Extracting Causal Rules from Spatio-temporal Data

Antony Galton[1], *Matt Duckham*[2], and *Alan Both*[2]

[1] Exeter University, Exeter, UK
(supported by EPSRC project EP/M012921/1)
[2] RMIT University, Melbourne, Australia
(supported by ARC projects DP120100072 and DP120103758)

COSIT 2015, Santa Fe, New Mexico

Overview

We report work with data concerning the movement of fish in the Murray River system in S. E. Australia.

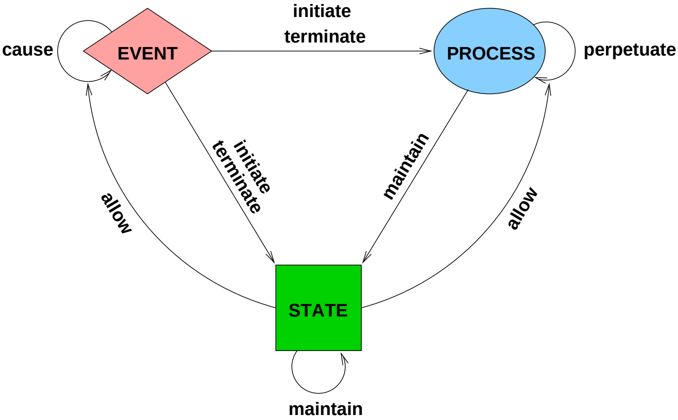
The data comprises a large number of records of individual fish movements, together with records of a number of environmental variables such as water temperature, water level, and salinity.

We are interested in discovering **causal rules** relating the variation in the environmental variables to fish movement upstream or downstream.

We searched for such rules systematically using an algorithm targeted towards discovering rules having a specified form.

States, Processes, Events, and Causation

We take the view (following Galton's paper in FOIS2012) that causal relations between **events** are different in character from causal relations between **processes**, and that **states** can play the role of **enabling conditions** for process and event causation.



Data

We investigate causal rules in relation to data in the form of one or more **history files**, which record the occurrences of events and the values of process variables over a time period $T = [0, t_{max}]$.

Events ($\mathcal{E}$) and processes ($\mathcal{P}$) are collectively **occurrents** ($\mathcal{O}$):

$$\mathcal{O} = \mathcal{E} \cup \mathcal{P}, \qquad \mathcal{E} \cap \mathcal{P} = \emptyset.$$

An event-type is represented as a function $e : T \to \mathbb{Z}^+ \cup \{0\}$, where $e(t)$ is the number of distinct occurrences of e at time t.

A process is represented as a function $p : T \to \mathbb{R}$ giving the **values** of the process, considered as the continuous variable of some quantity over time.

Causal Rules

We work with causal rules of the form

$$R : [\text{Causes}_R \mid \text{Conditions}_R] \Rightarrow \text{effect}_R \ \text{ after } \text{Delay}_R,$$

where

- $\text{Causes}_R \subset \mathcal{E}$ is a set of effects functioning as **causes**,
- Conditions_R is a set of **conditions**, where each condition is a triple $c = (p_c, v_c^-, v_c^+) \in \mathcal{P} \times \mathbb{R} \times \mathbb{R}$,
- $\text{effect}_R \in \mathcal{E} \setminus \text{Causes}_R$ is an event distinct from any of the causes, functioning as an **effect**,
- Delay_R is a **delay interval** $[d_R^-, d_R^+]$, where d_R^-, d_R^+ are integers such that $0 \leq d_R^- \leq d_R^+$.

In a condition, v_c^- and v_c^+ are the limits of a range within which the value of p_c must fall to satisfy it.

Rule Activation

- The causal rule

$$R = [\mathsf{Causes}_R \mid \mathsf{Conditions}_R] \Rightarrow \mathsf{effect}_R \ \text{ after } \mathsf{Delay}_R$$

  is **activated** at time t if and only if both:

  1. For every $e \in \mathsf{Causes}_R$, $e(t) > 0$.
  2. For every $c \in \mathsf{Conditions}_R$, $v_c^- \leq p_c(t) \leq v_c^+$.

