Predicting and evaluating how changes in exposures change health risks

Tony Cox

April 29, 2018

Goals

- *Predict* changes in public health effects caused by changes in exposures
  - Not associations or slopes, but changes over time
- *Evaluate* changes in effects caused by changes in exposures in hindsight (accountability)
  - Model data on changes, not just levels
- Use trustworthy methods, get objective answers
  - Do not rely on untested assumptions or counterfactual comparisons (Dublin)
  - Use automated algorithms to avoid p-hacking
  - Discover who benefits, how, and how much from reduced exposures to air pollution

Causal questions

- Statistical inference question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) depend on observed values of other variables?
    - P(deaths | tmin, PM2.5, etc.)
- Causal question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) *change* in response to changes in values of other variables?
    - P(deaths | tmin, do(PM2.5), etc.)
  - How would exogenously reducing PM2.5, tmax, etc. change elderly mortality, AllCause75?
- Seeing ≠ doing! (Pearl, 2009)
- This talk: Illustrate machine learning (ML) techniques for predicting causal impacts with minimal assumptions

?

year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

Causal questions

- Statistical inference question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) depend on observed values of other variables?
    - P(deaths | tmin, PM2.5, etc.)
- Causal question:
  - How does the conditional probability distribution for observed daily death count (AllCause75) *change* in response to changes in values of other variables?
    - P(deaths | tmin, do(PM2.5), etc.)
  - How would exogenously reducing PM2.5, tmax, etc. change elderly mortality, AllCause75?
- Seeing ≠ doing! (Pearl, 2009)
- This talk: Illustrate machine learning (ML) techniques for predicting causal impacts with minimal assumptions

?

year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

- Real data set for LA area (South Coastal Air Quality Management District)
- 1,461 days of data (1/1/07- 12/31/10)
- Data described by Lopiano et al., obtained from them, https://arxiv.org/abs/1502.03062
- Original data sources: CARB for PM2.5 (www.arb.ca.gov/aqmis2/aqdselect.php), CDPH for mortality counts, EPA for meteorological variables
- Download full data set from http://cox-associates.com/CausalAnalytics/ LA_data_example.xlsx

Alternative concepts of causality

- Probabilistic
- Associational
- Attributive
- Counterfactual
- Structural
- Predictive
- Manipulative
- Mechanistic/explanatory

- Associational/attributive/(counterfactual)
  - IARC: Regression, RR, burden-of-disease, PAR
    - Usually depends on untested assumptions
- Predictive: Causes help to predict their effects
  - Can be discovered and tested from data
    - Conditional independence tests, $X \rightarrow Y \rightarrow Z$
    - Granger tests, transfer entropy
- Manipulative: Changing causes changes effects
  - Randomized control trial (RCT)
  - Generalization/transportability
- Mechanistic: Changes propagate via networks of laws
  - Invariant laws (CPTs)
  - Composition of effects, well-behaved errors

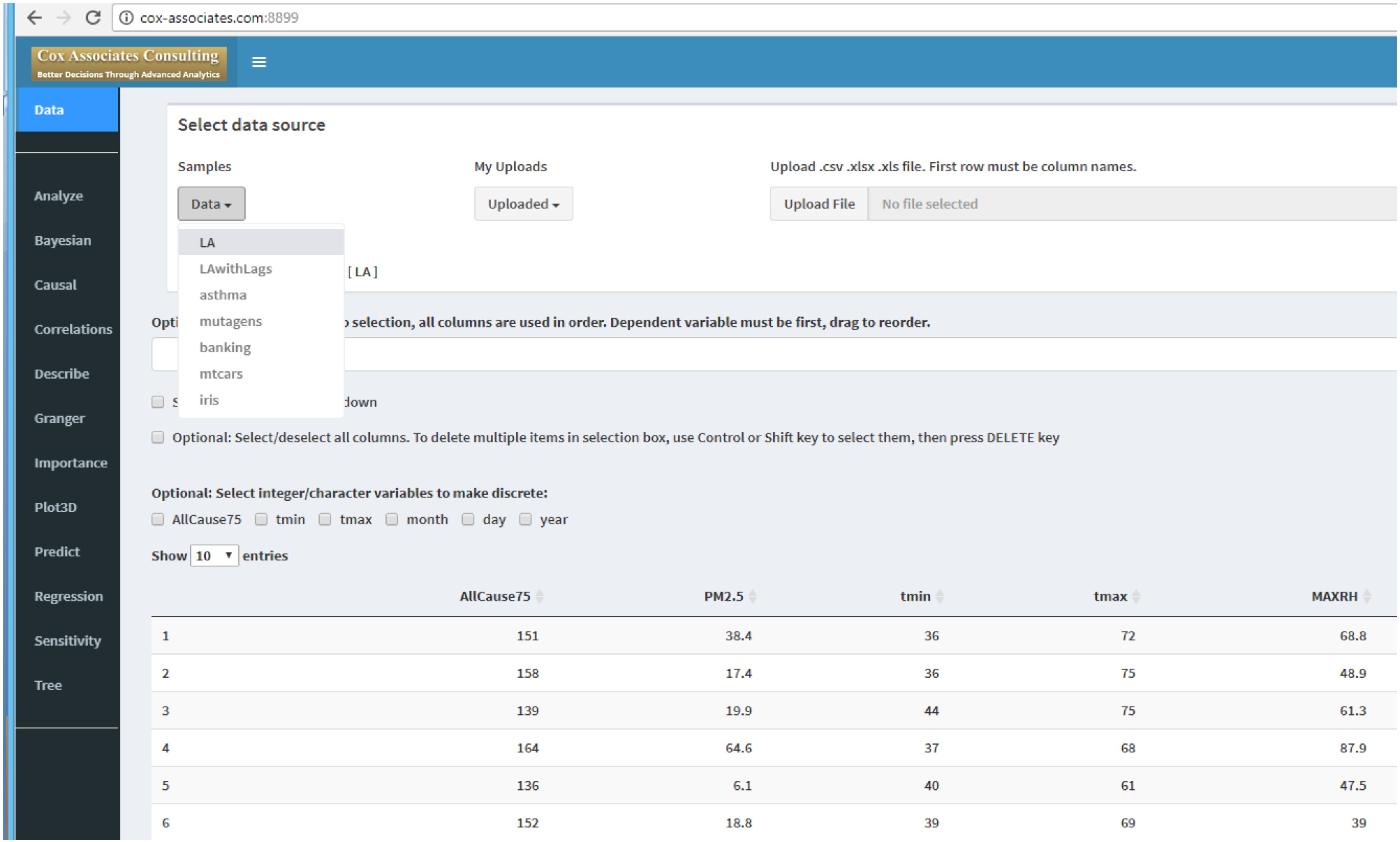
IARC = International Agency for Research on Cancer
RR = relative risk; PAR = population attributable risk; CPTs = Conditional probability tables or trees

