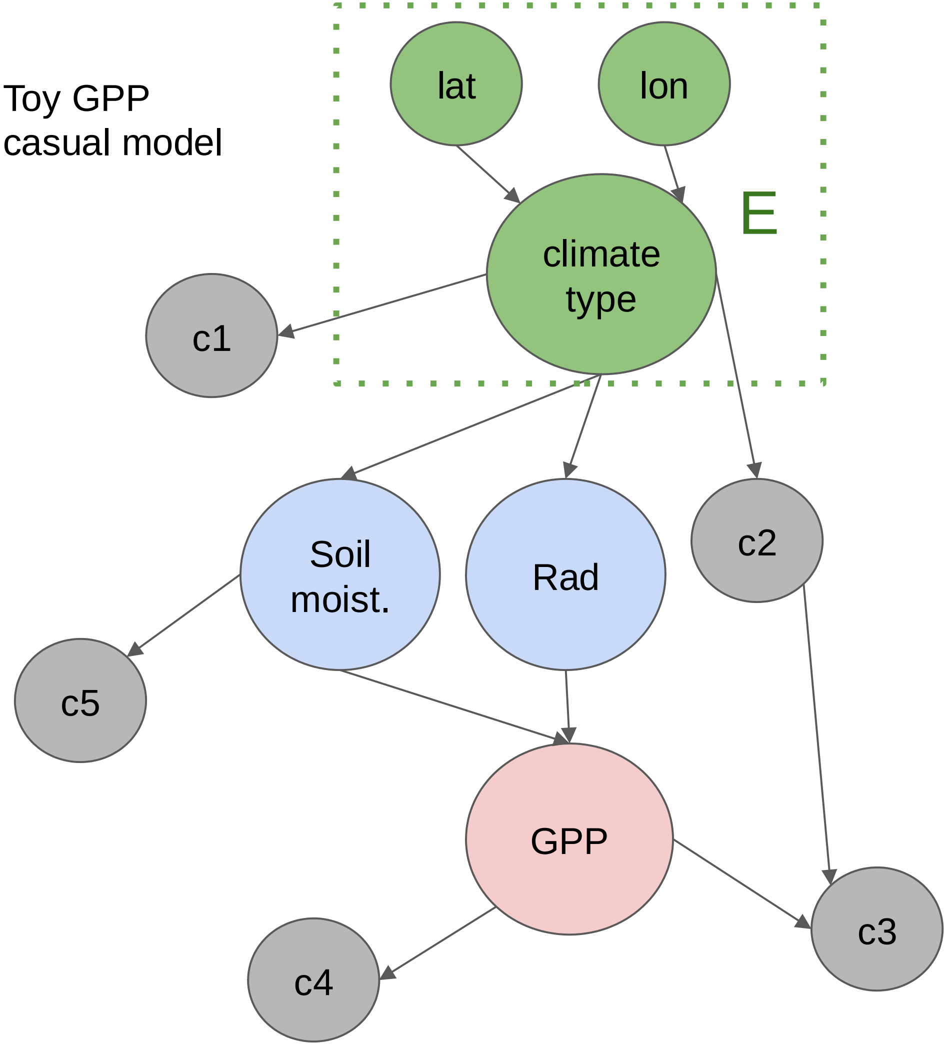
Can we use this heterogeneity to find causal drivers of phenomenon such as
extreme wildfires (PyroCb) and Photosynthesis (GPP).

Use cases:

Photosynthetic activity (toy model): can we separate direct causes of GPP from correlated variables (effects, shared common causes, indirect causes)?

PyroCb ocurrence (“real world” data): why do som large fires generate pyroCb and others do not?

