GFZ Helmholtz-Zentrum Potsdam

Aus: System Erde. GFZ-Journal (2018) 8, 1
„Data Science in den Geowissenschaften“
Alle Artikel verfügbar im Internet: http://systemerde.gfz-potsdam.de

Helmholtz-Zentrum Potsdam
Deutsches GeoForschungsZentrum

Big-Data-Ansätze für die schnelle Extraktion relevanter Informationen und Muster aus großen Datenmengen

Mike Sips, Daniel Scheffler, Tobias Rawald, Daniel Eggert, André Hollstein, Karl Segl
Deutsches GeoForschungsZentrum GFZ, Potsdam

Progress in sensor systems and computer simulation create large volumes of data with a variety of parameters. This development brings Big Data and the related challenges for data processing and data analysis also into the focus of geoscience. Computer science has developed concepts and technologies to handle Big Data. Geoscience can benefit from them since they facilitate efficient information extraction from big data, such as data from satellite-based remote sensing systems, or data from seismological or meteorological observation systems. To make use of computer science concepts and technologies, they have to be adapted into geoscience. Examples for this adaption are the development of efficient scalable geoscientific analysis methods by applying the divide-and-recombine concept, or the adaption of geoscientific methods to existing Big Data technologies.

Als Big Data werden Daten bezeichnet, die besondere Anforderungen an die Datenverarbeitung und die Datenanalyse stellen. In der Informatik werden Big Data durch „5 Vs" charakterisiert: *Volume* (Datenmenge), *Variety* (Vielfalt der Daten), *Velocity* (Geschwindigkeit der Erzeugung), *Veracity* (Zuverlässigkeit) und *Value* (Wert). Die 5 Vs lassen sich auch in geowissenschaftlichen Daten wiederfinden.

Der rasante technologische Fortschritt in der Sensorentwicklung und der Computersimulation ermöglicht es Wissenschaftlerinnen und Wissenschaftlern, immer größere Datenmengen mit einer Vielzahl unterschiedlicher Parameter zu generieren. Ein Beispiel für große Datenmengen am Deutschen GeoForschungsZentrum GFZ sind optische Satellitendaten. Diese Daten werden genutzt, um Zustand und Veränderung der Erdoberfläche, wie z. B. Landnutzung, schnell zu erfassen. Die Datenmengen dieser Anwendungen liegen im Petabyte-Bereich (PB), wobei 1 PB = 1000 Terabyte (TB) entsprechen. Für die Speicherung dieser Datenmengen werden schon heute in der Regel mehrere Hundert handelsübliche Festplatten benötigt (siehe Tab. 1). Die Datenmengen werden in Zukunft mit der Verfügbarkeit neuer Satellitensysteme enorm anwachsen. Die Geschwindigkeit, mit der geowissenschaftliche Daten generiert werden, steigt zudem stetig mit dem anhaltenden Fortschritt in der Sensor- und Rechentechnik.

Die durch die vielfältigen Sensorsysteme und Simulationsmodelle erzeugten Daten müssen zueinander in Beziehung gesetzt werden, um das komplexe System Erde mit seinen ablaufenden Prozessen verstehen zu können. Die gemeinsame Verarbeitung und Analyse der Daten ist eine anspruchsvolle Aufgabe, da die Daten sehr heterogen sind. Sie beschreiben die Prozesse des Erdsystems durch unterschiedliche Variablen, in verschiedenen räumlichen und zeitlichen Auflösungen, in unterschiedlicher räumlicher und zeitlicher Verteilung und mit unterschiedlicher Zuverlässigkeit. Die erfassten Daten sind Basis der wissenschaftlichen Arbeit und daher von großem Wert, was sich in den vielen Bestrebungen zu einem effektiven Management von Forschungsdaten mit entsprechenden Infrastrukturen wiederspiegelt.

Herausforderungen bei der Verarbeitung geowissenschaftlicher Big Data

Die Herausforderungen an die Datenverarbeitung und die Datenanalyse, die sich aus den 5 Vs ergeben, werden in Zukunft dringende Forschungsaspekte in den Geowissenschaften sein. Ein Beispiel dafür ist die schnelle Extraktion relevanter Informationen und Muster aus großen Datenmengen. Erforderlich dafür sind zum einen die Entwicklung effizienter und skalierbarer geowissenschaftlicher Analysemethoden und zum anderen die Anpassung geowissenschaftlicher Methoden an Big-Data-Technologien.

