Data Mining und visuelle Datenexploration zur Beantwortung geowissenschaftlicher Fragestellungen

Doris Dransch, Mike Sips, Andrea Unger, Patrick Köthur
Deutsches GeoForschungsZentrum GFZ, Potsdam

Geoscientists collect various data to study system Earth. To extract all the knowledge inherent in the data is a challenge geoscientists have to face. Methods for data mining and visual data exploration developed in computer science facilitate knowledge extraction from data. Although these methods are mostly applied to business data, they also offer potential to extract knowledge from geoscientific data. In our research we adapt and enhance methods from data mining and visual data exploration to geoscientific requirements. Two examples are given that show how the novel methods contribute to geoscientific research: The extraction of interesting spatiotemporal patterns from large data sets and the comparison of data from simulation models and real world observations.

System Erde (2013) 3, 1 | DOI: 10.2312/GFZ.syserde.03.01.11

Ziel der Geowissenschaften ist es u. a., Prozesse des Systems Erde, wie z. B. Veränderungen des Meeresspiegels oder Veränderungen im Wasserhaushalt, besser verstehen zu können. Dazu werden immer bessere Verfahren der Beobachtung und Datenerfassung entwickelt. Eine Herausforderung ist, aus den hierbei generierten vielfältigen Daten relevante Information zu extrahieren. Die Informatik hat im Bereich Data Mining und visuelle Datenexploration ein umfangreiches Methodenrepertoire zur Informationsextraktion entwickelt.

Unter Data Mining wird der Prozess verstanden, bei dem in Daten nach Mustern, Merkmalen und Zusammenhängen, wie z. B. geographische Regionen mit ähnlichem zeitlichem Verhalten, gesucht wird. Dazu werden Ansätze aus maschinellem Lernen, Statistik und Datenbanksystemen kombiniert. Visuelle Datenexploration ermöglicht es, ein umfassendes Verständnis der Daten und der darin enthaltenen Informationen zu erlangen. Sie umfasst die visuelle Repräsentation von Daten in Verbindung mit der Möglichkeit, in der Visualisierung Operationen auf den Daten durchzuführen, z. B. die Auswahl bestimmter Werte per „Mausklick“. Im Bereich der visuellen Datenexploration werden Ansätze aus Computergraphik, Informations- und wissenschaftlicher Visualisierung sowie Mensch-Computer-Interaktion kombiniert.

Die in der Informatik entwickelten Methoden zu Data Mining und visueller Datenexploration stellen allgemeine Ansätze dar, die für die spezifischen Anforderungen der Geowissenschaften angepasst und weiterentwickelt werden müssen. Diese Anpassung und Weiterentwicklung erfolgt in der Sektion „Geoinformatik“ am Deutschen GeoForschungsZentrum GFZ.

Auffindung räumlich-zeitlicher Muster in großen Datenmengen

In den Geowissenschaften werden oft sehr große Datenmengen erzeugt. Diese Datenmengen können z. B. durch Simulationsmodelle, welche die Prozesse der Erde abbilden, oder durch Satellitenbeobachtungen erzeugt werden. Ein Beispiel sind Satellitendaten zur Entwicklung der Meereshöhe, die Rückschlüsse auf Prozesse in den Ozeanen ermöglichen. Die Daten sind aufgrund ihrer Menge nicht mehr in ihrer Gesamtheit zu erfassen. Daher sind Verfahren erforderlich, die Teilmengen der Daten extrahieren, in denen interessante Muster und Merkmale zu finden sind. Diese können beispielsweise geographische Regionen mit ähnlichem zeitlichem Verhalten der Meereshöhe sein oder ein markanter räumlicher Zustand der Meereshöhe und dessen Vorkommen in der Zeit. In der Sektion Geoinformatik am GFZ wurde ein Verfahren entwickelt, das mit einer Kombination aus Data Mining-Methoden und visueller Datenexploration Geowissenschaftlerinnen und Geowissenschaftler dabei unterstützt, aus der Gesamtmenge von Daten die Teilmengen zu extrahieren, die interessante Muster und Merkmale aufweisen. Das Verfahren wurde an gut verstandenen Daten der Meereshöhe evaluiert. Dabei konnte gezeigt werden, dass das Verfahren in der Lage ist, bekannte Phänomene wie El Niño- und La Niña-Ereignisse, d. h. veränderte Strömungen und Temperaturen in Ozean und Atmosphäre, zu extrahieren. Die Testdaten liegen als Skalarwerte in einer räumlichen Auflösung von 194 x 96 Gitterpunkten und in 876 Zeitschritten vor.