Toy GPP
casual model



Invariant Causal Prediction (ICP) [Peters, J. et al 2016]:

Minimal conditional independence condition:

GPP independent of environment E given direct causes S*={soil moist., rad}

This is the minimal set S where this conditional independence holds


3. Clouds
4. Thunderstorm
2. Plume clouds
5. Downburst + lightning
1. Smoke plume
6. Unpredictable fire behaviour + new fires

28 variables total

atmospheric

fuel

thermal

~ 100 pyroCb events comprising ~6k hourly observations in North America and Australia

Variable	Description	Sensitive to
$ch1$	0.47 µm	smoke, haze
$ch2$	0.64 µm	terrain type
$ch3$	0.86 µm	vegetation
$ch4$	3.9 µm	thermal emissions & cloud ice crysta
$ch\{5,6\}$	$\{11.2, 13.3\}$ µm	thermal emissions & cloud opacity
$\{u,v\}$	$\{u,v\}$ comp. of wind at 250 hPa	upper-level dynamics which influenc motion
$\{u,v\}10$	10 m $\{u,v\}$ component of wind	change in fire intensity and spread
$fg10$	10 m gusts since prev. post-processing	(same as above)
blh	boundary layer height	height of turbulent air at the surface
$cape$	convective available potential energy	energy for air to ascend into atmospl
cin	convective inhibition	energy that will prevent air from risi
z	geopotential	energy needed for air to ascend int sphere as a function of altitude
$\{slhf, sshf\}$	surface {latent, sensible} heat flux	heat released or absorbed {from, neg phase changes
w	surface vertical velocity	ascent speed of the plume from the w
$cv\{h,l\}$	fraction of {high, low} vegetation	available fuel for the wildfire
$type\{H,L\}$	type of {high, low} vegetation	(same as above)
$r\{650,750,850\}$	rel. humidity at {650,750,850} hPa	condensation of vapour into clouds

From Tazi, K., et al 2022

ICP algorithm

To find the causes of Y:

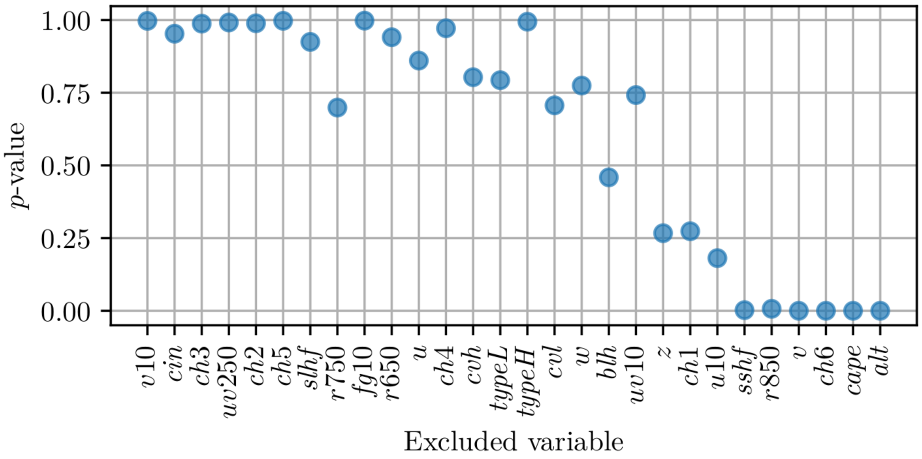
1. For each subset S_i of candidate predictors perform conditional independence test H_i:

$$\boxed{Y \perp\!\!\!\perp E \mid X_{S^*}.}$$

2. Take **intersection** of S_i where H_i is not rejected as causal predictors.

ICP: 28 variables in pyroCb dataset -> 250 million tests!

Greedy ICP: start with all candidate predictors and exclude one at a time -> 406 tests