Tab. 1: Optische Satellitendaten sind ein Beispiel für große Datenmengen am GFZ.

Tab. 1: Optical satellite data is an example for large data at GFZ.

Archiv	Zeitraum	Anzahl Szenen	Gesamtgröße	Anzahl Festplatten (à 4 TB)
Landsat TM (4-5)	1980-2012	2 765 783	0,4 PB	100
Landsat ETM (7)	1999-heute	2 465 602	0,5 PB	125
Landsat OLI/TIRS (8)	2013-heute	1 226 312	1,1 PB	275
Sentinel-2	2015-heute	3 127 130	2 PB	500

Links: Big Data in den Geowissenschaften – eine aktuelle Herausforderung

Left: Big Data in Geoscience – a current challenge

Kontakt: M. Sips
(mike.sips@gfz-potsdam.de)

Entwicklung effizienter und skalierbarer geowissenschaftlicher Analysemethoden

Die Laufzeit, die eine Analysemethode zur Berechnung von Daten benötigt, hängt sehr oft von der Größe der Eingangsdatenmenge ab. In der Informatik bezeichnet man diesen funktionalen Zusammenhang zwischen der Größe der Eingangsdatenmenge und der Anzahl von Rechenschritten (Laufzeit) als die Komplexität von Algorithmen. Die rasante technische Entwicklung von Rechnern hat zwar dazu geführt, dass viele Analysemethoden sehr schnell für kleine und mittelgroße Datenmengen Ergebnisse liefern, die Laufzeit aber mit zunehmender Größe der Eingangsdaten rapide zunimmt. Zum Beispiel benötigt der Random-Forest-Algorithmus (*Breiman, 2001*), der für die Klassifikation der Landnutzung in optischen Satellitendaten eingesetzt wird, nur wenige Minuten, um die Klassifikation in Satellitendaten mit geringer räumlicher Auflösung vorzunehmen. Für Satellitendaten mit mittlerer räumlicher Auflösung steigt die Laufzeit bereits auf Stunden an, und schließlich benötigt der Algorithmus Tage und Wochen für die Klassifikation von Daten mit hoher räumlicher Auflösung wie z. B. Sentinel-2-Daten (siehe Tab. 2 und Abb. 1).

Ein anderes Beispiel ist die Analyse großer Zeitreihen mit der Methode der Recurrence Quantification Analysis (RQA), die zur Analyse dynamischer Systeme, wie z. B. des Klimasystems, eingesetzt wird. Aufgrund der Komplexität der Methode nimmt die Berechnungszeit mit zunehmender Datenmenge in hohem Maße zu und steigt auf mehrere Stunden oder Tage. Daher ist es wichtig, die geowissenschaftlichen Analysemethoden mit Hilfe informatischer Methoden so anzupassen, dass sie skalierbar werden und damit auch mit großen Datenmengen effizient umgehen können.

Anpassung geowissenschaftlicher Methoden an Big-Data-Technologien

Eine weitere Möglichkeit relevante Informationen und Muster schnell aus großen Datenmengen zu extrahieren, ist die Verteilung der Daten und der Berechnungsschritte einer Analysemethode auf mehrere Rechner, um eine parallele Datenverarbeitung zu ermöglichen. Dazu werden verschiedene Big-Data-Technologien kombiniert. Eingesetzt werden zum einen schnelle und verteilte

