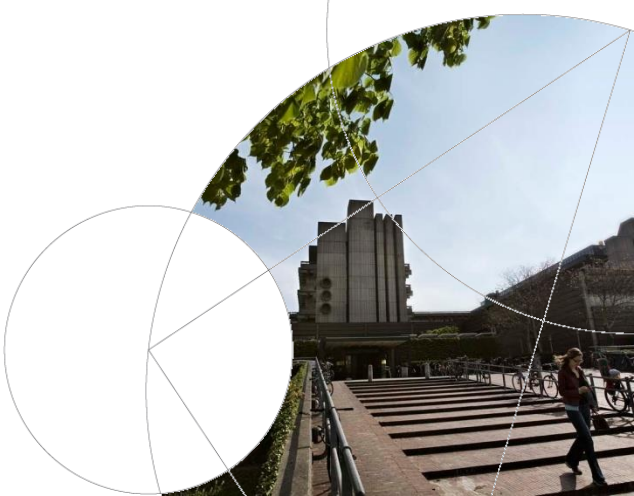
Machine Learning + Causal Inference: A new model building strategy for big data? Experiences from air-pollution research

Hefei - March 2019

HERMES
God of transportation. roads and travelers

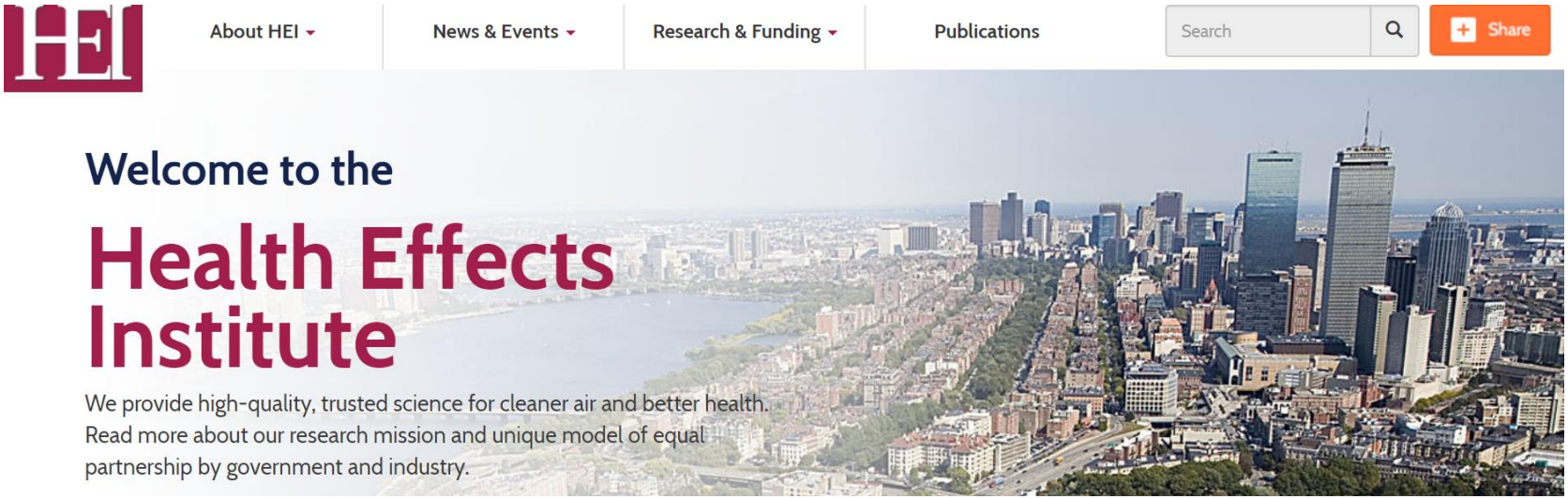
Theis Lange (thlan@sund.ku.dk)
Department of Biostatistics, University of Copenhagen
&
Center for Statistical Science, Peking University.

Funded by Health Effects Institute by grant #4960-RFA17-1/18-2



The background

- I was invited to join a group from the Danish Cancer Society in an application to HEI



- The goal was to inform on the relation between air-pollution and traffic noise and subsequent (long term) health effects.

