International Journal of Geographical Information Science
Vol. 00, No. 00, Month 2012, 1–20

Research Article

Mining candidate causal relationships in movement patterns

Susanne Bleisch[a,*], Matt Duckham[a], Antony Galton[b], Patrick Laube[c], Jarod Lyon[d]

[a]University of Melbourne, VIC, Australia; [b]University of Exeter, UK; [c]University of Zurich, Switzerland; [d]Arthur Rylah Research Institute, VIC, Australia
(*Received 00 Month 200x; final version received 00 Month 200x*)

In many applications, the environmental context for, and drivers of movement patterns are just as important as the patterns themselves. This paper adapts standard data mining techniques, combined with a foundational ontology of causation, with the objective of helping domain experts identify candidate causal relationships between movement patterns and their environmental context. In addition to data about movement and its dynamic environmental context, our approach requires as input definitions of the states and events of interest. The technique outputs causal and causal-like relationships of potential interest, along with associated measures of support and confidence. As a validation of our approach, the analysis is applied to real data about fish movement in the Murray River in Australia. The results demonstrate the technique is capable of identifying statistically significant patterns of movement indicative of causal and causal-like relationships.

Keywords: movement patterns, context-aware movement analysis, sequence mining, causation, geosensor networks, environmental monitoring

1. Introduction

Context is central to understanding movement. For example, in the field of movement ecology there is agreement that animal movement can only be understood through a study of both the movement and the embedding geographical context (Nathan *et al.* 2008).

However, to date, relatively few techniques help in the identification of the contextual *drivers* of movement (Andrienko *et al.* 2011, Gudmundsson *et al.* 2012). Instead, previous work has focused strongly on techniques for characterizing individual or group movement, including individual trajectory segmentation, clustering groups of trajectories, and even divining basic laws governing human mobility on a population level (Gudmundsson *et al.* 2012, Gonzalez *et al.* 2008).

The new technique proposed and evaluated in this paper aims to identify candidate *causal* relationships between movement data and the environmental context in which that

*Corresponding author. Email: susanne.bleisch@unimelb.edu.au

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

movement occurs. Our approach is based on an adaptation of established data mining techniques, and is applied to a specific example of long-term fish monitoring. The results, validated though comparison with patterns of random movement, demonstrate how the technique can be used to identify plausible environmental causes of fish movement.

2. Background

2.1. *Movement analysis*

Following the rapid development of tracking technologies, the study of moving individuals has received much attention, both theoretical and applied (Laube *et al.* 2011, Gudmundsson *et al.* 2012). The ability to track large numbers of individuals at previously unobservable spatial and temporal granularities requires new methods for the analysis of individual trajectories and collective motion. Methodological work includes techniques for segmenting individual trajectories (Buchin *et al.* 2011a, Pelekis *et al.* 2012) and assessing the similarity of trajectories for clustering (Buchin *et al.* 2011b, 2012). The study of moving groups has resulted in many different yet related definitions and algorithms for the detection of "flocks" (Laube *et al.* 2005, Gudmundsson *et al.* 2007), "convoys" (Jeung *et al.* 2008), "herds" (Huang *et al.* 2008), and "leaders" (Andersson *et al.* 2008, Nagy *et al.* 2010). In view of this diversity, the ontological foundations of collective motion have also received much-needed attention (Wood and Galton 2009a,b). Application domains have also embraced the new data sources. Whereas social physicists aim to discover basic laws governing human mobility at a population and even individual level (Gonzalez *et al.* 2008, Schneider *et al.* 2013), behavioral ecologists welcome the ability to follow their study subjects at every turn (Nathan *et al.* 2008).

Most movement analysis work so far focused on: (i) objects moving without constraints in a Euclidean two-dimensional space (e.g., migrating geese, Buchin *et al.* 2012); (ii) object trajectories monitoring positions over time (most tracked animals in movement ecology, Nathan *et al.* 2008); and (iii) on analyzing the shape of trajectories and by that largely ignoring the embedding of the movement in its geographic context (e.g., segmenting trajectories based on speed or sinuosity, Buchin *et al.* 2011b).

However, most objects moving in geographic space will in one way or another be constrained. Human movement is highly constrained, as we usually depend on transportation infrastructure best modeled as a network space (Gudmundsson *et al.* 2012, Duckham 2012). Even migrating geese will be constrained in their movement by winds and feeding sites along their route. Constrained movement furthermore offers an alternative to conventional GPS trajectory tracking, as moving objects can be monitored when passing checkpoints of cordons with fixed positions (Both *et al.* 2012). Application-driven research strongly suggests that a complete understanding of movement and the processes driving it can only be achieved when studying movement in combination with the geographic space in which the movement is embedded (Nathan *et al.* 2008, Andrienko *et al.* 2011). The approach presented in this article aims at exactly this: relating constrained movement with the changing geographical and environmental context for that movement.

2.2. *Causation*

Although causality has been actively investigated in philosophy for many centuries, with a literature far too voluminous for even a cursory survey here (but see Beebee *et al.* 2009, for some initial pointers), systematic treatments of causality in GIScience have been few and far between, although exceptions can be found.

Yuan (2007) introduces the notion of “geographic dynamics,” whereby directly observable changes and movements result from (presumably unobservable) “drivers,” which she characterizes as activities, events, and processes. The problem confronting the researcher is to infer the latter from the former. Clearly this cannot be done effectively in the absence of generalizable regularities in the behavior of the drivers and their relation to the observables—in effect a suite of causal laws governing the evolution of the geographical system under study (although this aspect seems to be rather under-emphasized in Yuan’s highly programmatic account). More details relating to this approach can be found in Yuan and Hornsby (2008).

Explicit reference to causality is similarly downplayed in works such as Claramunt and Thériault (1995, 1996), which provide detailed analyses of the possible forms of spatiotemporal evolution in the geographical domain. Such analysis is a prerequisite for causal modeling, but so long as this linkage remains implicit there is a danger of conflating distinct forms of causal and causal-like relationship which play different roles in our understanding of a phenomenon. These different relationships were addressed by Galton and Worboys (2005) and Galton (2012), on which we base the account in this paper.

An early paper by Allen *et al.* (1995) is particularly interesting for our purposes because of the importance accorded to *conditional causality*, by which “the cause ... must be interpreted as a ‘trigger’ of a process which cannot occur without certain external or internal conditions, and not as a necessary and sufficient producer of the effect” (op. cit, p.403, with reference to Bunge 1966). In this picture, it is natural to model the “trigger” as an event, and the condition which makes it causally efficacious as a state. In view of this, therefore, Allen et al. “do not consider one state to have been ‘caused’ by another, but rather one change of state in an object to have been ‘caused’ by another change of state of either the same object or a different object.” This fits in well with the approach of Galton (2012) which we follow in this paper, as discussed below (Section 3).

El-Geresy *et al.* (2002) likewise ascribe the cause and effect relation to events (“changes”) and not to states, and some of their remarks concerning the relative timing of cause and effect are of relevance to us here. Although one might expect an effect to occur as soon as its cause occurs, they note that the effect may be delayed, either because the cause must attain some intensity threshold before the effect can occur, or because the cause and effect are spatially separated, and it takes time for the influence of former to reach the latter. Rather than taking El-Geresy *et al.* at face value here, we would suggest that the proper description of cases like this is sensitive to the granularity at which the phenomena are described. One of their examples concerns a case where the release of pollutants into a river causes the death of vegetation at a certain point downstream. At this coarse level of description we do indeed appear to have a case of delayed causality. But at a finer granularity this appearance is dispelled: first, the release of the pollutants causes the pollutants to start flowing downstream; this leads, after a time, to the pollutants reaching the vegetation, an event which causes the vegetation to die. The two cases of “causes” here are, at this level of granularity, effectively instantaneous, while the delay between them results from the finite speed of the river’s flow (which *leads to*, but does not *cause*, the arrival of the pollutants at the vegetation).