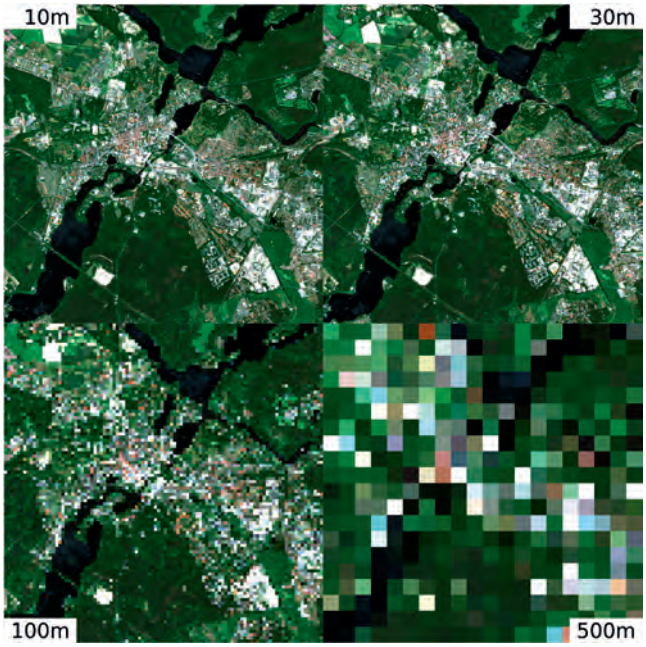

Abb. 1: Die zunehmende räumliche Auflösung von Satellitendaten durch verbesserte Sensoren erhöht wesentlich das Datenvolumen und damit die erforderliche Rechenzeit bei der Verarbeitung (siehe Tab. 2).

Fig. 1: The increasing spatial resolution of satellite data resulting from sensor improvement also increases data volume and computing times for data processing (see tab. 2).

Dateisysteme, welche die Daten für Speicherung und Zugriff automatisch verteilen, zum anderen werden parallele Analysesysteme genutzt, welche die Berechnungsschritte einer Methode automatisch auf verschiedene Rechner verteilen. HDFS und XtreemFS sind zwei Beispiele für schnelle und verteilte Dateisysteme. HADOOP, FLINK, SPARK und STORM sind Beispiele für parallele Analysesysteme. Diese Analysesysteme basieren auf dem Map-Reduce-Programmiermodell, dass von Google entwickelt worden ist (*Dean et al., 2004*). Die Map-Phase verteilt dabei automatisch die Berechnungsschritte auf die vorhandenen Rechner,

Tab. 2: Laufzeit des Random-Forest-Algorithmus steigt in Abhängigkeit von der räumlichen Auflösung von optischen Fernerkundungsdaten enorm an.

Tab. 2: Computing times of the Random Forest algorithm depend significantly on the spatial resolution of optical remote sensing data.

Satellitenmission	Auflösung	Anzahl Pixel für Europa (10 Mio. km² = 10^{13} m²)	Laufzeit für Klassifikation (20 min für 10^8 Pixel)
MODIS	500 m → 250 000 m² pro Pixel	$4 * 10^7$	8 min
Landsat (Thermalbänder)	100 m → 10 000 m² pro Pixel	10^9	3,3 h
Landsat	30 m → 900 m² pro Pixel	$1{,}1 * 10^{10}$	1,5 Tage
Sentinel-2	10 m → 100 m² pro Pixel	10^{11}	2 Wochen

System Erde (2018) 8,1 | DOI: 10.2312/GFZ.syserde.08.01.6

und die Reduce-Phase führt die einzelnen Zwischenergebnisse der Map-Phase zu einer globalen Gesamtlösung zusammen. Die Berechnungsschritte werden dabei so verteilt, dass diese automatisch auf dem Rechner ausgeführt werden, der die für den Berechnungsschritt benötigten Daten vorhält.

Eine Voraussetzung für die Anwendung von Big-Data-Technologien in den Geowissenschaften ist die Anpassung der existierenden geowissenschaftlichen Methoden an aktuelle Big-Data-Technologien. Eine zusätzliche Herausforderung bei der Entwicklung von Big-Data-fähigen Lösungen für die geowissenschaftliche Anwendung ist die notwendige Kooperation in interdisziplinären Teams aus den Geowissenschaften und der Informatik, in denen unterschiedliche Konzeptwelten und Fachsprachen überbrückt werden müssen.

Big-Data-Ansätze am GFZ

Am GFZ werden in enger Kooperation von Geowissenschaften und Informatik Lösungen für die spezifischen Big-Data-Aufgaben entwickelt. Die folgenden zwei Beispiele geben einen Einblick in diese Projekte.

Entwicklung eines effizienten Algorithmus für die Recurrence Quantification Analysis (RQA) auf der Basis des Informatikkonzepts Divide and Recombine

Recurrence Quantification Analysis (RQA) ist eine grundlegende Methode zur Analyse von Rekurrenzen in Zeitreihen und damit zur Analyse des dynamischen Verhaltens von Systemen. Sie bietet ein breites Anwendungsfeld, in den Geowissenschaften wird sie z. B. eingesetzt, um saisonale Veränderungen im Klimasystem oder die Dynamik von Erbeben zu untersuchen.