- But also: We need new fancy statistical methods…
- And then we got the money (~10M Rand) so here goes…

The project objectives

1. estimate exposure to NO2, NOx, black carbon (BC), ultrafine particles (UFP), PM2.5, PMcoarse and PM10 by the AirGIS dispersion modelling system with focus on both tailpipe and non-tailpipe contributions from traffic since 2005

2. identify the specific TRAP exposures most strongly related to myocardial infarction (MI), stroke and diabetes, when considering several pollutants/traffic indicators at the same time

3. investigate associations between TRAPs and a battery of biomarkers related to cardiovascular disease (CVD) and diabetes

4. disentangle how TRAPs and road traffic noise interact in relation to risk for MI, stroke and diabetes, as well as to a number of biomarkers relevant for these diseases

5. investigate how SES, co-morbidity and stress confounds or modifies the associations between TRAP and risk for MI, stroke and diabetes

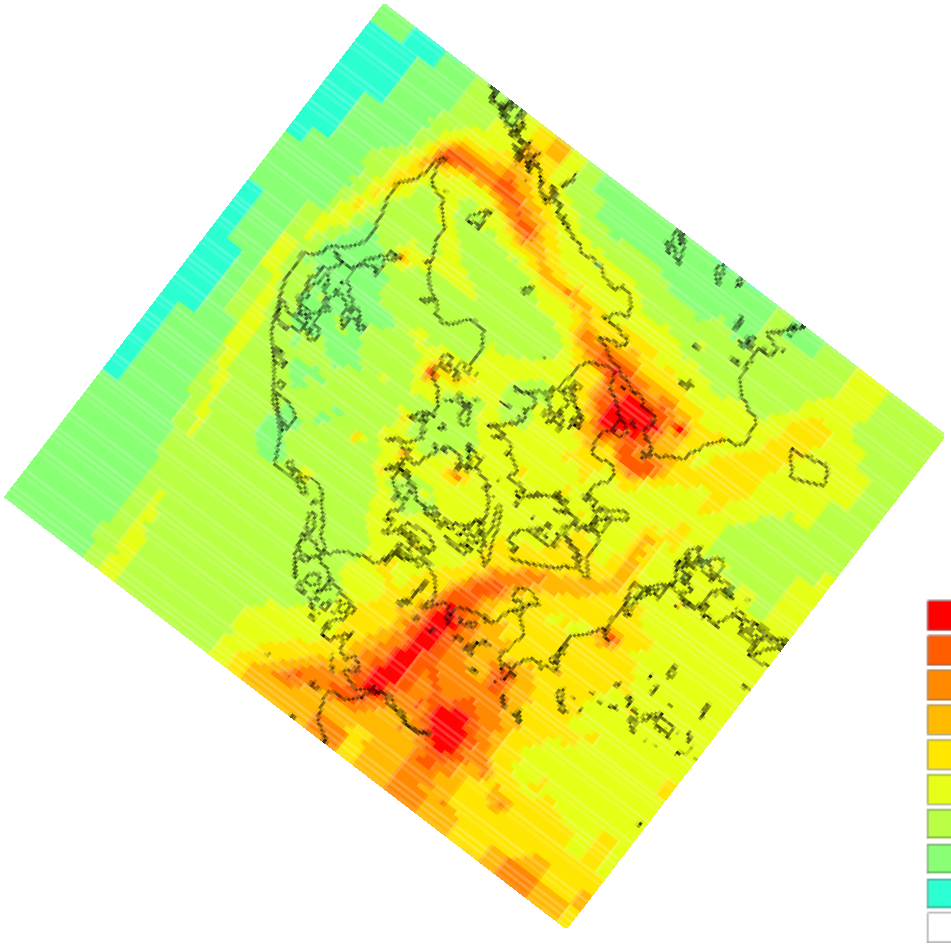
The unique data

- All Danish citizens have unique id-number.
- This number can be used to link between all public registers, including
  - All use of health care system
  - Deaths and hospital diagnosis
  - All use of prescription medicine
  - Income and education
  - AND addresses for all in last 20 years