3. Ontological model

We adopt for the ontological foundations of this work the approach of Galton (2012), summarized in Figure 1. By starting from solid ontological foundations, our aim is to ensure our approach is flexible enough to be useful in a range of applications beyond our specific example of fish movement.

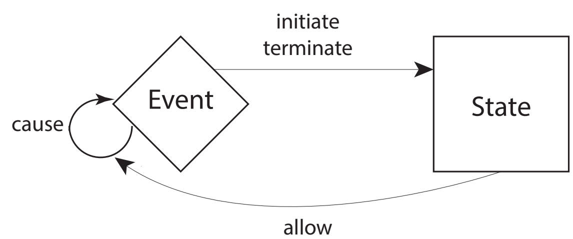

Figure 1. Causal-like relationships amongst states, events, and processes, after Galton (2012)

In summary, after Galton (2012):

- Only events may strictly *cause* other events.
- Events may *initiate* or *terminate* states.
- States (of the world) only affect causation in as much as they can *allow* events to cause other events.

Like Galton (2012), we refer to the relationships "initiate," "terminate," and "allows" as *causal-like* relationships to distinguish them from strict event-event causation. After Galton (2012), events are defined as temporally bounded "happenings" where one or more participants in that event change. Galton's ontology of causation additionally accounts for processes, defined as an "an open-ended homogeneous activity" akin to a "state of change." However, in this paper we are concerned solely with events, and leave an investigation into the causal role of processes as a matter for future work.

Events, states, and their inter-relationships can all play important roles understanding movement and its context. For example, Figure 2 shows diagrammatically a possible causal explanation of fish movement, our motivating application in this paper. In Figure 2, a full-moon event *causes* the start of a fish migration event. This cause is *allowed* by the state of high river flow, itself *initiated* by an event, the start of high river flow.

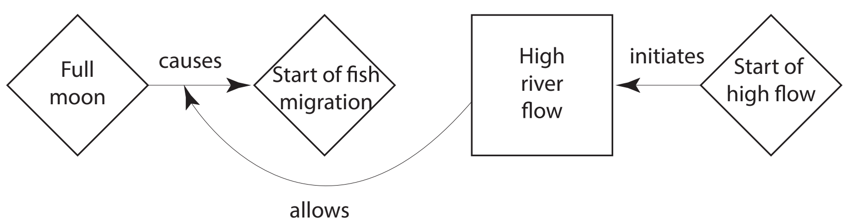

Figure 2. Example ontology of fish migration causation (cf. Figure 1)

3.1. *Co-location and granularity*

Armed with this (aspatial) ontology, we make one further (spatiotemporal) assumption: that spatiotemporal co-location is a prerequisite of causal and causal-like relationships. In order for one spatial event to cause another, those events must share at least one point in space and time. Similarly, a state can only allow causation of events in its spatiotemporal vicinity.

Counterexamples to the principle of co-location may appear to occur where an event in one spatiotemporal location is cited as the cause of an event in a remote location (cf. El-Geresy *et al.* 2002). However, as explained above (cf. Section 2.2) we interpret these apparent counterexamples as the result of describing the phenomena in a coarse-grained way that glosses over an intermediate chain of causal or causal-like connections. In detail, each causal link in the chain is expected to satisfy the spatiotemporal co-location requirement.

Specifically, we identify three distinct granularity effects that can lead to our principle of strict spatiotemporal co-location being obscured in practical observations and data:

- *Causal granularity* Causal relationships such as those in Figure 2 are at a relatively coarse level of granularity. As we have argued above, there may in practice be a chain of unobserved, finer-grained causation. For example, at a finer granularity river flow and temperature events may cause changes to a variety of physiological and chemical processes in fish. In turn, these may subsequently cause a migration event.
- *Temporal granularity* Even in cases where causation is instantaneous, the unavoidable temporal granularity in data may lead to an apparent lag between causing and caused event. Sampling granularity may mean that the start of a migration event, for example, may not in practice be observed until a short time after it began in actuality.
- *Spatial granularity* Similar to temporal granularity, spatial granularity may lead to an apparent lag between causing and caused event. An event such as an increase in river flow in a section of the river will not strictly and immediately hold over all of that section.

In summary, we assert that *events can only initiate or terminate the states with which they are co-located*. Apparent counter-examples to this principle we ascribe to one or more of the granularity effects discussed above.

3.2. *Caveat*

As was famously pointed out by Hume (1739), empirical data alone can never furnish conclusive evidence of causal relationships. Thus, the techniques developed in this paper cannot be claimed to identify causal or causal-like connections as such. Instead, their purpose is to reveal prima facie *candidates* for such connections: sequences of co-located events, or states and events, which are plausible to consider as exhibiting causal or causal-like relationships. It is to be expected that these *candidate* causal relationships may be corroborated or refuted by closer analysis that goes beyond the immediate data. In this way, we expect our technique could assist domain experts in exploring hypotheses about what causal relationships may explain patterns in dense movement data sets.

4. Analysis method

In this section we outline our analysis method for deriving candidate causal and causal-like relationships, discussed in the previous section, from the combination of movement and environmental data.

Our approach is based on the combination of two well-established data mining techniques: *association rule mining* and *sequence mining*. With reference to Figure 2, in

Section 4.1 we show how association rule mining can be used as the basis for inferring causal-like "allows" relations. Section 4.2 then applies sequence mining to the problem of inferring true causal relations between events. In Section 4.3 we discuss the definition of events and states, effectively yielding the causal-like relations "initiates" and "terminates." Finally, Section 4.4 briefly outlines our implementation, built on top of open-source R software packages.

The illustrative examples used to explain (and in Section 5 validate) our technique make reference to data concerning fish movements in the Murray River, in south eastern Australia. As part of a separate study of the effect of conservation activities upon fish populations (Lyon 2012), a major environmental monitoring project has tagged upwards of 1000 fish in the Murray River with radio transmitters. River-side radio receivers at 18 strategic locations along the course of river partitioned the river and its tributaries into 24 zones. Over a period of six years, the movement of tagged fish between different zones was tracked.

The project is providing important insights into the effects of a restoration intervention to the river to improve fish habitats (with the ultimate aim of bolstering fish populations). However, in addition to the statistical analyses of fish populations undertaken in the original study (see Lyon 2012), this data set could also help in understanding important causal relationships connected with the ecology of this environmentally and economically vital river system. For example, domain experts may be interested to know: Do moon phases or high river flow events cause certain patterns of long-range movement in tagged fish? Or do low water temperatures lead to fish staying within limited home ranges of the river? The technique developed and explained below aims to assist in the identification of such causal and causal-like relationships. In our experience, such tools can in turn assist ecological domain experts in formulating and testing different hypotheses about the (first-order) context of animal movement.

4.1. *Association rule mining*

Association rule mining is often explained using the example of the "market basket" analysis. Given a data set of supermarket shopping transactions, association rule mining is able to identify associations between sets of items that customers tend to buy together (Agrawal *et al.* 1993). For example, association rule mining can help in answering questions such as: "How frequently do customers that buy beer also buy crisps?"

Instead of shopping transactions, our analysis applies association rule mining to appropriately formatted spatiotemporal data, in order to identify candidate causal-like "allows," "initiates," and "terminates" relationships between environmental states and movement events. Table 1 shows an example of appropriately structured fish movement data. We assume a set I of moving-object identifiers (in our case, tagged fish IDs) and a set T of timestamps (days, in the case of our fish tracking example). Next, we assume a set S of observed environmental states of interest (for example, a state of moderate water temperature, labeled $wt4_s$, or high river flow, labeled $rf4_s$) together with a set M of movement events (such as upstream, u_e, or downstream, d_e, movement). To avoid any confusion, we use the subscripts "e" and "s" to distinguish events and states respectively.