Machine learning can help to avoid model-dependent conclusions and p-hacking

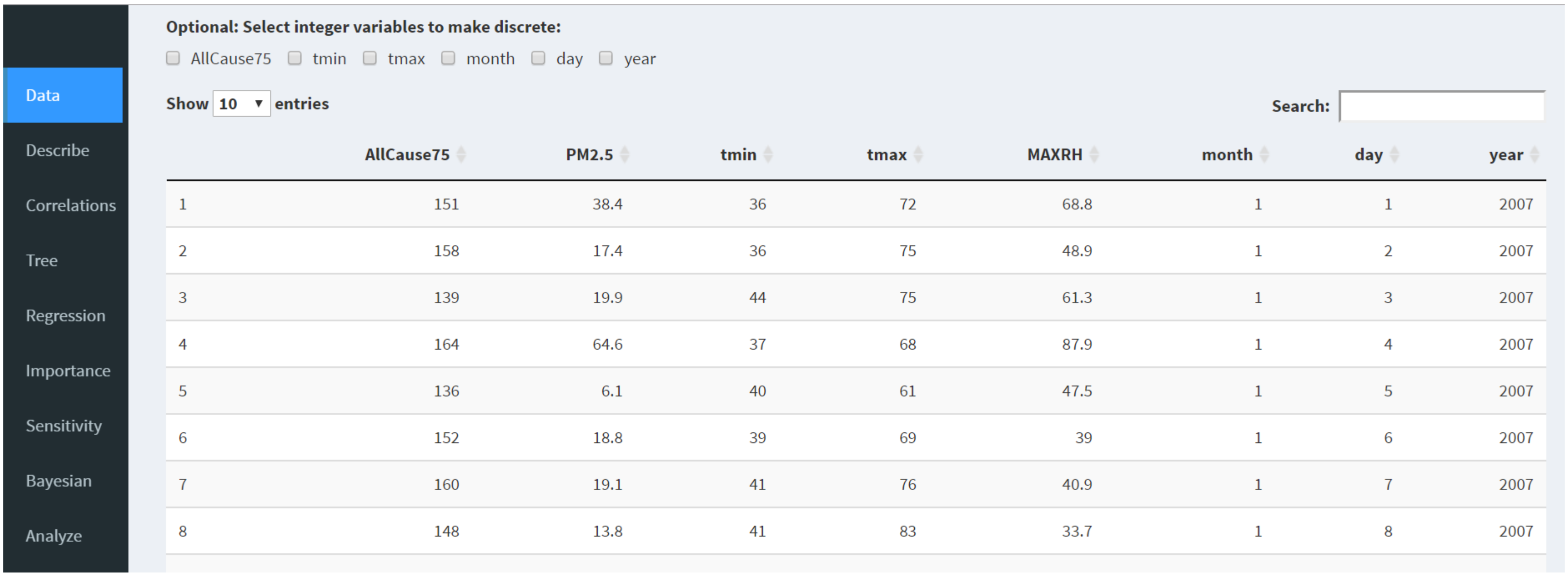
- Information-based algorithms: Automated, data-driven, minimal assumptions, empirically testable (usually)
  - Effects are *not conditionally independent* of their causes
  - Changes in causes *help to predict* changes in their effects
    - Granger causality for time series data; DAG models
  - Non-parametric methods minimize modeling assumptions
    - Trees
    - Bayesian networks
    - Causal directed acyclic graph (DAG) models
  - Model ensembles address model uncertainty
    - RandomForest algorithm
    - Causal partial dependence plots

DAG = directed acyclic graph

Automated analysis with these methods is now practical: Enter data, click to analyze



Automated analysis with these methods is now practical: Enter data, click to analyze



	AllCause75	PM2.5	tmin	tmax	MAXRH	month	day	year
1	151	38.4	36	72	68.8	1	1	2007
2	158	17.4	36	75	48.9	1	2	2007
3	139	19.9	44	75	61.3	1	3	2007
4	164	64.6	37	68	87.9	1	4	2007
5	136	6.1	40	61	47.5	1	5	2007
6	152	18.8	39	69	39	1	6	2007
7	160	19.1	41	76	40.9	1	7	2007
8	148	13.8	41	83	33.7	1	8	2007

Automated analysis is now practical for all of the foregoing methods

- Data
- Analyze
- Bayesian
- Causal
- Correlations
- Describe
- Granger
- Importance
- Plot3D
- Predict
- Regression
- Sensitivity
- Tree

Executive Report:

What are the potential causal drivers of < AllCause75 > in this data set?

The following were identified (by a Bayesian Network machine-learning algorithm) as potential causes of < AllCause75 > in this data set:
Neighbors of < AllCause75 > are: tmin, month, tmax

Potential causes of < AllCause75 > are defined as its neighbors in a Bayesian Network.

The exposure variable [PM2.5] is NOT a significant predictor for [AllCause75] (p = 0.10) in a Quasi-Poisson regression model.
[tmin] is a significant predictor for [AllCause75] (p = 0.00) in a Quasi-Poisson regression model.
[month] is a significant predictor for [AllCause75] (p = 0.00) in a Quasi-Poisson regression model.
Significant predictors of < AllCause75 > are defined here as those with regression coefficients significantly different from zero in a Quasi-Poisson regression model.

How important are these causal drivers?

From most to least important (using importance table), the relative importances of these potential causes are as follows:

Variable	Importance(%IncMSE)
month	168.28
tmin	62.27
tmax	34.07
PM2.5	5.83

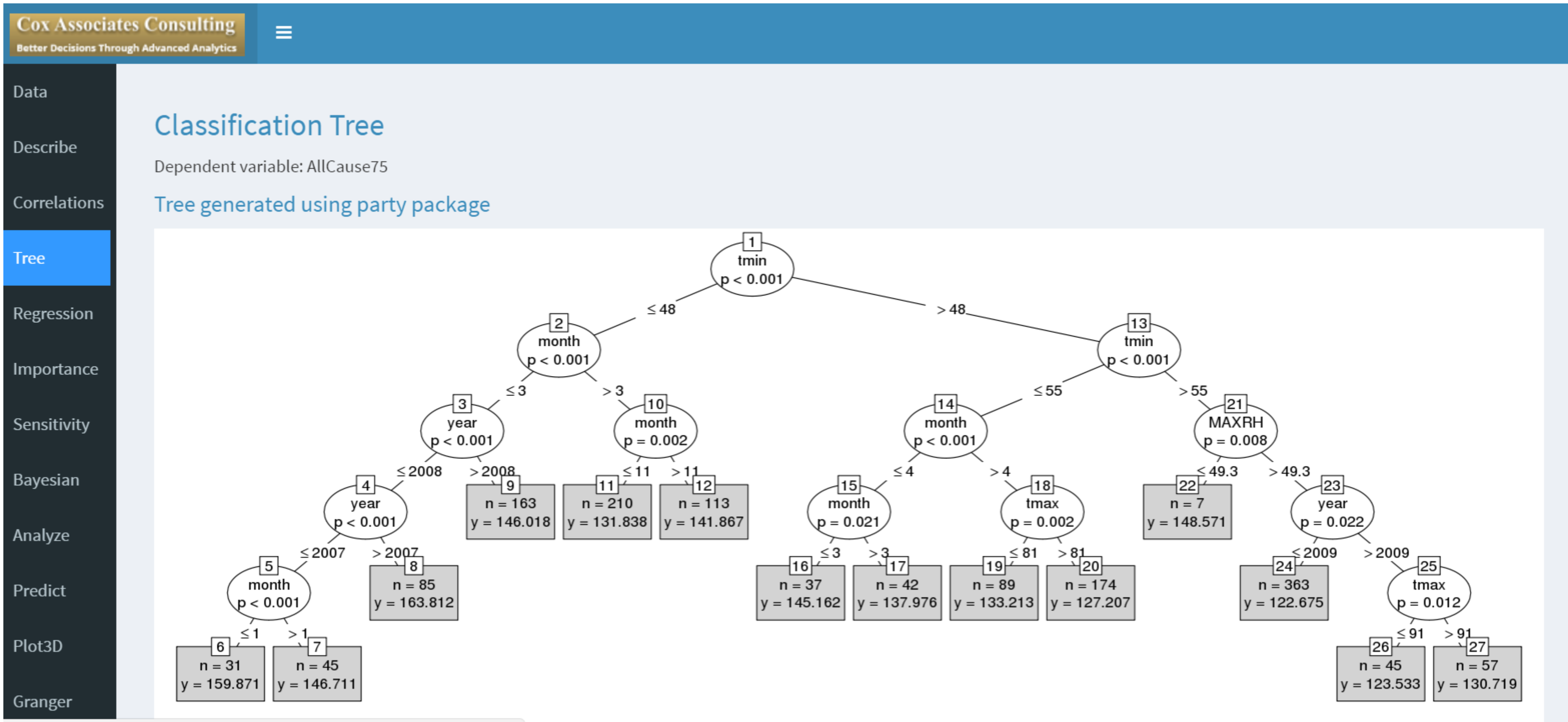
A variable's importance is measured here as the increase in mean squared error in predicting < AllCause75 > if the variable is dropped.

