R. Reussner, A. Koziolek, R. Heinrich (Hrsg.): INFORMATIK 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021 1035

Feature-basiertes Clustering von Umweltzeitreihen mit Self-Organizing-Map-Ensembles

Short Paper

Andreas Wunsch [1], Tanja Liesch [2], Stefan Broda [3]

Abstract: Die Zeitreihenanalyse ist für Umweltwissenschaften ein wichtiges Werkzeug, um Systeme zu charakterisieren, da sich in den Zeitreihen Signale, welche von unterschiedlichen Einflussgrößen herrühren, wiederfinden lassen. Ein Clustering kann helfen ähnliche Dynamiken zu gruppieren um so entsprechende Einflussgrößen zu erkennen und deren Einflussbereich zu charakterisieren. Wir stellen einen unüberwachten Ensemble-Modellierungsansatz für das Clustering von Umweltzeitreihen auf der Grundlage ihrer Dynamik vor. Der Feature-basierte Ansatz erlaubt es auch heterogene Datensätze zu nutzen, das Clustering der Features erfolgt schließlich auf der Basis von Self-Organizing-Maps. Der Ensemble-Ansatz reduziert die Willkür bei der Featureauswahl und erhöht die Robustheit des Endergebnisses. Die Ergebnisse einer beispielhaften Anwendung im Grundwasserbereich zeigen, dass die vorgestellte Methodik adaptiv in der Lage ist, homogene Gruppen von Zeitreihen-Dynamiken zu identifizieren.

Keywords: Time Series Clustering; Environmental Time Series; SOM; Ensemble Modelling; Feature Clustering; Groundwater

1 Einführung

Umweltzeitreihen sind meist das Ergebnis eines komplexen Zusammenspiels einer Vielzahl unterschiedlichster Prozesse, oft zusätzlich überlagert von einem gewissen Maß an Rauschen. Daher ist die Zeitreihenanalyse für Umweltwissenschaften ein wichtiges Instrument um natürliche Prozesse und ihre Dynamik zu beschreiben und zu verstehen. Zeitreihen-Clustering kann hilfreich sein, um gemeinsame räumliche und zeitliche Dynamikmuster zu identifizieren und zwischen Signalen, die aus externen beeinflussenden Faktoren resultieren sowie Rauschen zu unterscheiden. Dies gilt insbesondere für größere Datensätze, für die manuelle Methoden schnell an ihre Grenzen stoßen. Auf diese Art und Weise können Gruppen mit gemeinsamen Dynamik-Mustern, also gemeinsamen Systemeigenschaften, identifiziert und

[1] Karlsruher Institut für Technologie (KIT), Inst. f. Angewandte Geowissenschaften, Abt. Hydrogeologie, Kaiserstr. 12, 76131 Karlsruhe, Germany, andreas.wunsch@kit.edu, https://orcid.org/0000-0002-0585-9549
[2] Karlsruher Institut für Technologie (KIT), Inst. f. Angewandte Geowissenschaften, Abt. Hydrogeologie, Kaiserstr. 12, 76131 Karlsruhe, Germany, tanja.liesch@kit.edu, https://orcid.org/0000-0001-8648-5333
[3] Bundesanstalt für Geowissenschaften und Rohstoffe (BGR), Wilhelmstr. 25-30, 13593 Berlin, Germany, stefan.broda@bgr.de, https://orcid.org/0000-0001-6858-6368

doi:10.18420/inf2020_98

so besser verstanden werden. Clustering kann darüber hinaus auch als Grundlage für weitere Analysen dienen, wie zum Beispiel der Berechnung von Prognosen und dem Aufbau von Szenarientools. Typische Anwendungen von Zeitreihen-Clustering im Umweltbereich finden sich unter anderem in der Hydrometeorologie [GL16, Wa18], Hydrologie [Mi19, TS17], Fernerkundung [GWW16] oder bei Luftschadstoffen [DDM15].