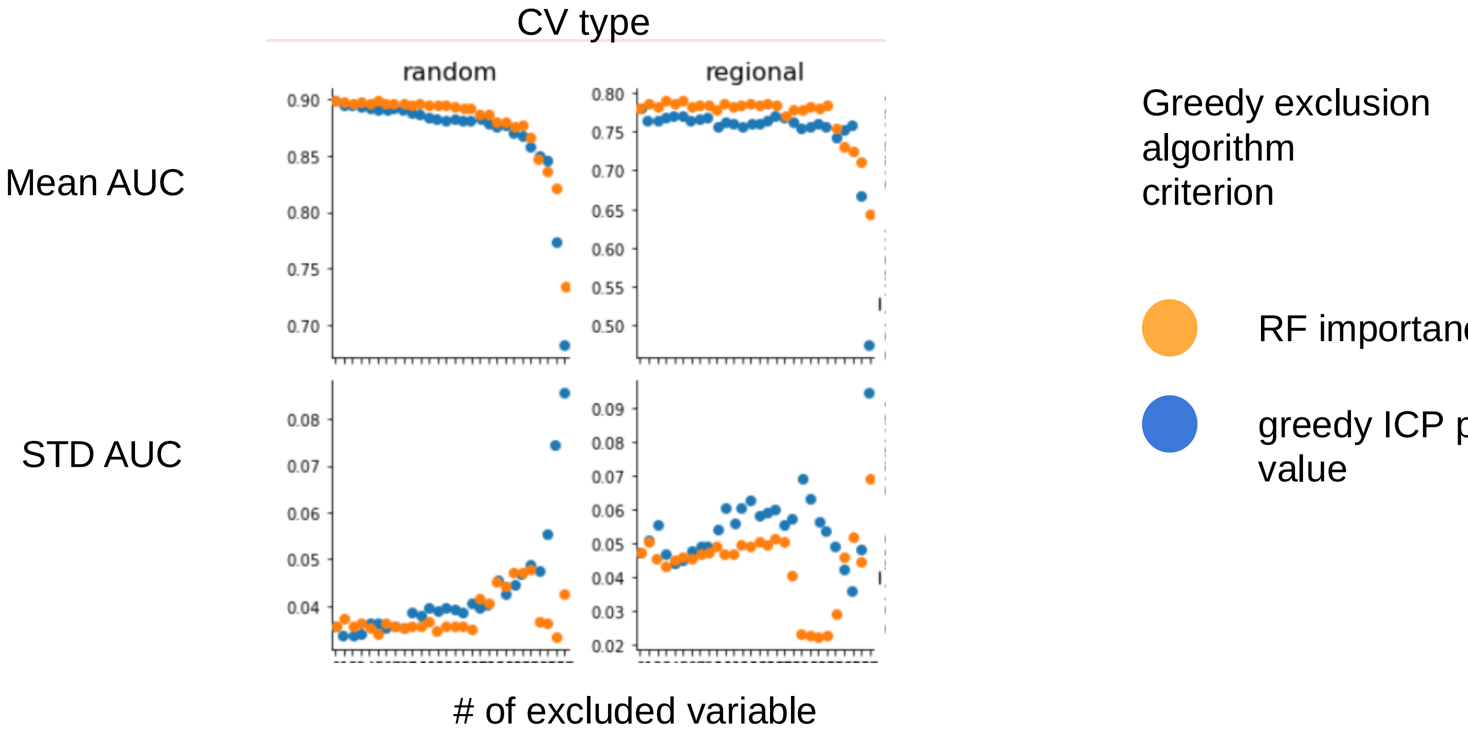


Conditional independence test based on difference between reduced (Random Forest) model (excluding E) and full model (including E).

Use DeLong, E.R, et al (1988) test for comparing AUCs

Plot shows p-value of H_i: $\boxed{Y \perp\!\!\!\perp E \mid X_{S^*}.}$ as we exclude variables with Greedy ICP

	variable	**proxy for…**
alt	altitude	energy needed to breach atmosphere
sshf	surface sensible heat flux	energy transferred by fire
ch6	13.3 μm reflectance	
r850	relative humidity at 850 hPa	potential for cloud formation in atmosphere
v	component of wind at 250 hPa	atmospheric instability
cape	convective available potential energy	


CV type
random
regional
Mean AUC
STD AUC
# of excluded variable
Greedy exclusion algorithm criterion
RF importanc
greedy ICP p value
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02

Limitations of the ICP approach

ICP :

- number of hypothesis tests needed very large
- Dependence among predictors results in empty set inference

Greedy ICP

- order dependent- variables chosen for exclusion in beginning affect inference.