In einem ersten Schritt wird ein Data Mining-Verfahren eingesetzt. Hierbei werden die Daten mit Hilfe eines hierarchischen Clusterings aggregiert, welches alle Zeitschritte anhand der Ähnlichkeit ihrer räumlichen Ausprägungen der Meereshöhe zu Gruppen zusammenfasst. Daraus resultiert eine Hierarchie von Aggregationen, die eine Vielzahl möglicher Gruppierungen ähnlicher Zeitschritte beinhaltet (Abb. 1a). Jede Aggregation in der Hierarchie ermöglicht es, eine bestimmte Menge von Zeitschritten durch ein einzelnes räumliches Muster näherungsweise zu beschreiben (Abb. 1b). Welche der Aggregationen relevant sind, hängt von der geowissenschaftlichen Fragestellung ab und kann daher nur von Experten bestimmt werden. Daher wurde das Clustering-Verfahren durch zusätzliche Komponenten (K1, K2, K3) zur visuellen Datenexploration ergänzt (Abb. 2). Damit können das Ergebnis des hierarchischen Clustering-Verfahrens wie auch die räumlichen Muster der einzelnen Aggregationen dargestellt werden. K1 zeigt die hierarchische Baumstruktur aller durch das Clustering gebildeten möglichen Aggregationen der Zeitschritte mit ähnlichen räumlichen Ausprägungen der Meereshöhe. Die Wissenschaftlerinnen und Wissenschaftler können die einzelnen Aggregationen direkt am Bildschirm auswählen und visuell explorieren. Für jede Aggregation lässt sich mittels der Visualisierungskomponente K1 und K2 die räumliche wie auch die zeitliche Verteilung der Meereshöhe anzeigen. Farben der Umrandung ordnen räumliche Verteilung und zeitliche Verteilung einander

Links: Die Weiterentwicklung von Methoden aus Data Mining und visueller Datenexploration für geowissenschaftliche Fragestellungen ist Forschungsgegenstand der Sektion „Geoinformatik“ am GFZ.

Left: The enhancement of methods from data mining and visual data exploration for geoscientific application is research topic of “Geoinformatics” Section at GFZ.

Kontakt: D. Dransch
(dransch@gfz-potsdam.de)