Pairs from the set $I \times T$ form our "transactions" in the association-rule mining terminology. For each pair $(i, t) \in I \times T$, it is then possible to list as an "itemset" the environmental states "experienced" by moving object i at time t. "Experienced" in this sense means specifically "spatially and temporally co-located with" (see Section 3). Finally, we also add to the "itemset" for the pair (i, t) any movement events that occurred to object i at time t. To stretch the market-basket analogy, a moving object at a particular time "buys" the environmental states it experiences along with the movement events it participates in.

Transaction ID		*Itemset*
Fish identifier (I)	Timestamp (T)	States and events ($2^{S \cup M}$)
41937-610-67	1753	$\{rf4_s, m_e, d_e\}$
43521-530-68	1754	$\{rf4_s, m_e, d_e\}$
41937-610-67	1755	$\{rf4_s, m_e, d_e\}$
41937-610-67	1756	$\{rf4_s, wt3_s, m_e, d_e\}$
41937-610-67	1771	$\{rf4_s, wt3_s, m_e, u_e\}$
43521-530-68	1772	$\{rf4_s, wt3_s, m_e, d_e\}$
41937-610-67	1779	$\{rf4_s, wt4_s, m_e, u_e\}$
...	...	...

Table 1. Example state table, showing the environmental states "experienced" and movement events "participated in" by two fish (IDs 41937-610-67 and 43521-530-68) in consecutive time stamps (1753, 1754, 1755, 1756, 1771, 1772, 1779). For example, the "itemset" $\{rf4_s, m_e, d_e\}$ (high river flow together with fish movement downstream) occurs five times in this table (see Section 4.3 for an explanation of the state definition).

The output of association rule mining is the frequency of specified rules. For example, for a rule such as "beer $\implies$ crisps" (that customers that purchase beer also purchase crisps) the output of association rule mining is the frequency with which beer and crisps appear in the same shopping transaction. In data mining terminology, the frequency of the co-appearance of items in a transaction relative to the total number of transactions is termed the *support* of a rule. In contrast, the frequency of co-appearance of items relative to the frequency of the transactions containing the antecedent in the rule is termed the *confidence* of that rule (Mohammad and Nishida 2010).

Turning back to our moving object data, we can interpret the support and confidence of a rule $m \implies s$, for movement event $m \in M$ and environmental state $s \in S$, as measures of the strength of evidence that state s "allows" movement event m to occur. This two-step mapping—first from movement and environmental data to association rule mining input and second from association rule mining output to causal-like relationship—is at the core of our approach to mining causal and causal-like relationships.

4.2. *Sequence mining*

The causal-like relationship "allows" is in fact not the main focus of our approach. Instead our primary focus is on true causal relationships between environmental and movement events. To identify candidate causal relationships, we take our approach a step further, and use the technique of frequent sequence mining.

Frequent sequence mining is an extension of association rule mining that additionally accounts for the *order* in which items were bought. For example, sequence mining can help in answering questions such as: "How often do customers that buy beer and crisps subsequently buy headache tablets and fruit?" (Zaki 2001).

Turning once more to spatiotemporal data about movement and its environmental context, we can now construct a table of movement and environmental events, such as that in Table 2, in a similar way to Table 1. The key difference between Tables 1 and 2 is that the itemsets in Table 2 are from the combination of the set M of movement events and a set of environmental events of interest, E (for example, the *start* of a high river flow state, labeled $rf4_e$, or a full moon event, labeled fm_e). By contrast, Table 1 combines the set states S with the set of movement events M.

Sequence ID Fish identifier (I)	*Time* Timestamp (T)	*Itemset* Events ($2^{E \cup M}$)
25598-350-66	1657	$\{rf4_e, m_e, d_e\}$
25598-350-66	1659	$\{fq_e\}$
25598-350-66	1665	$\{rf3_e\}$
25598-350-66	1667	$\{fm_e, rf2_e\}$
25598-350-66	1669	$\{m_e, d_e\}$
25598-350-66	1675	$\{lq_e, rf3_e\}$
25598-350-66	1676	$\{rf4_e\}$
...	...	...

Table 2. Example event table, showing the environmental events "experienced" by a fish (ID 25598-350-66) in seven consecutive time stamps (1657, 1659, 1665, 1667, 1669, 1675, 1676) during the first 20 days of monitoring this fish. For example, downstream (d_e) movement (m_e) at timestamp 1669 follows the full moon (fm_e) and start of lower river flow ($rf2_e$) events at timestamp 1667 (see Section 4.3 for an explanation of the event definition).

Sequence mining again outputs the frequency of co-appearance, but for specified *sequences* rather than rules. Nevertheless, we again interpret the frequency of a specified *sequence* of events, $v_e \to m_e$ for some $v_e \in E$ and $m_e \in M$, as an indication of the strength of evidence that environmental event v_e "caused" movement event m_e.

4.3. *States and events*

The previous two sections in essence set out a mapping first from spatiotemporal data about movement events and environmental states/events to the input of frequent-pattern mining techniques; and then from the output of those frequent-pattern mining techniques to inferences about candidate causal and causal-like relationships. An important unanswered question, then, is: How can one derive the required input information about movement events and associated environmental events and states?

In general we can identify four broad cases:

(1) Categorical data is supplied in the form of timestamped states (e.g., habitat classifications, such as "high habitat quality");
(2) Measured data is available that must then be categorized into timestamped states, for example using thresholding (e.g., < 10°C is classified as "low temperature" state $wt1_s$, ≥ 10 and < 15°C is classified as state $wt2_s$, and so forth...)[1];
(3) Data is supplied in the form of timestamped events (e.g., fish movement events m_e or moon phases, such as "full moon event" or event fm_e);
(4) Categorical data about states (whether supplied directly, see 1. above, or categorized from measured data, as in 2. above) must be further categorized into timestamped events, based on transitions between states (e.g., for water temperature a transition between state $wt2_s$ and $wt1_s$ may be classified as the event $wt1_e$ "start of a low temperature state"). Formally, an event v_e is a relation on the set of states S, $v_e \subseteq S \times S$.

[1]Note that the usual range of options exist for thresholding continuous data into qualitative categories, including equal interval, quantiles, k-means, and so forth.

In practice, all four cases are to be expected, and examples of each case were encountered in the course of our specific study, discussed in more detail below.

The import of these cases is that in addition to data about object movement and its environmental context, our analysis requires as input human commonsense or domain expert definitions about the events and states of interest. In some cases (specifically, cases 2 and 4), these commonsense or expert definitions will effectively encode "initiation" and "termination" relationships between states and events. Hence, while our technique outputs candidate "causes" and "allows" relationships, it typically requires "initiation" and "termination" relationships to be provided as input. Further, note that there exists a duality in initiation and termination of states. Any event that initiates a state will, necessarily, terminate another (the state that previously existed before the initiation). As a consequence, in the sequel we only discuss the initiation of states, and ignore their dual, termination.

Importantly, our analysis is entirely agnostic about whether the chosen definitions are sensible or "correct." Arbitrary or nonsense definitions are unlikely to yield meaningful candidate causes. However, state and event definitions may easily be changed (e.g., varying the thresholds used for state category boundaries) and the analysis re-run with new (and hopefully more salient) definitions.

4.3.1. Movement data set

We applied our technique to the raw fish movement data, which contains information about the location (river zone) of each fish on each day. Despite its limitations, (including the relatively coarse spatial granularity and the small sample of fish, when viewed relative to the total fish population in the river), the data is a remarkably rich source of fish movement patterns, such as up- and downstream movements in varying cycles and over different distances.

Figure 3 illustrates a small part of the raw movement data pictorially. The figure shows different river zones (differentiated using different colors of dots) in which 32 fish (out of the total of 1050 in our data set) were located over a period of 108 days (out of the total of 6 years).