Neben den klassischen Ansätzen wie der Cluster-Analyse (CA) und der Hauptkomponentenanalyse (PCA), die beide häufig zum Clustering von Umweltzeitreihen angewendet werden, bieten Künstliche Neuronale Netze (ANN) alternative Lösungen für den Umgang mit größeren mehrdimensionalen Datensätzen, z.B. durch die Verwendung von Self-Organizing-Maps (SOM) für unüberwachtes Clustering. Üblicherweise verwenden die meisten Methoden die Zeitreihen direkt für das Clustering und sind in großem Maße von qualitativ hochwertigen Daten abhängig (Zeitreihen mit gleicher Länge oder Periode, keine Datenlücken usw.). Umweltzeitreihen zeichnen sich jedoch häufig durch eine signifikante Anzahl an Datenlücken und Ausreißern aus. Größere Datensätze sind weiterhin oft sehr heterogen bez. Zeitreihenlänge und zeitlicher Auflösung der Einzelzeitreihen. Hierdurch reduziert sich der Anteil der verwendbaren Daten in der Regel erheblich. Feature-basierte Ansätze können dieses Problem potenziell lösen, da auch lückenhafte Eingabedaten unterschiedlicher Länge verwendet werden können und sich somit die Abhängigkeit von der Datenqualität reduziert. Bei der Anwendung eines Feature-basierten Ansatzes auf Umweltdaten ist es wichtig, dass die gewählten Features die Besonderheiten der jeweiligen Anwendung bzw. des betrachteten Parameters berücksichtigen. Während einige allgemeine statistische Maße und Merkmale wahrscheinlich für die meisten (Umwelt-) Zeitreihen als Features geeignet sind, sind für aussagekräftige Clusterergebnisse meist zusätzliche, an die Dynamik des jeweiligen Zeitreihenparameters angepasste Features notwendig.

Ein Feature-basiertes Zeitreihen-Clustering unter Verwendung von SOM wurde für Umweltzeitreihen beispielsweise von [NAV15] im Bereich der Hydrologie und von [Di13] im Kontext einer geophysikalischen Fragestellung angewendet. Erstere nutzen Wavelet-basierte Features, letztere Features auf Basis von PLR (Piecewise Linear Representation).

In der vorliegenden Arbeit wird ein robuster und halbautomatisierter Ansatz für das Feature-basierte Clustering von Umweltzeitreihen skizziert.

2 Methodik

Die entwickelte Methode wird beispielhaft anhand eines Datensatzes von ca. 1200 Grundwasserstands-Zeitreihen in wöchentlicher zeitlicher Auflösung demonstriert. Aufgrund von üblicherweise heterogener Datenqualität in Datensätzen von Umweltzeitreihen (Zeitreihenlänge, Sampling Intervall, Datenlücken) wurde ein Feature-basierter Ansatz gewählt, bei dem die unterschiedlichen Aspekte der Zeitreihendynamik auch für Zeitreihen variabler Datenqualität beschrieben werden können. Neben Standard-Maßen der deskriptiven Statistik wurden dazu Features entwickelt, die sich speziell für die Beschreibung von

Dynamik-Aspekten in Grundwasserstands-Zeitreihen eignen, aber durch entsprechende Modifikation auch für andere Arten von Umweltdaten angepasst werden können. Darüber hinaus wurden auch Features aus der Literatur, wie z.B. aus der Zusammenstellung von [He19] verwendet. Tab. 1 zeigt die Auswahl der Features, welche sich nach visueller Überprüfung für den vorliegenden Datensatz als aussagekräftig erwiesen haben und erläutert sowohl deren Hintergrund als auch die Herkunft.

Tab. 1: Liste der verwendeten Features für das Clustering von Grundwasserstands-Zeitreihen

Feature Name (Abkz.)	Zweck/Beschreibung	Ref*
Jährliche Periodizität (P52)	Stärke des Jahresgangs, berechnet durch Korrelation (Pearson) der mittleren jährlichen (52 Wochen) Periodizität mit der vollständigen Zeitreihe	sd
High Pulse Duration (HPD)	Durchschnittliche Dauer der Abschnitte über dem 80. Perzentil der Nichtüberschreitung, siehe [Ri96], entnommen aus [He19]	lit
Jumps	Inhomogenitäten/Brüche, teilweise auch Variabilität, berechnet als absolute und standardisierte maximale Änderung des Mittelwertes von zwei aufeinanderfolgenden Jahren	sd
Longest Recession (LRec)	(unnatürlich) lang abfallende Grundwasserstände / längste Sequenz ohne steigende Werte	sd
Low Pulse Duration (LPD)	Durchschnittliche Dauer der Abschnitte unter dem 20. Perzentil der Nichtüberschreitung, siehe [Ri96], entnommen aus [He19]	lit
Median[0,1] (Med01)	Begrenztheit, Median nach Skalierung auf [0,1], statistisches Standardmaß, entnommen aus [He19]	ss/lit
Range Ratio (RR)	Detektion von überlagernden langperiodischen Signalen, auch ausreißerempfindlich, berechnet als Verhältnis der mittleren Jahresspannweite zur maximalen Spannweite	sd
Richards-Baker Index (RBI)	Flashiness, Häufigkeit und Schnelligkeit kurzfristiger Änderungen, für detaillierte Erklärung siehe [Ba04]	lit
SD_{diff}	Flashiness, Häufigkeit und Schnelligkeit von kurzfristigen Änderungen, berechnet als Standardabweichung aller ersten Ableitungen	sd
Seasonal Behaviour (SB)	Position des Maximums im Jahresgang, Übereinstimmung mit der erwarteten durchschnittlichen Saisonalität (Min im September, Max im März)	sd
Schiefe (Skew)	Begrenztheit, Inhomogenitäten, Ausreißer, Schiefe der Wahrscheinlichkeitsverteilung	ss
Standard Error of the Mean (SEM)	standardisierte Standardabweichung der Zeitreihe	ss
Jährliche Varianz (Y_{var})	Variabilität, Periodizität, berechnet als Median der jährlichen Varianz	sd