- An activation of the rule at time t is **explanatory** if the effect predicted by the rule does indeed occur, i.e.:
  - For some $d \in \mathsf{Delay}_R$, $\mathsf{effect}_R(t + d) > 0$.
- An occurrence of effect_R at time t is **explained** by rule R if some activation of R is made explanatory by that occurrence of the effect, i.e.,
  - For some $d \in \mathsf{Delay}_R$, R is activated at $t - d$.

The Problem

Our algorithm is designed to solve the following problem:

- Given a data set as described, we seek a set of rules $\mathcal{R}$ which, **as nearly as possible**, accounts fully for the data, in the following sense:
  - *No false positives*: For each $t \in T$ and $R \in \mathcal{R}$, if R is activated at t then it is explanatory, i.e., effect_R occurs after an admissible delay.
  - *No false negatives*: For each occurrence of each effect f in the data, there is a rule $R \in \mathcal{R}$ which explains it, i.e., $f = \text{effect}_R$ and R is activated within an admissible delay time preceding the occurrence.

Evaluating a rule set

Because of the delay factor in our causal rules, sensitivity and precision are defined in a somewhat non-standard way:

- **Cause-based precision**

$$\textit{c-precision} = \frac{cTP}{cTP + cFP} \qquad \text{where}$$

  - cTP is the number of explanatory activations of R
  - cFP is the number of non-explanatory activations of R

- **Effect-based sensitivity**

$$\textit{e-sensitivity} = \frac{eTP}{eTP + eFN} \qquad \text{where}$$

  - eTP is the number of occurrences of effect_R which are explained by R.
  - eFN is the number of occurrences of effect_R that are not explained by R

Rough Outline of the Rule-Detection Algorithm

The algorithm searches systematically for explanatory rules which can account for the data. For details see the paper.

- For each effect f and each subset E of the available causes, we consider whether any of the data for f can be explained by a rule whose cause-set is E.
- For each subset E which passes certain tests we then determine a suitable delay interval $[d^-, d^+]$.
- If for every time at which all the causes in E occur, f occurs after a delay in the interval $[d^-, d^+]$, we have an **unconditional rule** $E \Rightarrow f$ after $[d^-, d^+]$.
- Otherwise, we look for processes that can provide conditions for **conditional rules** of the form $[E \mid v^- \leq p \leq v^+] \Rightarrow f$ after $[d^-, d^+]$.

Experiments

We performed three sets of experiments:

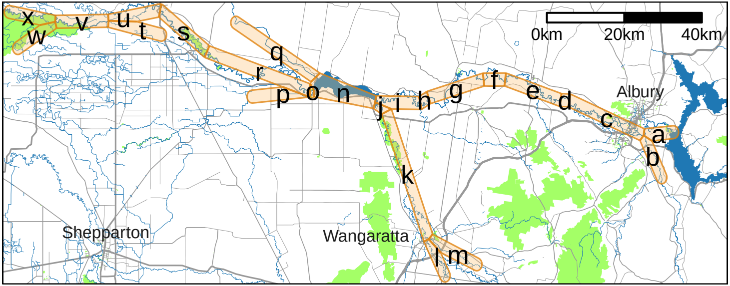
1. *Extracting causal rules from synthetic data generated using known rules.*
   The results (see paper) were very favourable, with 100% sensitivity and precision in most cases.
2. *Extracting unconditional rules from real-world data.*
   Here we used as candidate causes events defined in terms of the processes in the data; no conditions were looked for. The results were disappointing (see paper).
3. *Extracting "always" rules from real-world data.*
   Here we were looking for process-perpetuation, expressed by rules with a trivial "always" event as cause, and conditions using the processes in the data. The results were promising but equivocal.

