

Raum-zeitliches Data Mining – ein Werkzeug zur Analyse von Fischbewegungen im Murray River

Sensornetzwerke ermöglichen Umweltmonitoring in höchster räumlicher und zeitlicher Auflösung. Raum-zeitliches Data Mining bietet ein geeignetes Analysewerkzeug, um die in derartigen Messnetzen entstehenden grossen Datenmengen hinsichtlich interessanter Zusammenhänge und Muster zu untersuchen. Dieser Artikel berichtet über den Einsatz von Sequenzanalyse und Assoziationsregelsuche in einem Umweltmonitoring-Programm zum Gesundheitszustand von Fischhabitaten im Südosten Australiens. Erste Resultate liefern vielversprechende Hinweise auf kausale Zusammenhänge zwischen Umweltvariablen und beobachteten Fischbewegungen.

Geosensor networks allow for environmental monitoring at previously unseen spatial and temporal granularities. Spatio-temporal data mining offers a toolset for the analysis of the rich data sources produced by such systems. This article reports on a study using sequence and association rule mining on a data source emerging from a river health study in the Murray River, southeastern Australia. First results indicate that discovered sequence patterns and association rules might serve as proxies for causal relationship between environmental variables and monitored fish movements.

Susanne Bleisch, Patrick Laube,
Matt Duckham

Fischmonitoring im Murray River

In einem umfangreichen Umweltmonitoring Projekt im Murray River im Südosten Australiens wurde über die letzten zehn Jahre ein Netzwerk von Umweltsensoren aufgebaut. 18 Beobachtungsstationen (Abbildung 1, rechts) zeichnen die Bewegungen von über 1100 besenderten Fischen (Abbildung 1, links und Mitte) zwischen verschiedenen Flusszonen auf (Abbildung 2). Die gesammelten Daten dienen der Erfolgskontrolle verschiedener Management-Massnahmen zur Verbesserung der Habitate einheimischer Fischarten (Lyon et al., 2010). Die umfangreiche Datensammlung bietet aber auch eine umfassende Informationsquelle zur Ökologie dieses wichtigen Flusssystems. So haben die mit dem Projekt betrauten Fischbiologen des Arthur Rylah Institute

for Environmental Research in Melbourne eine Reihe von Hypothesen über potenzielle Ursachen von Fischbewegun-

gen. Dieser Artikel beschreibt, wie Data Mining-Methoden für die systematische Evaluierung solcher Hypothesen verwendet werden können. Das gewonnene Wissen fliesst in das Management zur nachhaltigen Nutzung des Murray Rivers mit ein.

Fischbewegungs- und Umweltdaten

Der am Murray River gesammelte und im Rahmen dieser Untersuchungen ausgewertete Datensatz besteht aus einem täglichen Aufenthaltsort (Flusszone) jedes beobachteten Fisches während der Lebensdauer der Senderbatterie. Der gesamte analysierte Zeitraum beträgt sechs Jahre von Anfang 2006 bis Ende 2011. Ausserdem werden zur Untersuchung bestehender Hypothesen über die möglichen Ursachen von Fischbewegungen weitere verfügbare Datensätze mit täglichen Messungen wie etwa Wassertemperatur, Wasserabflussmenge oder Mondphasen verwendet.

Als theoretische Grundlage dient der Studie eine Ontologie zu Kausalbeziehungen von Galton (2012). Darin wird zum Beispiel definiert, dass Ereignisse (engl.



Abb. 1: Murray cod (*Maccullochella peelii*) bereit zur Implantation des Radiosenders (links), verschiedene Radiosensoren (Mitte, Grösse der Münze entspricht einer Schweizer Fünffranken-Münze), Beobachtungsstation am Murray River (rechts). (Quelle: Arthur Rylah Institute, Melbourne.)