How strongly does < PM2.5 > predict or explain < AllCause75 >?

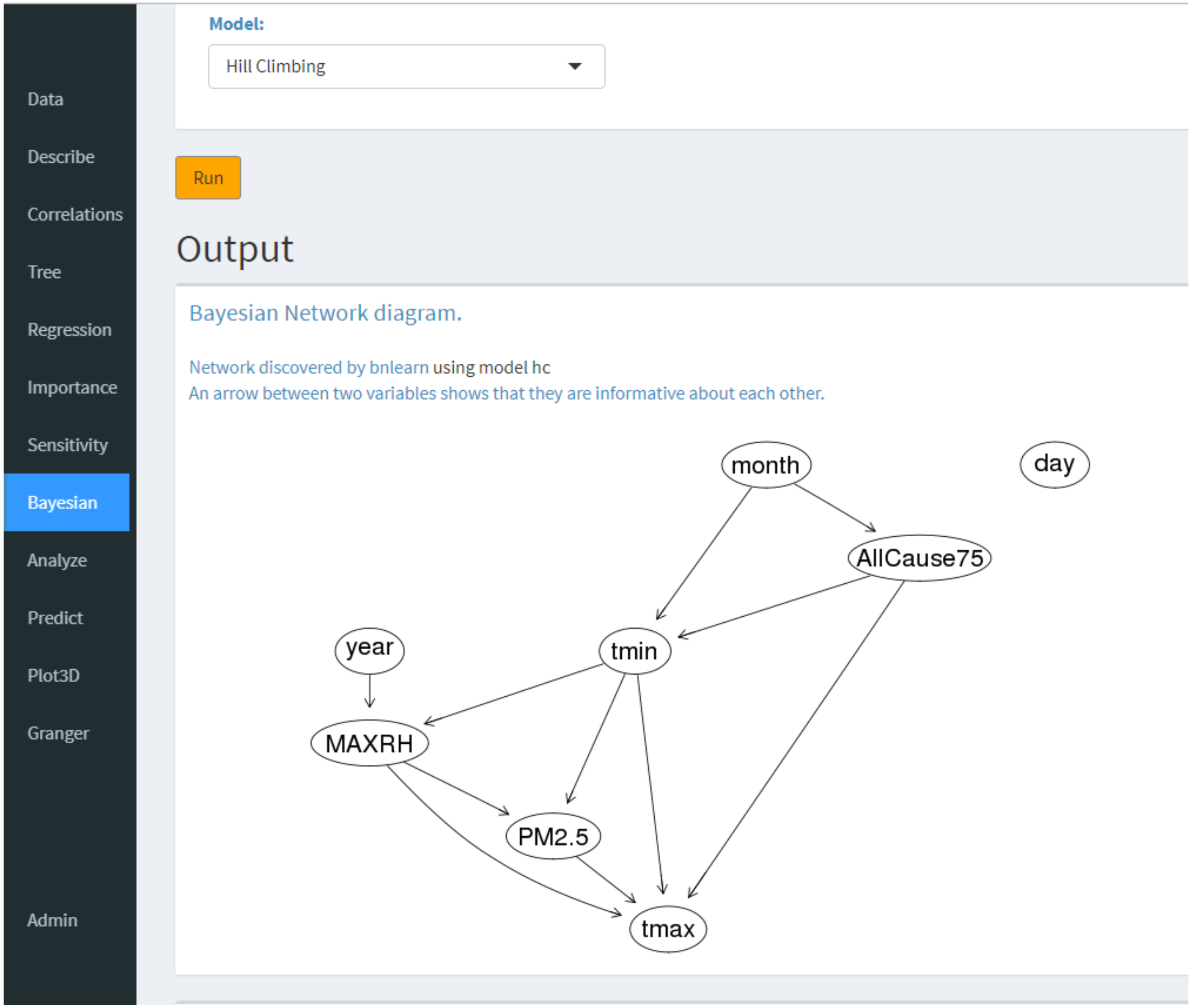
Including < PM2.5 > changes the percentage of explained variance in < AllCause75 > from 40.25 % to 40.80 % in a randomForest analysis. Thus, including < PM2.5 > as a predictor changes the

In multiple linear regression modeling, the percentage of explained variance (adjusted R-squared) in < AllCause75 > is 31.44 % when < PM2.5 > is included and is 31.37 % when < PM2.5 > is dı
by about 0.07 % in a multiple linear regression analysis.

Automated analysis is practical: trees and conditional probability tables



Automated analysis: Bayesian network



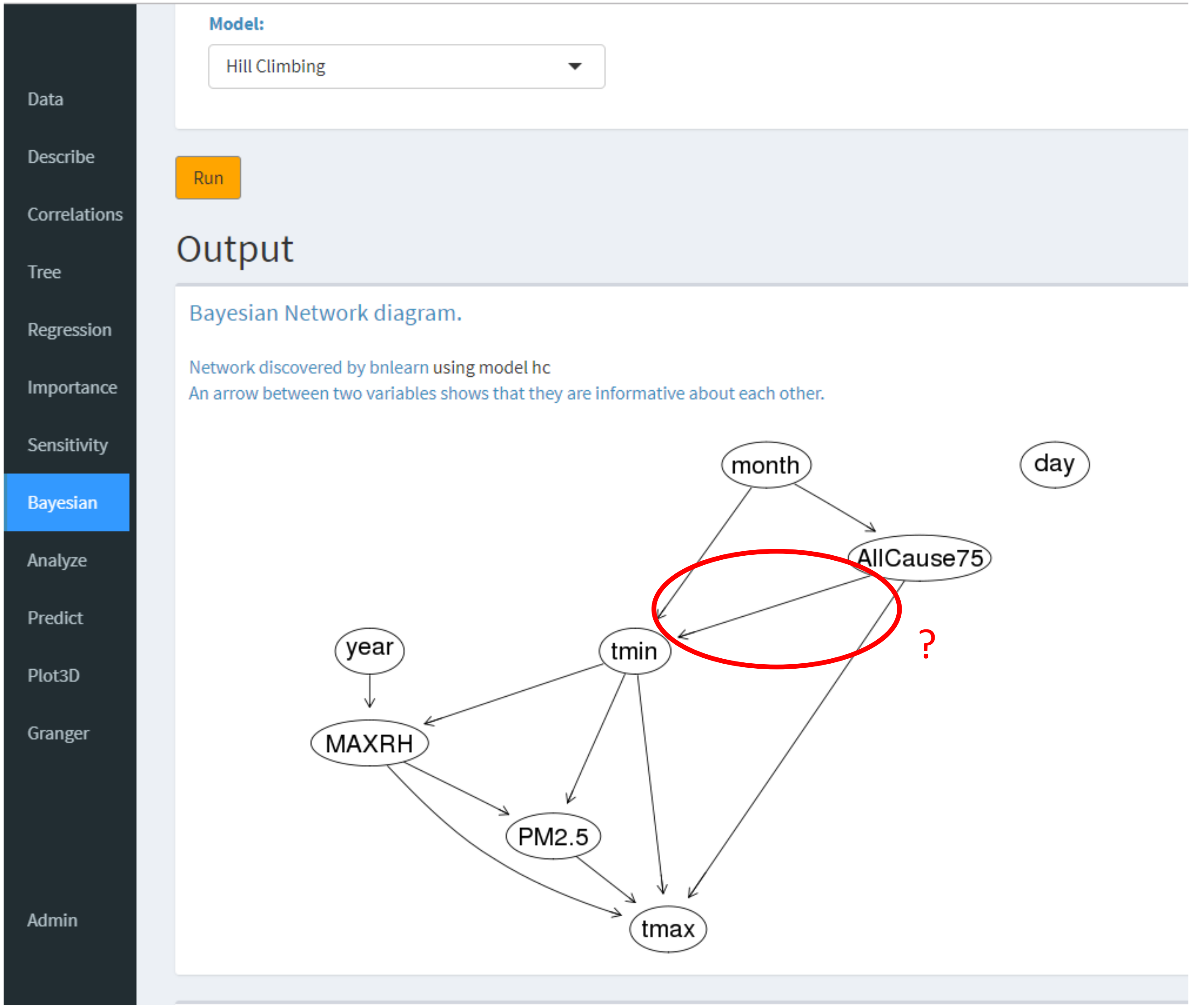
Conditional probability table (CPT) or tree at each node