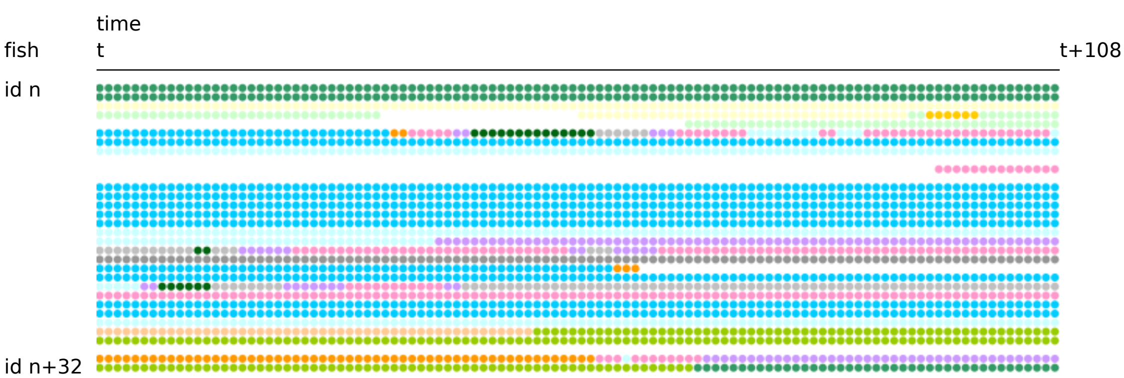

Figure 3. Pictorial illustration of a subset of the raw fish data. Each horizontal line represents a single fish. Each dot along the horizontal lines represent a day. The colors of the dots indicate different river zones in which the specific fish was located each day. Fish movement occurs when the horizontal line changes its color, i.e., in the 6th line the fish moves from river zone blue to river zone orange, stays there for two days and moves on to river zone pink.

In reference to Tables 1 and 2, the set of moving object identifiers I are the fish ids; the set of timestamps T is the set of days over which the monitoring occurred.

The set of states (locations) for our movement data are therefore already given (case

1. above), in terms of the set L of 24 river zones. Clearly, in some other data sets, such a categorization of states may not exist (e.g., in trajectory-based coordinate movement data). In those cases, other categorizations into location states (e.g., in cordon-structured data more generally, cf. Both *et al.* 2012) or movement states (e.g., Laube *et al.* 2005) are required.

The set of movement *events* must then be selected based on transitions between these states, as in case 4. above. In the example of our fish movement data set, a set of atomic movement events follow relatively naturally from those states. The key movement events chosen were upstream movement (u_e, i.e., movement of a fish from one river zone to another zone upstream); downstream movement (d_e, i.e., movement of a fish from one river zone to another zone downstream); as well as the coarser-grained event of movement (m_e, i.e., movement of a fish from one river zone to another zone).

It is worth noting that there is no requirement in our analysis that the events of interest chosen are pairwise disjoint or jointly exhaustive. As we shall see, our data mining technique operates whatever states and events are chosen. Indeed, the approach is "data hungry": we also experimented with a range of other states and events, such as *from* and *to zone* movement events (e.g., movement to or from river zone g, tg_e and fg_e, respectively).

In addition to the actual fish movement data set we generated a data set of randomized fish movements between the different river zones against which to validate our technique. The simulated data set is of the same size as the actual fish data set, i.e., comprising the same number of fish and movements over the six year time span.

4.3.2. Contextual environmental data sets

Contextual environmental data will most frequently be supplied as measured environmental parameters. In the specific example of environmental data relevant to fish movement, five separate contextual data sets were available: water temperature, maximum daily air temperature, moon phases, river flow, and water level.

Categorization of the environmental data into states therefore most often follows case 2. above. For example, observations of water temperature were initially classified into five categories based on equal intervals: $< 10°$C ($wt1_s$), $\geq 10°$C and $< 15°$C ($wt2_s$), $\geq 15°$C and $< 20°$C ($wt3_s$), $\geq 20°$C and $< 25°$C ($wt4_s$), and $\geq 25°$C ($wt5_s$). In the more complex case of river flow, we used categories based on quartiles, to enable comparison of high river flow in, say, a small tributary with high river flow in the main river. We return to a discussion of the effects of the choice of categories in Section 5.1.2. For now we assume that the domain expertise or general knowledge required to formulate categories of interest is available as an input to the analysis process.

Categorization of environmental events typically proceeds as in case 4. above. For example, the start of a state of high river flow ($rf4_e$) was defined as a change from one of the other three river flow states (recall, events are defined as relations on states, i.e., $rf4_e = \{(rf1_s, rf4_s), (rf2_s, rf4_s)\}, (rf3_s, rf4_s)\})$.

One exception is that moon phase data comprises events as first-class observations, case 3. above. We predict, observe, and record the occurrence of a full moon event directly, rather than inferring full moon event occurrence from observations of moon-fullness states (e.g., a transition between a state of 99% moon visibility to the state of 100% moon visibility).

In most cases these environmental data sets exhibit some spatial variation (e.g., at some given time, water temperature in zone f may be different to that in zone g). However, it is also allowable that data may on occasion not vary spatially over the study area, either because of lack of available detail, or simply that no spatial variation is found over the study area (such as in the case of moon phases, which vary temporally but not spatially over our study area).

4.4. *Implementation*

The analysis procedure outlined above was implemented using TraMineR sequence mining package for the R statistical language (Gabadinho *et al.* 2011). Some customized Python script was also generated to automate the data preprocessing. In short:

(1) The raw input data can be transformed into tables of atomic movement events using the standard R commands and the TraMineR `seqdef` command. As an intermediate stage in this step, TraMineR allows the definition of a transition matrix, which specifies movement event tokens for each zone transition. For example, transition between zones f and g may be classified in the matrix with a range of different event tokens: "movement" (m_e), "from f" (ff_e); and/or "downstream" movement (d_e).

(2) Custom-written Python scripts were developed to annotate fish movements with contextual, environmental events. The script requires as input the definitions of contextual events of interest (e.g., low temperatures, high flows, etc.) The script then looks up for each fish what, if any, contextual events that fish "experienced."
   a) In the case of mining candidate causal-like "allows" relationships, the script associates with each fish identifier the time and type of any movement events that fish participated in, and the environmental states that were spatiotemporally co-located with that fish at the beginning of a movement (see Table 1).
   b) In the case of mining true candidate causal relationships, the script associates with each fish identifier the time and type of any movement events that fish participated in and/or any environmental events that were experienced by that fish (i.e., spatiotemporally co-located when the event began).

(3) The command `seqefsub` provided by TraMineR was used to mine the state/event tables for sequences. In addition to the tabulated state/event data created in the previous step, the `seqefsub` command accepts as additional input a maximum time-lag between events. A time lag of zero ensures only strictly contemporaneous items are mined (i.e., conventional association rule mining) and so in combination with the state table outputs candidate "allows" relationships. A non-zero time-lag, in combination with the event table, results in mining of candidate causal relationships, allowing for the granularity effects discussed in Section 3.1.

5. Results

This section evaluates our candidate causal mining technique through application to our specific example data set of fish movement events in the Murray River. The results of three distinct analyses are presented. First, an analysis of the output of mining candidate causal relationships between atomic environmental and movement events is presented in Section 5.1. Second, an analysis of the candidate causal-like "allows" relationship between environmental state and movement events is presented in Section 5.2. Third, Section 5.3 examines candidate causal relationships for aggregate movement event, involving complex sequences of events.

For validation purposes, these analyses were also repeated on a simulated data set of randomized fish movement events. This approach did not in any cases identify candidate causal relationships between the environmental data and the randomized fish movements. Hence, we do not report in detail the results of those analyses here. However, where sufficient numbers of observations exist (Sections 5.1 and 5.2) we do compare the patterns identified by our analysis with patterns of random movement, using statistical hypothesis tests to check the likelihood that patterns identified by our technique could have occurred by chance.