* lit: literature, sd: self-designed, ss: standard statistics

Das Clustering selbst erfolgt mit Self-Organizing Maps (SOM) in Verbindung mit dem DS2L-Algorithmus [CBF12], welcher speziell für prototypenbasiertes Clustering (z.B. SOM oder Neural-Gas) entwickelt wurde. Vorteil gegenüber etablierten Cluster-Algorithmen wie k-means und verschiedenen Arten von hierarchischen Methoden ist vor allem die

automatisierte Bestimmung der Clusteranzahl, aber auch die große Flexibilität bez. der möglichen Clustergrößen. So tendiert der DS2L-Algortihmus im Gegensatz zu k-means bspw. nicht zu ähnlich großen Clustern. Im Gesamtablauf der entwickelten Methodik werden zudem mehrere Ensembleansätze verfolgt. So dient ein erstes Ensemble zur Auswahl einer Feature-Kombination, mit deren Hilfe sich im mathematischen Sinne und mittels interner Clustervalidierungsindizes beurteilt, möglichst optimale Clusterergebnisse erzielen lassen. Im mathematischen Sinn werden demnach also möglichst kompakte und möglichst gut separierte Cluster gesucht, wodurch in diesem Schritt auch der Anteil an anwenderspezifischer Subjektivität im Feature-Auswahlprozess reduziert wird. Weiterhin erhöht ein nachgeschaltetes resampling-basiertes Ensemble die Robustheit des Cluster-Ergebnisses. Eine abschließende visuelle Beurteilung der Clusterqualität, also ob Cluster mit homogener Zeitreihendynamik gefunden und Zeitreihen mit charakteristischen Eigenschaften (z.B. Inhomogenitäten) zusammen gruppiert wurden, ist nötig um im Workflow korrigierend eingreifen zu können. So kann bei vorhandenem Vorwissen über das System und dessen Eigenschaften bspw. die Verwendung eines Features erzwungen werden, oder begrenzt Einfluss auf die Feinheit des Clusterings (mehr oder weniger Cluster) genommen werden. Zur Berechnung wurden Matlab 2019b sowie die SOM-Toolbox [Ve00] genutzt. Für eine detaillierte Beschreibung und Evaluierung der Methodik wird an dieser Stelle auf [WLB20, Wu20] verwiesen.

3 Ergebnisse und Ausblick

Die beispielhafte Anwendung der entwickelten Methodik auf 1196 Zeitreihen von Grundwasserständen aus Baden-Württemberg und Hessen im rechtsrheinischen Bereich des Oberrheingrabens zeigt generelle Dynamik-Unterschiede zwischen dem nördlichen und mittleren bzw. südlichen Oberrheingraben. Durch die reine räumliche Verteilung der Cluster, aber auch durch die Korrelation der Cluster mit räumlichen Daten zu mutmaßlich wichtigen Einflussfaktoren, können die Cluster zudem hydrogeologisch interpretiert werden. Mutmaßlich mit intensiverer Grundwasserbewirtschaftung und geringeren Grundwasserneubildungsraten in Verbindung stehend, finden sich im nördlichen Teil nur geringe ausgeprägte Jahresgänge und allgemein kleinere Zeitreihenvariabilität. Im mittleren und südlichen Oberrheingraben, einem Bereich mit höherer Grundwasserneubildung und weniger intensiver Grundwassernutzung treten hingehen vermehrt Ganglinien mit deutlich ausgeprägtem und eher gleichförmigem Jahresgang auf (Abb. 1a). Weiterhin lassen sich (Klein-) Gruppen differenzieren, die mutmaßlich stark durch externe Faktoren wie Zuflüssen vom Grabenrand aus Schwarzwald und Odenwald, durch den Rhein oder auch durch einzelne Staustufen bestimmt sind. Sowohl die Feature-Werteverteilungen der Cluster als auch die Ganglinien in den Clustern selbst zeigen, dass trotz Datenlücken und unterschiedlicher Zeitreihenlängen mit diesem Ansatz eine homogene Gruppierung erreicht werden konnte (Abb. 1a). Mittels bestimmter Features ist es darüber hinaus möglich, Ganglinien unterschiedlicher Dynamik, die jedoch durch andere gemeinsame Merkmale wie Ausreißer oder Sprünge charakterisiert sind, in einem Cluster zusammenzufassen (Abb. 1b).