Assumptions:

- Causes are informative about their effects (Arrow directions here do not yet indicate causality)
- Effects are not conditionally independent of causes
- Direct causes are adjacent to their effects
- Arrows reflect information, possible direct and indirect causal pathways

Automated non-parametric model: Bayesian network (BN)

Automated discovery: Arrows unclear

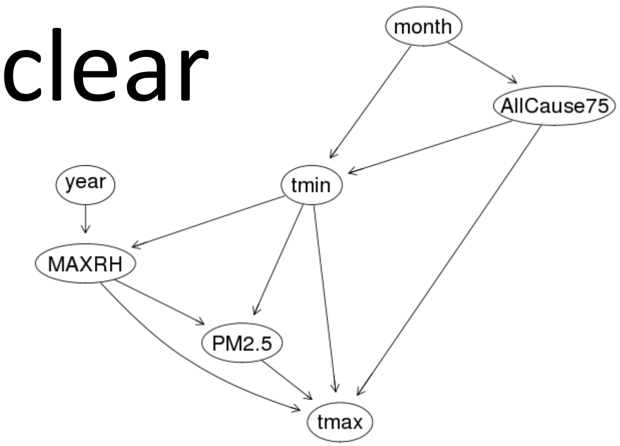


Conditional probability table (CPT) or tree at each node

Assumptions:
- Causes are informative about their effects. (Arrow directions here do not yet indicate causality)
- Effects are not conditionally independent of causes
- Direct causes are adjacent to their effects
- Arrows reflect information, possible direct and indirect causal pathways

Automated non-parametric model: Bayesian network (BN)

Regression coefficients unclear



- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger
- Admin

Regression

```
Quasi-Poisson
```

```
Dependent variable: AllCause75
```

Quasi-Poisson regression model

Estimated Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	Signif
(Intercept)	3.682	4.997	0.737	0.46133	
PM2.5	0.001	0.000	2.928	0.00347	**
tmin	-0.004	0.001	-6.092	< 0.001	***
tmax	-0.002	0.000	-3.977	< 0.001	***
MAXRH	-0.001	0.000	-4.098	< 0.001	***
month	-0.010	0.001	-11.972	< 0.001	***
day	-0.000	0.000	-0.112	0.91102	
year	0.001	0.002	0.335	0.73756	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 3148.4 on 1460 degrees of freedom

Residual deviance: 2126.7 on 1453 degrees of freedom

AIC: NA

Exposure-response regression coefficient for PM2.5 as predictor of AllCause75 is significantly positive. Q: *Why?*

A: PM2.5 helps to correct model specification errors (errors in variables, month treated as a continuous predictor, omitted lagged daily temperatures)

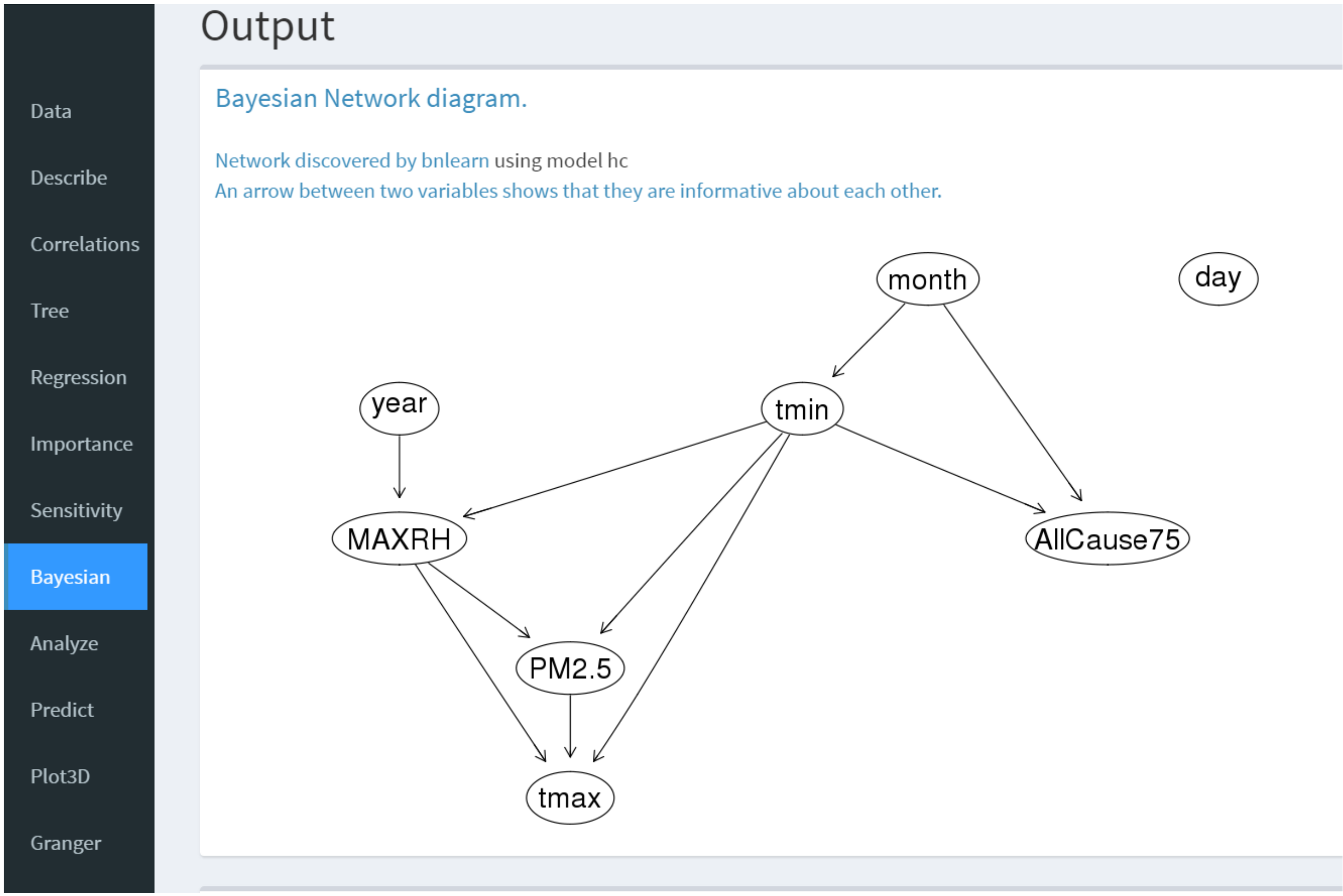
Regression coefficients (and associations) mix direct, indirect, selection, confounder, and non-causal effects

Strong, consistent association ≠ evidence for predictive or manipulative causation