5.1. *Results #1: Atomic Events*

Our first analysis aims to validate those atomic event pairs that relate an environmental event to a subsequent fish movement event. Our sequence mining technique was applied to the entire fish data set, with a time lag of two days, to allow for granularity effects (see Section 3.1). The lag time of two days was chosen in discussion with domain experts, who indicated this was a reasonable lag from the perspective of fish biology. The outputs of our frequent sequence mining procedure were filtered to include only those binary sequences that began with an environmental event and were followed (within two days) by fish movement.

Figure 4 summarizes the results of this analysis. The rows in Figure 4 show three environmental variables: water temperature (wt), river flow (rf), and moon phase (mp). The columns show the different types of movement events either upstream (u_e) or downstream (d_e) movement, or up- or downstream movement (m_e). The histogram in each cell compares the observed and expected frequency of the corresponding binary sequences of the environmental event followed by the specified movement event (e.g., the top right histogram in Figure 4 shows the observed and expected frequency of the specified water temperature events being followed by downstream fish movement).

As introduced above, the expected frequency was computed by assuming that fish movements were random, and so causally unrelated to environmental events and equally likely to occur within two days of any environmental event. Thus, the expected frequency reflects the underlying frequency of different environmental events.

The figure also summarizes the results of a chi-square statistic, to test the null hypothesis that there is no significant difference between the observed and expected frequencies of event sequences.

For example, looking at the histogram in row rf (river flow) and column d_e (downstream movement), we can see that a high river flow event was followed (within two days) by downstream movement more often than would be expected. Conversely, low river flow events were followed less frequently than expected by downstream movement. The difference between the observed and expected frequencies is significant at 95% confidence level in this case.

The environmental variables maximum daily air temperature and water level are omitted from Figure 4 because it was found that those results followed very closely water temperature and river flow respectively.

5.1.1. Discussion

We interpret these results as an indication that changes in environmental water temperature are associated with changes to subsequent fish movement. Extreme water temperatures (high or low) tend to be followed by decreased movement compared with that expected; moderate water temperatures tend to be followed by increased movement.

Similar effects are observable in river flow events, albeit to a lesser extent. Higher flows are associated with greater than expected movement; lower flows with less fish movement.

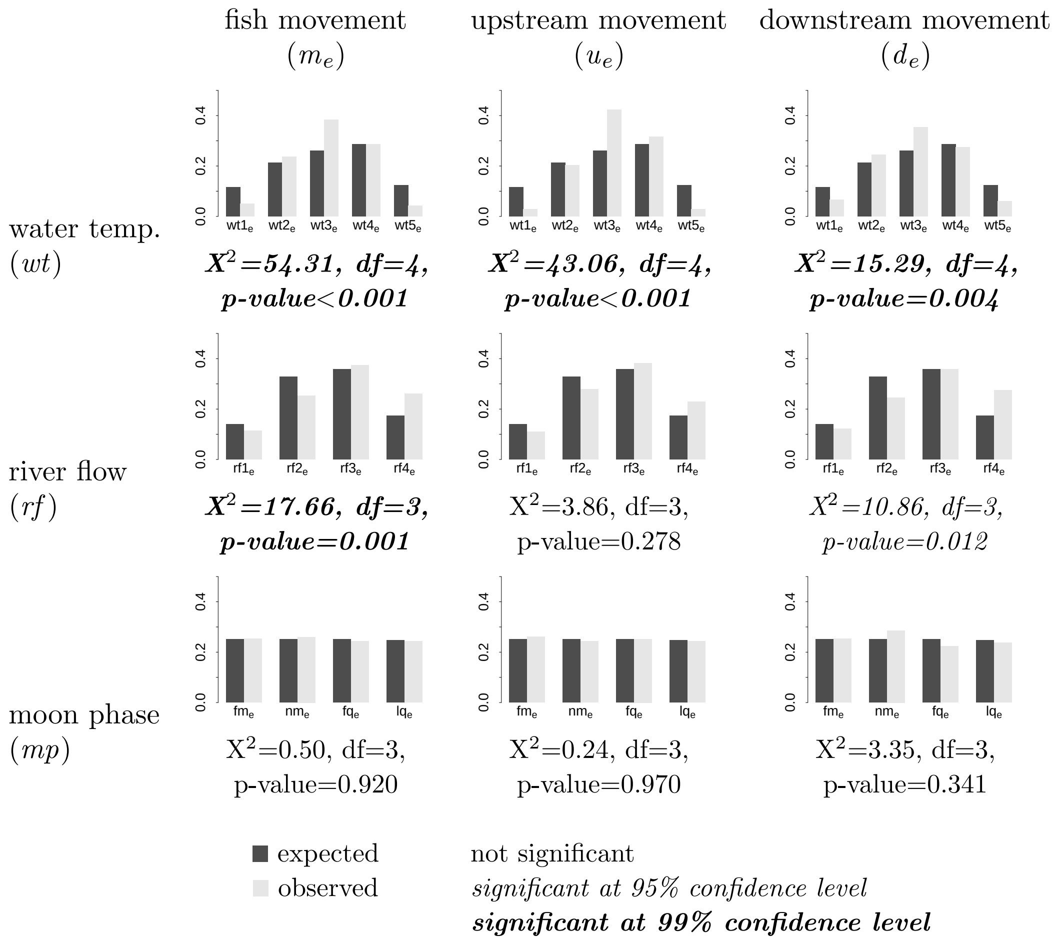

Figure 4. Summary of expected and observed frequency of general movement (m_e), or specific upstream (u_e) or downstream (d_e) movement following water temperature (wt, equal intervals from $wt1_e$ ($<10°$) to $wt5_e$ ($>25°$)), river flow (rf, first quartile $rf1_e$ to fourth quartile $rf4_e$), or moon phase (mp, full moon fm_e, new moon nm_e, first quarter fq_e, last quarter lq_e) events within two days.

However, while this effect is significant for movement in general, it is not significant at the 99% level for upstream movement or for downstream movement specifically (although it is significant at the 95% level for downstream movement). This is on one hand due to the smaller sample sizes for up- and downstream movement. However, it can also be attributed to high river flow events being followed by increased *downstream* movement, but not similarly influencing *upstream* movement. The discussion in Section 5.2.1 shows that upstream movement is more strongly influenced by the "allowing" state high river flow.

Surprisingly, from a domain expert's point of view, moon phase was not associated with significant changes in movement patterns. None of the differences between observed and expected movement following moon phase changes were significant. However, there is a slight apparent trend towards an increase of downstream movement following a new moon. Even though not statistically significant, such observations may warrant further investigation.

From the domain expert's perspective, we take these patterns as indicative of potential causal relationships. We may wish to infer from the results that water temperature and river flow events may *cause* the associated movement events, while moon phase events may not.

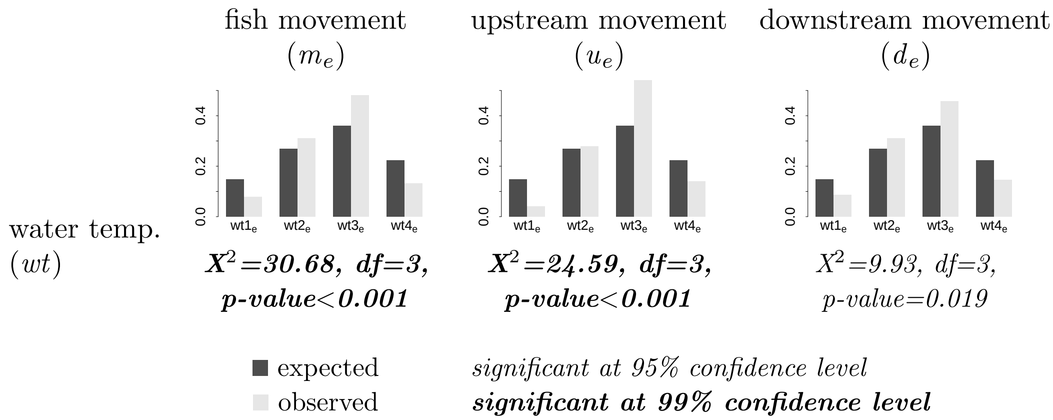

Figure 5. Summary of expected and observed frequency of movement with revised classification of water temperature (wt), highlighting the biologically significant temperature range 16–23° ($wt3_e$).