Invariant Causal Features

Can we use Neural Networks to:

1. learn a causal representation (get around ICP and Greedy ICP problems)

2. Learn latent environment -> identify our “quasi-experiments” (climatic type in GPP toy model)

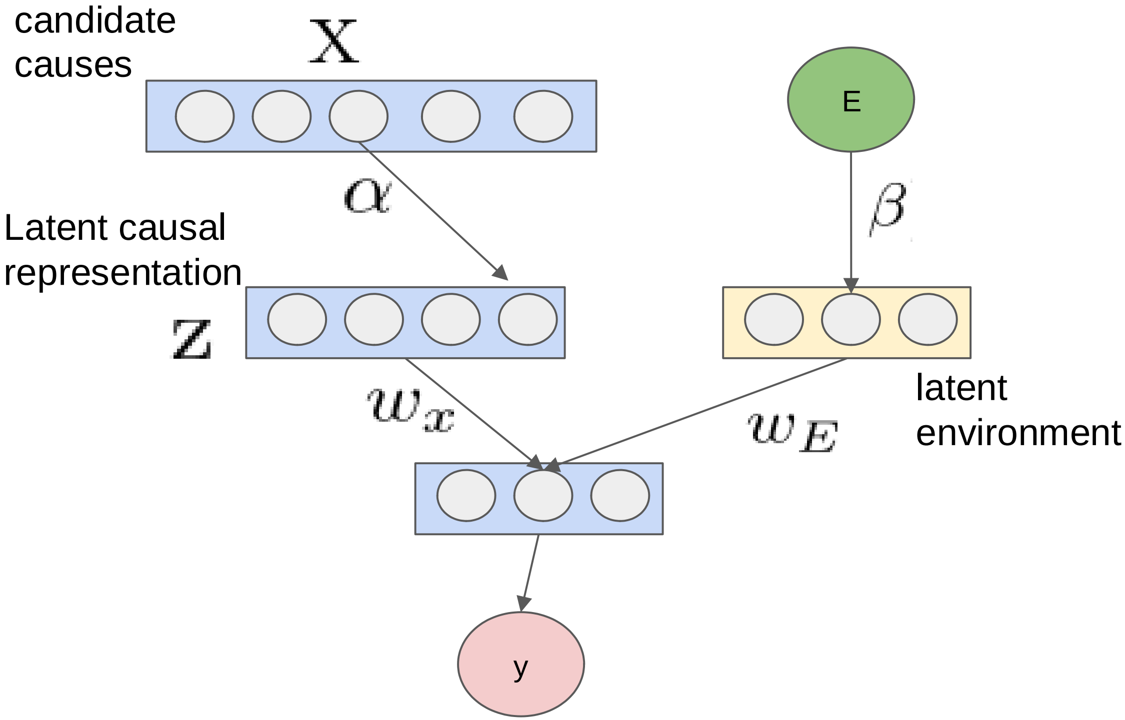
$$\mathrm{L}(\mathrm{y},\ \mathrm{x};\ \mathrm{w}_E, w_x, \alpha, \beta) = L_1(y, x; w_E, w_x, \alpha, \beta) + \lambda || \nabla_{w_E} L_1(y, x; w_E, w_x, \alpha, \beta)$$

Prediction Loss:

First term the usual MSE or Cross Entropy loss

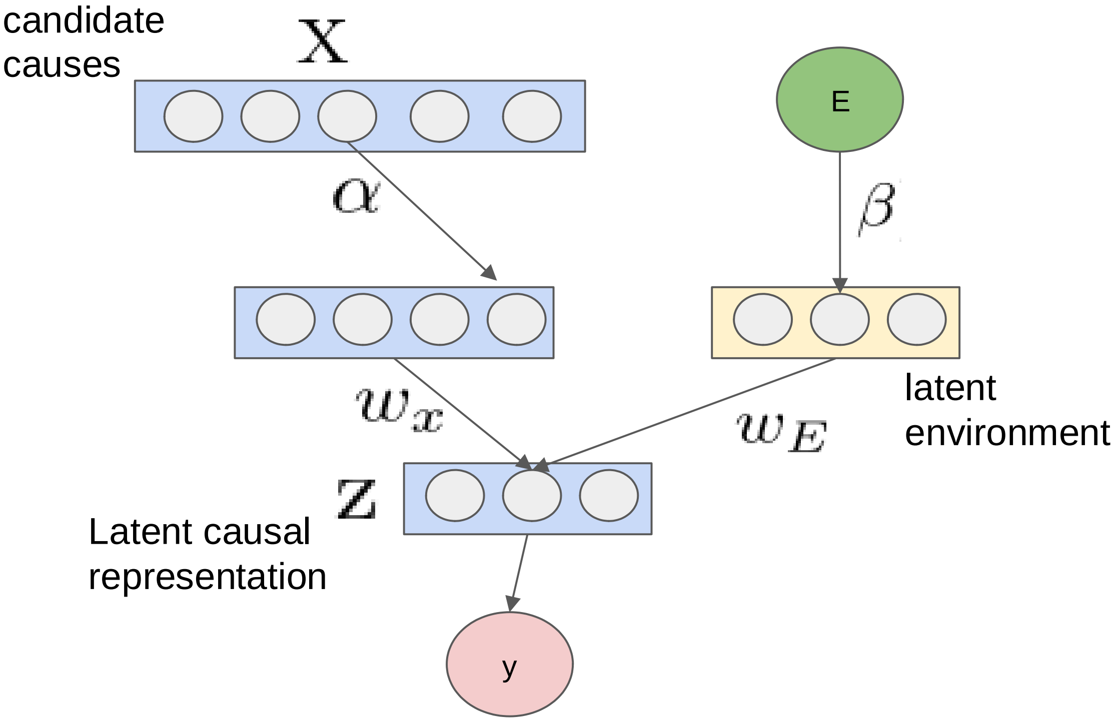
Second term in loss conditional independence proxy

$$\mathrm{L}(\mathrm{y},\, \mathrm{x};\, \mathrm{w}_E, w_x, \alpha, \beta) = L_1(y, x; w_E, w_x, \alpha, \beta) + \lambda||\nabla_{w_E} L_1(y, x; w_E, w_x, \alpha, \beta)$$



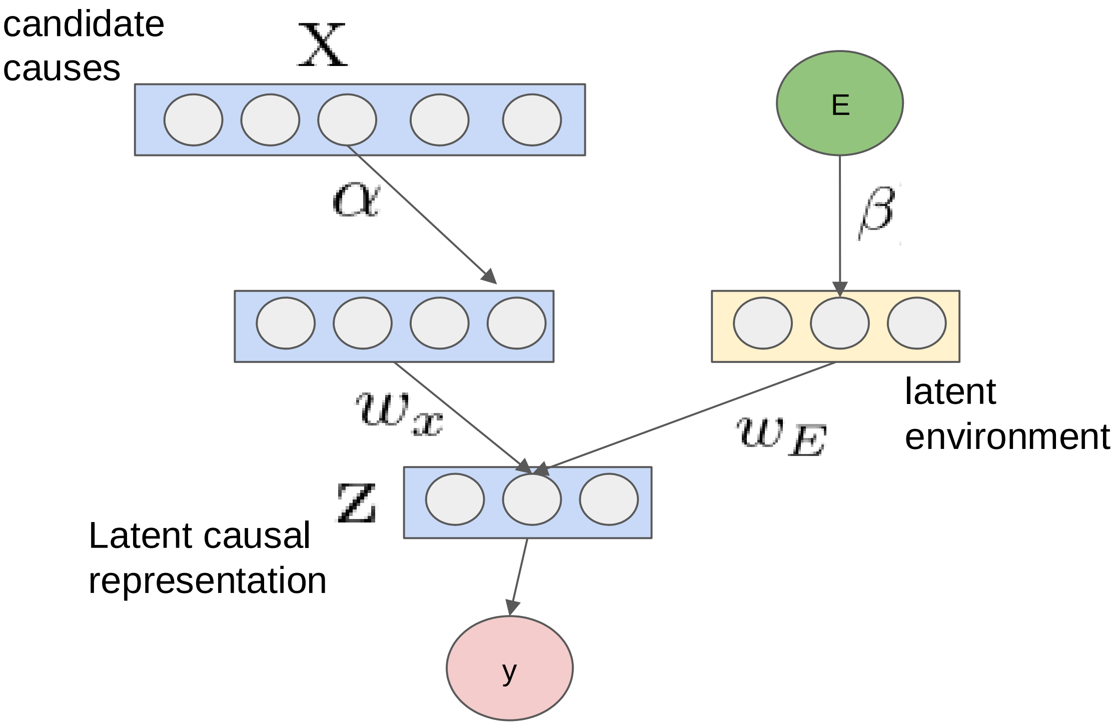
Each rectangle represents a fully connected (possibly deep) NN

$$\mathrm{L}(\mathrm{y},\, \mathrm{x};\, \mathrm{w}_E, w_x, \alpha, \beta) = L_1(y, x; w_E, w_x, \alpha, \beta) + \lambda||\nabla_{w_E} L_1(y, x; w_E, w_x, \alpha, \beta)$$