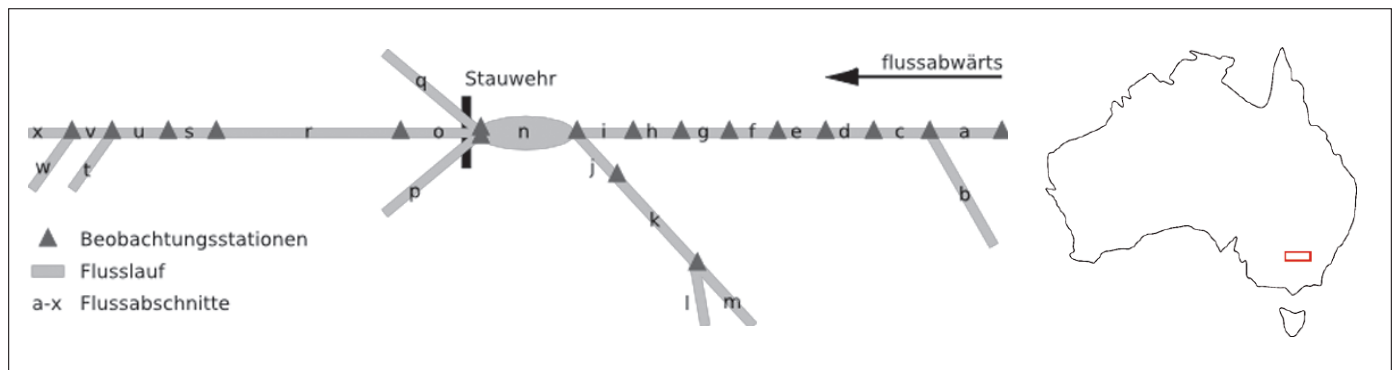


Abb. 2: Schematische Skizze des beobachteten Flussabschnitts des Murray Rivers (links) im Südosten Australiens (rechts).

«events») nur von anderen Ereignissen verursacht werden können. So könnte zum Beispiel das Ereignis «Vollmond» ein Ereignis «Start Fischbewegung» auslösen. Die Ontologie definiert ausserdem Zustände (engl. «states»), die es Ereignissen erst erlauben, andere Ereignisse zu verursachen. Galton (2012) veranschaulicht dieses Konzept mit dem Beispiel eines Autofahrers, der auf vereister Strasse versucht zu bremsen und dadurch einen Unfall verursacht. Das Ereignis «Bremsen» hätte eventuell nicht zum Ereignis «Unfall» geführt, wenn nicht zu diesem Zeitpunkt die Strasse vereist gewesen wäre (ermöglichender Zustand). Auf die hier betrachteten Daten übertragen hiesse dies, dass zum Beispiel das Ereignis «Start Fischbewegung» durch den Zustand «hohe Wassertemperatur» begünstigt werden könnte. Diese Definitionen erlauben Klarheit über mögliche Analysen und Schlüsse und vereinfachen die Übertragbarkeit der Methodik auf andere Anwendungen.

Auf der Basis dieser theoretischen Grundlage wurden Fischbewegungs-Ereignisse sowie Umgebungs-Ereignisse und -Zustände definiert. Wenn zum Beispiel ein Fisch an einem bestimmten Tag von einer Flusszone in eine andere Flusszone wechselt, definiert dies ein Ereignis. Es ist anzumerken, dass eine Fischbewegung grundsätzlich ein seltenes Ereignis ist. Die beobachteten Fische haben über den Beobachtungszeitraum total nur knapp 2000 Fischbewegungs-Aufzeichnungen generiert. Aus den zum Untersuchungszeitraum vorliegenden Umgebungsdatensätzen werden weiter Umgebungszu-

stände definiert. Diese Kategorisierung erfolgt durch von den Fischbiologen vorgegebene Schwellenwerte. Dabei definiert jeder Beginn eines Zustands ein Ereignis (z.B. «Start des Zustands hohe Wassertemperatur»). Des Weiteren wird definiert, dass ein Umgebungszustand von einem Fisch durch raum-zeitliche Nähe «erlebt» werden muss, damit dieser Zustand in den Data Mining-Prozess eingeführt wird. Das heisst, zum Startzeitpunkt des Zonenwechsels muss der Fisch sich in der Zone mit dem besagten Zustand befinden. Abbildung 3 zeigt schematisch, wie fiktive Fischereignisse vor dem Hintergrund verschiedener Umgebungszustände stattfinden und damit potenziell von diesen ausgelöst und/oder erlaubt werden.