5.1.2. *Choice of categories*

As already noted, our analysis is dependent on the definition of the categories of interest. The category boundaries define the different environmental states, and transitions between environmental states provides our environmental event definitions. Poor or unfavorable choices of category boundaries may mask certain effects; conversely, stronger effects may be less sensitive to changes in category boundaries.

For example, imagine that a only narrow temperature range of, say, 16–18° in actuality causes upstream movement events; other temperature ranges have no impact on movement. In this case, choosing category boundaries that do not correspond well to this actual causal effect (e.g., below 10°, 11–30°, and above 30°) may prevent our analysis identifying this relationship.

In the context of our application, a discussion with domain experts about the results in Figure 4 revealed that the temperature range 16–22° is especially significant for fish biology. In this way, the results of our analysis using one choice of category boundaries can provide feedback that may prompt experts to link to other relevant knowledge. Figure 5 shows the results of repeating our analysis with revised water temperature categories, taking into account this additional expert knowledge. Compared to row wt in Figure 4, Figure 5 shows a similar size of effect in the category 16–22° ($wt3_e$), as in the original categories 15–19° ($wt3_e$) and 20–24° ($wt4_e$). The revision in this case actually leads to a slightly lower statistical significance: downstream movement is only significant at the 95% level in Figure 5 (as opposed to the the 99% level in Figure 4). However, this difference is, at least in part, likely to be due to the reduction in the number of categories from five to four: the chi squared test is sensitive to changes in the degrees of freedom.

Thus, in this case it seems the effect of water temperature upon fish movement is relatively robust to changes in the category boundaries. Be that as it may, the example illustrates that the objective of our analysis is not to *find* good categorizations, only to support domain expertise in exploring evidence for or against *chosen* categorizations.

5.2. *Result #2: State associations*

Turning now to the role of *states*, rather than events, Figure 6 summarizes impact of environmental states upon fish movement (the output of our association rule mining upon the state-table for the causal-like "allows" relationship). Like Figures 4 and 5, Figure 6 compares histograms showing the expected and observed frequency of fish movement events. Unlike Figures 4 and 5, Figure 6 classifies this movement according to environ-

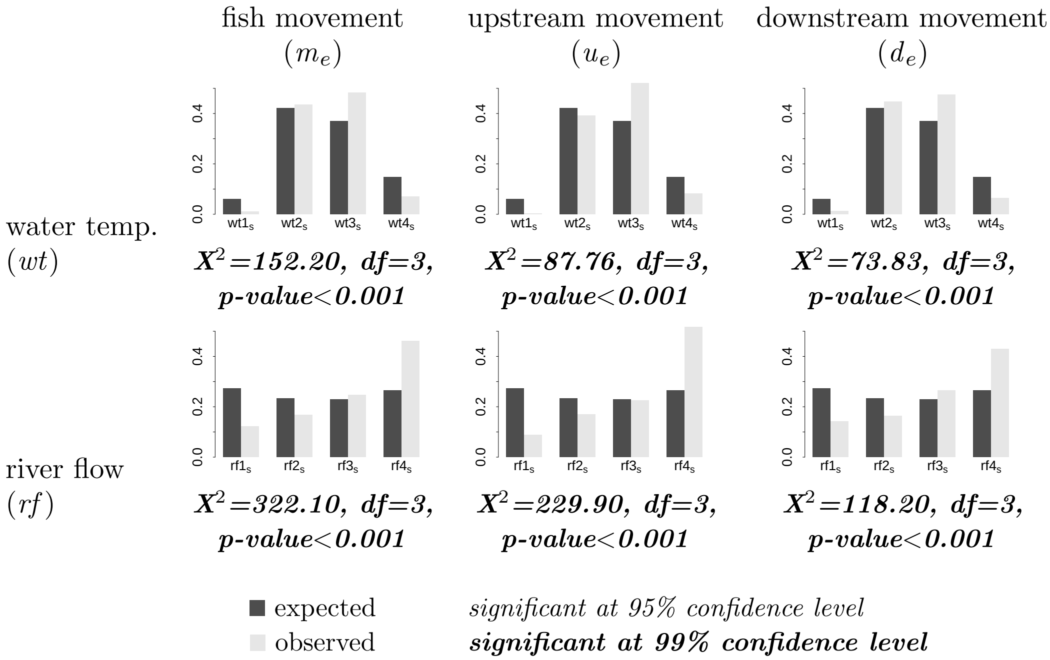

Figure 6. Summary of expected and observed frequency of general movement (m_e), or specific upstream (u_e) or downstream (d_e) movement co-occurring with specified water temperature (wt, equal intervals from $wt1_e$ (<9°) to $wt4_e$ (>23°)), river flow (rf, first quartile $rf1_e$ to fourth quartile $rf4_e$) states, persisting for more than two days.

mental *states* that had persisted for more than two days. States that had not persisted for more than two days would imply an event (change in state) had occurred within that period. Such events were excluded from analysis of state associations—instead they were captured within the analysis of events (in Sections 5.1 and 5.3). Thus, the labels on the x-axes of Figure 6 refer to the state itself (e.g., the persistent state of "low river flow") as opposed to the associated event used in Figures 4 and 5 (e.g., "the *start* of a state of low river flow").

The results indicate that both water temperature and river flow states have a statistically significant (at the 99% level) association with all fish movement, both upstream and downstream. The histograms show that a persistent state of 16–23°C water temperature is most strongly associated with greater than expected fish movement; below 9° with less than expected movement. Similarly, a persistent state of highest river flow is most strongly associated with greater than expected fish movement; lowest river flow is most strongly associated with less than expected movement.

Figure 6 only presents histograms related to water temperature and river flow states, but not moon phase. Moon phase cannot be analyzed in this way, because as already discussed in Section 4.3.2 moon phase data comprises events as first-class observations, not derived from associated states.

5.2.1. Discussion

We interpret the results as indicative that persistent moderately high water temperatures and high river flows are enabling states that *allow* fish movement events to occur. These enabling states show slightly different causal patterns to those of their associated events. For example, while the occurrence of a high river flow event is not significantly associated with upstream fish movements at the 99% level (see Figure 4), the persistence of a state of high river flow is associated with upstream fish movements at this significance level. Thus, as already discussed in Section 5.1.1, we may infer that high river flow is an enabling state for upstream fish movement but a potentially causing event for downstream fish movement. Of course, care is needed to acknowledge the inextricable relationship between the persistence of a state and the event that originally initiated that state. An alternative interpretation might posit that if the start of a high river flow event is a cause of downstream movement, a fish's subsequent and unobserved need to return (upstream), even in the face of persisting high river flow, may be the true cause of the upstream movement. In short, it is possible that the coincidence of upstream movement and a persistent state of high river flow might be instead the result of a chain of unobserved causes that began with the start of high flow event.

5.3. *Result #3: Aggregate events*

After looking at enabling states and atomic events we examine more complex, aggregate sequences of events and states. Aggregate events can consist purely of fish movement events, such as a sequence of several upstream movements of a fish (each within the defined time gap from its predecessor). Further, it is also possible to look for the environmental events that are candidate causes of a longer sequence of fish movement events. Naturally, the more complex a sequence of events, the less frequently it will occur. Thus, in our data set we can no longer rely on statistical hypothesis testing, since the number of samples of longer movement events is too small. However, we can still use other measures of importance, in particular *support* and *confidence*.