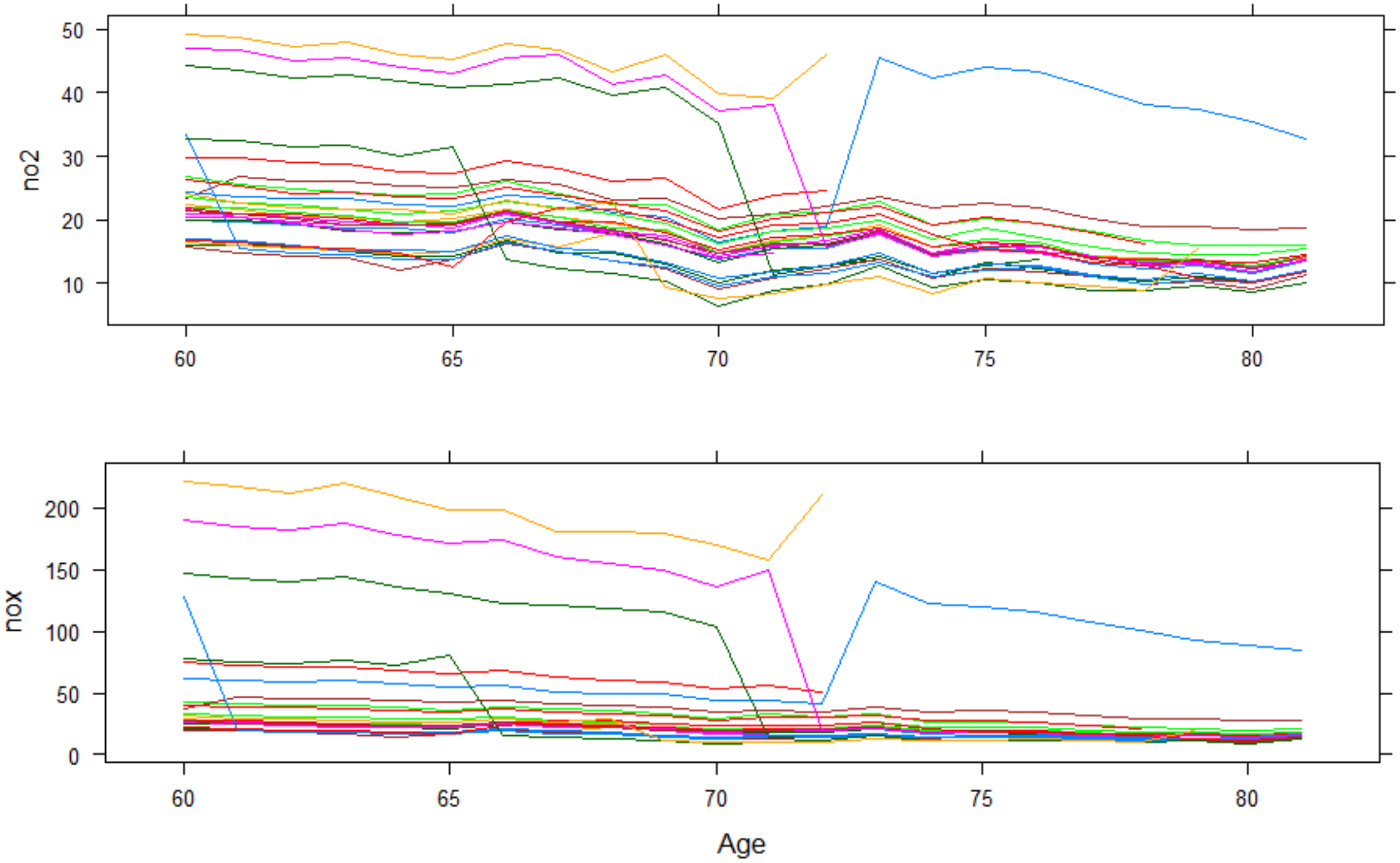
- Other registers (BBR) have precise location and physical dimensions of all buildings in DK.

- Combing these with pollution and traffic measurements using climatic models yields a person specific exposure trajectory for air-pollutants and noise for each Dane.

And example of pollution data



And example of pollution data



Statistical challenges

- Air pollution data is a vector (~20 dim) of highly correlated data.
  - Can we deduce health effects of the single components?
  - Ordinary regressions do not respond well to highly correlated measures.
  - We doubt if linear effects are reasonable.
- Restricting to persons above 18 years we have about 4.5M persons where we could get daily data for pollution.
  - That would give data set of about 120GB
- We would like effect measures that are easy to communicate.
- We need to take account of confounding by social status (rich people do not live next to highways…)
  - Details here that I omit for this discussion.