Data-Mining Prozess

Für den Data-Mining Prozess werden etablierte Techniken wie Assoziationsregeln und Sequenzanalyse verwendet (z.B. Han, Cheng, Xin, & Yan, 2007). Assozia-

tionsregeln suchen häufig zusammen auftretende, «assoziierte» Elemente oder Ereignisse (z.B. «Start von Fischbewegungen» an «Vollmondtagen»). Die Sequenzanalyse sucht häufige Sequenzen von Ereignissen (z.B. «Start von Fischbewegungen» folgt ein bis zwei Tage nach «Leermondtagen»). Die Toolbox TraMineR für R (Gabadinho, Ritschard, Müller, & Studer, 2011) kombiniert die beiden Verfahren. Mit TraMineR können die Fisch- und Umgebungsereignisse und -zustände auf das Vorkommen bestimmter Ereignissequenzen hin untersucht und die Häufigkeit der Vorkommen gezählt sowie generell häufig vorkommende Ereignissequenzen gesucht werden. Die Suche nach bestimmten Ereignissequenzen wird durch die Hypothesen der Fischbiologen vorbestimmt und erlaubt die Evaluation der Hypothese mittels statistischer Methoden. Die generelle Suche nach häufig vorkommenden Eventsequenzen erlaubt es, den Datensatz ohne Hypothesen zu durchsuchen und potenziell interessante häufig vorkommende Eventsequenzen zu finden.

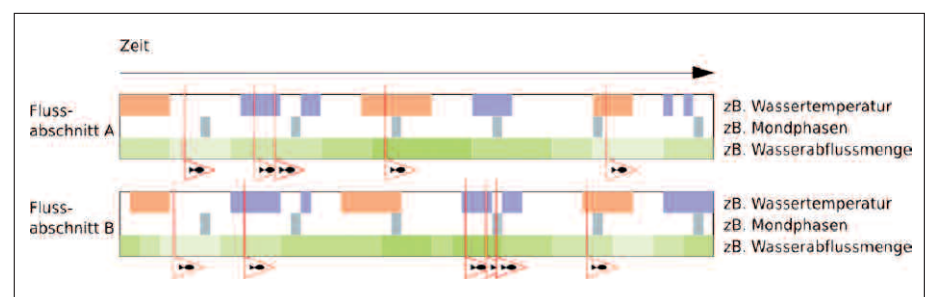


Abb. 3: Fiktive Zeitreihen von Fischbewegungen und Umgebungsvariablen (z.B. Wassertemperatur), rote Flaggen mit Fischsymbol markieren Startpunkte von Fischbewegungen.

Evaluation der Umgebungsbedingungen für Fischbewegungen im Murray River

Die Resultate der Data-Mining Prozesse sind häufige Eventsequenzen und deren relatives Vorkommen. Die Interpretation der Resultate bezüglich Kausalität wird den Fischbiologen und weiterführenden Studien überlassen. Es konnte jedoch festgestellt werden, dass beispielsweise in bestimmten Bereichen der Wassertemperatur mehr oder weniger Fischbewegungen stattfinden, als erwartet werden könnte aufgrund der Häufigkeit des Vorkommens dieser Temperaturbereiche (siehe Abbildung 4).

Aus den Resultaten des Data Minings lässt sich ableiten, welche Umgebungsvariablen mögliche Auslöser oder unterstützende Zustände für Fischbewegungen sind. Dies lässt sich insbesondere auch für ganz bestimmte Fischbewegungen feststellen, zum Beispiel Fischwanderungen zum Laichen über mehrere Flussabschnitte flussaufwärts (vgl. Tabelle 1). Im Falle der idealen Wassertemperatur für Laichwanderungen sind diese Werte den Fischbiologen schon aus anderen Studien bekannt, was eine gewisse Überprüfung der hier vorgestellten Ansätze erlaubt. Interessant ist vor allem der Detaillierungsgrad der Resultate sowie die Tatsache, dass der Einfluss der Wassertemperatur auf Fischbewegungen noch nie mittels eines so umfangreichen Datensatzes überprüft werden konnte. Aus methodischer

Sicht hat die Studie aufgezeigt, das raumzeitliches Data Mining zur Untersuchung von potentiellen Kausalitätsbeziehungen im Umweltmonitoring ein beträchtliches Potential hat. Der Prozess kann mit fast beliebig vielen Umgebungsdaten durchgeführt werden. So können nicht nur bestehende Hypothesen untersucht werden, sondern auch bisher unbekannte Zusammenhänge aufgedeckt werden. Während der Murray River ein weitgehend lineares System bildet und die Fische auf Bewegungen entlang des Flusslaufs beschränkt sind, liesse sich die vorgestellte Methodik aber mit einigen Anpassungen auch auf andere Tierarten und komplexere Lebensräume übertragen.