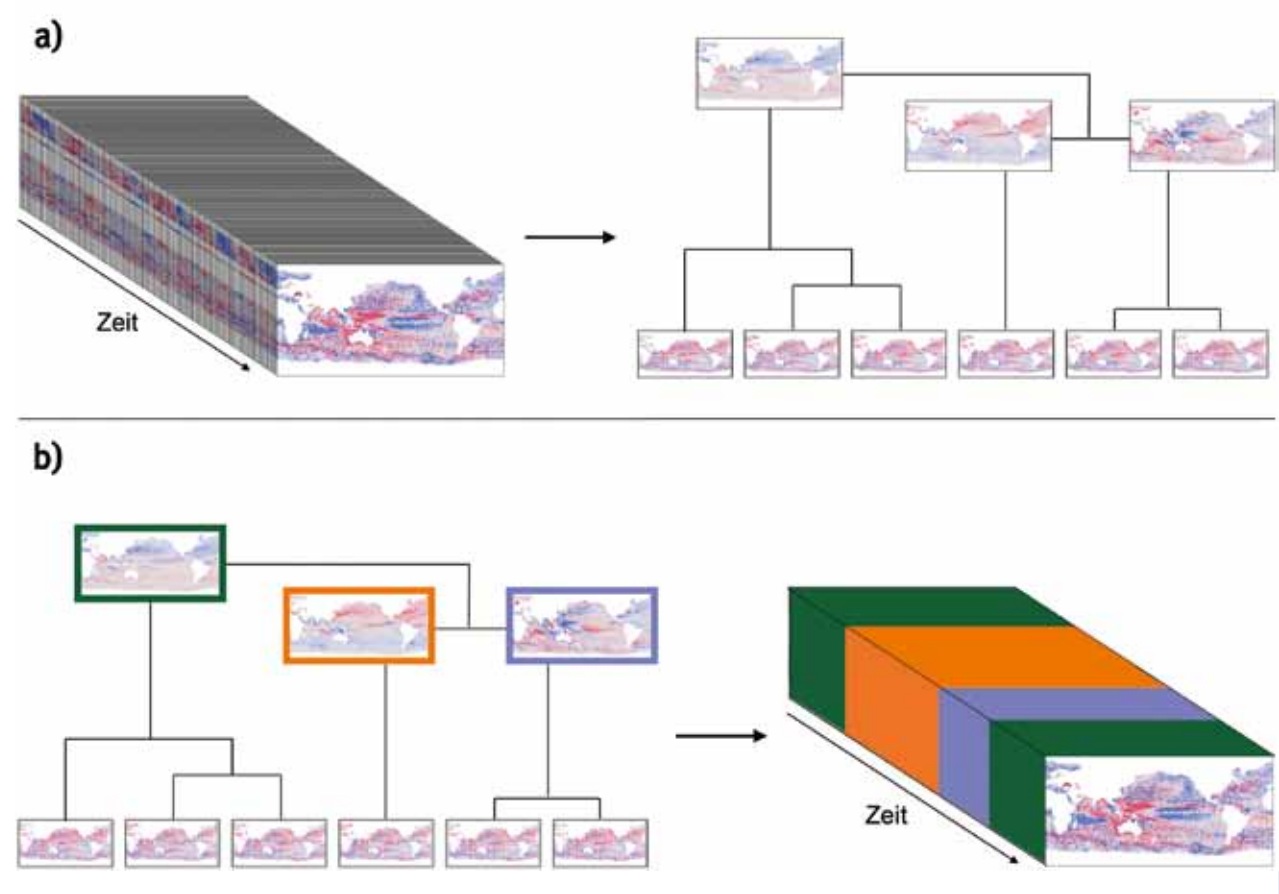

Abb. 1: Data Mining am Beispiel von El Niño- und La Niña-Ereignissen: Durch Aggregation werden alle Zeitschritte anhand der Ähnlichkeit ihrer räumlichen Ausprägungen der Meereshöhe zu Gruppen zusammengefasst (a). Jede Aggregation in der Hierarchie beschreibt näherungsweise eine bestimmte Menge von Zeitschritten durch ein einzelnes räumliches Muster (b).

Fig. 1: Data Mining using the example of El Niño and La Niña phenomena: The clustering groups all time steps with similar spatial patterns of sea level height (a). Each cluster approximates a set of time steps with one spatial pattern (b).

zu. Zusätzlich wird ein statistisches Maß durch die Komponente K3 visualisiert, das die Homogenität bzw. Heterogenität einer Aggregation durch einen Balken deutlich macht. Je länger der Balken ist, umso heterogener ist die Aggregation. Durch die visuelle Datenexploration werden Wissenschaftlerinnen und Wissenschaftler in die Lage versetzt, das automatisierte Clustering-Verfahren nachzuvollziehen und die einzelnen Aggregationen zu bewerten. Auf dieser Basis können die für die Fragestellung relevanten Aggregationen ausgewählt werden.

Im Testbeispiel konnten die Wissenschaftlerinnen und Wissenschaftler mit Hilfe unseres Verfahrens die bekannten charakteristischen Muster der El Niño- und La Niña-Ereignisse aus den Daten der Meereshöhe extrahieren. Dies zeigt, dass mit dem Verfahren relevante Muster in großen Datenmengen erkannt werden können. Aufgrund des positiven Resultats wird das Verfahren derzeit erweitert, um auch Regionen mit gleichem zeitlichem Verhalten aus großen raum-zeitlichen Datenmengen extrahieren zu können. Das Verfahren bietet die Möglichkeit, Teilmengen objektiv auf Basis definierter Kriterien zu bilden.

Vergleich von Daten aus Beobachtungen und Simulationsmodellen

Simulationsmodelle müssen daraufhin überprüft und bewertet werden, wieweit sie Prozesse der realen Welt korrekt abbilden. Ein Verfahren für diese Validierung ist der Vergleich von Daten aus Beobachtungen mit Daten aus Simulationen. Eine Möglichkeit ist, mit Hilfe statistischer Verfahren eine Maßzahl für die Übereinstimmung der Simulations- und Beobachtungsdaten zu berechnen. Diese Maßzahl kann einen Wert zwischen Null (keine Übereinstimmung) und Eins (volle Übereinstimmung) annehmen. Werte zwischen Null und Eins stellen unscharfe Ergebnisse (Fuzzylogik) dar, wie z. B. „stimmt wenig überein“ oder „stimmt stark überein“.