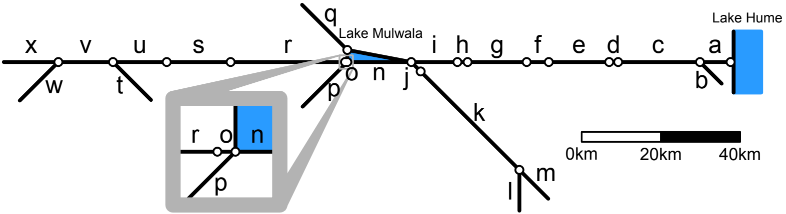
The Real-World Data

The data-set concerns fish movement in the Murray River system (Lyon *et al.* 2011).

- Over 1000 individual fish were tagged with radio transmitters.
- Their movements were monitored by 18 river-side radio receivers, defining 24 *zones* in the river system, labelled a–x.
- The movement of tagged fish between zones was tracked over six years.
- At the same time water temperature, water level, and salinity data were collected.

Map of the study area, showing zones a–x





The Data

The data consisted of records of the following types:

- For each environmental variable, a record of its value at each recording station on each day of the period of study;
- A collection of records of zone-boundary crossings by individual fish, where each record takes the form "fish i moves from zone z_1 to zone z_2 on day d".

The aim of the study was to determine to what extent the movement of fish was causally influenced by the variations in the environmental variables.

Fish-Movement Events

The fish-movement event types were defined using the data as follows:

For each pair z_1, z_2 of adjacent zones, where z_2 is downstream from z_1,

- event $\mathrm{z}_1\backslash\mathrm{z}_2$ occurs whenever a fish moves from z_1 to z_2,
- event $\mathrm{z}_2/\mathrm{z}_1$ occurs whenever a fish moves from z_2 to z_1.

Note that it is possible for there to be several occurrences of any one of these events on any given day.

Experiment 3: Perpetuation

For this experiment we used the algorithm to look for "always" rules of the form

$$[\text{always} \mid \text{Conditions}] \Rightarrow \text{effect} \ \text{ after Delay}$$

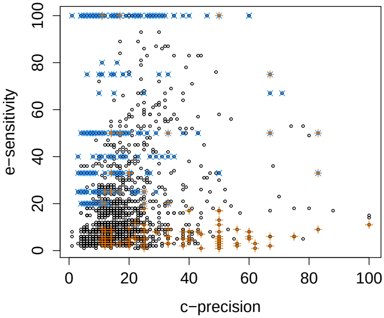
which for clarity we write in shorter form as

$$\text{Conditions} \Rightarrow \text{effect} \ \text{ after Delay}.$$

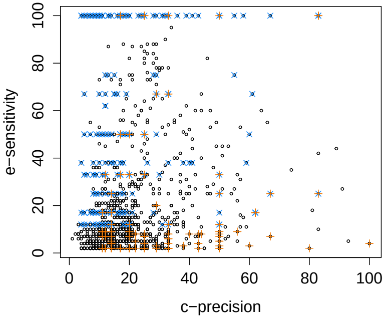
The conditions are expressed in terms of value-ranges for the environmental variables.

The effects to be explained were fish-movement events of the forms $x \backslash y$ and x/y. Understanding these as proxy indicators for fish-movement *processes*, these rules identify possible *perpetuation* relations between environmental processes and fish movements.

Support



a. Upstream movement



b. Downstream movement

< 10 condition instances

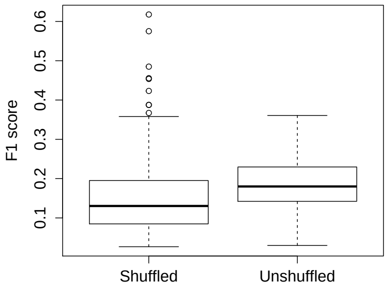
< 10 effect instances

Spatial coincidence

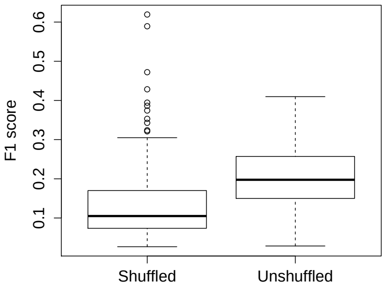
Is an effect in any way spatially related to its condition?