Die Idee der RQA ist es, Vektoren aus einer Zeitreihe zu extrahieren, die paarweise Ähnlichkeit zwischen diesen Vektoren zu bestimmen und anschließend mit einem Schwellenwert zu normieren. Liegt die Ähnlichkeit zwischen zwei Vektoren unterhalb des Schwellenwerts, wird diesem Paar eine 1 zugewiesen (Rekurrenz). Ist diese Bedingung nicht erfüllt, wird einem Paar von Vektoren eine 0 zugeordnet. Das Ergebnis ist eine Rekurrenz-Matrix, wobei 1 mit schwarz und 0 mit weiß repräsentiert werden (Abb. 2a). Für die Analyse von Rekurrenzen spielen die Diagonalen und vertikalen schwarzen Linien eine wichtige Rolle.

In diesem Beispiel wurde die RQA-Methode verwendet, um die Potsdamer Klima-Messreihe, in der seit mehr als 100 Jahren kontinuierlich Wetterdaten erfasst werden, auf saisonale Veränderungen hin zu untersuchen. Bestehende Berechnungsansätze zur Durchführung der RQA können entweder nur Zeitreihen bis zu einer bestimmten Länge verarbeiten oder benötigen viel Zeit zur Analyse von sehr langen Zeitreihen. Die Big-Data-Herausforderung in diesem Projekt bestand darin, die Laufzeit der RQA-Methode zu reduzieren; Ziel war es, die Detektion von Diagonalen und Vertikalen für die Potsdamer Klima-Messreihe mit 1 Mio. Datenpunkten von mehr als sechs Stunden auf wenige Minuten zu reduzieren. Damit ist es möglich, die RQA-Methode auch für eine große Eingabedatenmenge effizient zu parametrisieren.

Um die Laufzeit der RQA signifikant zu verringern, wurde die RQA-Methode mit Hilfe des Informatik-Konzepts „Divide and Recombine“ (D&R), sowie unter Nutzung einer handelsüblichen Grafikkarte (GPU = graphics processing unit) parallelisiert. Die Grundidee von D&R besteht darin, eine große Datenmenge in kleinere Teilmengen zu zerlegen (Divide). Danach wird für jede Teilmenge ein Teilergebnis berechnet. Die Berechnung dieser Teilmengen wird meist parallel ausgeführt. Zum Schluss werden die Teilergebnisse zu einem globalen Gesamtergebnis kombiniert (Recombine). Im Gegensatz zum Map-Reduce-Programmiermodell, welches Berechnungsschritte anhand technischer Metriken verteilt, wird die Datenmenge anhand semantischer Bedingungen zerlegt. In dem hier vorgestellten Ansatz wird die Rekurrenz-Matrix zunächst in mehrere Teilmatrizen zerlegt (Abb. 2b). Innerhalb jeder Teilmatrix werden alle vertikalen und diagonalen Linien mit Hilfe einer GPU detektiert und die jeweilige Anzahl und Länge in lokalen Histogrammen gespeichert. Die GPU erlaubt es, eine große Anzahl von vertikalen und diagonalen Detektionsaufgaben gleichzeitig auszuführen (in Abb. 2c, gestrichelte Pfeile). In der Recombine-Phase werden die lokalen Histogramme zu einem globalen Histo-