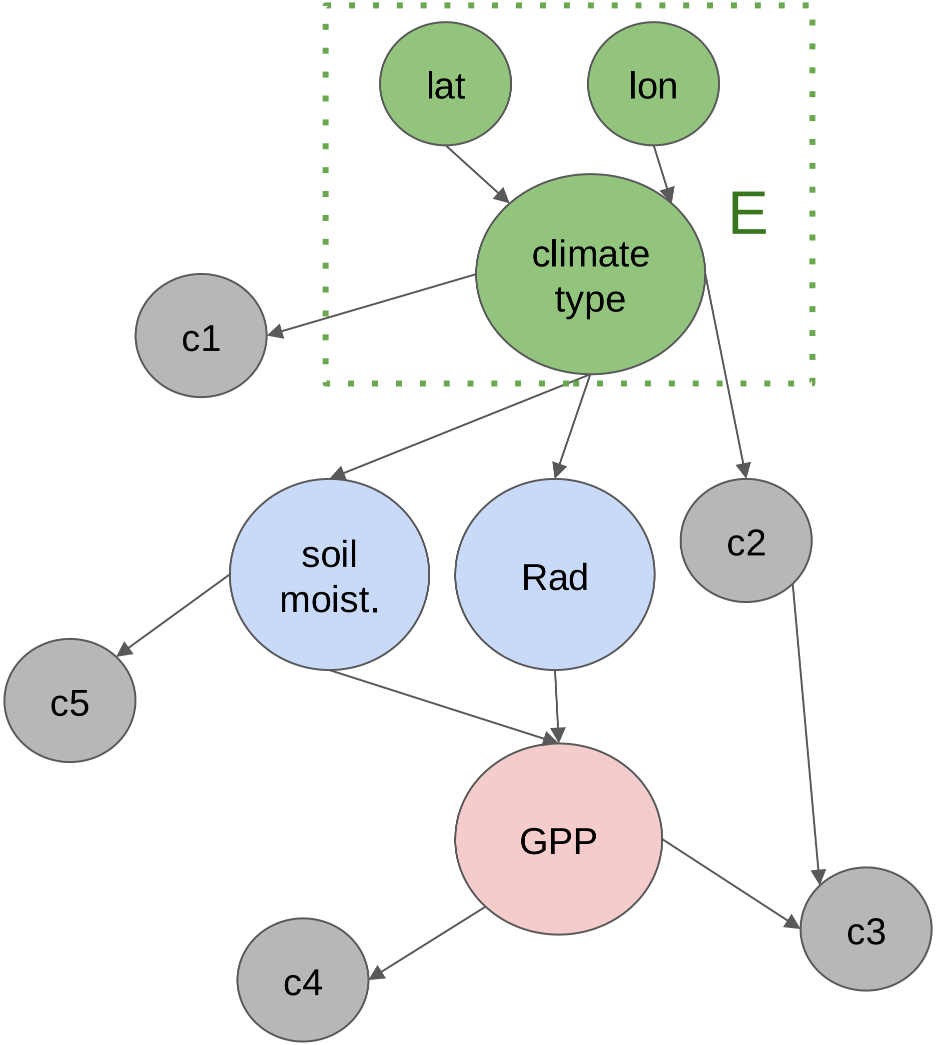
- Learn causal representation
- Learn latent environment

$$\mathrm{L}(\mathrm{y}, \mathrm{x}; \mathrm{w}_E, w_x, \alpha, \beta) = L_1(y, x; w_E, w_x, \alpha, \beta) + \lambda||\nabla_{w_E} L_1(y, x; w_E, w_x, \alpha, \beta)$$



We don’t want to use environment info for prediction. Use it to:

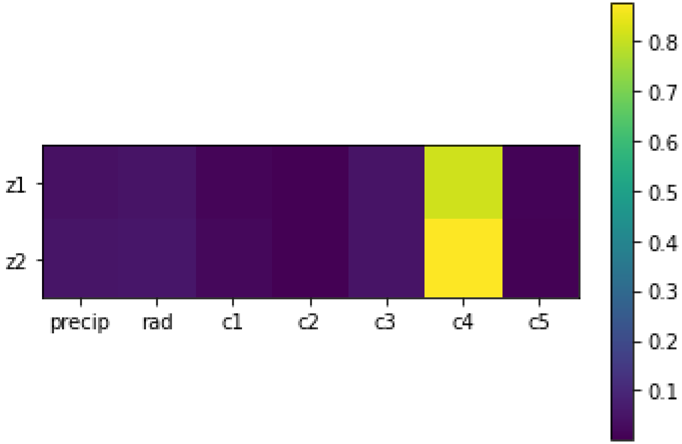
- enforce conditional independence proxy
- estimate latent environment



With toy GPP causal model, with **kno**
ground truth we test if we can learn:

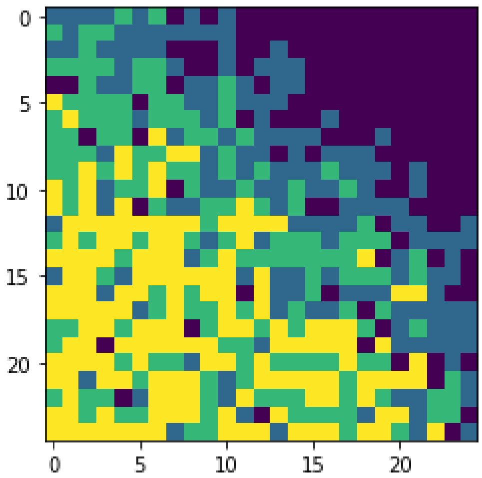
1. causal representation

2. climatic type (latent environmen

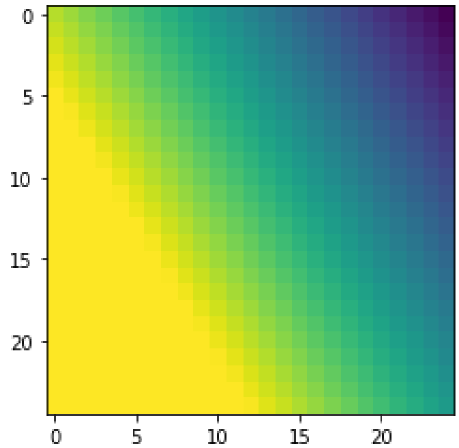


The representation is using c4 as a proxy for GPP

Ground truth climatic region



Estimated climatic region



This might be a way of investigating when environments create different conditions that can be exploited in causal discovery.

Take aways:

1. ICP unfeasible when large number of candidate predictors.

2. Greedy ICP finds a plausible set of causes for pyroCb but inference is unstable

3. Unclear if NN are effective in finding causal representation but may help to identify natural interventions which could help in causal discovery .

Next Steps:

1. Can we get NN to learn correct causal representation.
2. Can we use learnt environment in causal discovery with mixed data

Acknowledgements

This project was first developed as part of the Frontier Development Lab (FDL) Aerosols-Europe 2022 challenge

European Research Council (ERC) Synergy Grant “Understanding and Modelling the Earth System with Machine Learning (USMILE)” has provided funding for research subsequently.

References

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. Journal of the Royal Statistical Society Series B, 78(5):947–1012, 2016. URL https://EconPapers.repec.org/RePEc:bla:jorssb:v:78:y:2016:i:5:p:947-1012.

Kenza Tazi, Emiliano Díaz Salas-Porras, Ashwin Braude, Daniel Okoh, Kara Lamb, Duncan Watson-Parris, Paula Harder, and Nis Meinert. Pyrocast: a machine learning pipeline to forecast pyrocumulonimbus (pyrocb) clouds. In NeurIPS 2022 Workshop-Tackling Climate Change with Machine Learning, 2022.