R packages and principles for identifying causal DAGs from data

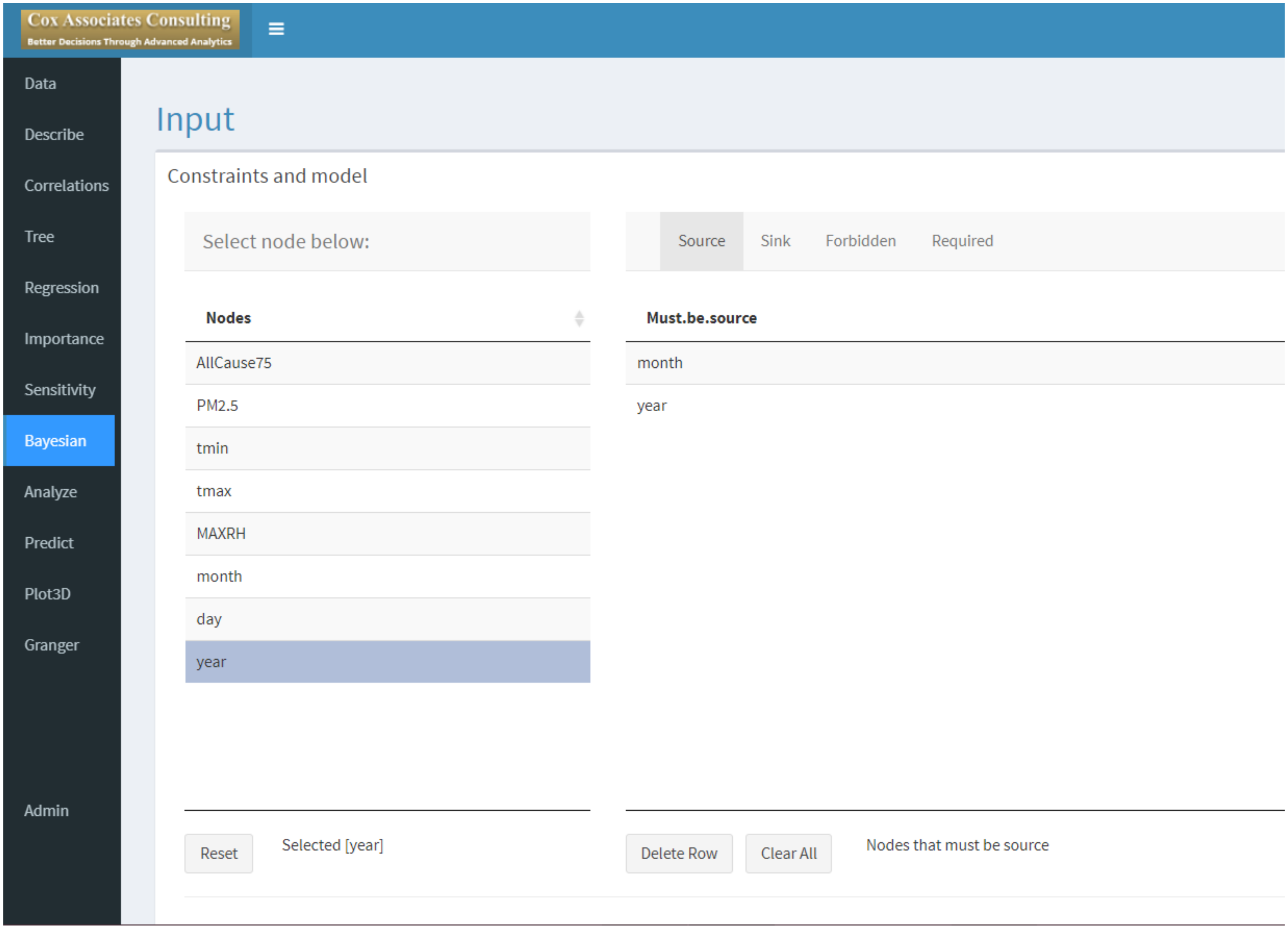
- Conditional independence (constraint-based algorithms)
  - *bnlearn, Tetrad, CompareCausalModels, dagitty* packages
- Likelihood principle (score-based algorithms)
  - Choose DAG model to maximize likelihood of data
  - Included among the algorithms in *bnlearn* package
- Composition principle: If $X \rightarrow Y \rightarrow Z$, then $dz/dx = (dz/dy)*(dy/dx)$
- Granger/transfer entropy principle: Predictively useful information flows from causes to their effects over time
  - Transfer entropy, Yin & Yao, 2016, www.nature.com/articles/srep29192
- Model error specification principle
  - effect = f(cause) + error
  - LiNGAM software, https://arxiv.org/ftp/arxiv/papers/1408/1408.2038.pdf
- Homogeneity and invariance principles for causal CPTs
  - Li et al., 2015, https://pdfs.semanticscholar.org/a051/9a2c6b85ca65d0df037142f550cf87d4e43f.pdf
  - Peters et al., 2015, *InvariantCausalPrediction* package
    http://stat.ethz.ch/~nicolai/invariant.pdf

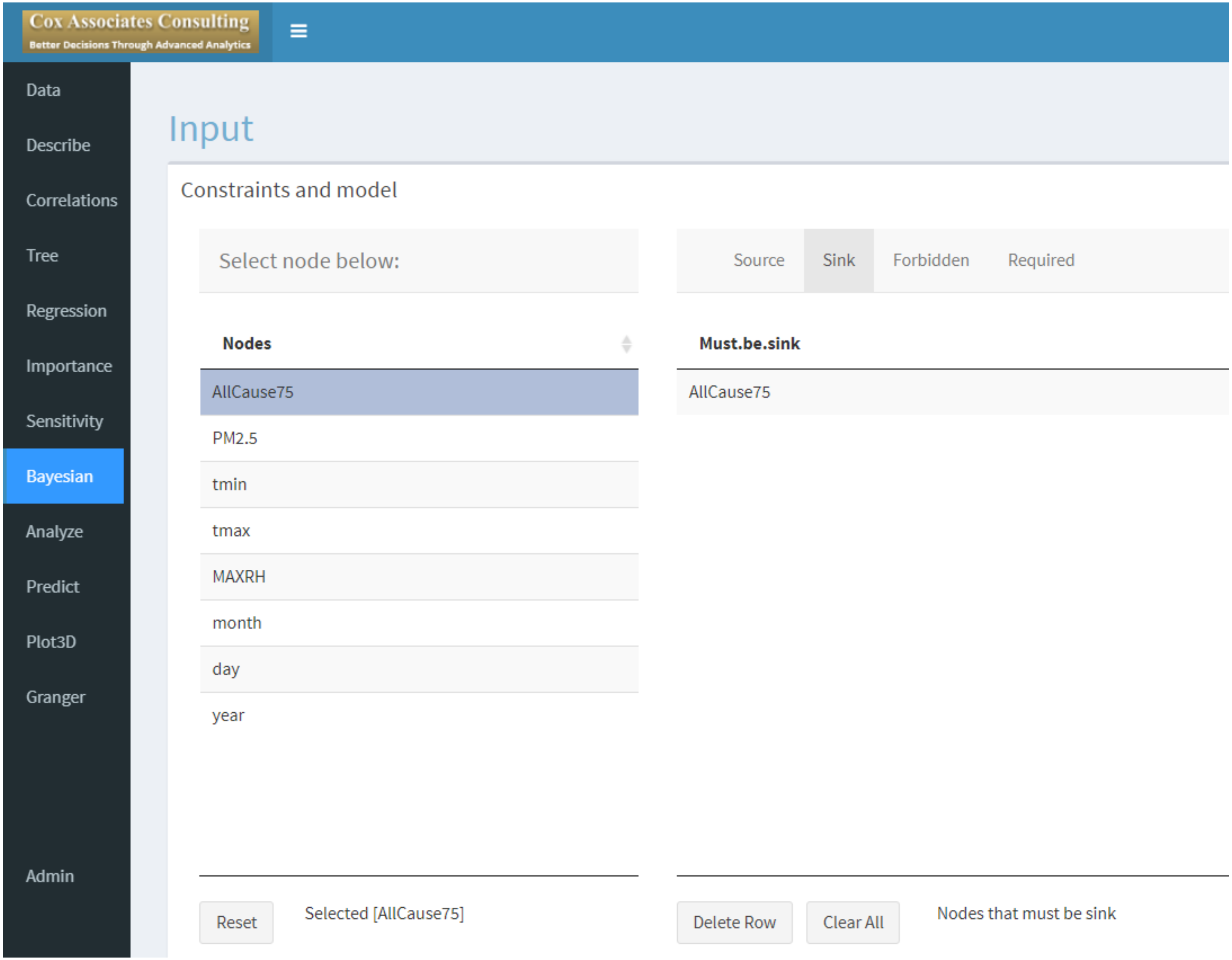
Automated analysis can be improved with causal knowledge, if available