- No evidence to support the hypothesis that rules that relate spatially proximal conditions and effects are associated with higher F_1 scores ($p = 0.39$ upstream, $p = 0.89$ downstream).
- May be a consequence of spatial autocorrelation and granularity effects.

Shuffled data

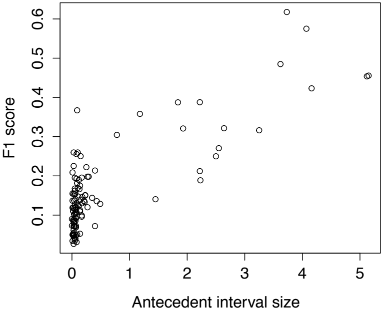


a. Upstream movement

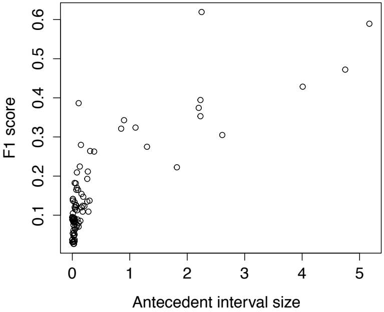


b. Downstream movement

Condition value ranges



a. Upstream movement
(Pearson's $R = 0.807$)



b. Downstream movement
(Pearson's $R = 0.813$)

Condition value ranges

Best rule found	F_1 score
$2.79 \leq \text{wl(cd)} \leq 4.72 \Rightarrow \text{d/c after } [0,5]$	0.32
$2.39 \leq \text{wl(efgh)} \leq 5.03 \Rightarrow \text{e/d after } [0,5]$	0.32
$2.81 \leq \text{wl(efgh)} \leq 5.03 \Rightarrow \text{f/e after } [0,5]$	0.38
$1.77 \leq \text{wl(efgh)} \leq 1.92 \Rightarrow \text{g/f after } [0,5]$	0.25
$0.77 \leq \text{wl(efgh)} \leq 1.55 \Rightarrow \text{h/g after } [0,5]$	0.30
$126.41 \leq \text{wl(ijklm)} \leq 131.53 \Rightarrow \text{i/h after } [0,5]$	0.45
$126.85 \leq \text{wl(ijklm)} \leq 128.69 \Rightarrow \text{i/j after } [0,5]*$	0.39
$126.98 \leq \text{wl(ijklm)} \leq 128.16 \Rightarrow \text{j/i after } [0,5]*$	0.36
$126.89 \leq \text{wl(ijklm)} \leq 126.92 \Rightarrow \text{j/k after } [4,5]$	0.26
$124.67 \leq \text{wl(np)} \leq 124.75 \Rightarrow \text{n/i after } [0,5]$	0.26
$1.60 \leq \text{wl(or)} \leq 6.75 \Rightarrow \text{o/n after } [0,5]$	0.46
$3.02 \leq \text{wl(or)} \leq 6.75 \Rightarrow \text{r/o after } [0,5]$	0.62
$2.24 \leq \text{wl(stuv)} \leq 6.40 \Rightarrow \text{s/r after } [0,5]$	0.42
$2.33 \leq \text{wl(stuv)} \leq 6.40 \Rightarrow \text{u/s after } [0,5]$	0.58
$2.78 \leq \text{wl(stuv)} \leq 6.40 \Rightarrow \text{v/u after } [0,5]$	0.48

The best rules discovered for each upstream movement effect.

Summary and conclusions

We developed an algorithm capable of mining causal rules of a particular logical form that best fit the data.

- Performance with synthetic data is very good.
- Performance with real data failed to identify strict causation (possibly granularity effects).
- Perpetuation rules were more successful with real data, including:
  - Perpetuation rules appear to relate to meaningful structure.
  - Top-ranked rules compactly describe approximately 20% of fish movements.