The two of the most common measures of the strength of association rules are *support* and *confidence*. Support is generally defined as the frequency of an association rule in a data set; while confidence expresses the prediction strength of the rule (Mohammad and Nishida 2010). Following previous work, the support and confidence for an event sequence can be similarly defined:

$$\text{Support (object)} : s_o(A) = \frac{\sigma(A)}{n} \tag{1}$$

$$\text{Confidence (object)} : c_o(A \rightarrow B) = \frac{\sigma_o(A \rightarrow B)}{\sigma_o(A)} \tag{2}$$

where A and B are atomic events or aggregate event sequences; $\sigma_o(A)$ is the number of objects (i.e., fish) that exhibit the pattern A at least once; and n is the total number of objects in the data set.

However, these standard definitions are primarily designed for association rule mining, and consequently capture information on re-occurring event sequences. Thus, for sequence mining it is important additionally to know the *event* frequency as well as the *event* confidence, defined as follows (Das *et al.* 1998):

$$\text{Confidence (event)} : c_e(A \rightarrow B) = \frac{\sigma_e(A \rightarrow B)}{\sigma_e(A)} \tag{3}$$

where $\sigma_e(A)$ is the absolute frequency of event or event sequence A in the data set (possibly occurring multiple times for the same object).

To illustrate, Table 3 shows some examples of aggregate events (E_i) and their respective support and confidence on fish and event levels. In Table 3, the sequence of two consecutive atomic upstream movement events $A = (\{u_e\}, \{u_e\})$ (within two days) is exhibited by 45 out of 1050 fish[1]. Thus the per-object support for this aggregate event is $s_o(E_4) = 45/1050 = 0.042$. Further, five of those 45 fish engage in the rapid upstream movement within 3 days of a moderate water-temperature event, i.e., $c_o(E_4 \rightarrow E_5) = 5/45 = 0.111$.

	Aggregate events	$s_o(E_i)$	$c_o(E_{i-1} \rightarrow E_i)$	$\sigma_e(E_i)$	$c_e(E_{i-1} \rightarrow E_i)$
E_1	$(\{u_e\}, \{d_e\}, \{u_e\}, \{d_e\})$	0.014		30	
E_2	$(\{d_e\}, \{u_e\}, \{d_e\}, \{u_e\}, \{d_e\})$	0.005	0.333	14	0.467
E_3	$(\{u_e\}, \{d_e\}, \{u_e\}, \{d_e\}, \{u_e\}, \{d_e\})$	0.003	0.600	5	0.357
E_4	$(\{u_e\}, \{u_e\})$	0.042		84	
E_5	$(\{wt3_e\}, \{u_e\}, \{u_e\})$	0.005	0.111	6	0.071
E_6	$(\{u_e\}, \{u_e\})$	0.042		84	
E_7	$(\{rf4_s, u_e\}, \{u_e\})$	0.001	0.022	1	0.012

Table 3. Example aggregate event table, showing six aggregate event sequences (E_1, E_2, E_3, $E_4 = E_6$, E_5, E_7) composed of up- and downstream movement events, environmental water temperature event, and a environmental river flow state (see Section 4.3 for an explanation of the event and state definitions). The support (s_o, per-object) indicates the proportion of fish for which the specified aggregate event occurs at least once. The confidence (c_o, per-object) indicates the number of fish for which the specified event happens as a proportion of those fish that also exhibit the simpler aggregate event (shown on the preceding row). The event frequency (σ_e reports how often a specified aggregate event occurs (including possibly multiple times for the same fish). The confidence (c_e, per-event) indicates the number of specified events as a proportion of the number of simpler aggregate events (preceding row).

Turning to event confidence, Table 3 shows that of the 84 occurrences of rapid upstream movement (spread amongst 45 fish, see above), 6 occurrences are immediately preceded (within two days) by a moderate water-temperature event, i.e., $c_e(E_4 \rightarrow E_5) = 6/84 = 0.071$.

In this way, the per-object and per-event measures provide different and complementary information about the relative strengths of the rules. Higher per-object support and confidence ($s_o(E_i)$ and $c_o(E_{i-1} \rightarrow E_i)$) indicate a rule applies to most fish; but ignores the frequency with which that rule occurs over time (e.g., it may only occur once in 6 years for each fish). Higher event frequency and per-event confidence ($\sigma_e(E_i)$ and $c_e(E_{i-1} \rightarrow E_i)$) indicate that rule holds over time; but ignores the number of fish for which that rule holds (e.g., the rule may occur repeatedly, but only for a small number of "unusual" fish).

5.3.1. *Discussion*

While complex aggregate events occur only rarely in our data set, reporting support, confidence, and frequency does allow some assessment of the strength of different rules. For example, looking at oscillating up- and downstream movement patterns of fish (see aggregate events E_1, E_2, and E_3 in Table 3) we find that 33.3% of fish that exhibit the first movement pattern (E_1) also exhibit the second pattern (E_2). Of those, 60% also exhibit the third longer pattern (E_3). Thus, it appears previous movement oscillation is a relatively good predictor of future oscillation for individual fish. However, the per-event confidence does not show such strength of pattern, reducing from 46% for E_2 to 35% for E_3. This can be interpreted as an indication that while at *some* point fish that perform

[1]In fact, a surprisingly small number of fish in the study *ever* move far. Of the 1050 tagged fish, only around 260 are ever recorded moving between river zones. Consequently, the support and confidence reported arguably underestimates the strength of causal relationships: although only 4% of fish engage in rapid upstream movement (E_4/E_6), this translates to approximately 17% of the fish that *ever* move.

oscillation E_2 are likely to also perform oscillation E_3, this pattern is less likely to hold for *every* occurrence of pattern E_2 for a fish.

A further example in Table 3 concerns the aggregate two-zone upstream movements of fish (E_4/E_6). The aggregate event E_5 is a two-zone upstream movement preceded (candidate cause) by a moderate water-temperature event ($wt3_e$). By contrast, the aggregate event E_7 is a two-zone upstream movement accompanied (candidate causal-like allows relation) by a high river-flow state ($rf4_s$). Based on the available data, the per-object and per-event confidence provide stronger evidence that the water temperature event may cause this movement, rather than the high river flow may allow this movement.

6. Summary and conclusions

Based on a foundational ontology of causation we successfully used association-rule and sequence mining to identify candidate causal relationships between fish movement patterns and their environmental drivers. Our focus has been on environmental states ("allows" relationship) and events ("causes" relationship) that influence fish movement events.

In this paper, we focus primarily on the technique itself, and use the specific application to fish movement in validating our approach. Discussions with fish ecologists have given tentative indications of the potential usefulness of the approach. For example, the result that moderate water temperature is a candidate cause for increased movement elicited the further information from the domain experts that the temperature range 16–23°C is important for fish physiology. Conversely, the absence of moon phase as a candidate cause of fish movement was counter to the ecologists expectations, potentially warranting further investigations. However, detailed interpretation of the results in the application domain is left as a topic for further study in collaboration with fish ecologists.

Our adoption of a solid ontological foundation gives us high confidence that the approach should be transferable to many different application domains. The increasing commonality of movement data guarantees a very wide range of potential further applications, including studying human movement via mobile phone logs; traffic movements via GPS tracking or electronic tolling; as well as other studies of animal movements, including other studies of fish ecology (e.g., Johnston and Bergeron 2010). For example, ongoing extensions to this work are currently investigating the application of this new technique to human activity and travel logs in the domain of health and epidemiology monitoring. Further, this paper has focused specifically on the important relationship between movement and its environmental context. However, exactly the same approach can also be applied to inferring candidate causal relationships between any type of context for movement, such as time or day of the week (e.g., in traffic monitoring), other known events (such as football matches, roadworks), or even other movements (such as bus or train schedules).