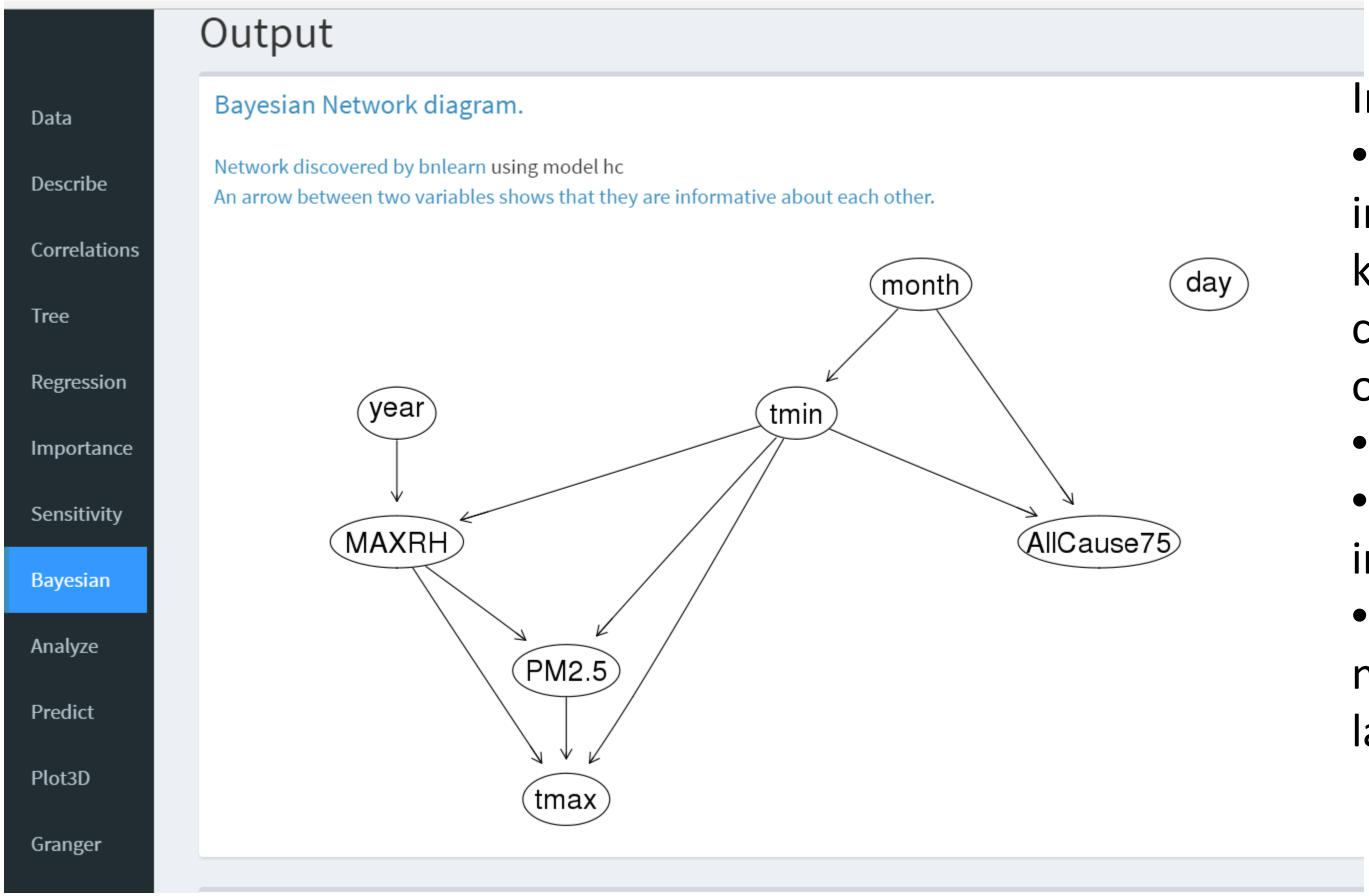
Knowledge-based constraints

Potential p-hacking point, but controllable




Cox Associates Consulting
Better Decisions Through Advanced Analytics
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
Admin
Input
Constraints and model
Select node below:
Source
Sink
Forbidden
Required
Nodes
AllCause75
PM2.5
tmin
tmax
MAXRH
month
day
year
Must.be.sink
AllCause75
Reset
Selected [AllCause75]
Delete Row
Clear All
Nodes that must be sink

Automated analysis can be improved with causal knowledge, if available

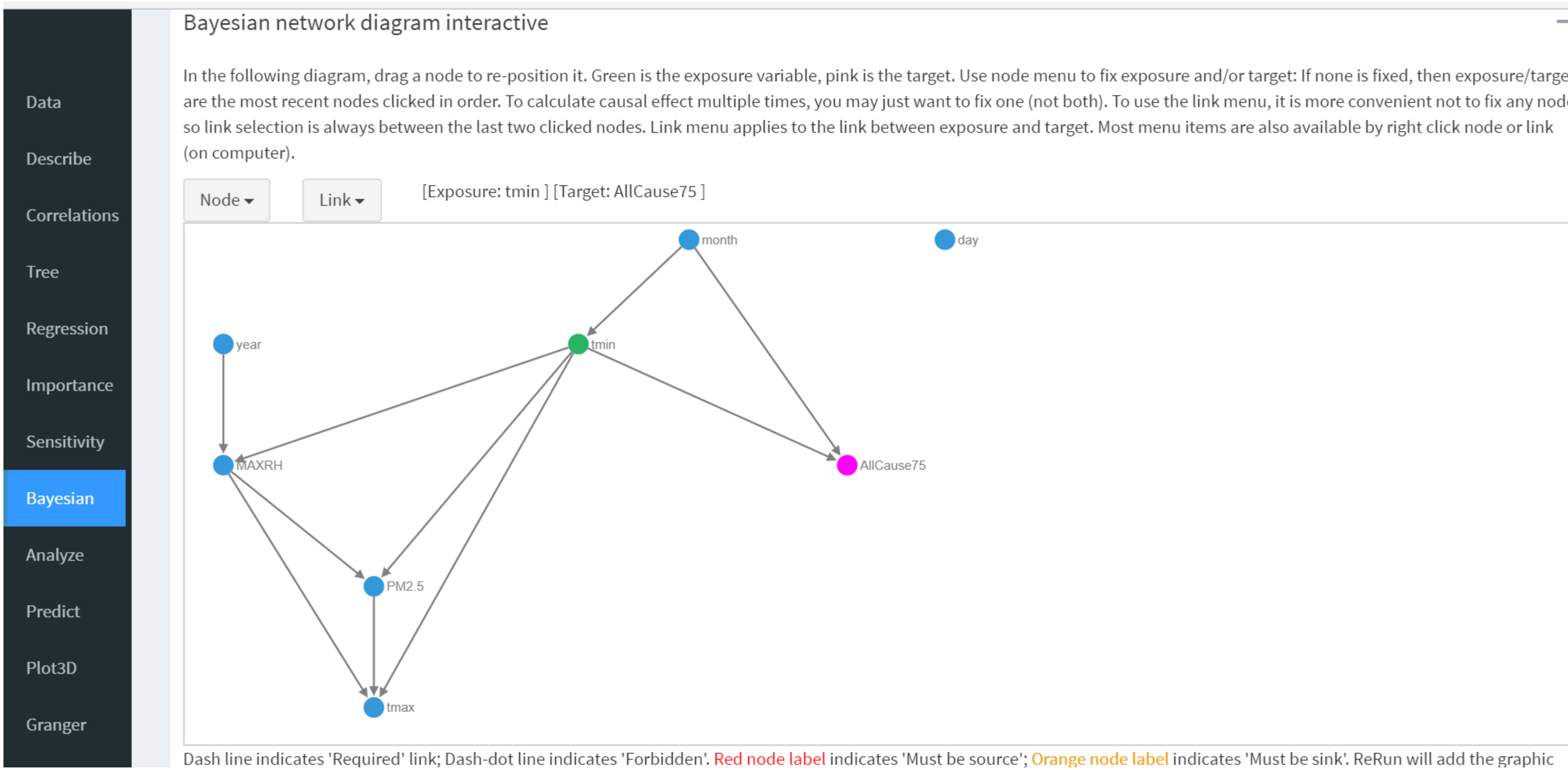


Interpretation:

- Now, arrows reflect information *and* knowledge-based constraints for predictive or manipulative causality
- CPT or tree at each node
- Causal CPTs or laws are invariant across studies
- Dynamic Bayesian networks (DBNs) include lagged values

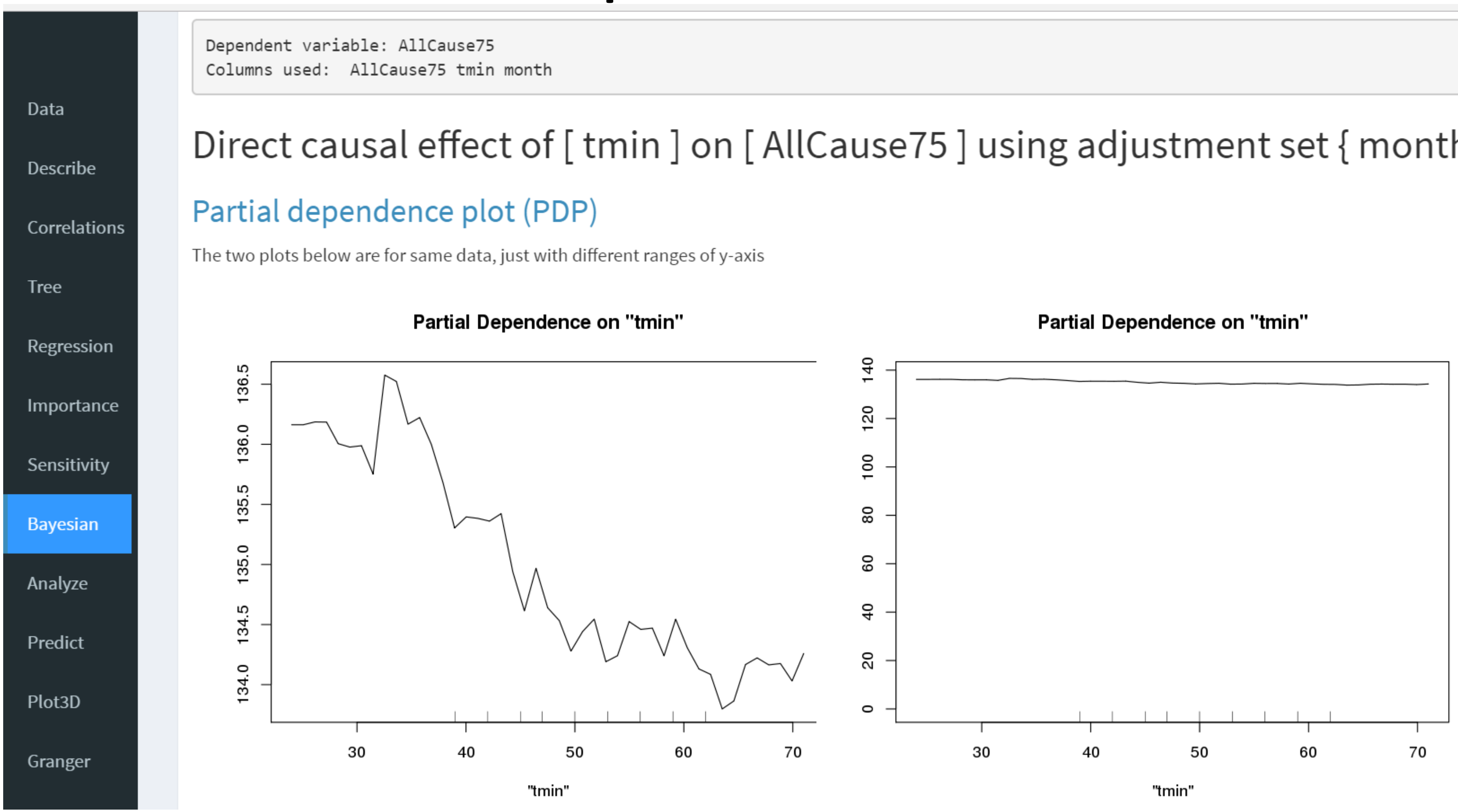
Constrained automated non-parametric causal model

Final models can be used to estimate causal input-output relations



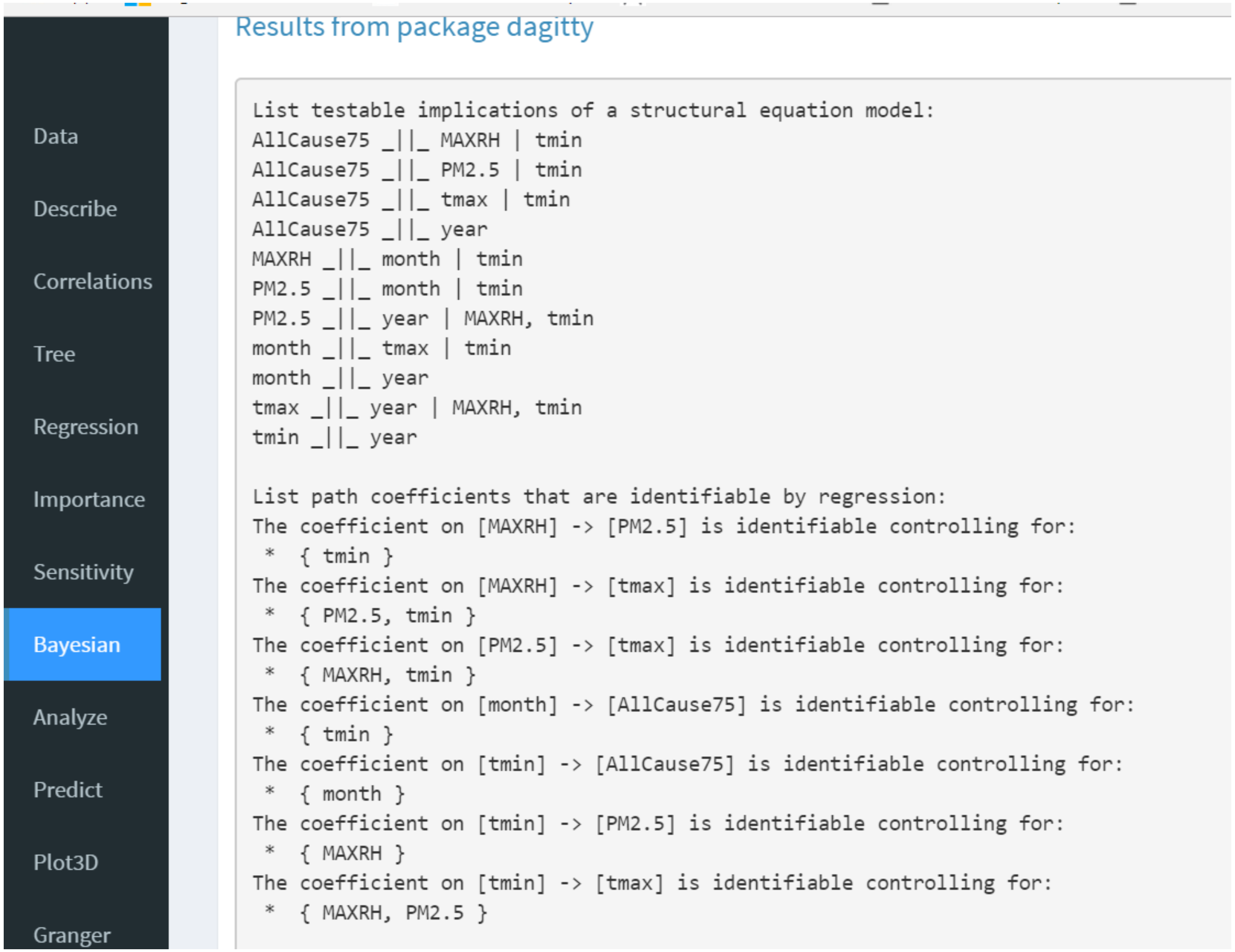
Constrained automated non-parametric causal model ensembles can quantify causal relations