Dank

Diese Studie wurde durch das Australian Research Council Discovery Projekt DP120100072 «From environmental monitoring to management: Extracting knowledge about environmental events from sensor data» finanziert.

Literatur:

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4).

Galton, A. (2012). States, Processes and Events, and the Ontology of Causal Relations. *FOIS 2012: 7th International Conference on Formal Ontology in Information Systems*. 24–27 July, Graz, Austria.

Sequenzen	Häufigkeit
(U)-(D)-(U)-(D)	30
(U)-(D)-(U)-(D)-(U)-(D)	5
(GF,U)-(D)-(U)-(D)	9
(KF,U)-(D)-(U)-(D)	3
(GF,D)-(D)-(D)	17
(KF,D)-(D)-(D)	0
(VM)-(D)-(D)	5
(NM)-(D)-(D)	1

Tab. 1: Zusammenstellung einiger Fisch- und Umgebungs-Ereignisse-sequenzen; U – Fischbewegung flussaufwärts, D – Fischbewegung flussabwärts, GF – Wasserabflussmenge gross, KF – Wasserabflussmenge gering, VM – Vollmond, NM – Neumond, (,) – Ereignisse finden am gleichen Tag statt (erlaubender Zustand), () – max. zwei Tage Abstand zwischen den Ereignissen.

Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1), 55–86. doi: 10.1007/s10618-006-0059-1

Lyon, J., Nicol, S., Kearns, J., Bird, T., Stuart, I., Todd, C., O'Mahony, J., et al. (2010). *Monitoring of Resnagging between Lake Hume and Yarrawonga*. Milestone 4 September 2010. Confidential Client Report for the Murray-Darling Basin Authority (TLM Program), Arthur Rylah Institute for Environmental Research. DSE, Heidelberg, Victoria.

Dr. Susanne Bleisch
Department of Infrastructure
Engineering, University of Melbourne,
Australia
susanne.bleisch@unimelb.edu.au

Dr. Patrick Laube
Geographisches Institut,
Universität Zürich
patrick.laube@geo.uzh.ch

A/Prof Matt Duckham
Department of Infrastructure
Engineering, University of Melbourne,
Australia
matt@duckham.org

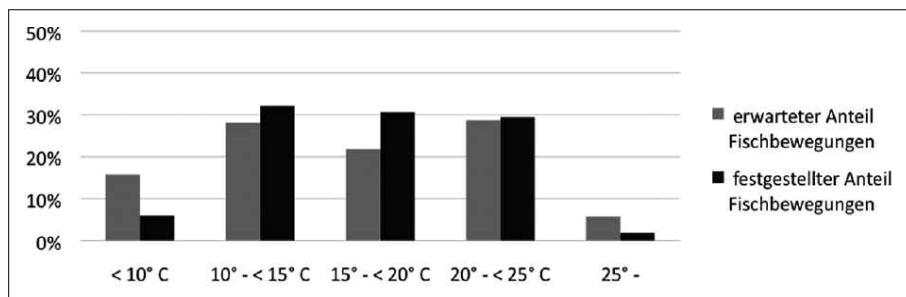


Abb. 4: Statistisch signifikante ($\chi^2 = 195.9$, $df = 4$, $p\text{-value} < 0.01$) Unterschiede zwischen den erwarteten und den festgestellten Fischbewegungen in den unterschiedlichen Wassertemperaturbereichen. Die beobachteten Fische bewegen sich deutlich weniger, wenn die Wassertemperaturen unter 10°C oder über 25°C betragen, jedoch deutlich mehr im Bereich 10–20°C.