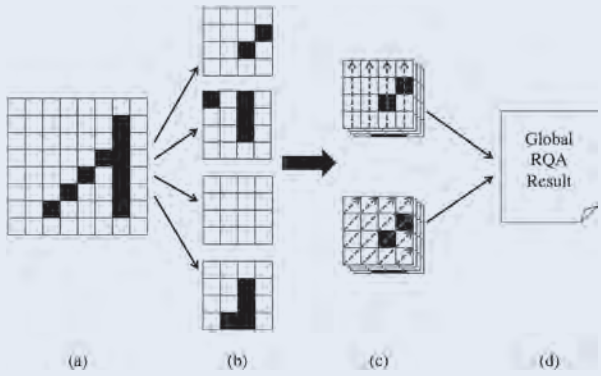

Abb. 2: Parallelisierung der RQA-Methode mit Hilfe von „Divide and Recombine“. (a) Rekurrenz-Matrix, die die Ähnlichkeit zwischen zwei Vektoren wiedergibt. Ähnlichen Werten wird schwarz zugewiesen, nicht ähnlichen Werten weiß. (b) Zerlegung der Rekurrenz-Matrix in Teilmatrizen. (c) Parallele Bearbeitung der Teilmatrizen. (d) Zusammenführen der Ergebnisse aus den Teilmatrizen. (Adaptiert und übersetzt auf Basis der Genehmigung durch Springer Nature, Springer Proceedings in Mathematics and Statistics; 103, pp. 17-29. Fast Computation of Recurrences in Long Time Series, Rawald, T. et al., © 2014.)

Fig. 2: Parallelizing RQA method by utilizing “Divide and Recombine.”(a) Recurrence-matrix, it shows the similarity between two vectors. Similar vectors are depicted by black, non-similar vectors by white. (b) Recurrence-matrix is divided in sub-matrices. (c) Parallel processing of sub-matrices. (d) Results of sub-matrices are combined in one final result

gramm addiert (Abb. 2d). Die Erweiterung der RQA-Methode mit D&R ermöglicht eine Verringerung der Laufzeit der RQA-Methode von sechs Stunden auf zehn Minuten. D&R ist dabei nicht auf eine GPU beschränkt. Durch die Nutzung mehrerer GPU kann die Laufzeit bis auf eine Minute gesenkt werden. Dieser Ansatz steht als Open-Source-Softwarepaket zur Verfügung, er ist für alle Anwendungsfelder einsetzbar (*Rawald, 2015; Rawald et al., 2017*).

Entwicklung von Algorithmen zur multisensoralen Analyse von optischen Satellitendaten in der Big-Data-Analyseumgebung Flink
Die multisensorale Analyse von optischen Satellitendaten, bei der Daten aus verschiedenen Satelliten-Beobachtungssystemen gemeinsam ausgewertet werden, ist Voraussetzung für verschiedene Fragestellungen, wie z. B. die Veränderung der Landbedeckung auf der Erde in den letzten 30 Jahren. Ein zentrales Problem dabei ist die Heterogenität der Daten aufgrund der unterschiedlichen Sensorspezifikationen der einzelnen Satellitensysteme. Zum Beispiel variieren die räumliche und zeitliche Auflösung der Daten, die erfassten spektralen Bereiche oder die Aufnahmegeometrie. Big-Data-Herausforderungen bei der Analyse dieser heterogenen Daten sind zum einen, die verschiedenen Satellitendaten zu homogenisieren, und zum anderen die große Menge an Daten, die dabei zu verarbeiten ist, effizient in angemessener Laufzeit zu verarbeiten.