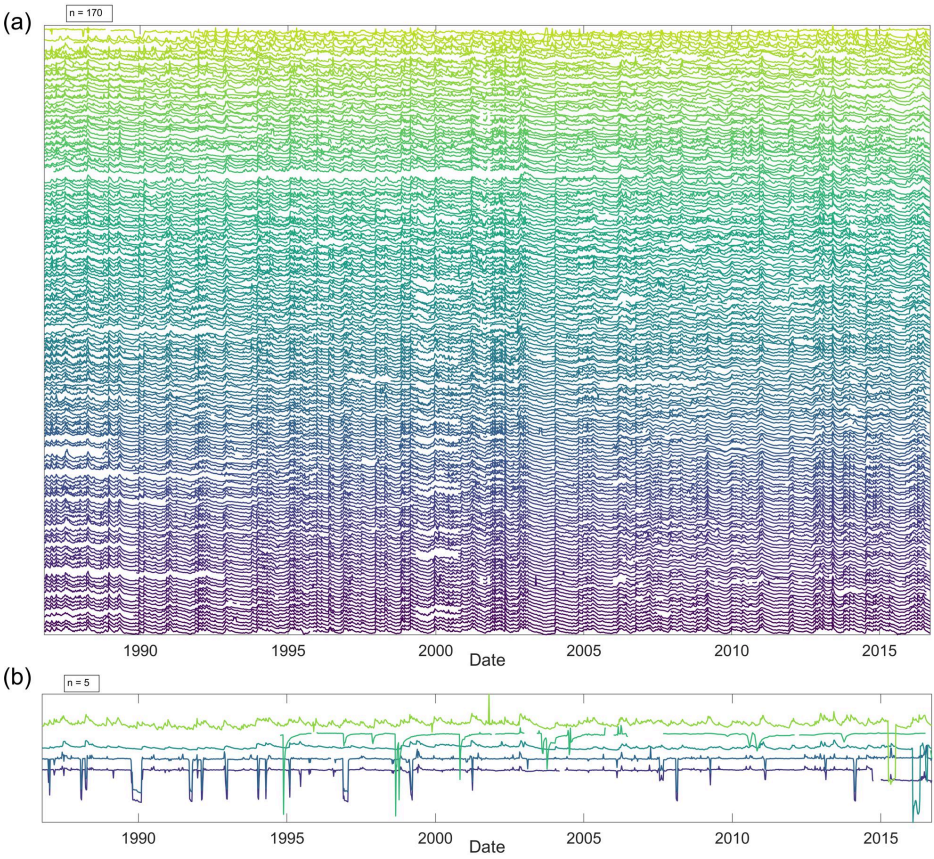

Abb. 1: Standardisierte, gestapelte Zeitreihen der Cluster 3 (a) charakterisiert durch stark korrelierte Verläufe sowie Cluster 21 (b) charakterisiert durch Ausreißer und Inhomogenitäten in der Zeitreihe.

Die vorgestellte Methode liefert damit ein solides Clustering-Framework für Umweltzeitreihen mit Vorteilen in Bezug auf (i) die Verwendung heterogener Daten, (ii) die vergleichsweise hoch automatisierte Arbeitsweise und die Möglichkeit, sich an bestimmte Datensatzmerkmale und Analyseziele anzupassen sowie (iii) robuste Ergebnisse. Neben dem verbesserten Systemverständnis und dem Erkennen lokaler und (über-) regionaler Zusammenhänge zwischen den betrachteten Zeitreihen, eignen sich die Cluster bei ausreichend guter Homogenität auch als Grundlage für weitergehende Analysen und Anwendungen. Hierzu zählen sowohl die Übertragung von z.B. Vorhersagen einer repräsentativen Zeitreihe des Clusters auf die übrigen Clusterzeitreihen, als auch das Schließen längerer Datenlücken oder eine Messwertplausibilisierung mit Hilfe hoch korrelierter Zeitreihen des gleichen Clusters. Auch die Umsetzung von Deep-Learning Ansätzen auf Zeitreihen mit für gewöhnlich zu wenigen Werten durch eine künstlich vergrößerte Datenbasis (z.B. Training auf Clusterdatensatz statt einzelne Zeitreihe) sind denkbar. Der Ansatz ist dabei auf andere Umweltzeitreihen übertragbar, wobei der Feature-Satz an die spezifische Dynamik des betrachteten Parameters angepasst werden sollten.