In enger Kooperation von Geowissenschaften und Informatik wurde am GFZ mit Hilfe interaktiver Visualisierung ein Verfahren und Werkzeug zum Vergleich von Beobachtungs- und Simulationsdaten entwickelt. Dieses Verfahren wurde exemplarisch für die Validierung eines Modells der glazial-

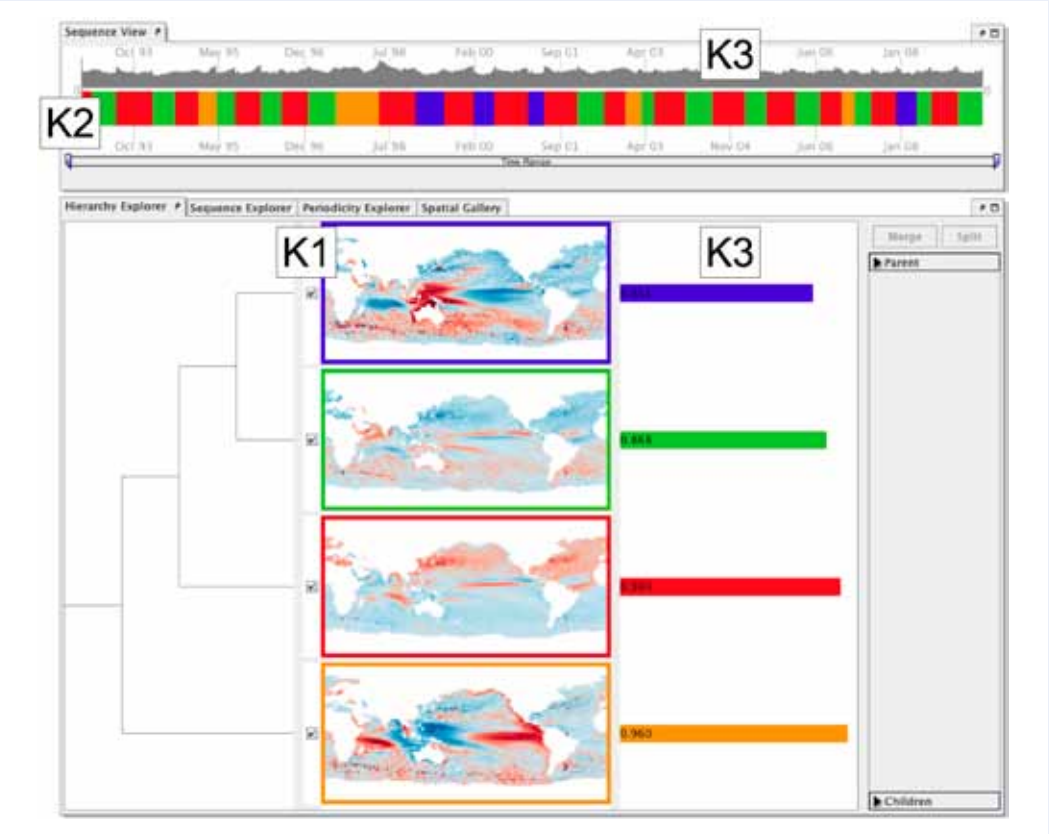

Abb. 2: Werkzeug zur visuellen Exploration der durch das Clustering erzeugten Aggregationen von Zeitschritten mit ähnlichen räumlichen Ausprägungen der Meereshöhe. Mit Hilfe der Komponente K1 kann die hierarchische Struktur der Aggregation ähnlicher Zeitschritte exploriert werden. Für jeden Knoten der Hierarchie lässt sich die räumliche Ausprägung der Meereshöhe darstellen. Komponente K2 zeigt die zeitliche Verteilung der Aggregationen. Komponente K3 gibt Auskunft über die Güte der Aggregation.

Fig. 2: Tool to visually explore the clustering results of time series data of sea level height. K1 presents the resulting cluster hierarchy. It also allows for depicting the spatial pattern of sea level height for each cluster. K2 visualizes the temporal distribution of clusters. K3 shows the quality of single aggregations.

isostatischen Ausgleichsbewegung implementiert. Das Modell bildet das Heben und Senken der Erdkruste ab, das durch das Abtauen der Eismassen der letzten Kaltzeit hervorgerufen wurde. Dieser Prozess hat Auswirkungen auf die Veränderung des Meeresspiegels und ist daher von entscheidender Bedeutung für die Vulnerabilität von Küstenregionen.