Das Forschungsprojekt „GeoMultiSens: Skalierbare multisensorale Analyse von Geofernerkundungsdaten" (www.geomultisens.de), das von 2015 bis 2017 am GFZ bearbeitet wurde, befasst sich mit diesen Herausforderungen. Das Projekt wurde im Rahmen der BMBF-High-Tech-Strategie „IKT 2020 – Forschung für Innovation" gefördert. In einem interdisziplinären Konsortium mit Wissenschaftlerinnen und Wissenschaftlern aus der Satellitenfernerkundung und der Informatik/Geoinformatik wurde ein Big-Data-System entwickelt, das Nutzerinnen und Nutzer dabei unterstützt, aus der großen Menge an Satellitendaten die geeigneten zu selektieren, sie zu homogenisieren, entsprechend der Fragestellung zu analysieren und die Ergebnisse zu explorieren (*GeoMultiSens Consortium, 2018*). Damit wurden Expertisen zur Auswertung von Satellitendaten, zu Big-Data-Datei- und -Analysesystemen und zur visuellen Exploration großer Datenmengen kombiniert.
Eine zentrale Komponente ist das Homogenisierungsverfahren, das am GFZ entwickelt wurde (Abb.3). Dieses Verfahren behebt in einem ersten Schritt Inhomogenitäten bezüglich Datenformaten und Metadaten-Standards, vereinheitlicht Aufnahme- und Beleuchtungsgeometrien, quantifiziert Lageverschiebungen gegenüber einer geometrischen Referenz und führt eine Korrektur von atmosphärischen Effekten durch. Im zweiten Schritt werden sämtliche Satellitendaten hinsichtlich ihrer geometrischen und spektralen Charakteristika homogenisiert und an die Spezifikationen eines von der Nutzerin/dem Nutzer definierten Zielsensors angepasst. Zudem werden Genauigkeitsinformationen für die einzelnen Datenpunkte berechnet, die die Homogenisierungsqualität quantifizieren. Damit kann in späteren Analysen der homogenisierten Daten eine Gewichtung jedes einzelnen Datenpunkts in Abhängigkeit von dessen Genauigkeit vorgenommen werden. Um die riesigen Datenmengen, die mittlerweile von den verschiedenen Satellitensystemen erzeugt werden, effizient mit dem Homogenisierungsverfahren verarbeiten zu können, wurde das Verfahren in die parallele Analyseumgebung Apache Flink (*Apache Software Foundation, 2017*) integriert. Sämtliche Module wurden konform zum Map-Reduce-Paradigma in der Programmiersprache Python implementiert. Gesteuert und ausgeführt werden die einzelnen Rechenprozesse durch einen Controller, der in enger Zusammenarbeit mit Projektpartnern der Humboldt-Universität zu Berlin und des Konrad-Zuse-Instituts Berlin entwickelt wurde. Der Controller wird auf einem Master-Rechner ausgeführt. Er splittet große Rechenprozesse vollautomatisch in sehr viel kleinere Berechnungsschritte. Diese kleinen Berechnungsschritte werden dann auf alle zur Verfügung stehenden Rechner verteilt (Map-Phase) und die Ergebnisse in der Reduce-Phase wieder zusammengefügt. Ergebnis ist ein hochperformantes Big-Data-System zur vollautomatischen Homogenisierung von Satellitenaufnahmen. Die Laufzeit dieses Systems verhält sich linear zur Anzahl der genutzten Rechner. Im Rahmen von GeoMultiSens konnten mehrere Terabyte Daten aus den Satellitensensoren Sentinel-2A, Sentinel-2B, Landsat-5, Landsat-7 und Landsat-8 in der Ausdehnung von Zentraleuropa homogenisiert und gemeinsam ausgewertet werden. Das System steht als Open-Source-Softwarepaket zur Verfügung (*GeoMultiSens Consortium, 2018*).

Zusammenfassung und Ausblick

Die Entwicklungen in der Sensortechnologie und Computersimulation erzeugen eine immer größere Datenmenge mit einer Vielzahl unterschiedlicher Parameter. Dies führt zu Herausforderungen in der Datenverarbeitung und -analyse. Große Datenmengen müssen effizient verarbeitet und die Daten aus verschiedenen Quellen miteinander verknüpft werden. Am GFZ werden dazu neue Verfahren entwickelt und existierende Methoden an Big-Data-Konzepte und -Technologien angepasst.

Der Begriff Big Data geht jedoch weit über die Betrachtung von Daten hinaus. Er umfasst auch den wissenschaftlichen Arbeitsprozess, der für datenintensive Forschung erforderlich ist. Er besteht aus fünf Hauptschritten: 1) Schnelles Finden von Daten, 2) Integration verschiedener Daten, 3) Bewertung der Qualität der Daten, 4) Analyse der Daten und 5) gemeinsame Veröffentlichung der Ergebnisse, Daten und Analysemethoden in interaktiven Publikationen. Zukünftige Big-Data-Technologien müssen Werkzeuge und Methoden für alle Schritte zur Verfügung stellen, um den Prozess digital zu unterstützen. Damit sollen zeitaufwändige Schritte, wie z. B. das Finden geeigneter Daten und ihre Integration, beschleunigt und die Bewertung und Analyse der Daten verbessert werden. Auch sollen dadurch kontinuierliche und reproduzierbare wissenschaftliche Arbeitsprozesse möglich werden. Die Helmholtz-Gemeinschaft hat diese Herausforderung aufgegriffen und einen Inkubator „Information and Data Science" ins Leben gerufen, der die Entwicklung und Anwendung von Methoden und Technologien für datenintensive Wissenschaft zum Ziel hat. Das GFZ beteiligt sich aktiv an diesen Entwicklungen.