The sales pitch: Statistical methods in Hermes

- There is a lot of traditional register epidemiology!
  This is important, but not the focus of this section.

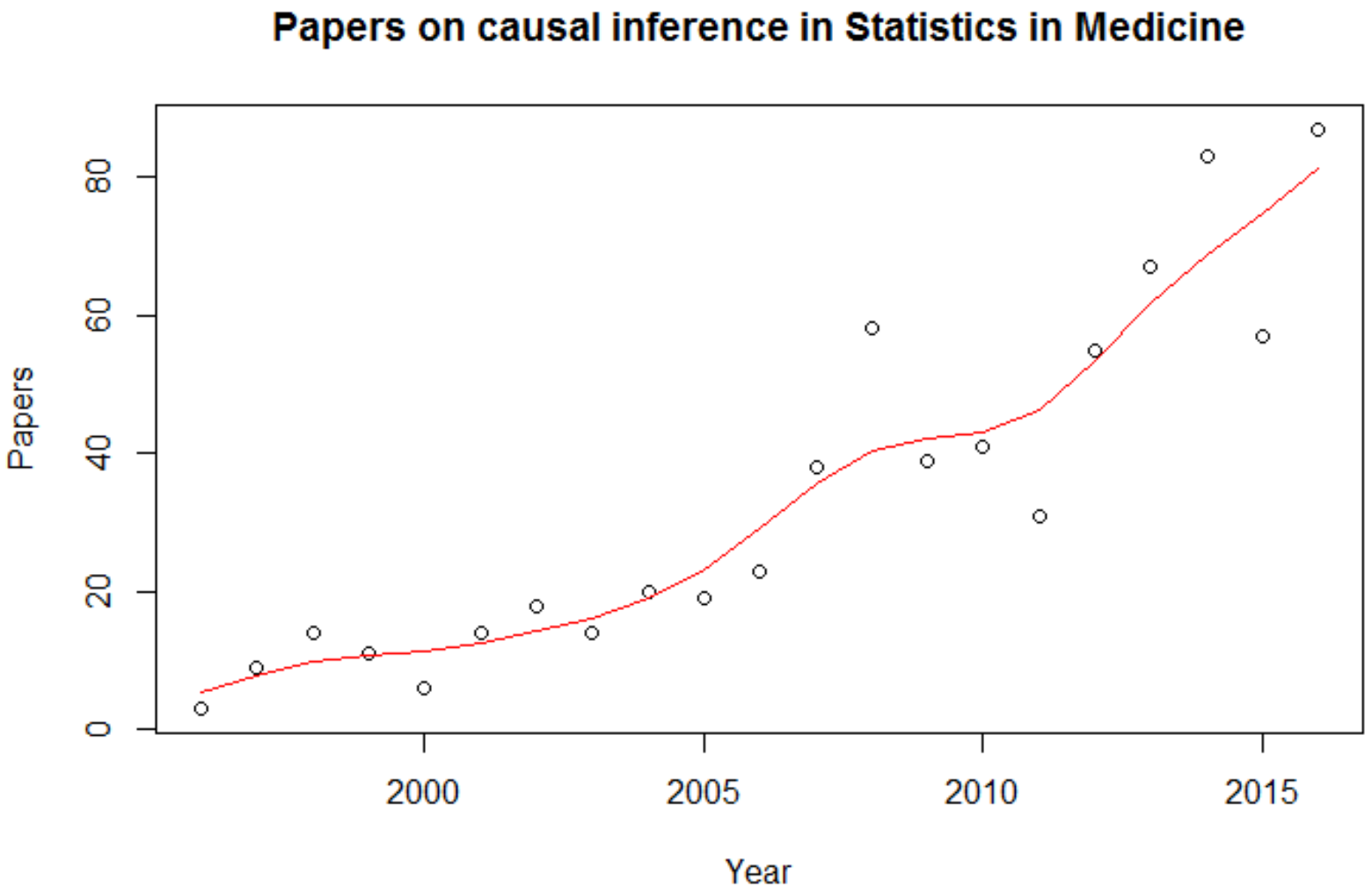
- What is the focus is this promise:
  - **Developing multi-pollutant model.** The traditional approach for multi-pollutant models has been to include only a limited number of pollutants using standard regression methodology.
  - While these approaches each have their merits, they share the same underlying weakness: they depend on a number of parametric assumptions.