Für die Validierung des Simulationsmodells stehen überwiegend paläontologische Daten zur Verfügung, wie z. B. fossile Muscheln oder Pflanzenbestandteile. Die paläontologischen Daten dienen als Indikator für historische Meeresspiegelhöhen. Sie liefern keine exakten Werte für die Meeresspiegelhöhe sondern unscharfe Werteintervalle. Je nach Größe der Intervalle ist ihre Aussagekraft für die Bewertung der Güte eines Simulationsmodells unterschiedlich. Je kleiner die Intervalle, umso besser die Aussagekraft. Um das Maß der Übereinstimmung der exakten Simulationsdaten und der unscharfen paläontologischen Beobachtungsdaten berechnen zu können, wurde mittels des Fuzzylogik-Ansatzes für jede paläontologische Beobachtung ein Wert zwischen Null und Eins errechnet. Durch diesen Vergleich sollen folgende Fragen beantwortet werden:

A Welche Wertebelegung der Parameter des Modells der glazial-isostatischen Ausgleichsbewegung repräsentiert am besten den realen Prozess? Die Werte der Modellparameter, wie die Dicke der Lithosphäre, Viskosität des oberen Erdmantels und Viskosität des unteren Erdmantels, sind nicht bekannt, sondern nur geschätzt. Es ist zu prüfen, welche Parameterwerte zur besten Übereinstimmung von Beobachtungs- und Simulationsdaten in Raum und Zeit führen.

B Gibt das Modell die reale raum-zeitliche Variation der glazial-isostatischen Ausgleichbewegung wieder? Dazu ist zu überprüfen, wie hoch die Übereinstimmung von Beobachtungs- und Simulationsdaten in verschiedenen Regionen und Zeiträumen ist.

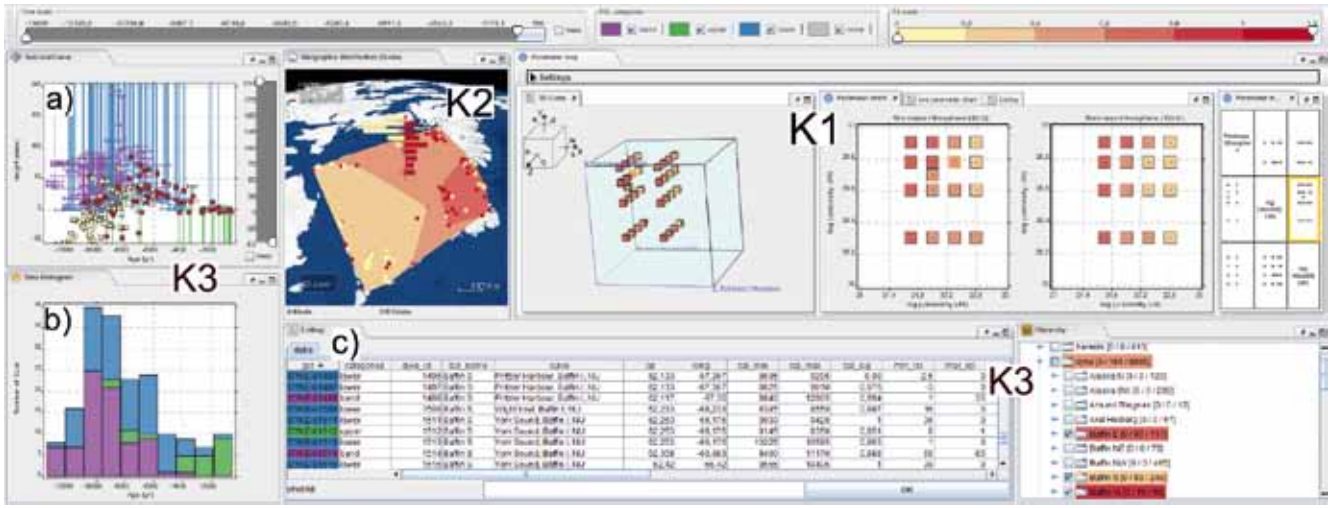

Abb. 3: Werkzeug zur visuellen Datenexploration für den Vergleich von Daten aus Beobachtungen und Simulationsmodellen. K1 erlaubt die Exploration der Übereinstimmung von Beobachtungs- und Simulationsdaten bezogen auf verschiedene Werte der Modellparameter. K2 zeigt die Übereinstimmung der Beobachtungs- und Simulationsdaten in ihrer räumlichen und zeitlichen Verteilung. K3 dient der Exploration der Beobachtungsdaten. Gezeigt werden für die einzelnen Beobachtungsdaten a) das Alter und ihre Gültigkeitsintervalle (links oben), b) die Häufigkeitsverteilung nach Gruppen und Alter (Diagramm darunter), sowie c) Detailinformationen zu jeder Beobachtung aus der Datenbank (Tabellen rechts unten).