Additionally, there is potential to extend this approach in at least two directions. First, the approach could benefit from integration within a broader visual analytics process. Remapping the patterns found in space and time, using geovisualization methods, could allow for visual exploration of the now enriched and condensed knowledge base. In collaboration with domain experts, such techniques could be used to help create knowledge about, for example, the effects of different state and event definitions, which in turn could be fed back in the data mining process. Second, our technique might usefully be extended to search automatically for different parameterizations, including varying time-lags and thresholds used for state/event definitions, that lead to the strongest results in regard to candidate causal relationships. Such an approach might reveal unexpected contextual drivers of movement, that in turn call for further exploration in collaboration with domain experts.

References

Agrawal, R., Imieliski, T., and Swami, A., 1993. Mining association rules between sets of items in large databases. *In*: *Proc. ACM SIGMOD International Conference on Management of Data* Washington, D.C., United States: ACM.

Allen, E., Edwards, G., and Bédard, Y., 1995. Qualitative causal modeling in temporal GIS. *In*: A. Frank and W. Kuhn, eds. *Spatial Information Theory (COSIT)*, Vol. 988 of *Lecture Notes in Computer Science* Berlin: Springer, 397–412.

Andersson, M., *et al.*, 2008. Reporting Leaders and Followers among Trajectories of Moving Point Objects. *GeoInformatica*, 12 (4), 497–528.

Andrienko, G., Andrienko, N., and Heurich, M., 2011. An event-based conceptual model for context-aware movement analysis. *International Journal of Geographical Information Science*, 25 (9), 1347–1370.

Beebee, H., Hitchcock, C., and Menzies, P., eds. , 2009. *The Oxford Handbook of Causation*. Oxford University Press.

Both, A., *et al.*, 2012. Decentralized monitoring of moving objects in a transportation network augmented with checkpoints. *The Computer Journal*.

Buchin, K., *et al.*, 2011a. Finding long and similar parts of trajectories. *Computational Geometry*, 44 (9), 465–476.

Buchin, M., *et al.*, 2011b. Segmenting Trajectories: A framework and algorithms using spatiotemporal criteria. *JOSIS*, 3, 33–63.

Buchin, M., Kruckenberg, H., and Klzsch, A., 2012. Segmenting Trajectories by Movement States. Advances in Geographic Information Science, *In*: S. Timpf and P. Laube, eds. *Advances in Spatial Data Handling*. Springer Berlin Heidelberg, 15–25.

Bunge, M., 1966. *Causality*. New York: Dover.

Claramunt, C. and Thériault, M., 1996. Toward semantics for modelling spatio-temporal processes within GIS. *In*: M.J. Kraak and M. Molenaar, eds. *Advances in GIS II*. Deflt, The Netherlands: Taylor and Francis, 47–64.

Claramunt, C. and Thériault, M., 1995. Managing Time in GIS: An event-oriented approach. *In*: J. Clifford and A. Tuzhilin, eds. *Recent Advances on Temporal Databases*. Zürich, Switzerland: Springer-Verlag, 23–42.

Das, G., *et al.*, 1998. Rule discovery from time series. *In*: R. Agrawal and P. Stolorz, eds. *Proc. Knowledge Discovery and Data Mining (KDD)* AAAI Press, 16–22.

Duckham, M., 2012. *Decentralized Spatial Computing, Foundations of Geosensor Networks*. Heidelberg: Springer.

El-Geresy, B.A., Abdelmoty, A.I., and Jones, C.B., 2002. Spatio-Temporal Geographic Information Systems: A Causal Perspective. *In*: *Proc. 6th East European Conference on Advances in Databases and Information Systems (ADBIS)* Berlin: Springer, 191–203.

Gabadinho, A., *et al.*, 2011. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40 (4), 1–37.

Galton, A., 2012. States, Processes and Events, and the Ontology of Causal Relations. *In*: M. Donnelly and G. Guizzardi, eds. *Proc. 7th International Conference on Formal Ontology in Information Systems (FOIS)*, Vol. 239 of *Frontiers in Artificial Intelligence and Applications* Amsterdam, NL: IOS Press, 279–292.

Galton, A. and Worboys, M., 2005. Processes and Events in Dynamic Geo-networks.

In: M.A. Rodriguez, I.F. Cruz, S. Levashkin and M.J. Egenhofer, eds. *Proc. GeoS: GeoSpatial Semantics* Springer, 45–59.

Gonzalez, M.C., Hidalgo, C.A., and Barabasi, A.L., 2008. Understanding individual human mobility patterns. *Nature*, 453 (7196), 779–782.

Gudmundsson, J., Laube, P., and Wolle, T., 2012. Computational Movement Analysis. *In*: W. Kresse and D.M. Danko, eds. *Springer Handbook of Geographic Information.* Berlin Heidelberg: Springer, 423–438.

Gudmundsson, J., van Kreveld, M., and Speckmann, B., 2007. Efficient Detection of Patterns in 2D Trajectories of Moving Points. *GeoInformatica*, 11 (2), 195–215.

Huang, Y., Chen, C., and Dong, P., 2008. Modeling Herds and Their Evolvements from Trajectory Data. *In*: T.J. Cova, H. Miller, K. Beard, A.U. Frank and M.F. Goodchild, eds. *Geographic Information Science*, GIScience 2008, 90–105.

Hume, D., 1739. *A Treatise of Human Nature.* 2003 Dover.

Jeung, H., *et al.*, 2008. Discovery of convoys in trajectory databases. *Proc. VLDB Endow.*, 1 (1), 1068–1080.

Johnston, P. and Bergeron, N., 2010. Variation of juvenile Atlantic salmon (Salmo salar) body composition along sedimentary links. *Ecology of Freshwater Fish*, 19 (2), 187–196.

Laube, P., *et al.*, 2011. Report on the first workshop on Movement Pattern Analysis MPA10. *JOSIS*, 1 (2), 127–133.

Laube, P., van Kreveld, M., and Imfeld, S., 2005. Finding REMO - Detecting Relative Motion Patterns in Geospatial Lifelines. Proc. 11th International Symposium on Spatial Data Handling, *In*: P.F. Fisher, ed. *Developments in Spatial Data Handling.* Berlin Heidelberg, DE: Springer, 201–214.

Lyon, J.P., 2012. Snags underpin Murray River restoration plan. *ECOS*, 2012 (177).

Mohammad, Y. and Nishida, T., 2010. Mining Causal Relationships in Multidimensional Time Series. *In*: E. Szczerbicki and N. Nguyen, eds. *Smart Information and Knowledge Management.*, Vol. 260 of *Studies in Computational Intelligence* Springer Berlin Heidelberg, 309–338.

Nagy, M., *et al.*, 2010. Hierarchical group dynamics in pigeon flocks. *Nature*, 464 (7290), 890–893.

Nathan, R., *et al.*, 2008. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences*, 105 (49), 19052–19059.

Pelekis, N., *et al.*, 2012. Visually exploring movement data via similarity-based analysis. *Journal of Intelligent Information Systems*, 38 (2), 343–391.

Schneider, C.M., *et al.*, 2013. Unravelling daily human mobility motifs. *Journal of the Royal Society: Interface*, 10 (84).

Wood, Z. and Galton, A., 2009a. Classifying Collective Motion. *In*: B. Gottfried and H. Aghajan, eds. *Behaviour Monitoring and Interpretation - BMI - Smart Environments.*, Vol. 3 of *Ambient Intelligence and Smart Environments* Amsterdam, NL: IOS Press, 129–155.

Wood, Z. and Galton, A., 2009b. A taxonomy of collective phenomena. *Applied Ontology*, 4 (3), 267–292.

Yuan, M., 2007. GIS Approaches for Geographic Dynamics Understanding and Event Prediction. *In*: R. Suresh, ed. *Proc. of SPIE, Vol. 6578, 65781A.*

Yuan, M. and Hornsby, K.S., 2008. *Computation and Visualisation for Understanding Dynamics in Geographic Domains: A Research Agenda.* Taylor and Francis (CRC Press).

Zaki, M.J., 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning*, 42, 31–60.