  - We will solve this problem by a fundamentally different modeling approach inspired by the principles guiding the rapidly growing scientific field of **causal inference** (Pearl 2009; VanderWeele 2015) in combination with the tools of **machine learning**.

What do we mean by "causal inference"?

- Much of the analysis of data in health and social sciences has as its central aim the quest to learn about cause-effect relationships.
- Does this treatment work? How harmful is the exposure?
- These are causal questions.
- Randomised studies can answer such questions.
- We will keep the randomization idea in the sense that our ultimate goal is to compare **pollutant scenarios**.

Do others care about causal inference?



Counterfactuals = a mathematical formulation of cause and effect

- So what are these counterfactuals?
- The idea is that each person has a (potential) outcome such as death time for any possible configuration of pollutants.
- This is denote *Y(a)* where *a* is the considered pollutant level.
- At most one of these counterfactuals will ever be realized.
- Traditionally "*a*" is binary, but there is actually nothing preventing a high-dimensional *a*.

Effect measures

- No causal effect, $P(Y^1 = 1) = P(Y^0 = 1)$ can be formulated via various effect measures
- $P(Y^1 = 1) - P(Y^0 = 1) = 0$ (risk difference)
- $\frac{P(Y^1=1)}{P(Y^0=1)} = 1$ (risk ratio).
- $\frac{P(Y^1=1)/P(Y^1=0)}{P(Y^0=1)/P(Y^0=0)} = 1$ (odds ratio).
- So we now have some causal estimands!
- For continuous response, it could be $E(Y^1) - E(Y^0)$

Here we again act is if a is binary, but the comparison could be between any two different levels of a (ie. pollutions)

G-FORMULAS

Using counterfactuals in observational studies

- Of course counterfactuals needs to be connected to observed data.

- In randomized studies this is easy
  Mean of counterfactual Y(a) = Mean in treatment grp a

- In observational studies need control for confounding.
- One solution is the g-formula

Mean of counterfactual Y(1) = $\frac{1}{n}\sum_i \widehat{E}(Y_i \mid L_i, A_i = 1)$

- The model used on left side needs to predict outcome for given values of *L* and *A*

How to do in practice

1. Expand the data set with 3 sections below each other (see next slide where the sections, however, are put side by side):
   (a) The original data set with observed L, A, Y
   (b) A data set keeping L, setting $A = 0$ and $Y =$ missing
   (c) A data set keeping L, setting $A = 1$ and $Y =$ missing
2. Fit the Q-model to the data (only part 1 will be used)
3. Based on the Q-model, predict Y from L, A in the second and third sections
4. Average the predicted outcomes separately in sections 2 and 3

G-formula – general outcome

- What we really need to do is the following:

  1. We decide on a pollution scenario
  2. We take our whole data base and change the pollution values to the value from 1.
  3. Delete the actual outcomes and replace with predictions made from the Q-model.
  4. Repeat 2-3 with a different pollution scenario.
  5. Compare the outcomes (eg. Survivals) from 3 and 4 with no adjustment. Eg. a Cox model with a single binary covariate.

- The challenge is how to predict outcomes… Our solution

Machine Learning

Machine Learning – the technicalities

- While the idea is the same in all implementations of Machine Learning the way to realize the ideas wary widely.
- Some keywords:

- We will likely focus on decision trees inspired solutions.

The scenarios we want to compare

- I tried to get the epidemiologists at Danish Cancer Society to formulate which scenarios they wanted to compare.

- Was not easy. They kept using regression-like-wording…

- Maybe is only causal inference people that find the "imagine that you could intervene wording" logical??