Fig. 3: Tool to visually compare data from observation and simulation models and their goodness of fit. K1 facilitates exploration the goodness of fit of observation and simulation data with respect to different values of model parameters. K2 shows goodness of fit in space and time. K3 allows for exploring observation data. It depicts a) age and interval of value (top left), b) frequency distribution (bottom left), and c) detail information for each single observation from the data base (tables bottom right).

C Wie hoch ist die Aussagekraft der einzelnen paläontologischen Beobachtungsdaten für die Bewertung der Güte des Simulationsmodells? Es ist zu beurteilen, wie groß die Werteintervalle der einzelnen unscharfen paläontologischen Beobachtungsdaten sind. Aus der Größe der Werteintervalle lässt sich ableiten, wie zuverlässig die einzelnen paläontologischen Beobachtungen sind; je kleiner die Intervalle, umso zuverlässiger ist die Beobachtung.

Das gemeinsam mit Geowissenschaftlerinnen und Geowissenschaftlern entwickelte Verfahren und Werkzeug hilft bei der Beantwortung dieser drei Fragen. Das Verfahren erlaubt es, visuell zu explorieren, wie Beobachtungs- und Simulationsdaten in Raum und Zeit übereinstimmen. Dafür wurden verschiedene interagierende Komponenten entwickelt. Basis ist eine Datenbank, die die Beobachtungsdaten, Simulationsdaten und verschiedene Modellbeschreibungen mit unterschiedlichen Parametrisierungen integriert und für weitere Berechnungen und Visualisierungen zur Verfügung stellt. Die Datenbank ist eng mit drei interaktiven Visualisierungskomponenten (K1, K2 und K3) verknüpft (Abb. 3).

Die Visualisierungskomponente K1 unterstützt die Beantwortung der Frage A: Welche Wertebelegung der Parameter des Modells der glazial-isostatischen Ausgleichsbewegung repräsentiert am besten den realen Prozess? Sie zeigt die Übereinstimmung von Beobachtungs- und Simulationsdaten für bestimmte Wertebelegungen der Parameter des glazial-isostatischen Simulationsmodells. Der Würfel spannt den Modellparameterraum auf. In diesem Beispiel sind das folgende Parameter: Dicke der Lithosphäre, Viskosität des oberen Erdmantels und Viskosität des unteren Erdmantels. Jeder kleine Würfel zeigt ein konkretes Modell mit einer konkreten Parameterwertbelegung. Die Farbe gibt an, wie hoch die Übereinstimmung von Beobachtungs- und Simulationsdaten für dieses konkrete Modell ist: je röter, umso höher die Übereinstimmung. Die Darstellungen rechts neben dem Würfel zeigen einzelne Schnitte durch den Modellparameterraum, die interaktiv selektiert werden können.

Die Visualisierungskomponenten K2 und K3 tragen zur Beantwortung der Fragen B und C bei: Gibt das Modell die reale raum-zeitliche Variation der glazial-isostatischen Ausgleichsbewegung wieder? Wie hoch ist die Aussagekraft der einzelnen paläontologischen Beobachtungsdaten für die Bewertung der Güte des Simulationsmodells? K2 zeigt für einzelne oder Gruppen von Beobachtungen die Übereinstimmung mit den Simulationsdaten. Die Farbe ist analog zu der Farbe der Würfel gewählt, je röter die Färbung, je besser die Übereinstimmung. Durch ein Diagramm wird zudem dargestellt, wie die paläontologischen Beobachtungsdaten über die Zeit verteilt sind, und wie sie mit den Simulationsdaten übereinstimmen. Die Vertikale ist die Zeitachse, die Horizontale gibt an, wie viele paläontologische Beobachtungsdaten in jeder Altersgruppe vorhanden