Automated estimation of causal input-output relation



Constrained automated non-parametric causal model ensemble result:
Mortality risk decreases slightly with same-day minimum temperature

Automated interpretation of statistical implications of a DAG model



Interpretation:

- Sound and complete inference algorithms generate all testable implications of DAG model learned from data
- Algorithms compute adjustment sets for estimating direct and total effects of changes in one variable on another for a given BN/DAG *if* its arrows and CPTs are causal

BN = Bayesian network

Summary: Machine learning helps avoid p-hacking and discover predictive causal relations

- Automated (but appropriate/intelligent) analyses can be carried out with current ML software for many real air pollution health effects data sets
  - Non-parametric
  - Information-based
  - Causal knowledge-constrained
  - Ensembles
  - Enabled by existing R packages: randomForest, bnlearn, dagitty, CompareCausalNetworks, etc.

Some useful extensions

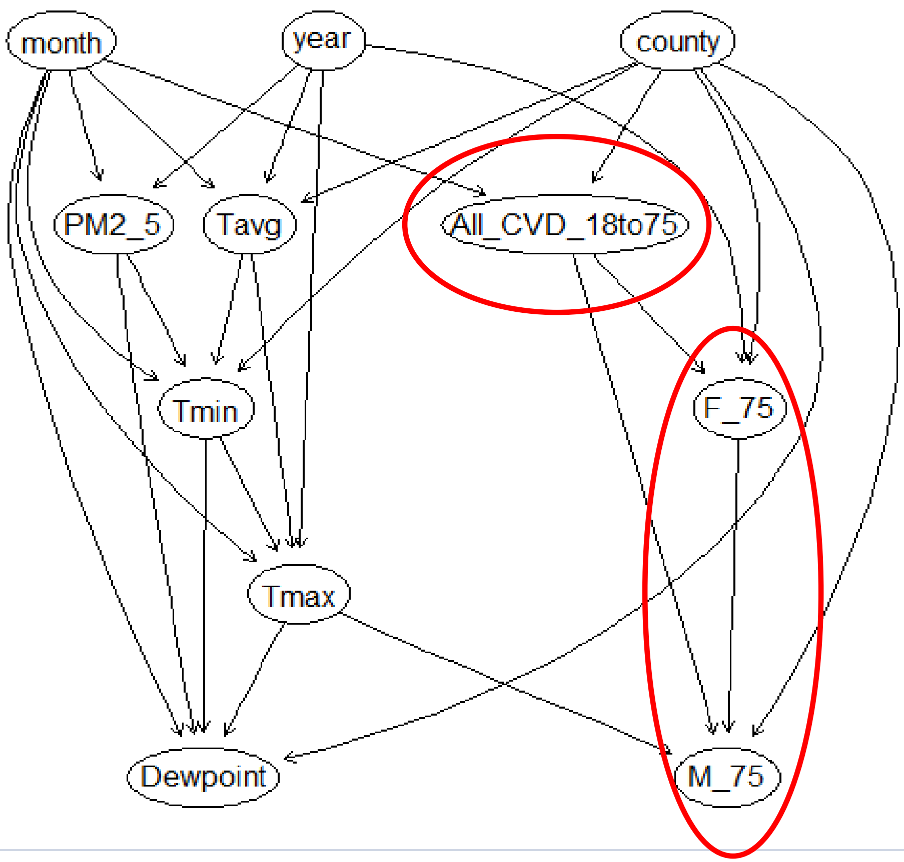
- Detecting omitted confounders
- Beyond DAGs
  - Allow for undirected arcs, cycles
- Transportability of results across settings
  - Appropriate generalization: Causal conditional probability tables (CPTs) are invariant, distributions of risk factors are not
- Combining results across studies
  - Different constraints from different studies
  - Causal CPTs are invariant across studies

Detecting hidden/omitted variables

CAT_bnLearn (M_75,PM2_5,F_75,month,year,All_CVD_18to75,Tavg,Tmin,Tmax,Dewpoint,county)

Bayesian Network diagram.

An arrow between two variables shows that they are informative about each other.



Network discovered by bnlearn

Boston data

- Daily death counts in disjoint subpopulations are correlated
- Latent (hidden) variables affect both
- F_75 = daily deaths among women 75 or older predicts
- F_75 predicts M_75
- All_CVD_18to75 predicts (is informative about) both

Information-based causal discovery algorithms in perspective

- Philosophical underpinnings
  - Information flows from causes to effects over time
  - Tracking information flows enables data-driven causal discovery
  - Discovery = empirical constraints on possible models from observed information patterns in data
    - Differs from formulating a hypothesis and then testing it: Causal discovery imposes no *a priori* hypotheses
    - Causal interpretation and orientation of arrows may require weak knowledge-based constraints

Practical aspects

- *Study design:* Ideally, track changes in exposures, covariates, and outcomes over time
  - *Data requirements* for causal discovery algorithms: Flexible (panel, time series, cross-sectional, etc.)
- *Assumptions:* Predictive causation + knowledge-based constraints provide a useful surrogate for manipulative causation
- *Model choices:* Learn tree ensembles, networks
  - Minimal assumptions, non-parametric, learned from data rather than assumed a priori
  - Use/compare multiple algorithms and principles
- *Sensitivity* to modeling choices: So far, causal model structure and estimates are robust to choice of algorithms
  - *CompareCausalNetworks* package
  - Model cross-validation

Caveats for information-based causal discovery algorithms

- Key assumptions:
  - *Data are available* to reveal information patterns and flows
    - Can be longitudinal or cross-sectional, many epidemiological and quasi-experimental (QE) designs suffice
  - *Effects are large enough to be detected* using non-parametric algorithms.
    - Power calculations reveal detection limits
    - (Causal Markov Condition, faithfulness, etc. useful but not essential)
- Limitations:
  - Unique identifiability from data not always possible $\rightarrow$ Must use multiple plausible models (model ensemble)
    - Arrow directions may be unclear, even in principle
    - Example: Income and air pollution
  - Predictive causation ≠ manipulative causation
  - Not yet well vetted for air pollution health effects research
    - Well vetted via Kaggle and other competitions in machine learning and causal learning communities

Conclusions

- Advice
  - Machine learning/information-based causal discovery is ready to apply to air pollution health effects data
    - Current software makes causal discovery relatively easy
  - Focus on predictive and manipulative causation (vs. other, e.g., associational/attributive or counterfactual, causation)
  - Focus on how well *changes* over time predict each other
    - Include at least 2 weeks of daily temperatures as lagged confounders in time series studies of daily mortality/morbidity
  - Use non-parametric model ensembles to avoid model specification errors, p-hacking, etc.
- Future research
  - Vet for air pollution health effects research
  - Compare information-based to potential outcomes methods in Kaggle-type competitions

Suggested readings

www.cox-associates.com/CausalAnalytics/

- Pearl J, 2009. Causal inference in statistics: An overview.
  - https://projecteuclid.org/euclid.ssu/1255440554
- Laganu V et al., 2016. Probabilistic Computational Causal Discovery for Systems Biology.
  - www.cox-associates.com/CausalAnalytics/CausalDiscoverySystemsBiologyLagani2016.pdf
- Cox LA Jr., 2017. Do causal concentration-response functions exist? A critical review of associational and causal relations between fine particulate matter and mortality
  - www.ncbi.nlm.nih.gov/pubmed/28657395
- Cox LA Jr., 2017. Socioeconomic and air pollution correlates of adult asthma, heart attack, and stroke risks in the United States, 2010-2013.
  - https://www.ncbi.nlm.nih.gov/pubmed/28208075

Thanks!