- Final result:

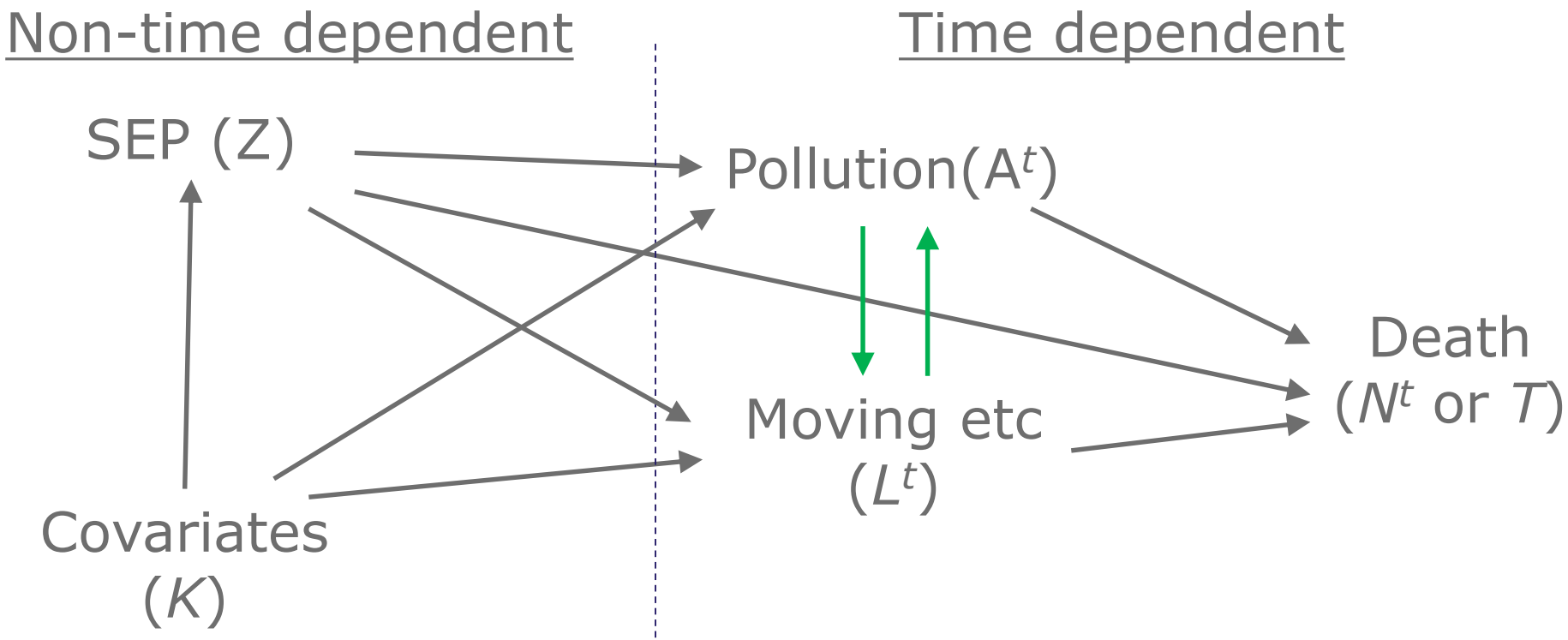
Scenario 0: Assume all pollutants were as observed.

Scenario 1: Assume pollutant A was increased X units (or Y%), while all other pollutants were as observed.

These scenarios can be further investigated for different baseline pollutant profiles

- *A high load city apartment.*
- *A low load city apartment.*
- *A suburban single-family house.*
- *A rural dwelling.*

Our assumed causal structure



My model plan

- Discretize time into 2 month periods.

- Let r_{it} denote if person i had an event (type) in period t, $i= 1,\ldots,N$ and $t=T_{min,i},\ldots, T_{max,i}$.
- Let x_{it} denote measured covariates.

- Build a random forest for r_{it} using x_{it} as predictor.
- If we have competing risk we need model for those events also.

- Challenges:
  - machine learning methods not so happy about massively unbalanced outcomes (practically all r_{it} will be zero)
  - how will machine learning methods cope with the cluster structure in the data?
    However, recall we just need $E[r_{it} \mid x_{it}]$ to be consistent.

Interpreting the random forests

- Create a whole population of x_{it} with maximal follow-up and pollution levels set according to the scenario considered.
- Simulate r_{it} by the constructed random forests.
- Remove parts of follow-up that happen after a simulated event.
- Provide simple summaries (ie. expected life times or HRs when comparing the scenarios).
- Note only one covariate here!
- Bootstrap the whole thing to obtain CIs.
  Is this computational feasible?

Comments?
Suggestions?