sind. Die Anordnung der Balken zur Mittelachse zeigt an, wie viele Observationsdaten mit Simulationsdaten übereinstimmen (rechte Seite) bzw. nicht übereinstimmen (linke Seite).
Die Komponente K3 gibt Informationen zu einzelnen paläontologischen Beobachtungsdaten. Das Diagramm links oben in Abb. 3 zeigt das Alter und das Werteintervall der einzelnen Daten. Die Farben Grün, Blau und Violett differenzieren die verschiedenen Typen von Meeresspiegelindikatoren. Die gelb- und rotgefärbten Quadrate stellen wie in K1 die Übereinstimmung von Beobachtungs- und Simulationsdaten dar. Das Diagramm unterhalb der Abbildung macht deutlich, wie viele paläontologische Beobachtungsdaten pro Typ und Alter vorhanden sind. Die Tabellen rechts von dem Diagramm enthalten Sichten auf die Datensätze der Datenbank; sie geben zusätzliche Informationen zu den paläontologischen Beobachtungsdaten.

Die einzelnen Komponenten sind über ein Interaktionskonzept miteinander verbunden. Wird in K1 interaktiv ein konkretes Modell ausgewählt, wird die Übereinstimmung von Beobachtungs- und Simulationsdaten berechnet und das Ergebnis sofort in den Ansichten von K2 und K3 dargestellt. Umgekehrt kann in K2 oder K3 interaktiv eine Teilmenge von Daten ausgewählt werden, um für diese Teilmenge die Übereinstimmung bezüglich eines konkreten in K1 ausgewählten Modells zu berechnen und anzuzeigen. Wissenschaftlerinnen und Wissenschaftler können mit diesem Verfahren sukzessive die Übereinstimmung der Beobachtungs- und Simulationsdaten explorieren und dabei ein umfassendes Verständnis dafür gewinnen, wie gut ihr Simulationsmodell den realen Prozess wiedergibt.

Ausblick

Data Mining und visuelle Datenexploration ergänzen traditionelle geowissenschaftliche Methoden. Sie ermöglichen, Daten ohne Annahmen zu explorieren und damit auch unbekannte Informationen aus Daten zu extrahieren. Visuelle Datenexploration gibt zudem einen schnellen, intuitiven und umfassenden Blick über die Daten und trägt dadurch zu einem besseren Verständnis der Daten bei. Zukünftig ist geplant, Data Mining und visuelle Datenexploration für weitere geowissenschaftliche Anwendungsfelder zu erschließen. Ein Beispiel hierfür ist die Extraktion von Informationen aus Daten verschiedener Simulationsläufe von geochemischen Modellen der Wasser-Gesteins-Interaktion. Ein anderes Beispiel ist die Extraktion von Informationen zu Auswirkungen und Schäden von Naturgefahren aus sozialen Netzwerken wie „Twitter“.

Literatur

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Fabrikant, S.I., Jern, M., Kraak, M.-J., Schumann, H., Tominski, C. (2010): Space, time and visual analytics. - International Journal of Geographical Information Science, 24, 10, 1577-1600, 10.1080/13658816.2010.50804

Dransch, D., Köthur, P., Schulte, S., Klemann, V., Dobslaw, H. (2010): Assessing the quality of geoscientific simulation models with visual analytics methods - a design study. - International Journal of Geographical Information Science, 24, 10, 1459-1479, 10.1080/13658816.2010.510800.

Keim, D., Kohlhammer, J. J., Ellis, G., Mansmann, F. (Eds.) (2010): Mastering the information age : solving problems with visual analytics, Goslar, 168 p.

Köthur, P., Sips, M., Unger, A., Kuhlmann, J., Dransch, D. (2013 online first): Interactive visual summaries for detection and assessment of spatiotemporal patterns in geospatial time series. - Information Visualization, 10.1177/1473871613481692.

Sips, M., Köthur, P., Unger, A., Hege, H.-C., Dransch, D. (2012): A Visual Analytics Approach to Multiscale Exploration of Environmental Time Series. - IEEE Transactions on Visualization and Computer Graphics, 18, 12, 2899-2907, 10.1109/TVCG.2012.191.

Unger, A., Schulte, S., Klemann, V., Dransch, D. (2012): A Visual Analytics Concept for the Validation of Geoscientific Simulation Models. - IEEE Transactions on Visualization and Computer Graphics, 18, 12, 2216-2225, 10.1109/TVCG.2012.190.