Literaturverzeichnis

[Ba04] Baker, David B.; Richards, R. Peter; Loftus, Timothy T.; Kramer, Jack W.: A New Flashiness Index: Characteristics and Applications to Midwestern Rivers and Streams. Journal of the American Water Resources Association, 40(2):503–522, April 2004.

[CBF12] Cabanes, Guénaël; Bennani, Younès; Fresneau, Dominique: Enriched Topological Learning for Cluster Detection and Visualization. Neural Networks, 32:186–195, August 2012.

[DDM15] D'Urso, Pierpaolo; DeGiovanni, Livia; Massari, Riccardo: Time Series Clustering by a Robust Autoregressive Metric with Application to Air Pollution. Chemometrics and Intelligent Laboratory Systems, 141:107–124, Februar 2015.

[Di13] Di Salvo, Roberto; Montalto, Placido; Nunnari, Giuseppe; Neri, Marco; Puglisi, Giuseppe: Multivariate Time Series Clustering on Geophysical Data Recorded at Mt. Etna from 1996 to 2003. Journal of Volcanology and Geothermal Research, 251:65–74, Februar 2013.

[GL16] Guo, Hongyue; Liu, Xiaodong: Dynamic Programming-Based Optimization for Segmentation and Clustering of Hydrometeorological Time Series. Stoch Environ Res Risk Assess, 30(7):1875–1887, Oktober 2016.

[GWW16] Gómez, Cristina; White, Joanne C.; Wulder, Michael A.: Optical Remotely Sensed Time Series Data for Land Cover Classification: A Review. ISPRS Journal of Photogrammetry and Remote Sensing, 116:55–72, Juni 2016.

[He19] Heudorfer, B.; Haaf, E.; Stahl, K.; Barthel, R.: Index-based Characterization and Quantification of Groundwater Dynamics. Water Resour. Res., 55(7):5575–5592, Mai 2019.

[Mi19] Mihailović, Dragutin T.; Nikolić-Đorić, Emilija; Malinović-Milićević, Slavica; Singh, Vijay P.; Mihailović, Anja; Stošić, Tatijana; Stošić, Borko; Drešković, Nusret: The Choice of an Appropriate Information Dissimilarity Measure for Hierarchical Clustering of River

Streamflow Time Series, Based on Calculated Lyapunov Exponent and Kolmogorov Measures. Entropy, 21(2):215, Februar 2019.

[NAV15] Nourani, Vahid; Alami, Mohammad Taghi; Vousoughi, Farnaz Daneshvar: Wavelet-Entropy Data Pre-Processing Approach for ANN-Based Groundwater Level Modeling. Journal of Hydrology, 524:255–269, Mai 2015.

[Ri96] Richter, Brian D.; Baumgartner, Jeffrey V.; Powell, Jennifer; Braun, David P.: A Method for Assessing Hydrologic Alteration within Ecosystems. Conservation Biology, 10(4):1163–1174, August 1996.

[TS17] Tongal, Hakan; Sivakumar, Bellie: Cross-Entropy Clustering Framework for Catchment Classification. Journal of Hydrology, 552:433–446, September 2017.

[Ve00] Vesanto, Juha: , Neural Network Tool for Data Mining: SOM Toolbox, März 2000.

[Wa18] Walz, Michael A.; Befort, Daniel J.; Kirchner-Bossi, Nicolas Otto; Ulbrich, Uwe; Leckebusch, Gregor C.: Modelling Serial Clustering and Inter-Annual Variability of European Winter Windstorms Based on Large-Scale Drivers. International Journal of Climatology, 38(7):3044–3057, 2018.

[WLB20] Wunsch, Andreas; Liesch, Tanja; Broda, Stefan: Feature-Based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing-Map-Ensembles. Water Resources Management (submitted), 2020.

[Wu20] Wunsch, Andreas: , Groundwater-Dynamic-Clustering. GitHub repository, Zenodo, 2020.