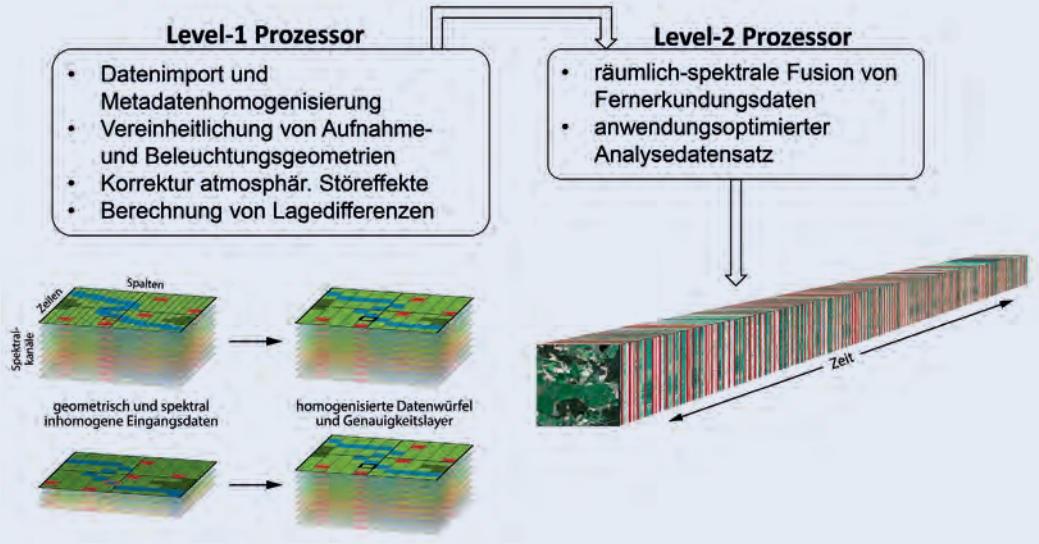

Abb. 3: Homogenisierungspipeline zur geometrischen und spektralen Vereinheitlichung von multisensoralen und multitemporalen Satellitenbilddaten

Fig. 3: Pipeline for the geometrical and spectral homogenization of multisensoral and multitemporal satellite image data

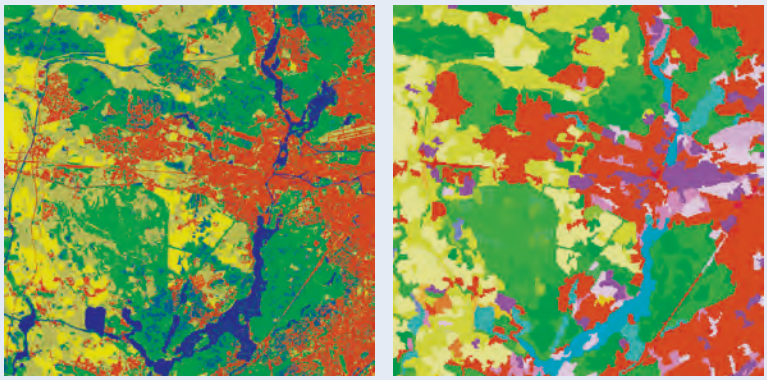

Abb. 4: Klassifikation der Landnutzung Raum Berlin-Potsdam erzeugt mit GeoMultiSens (links) und Corine (rechts)

Fig. 4: Classification of land use in Berlin-Potsdam area created by GeoMultiSens (left) and Corine (right)

Literatur

Breiman, L. (2001): Random Forests. – Machine Learning, 45, 1, pp. 5—32. DOI: https://doi.org/10.1023/A:1010933404324

Dean, J., Ghemawat, S. (2004): MapReduce: simplified data processing on large clusters. - In: OSDI'04 Proceedings of the 6th Symposium on Operating System Design and Implementation - Vol. 6, Berkeley, CA. : USENIX Association, pp. 137—150

GeoMultiSens Consortium (2018): GeoMultiSens: Scalable Analysis of Big Remote Sensing Data, verfügbar unter https://www.gfz-potsdam.de/en/section/geoinformatics/projects/geomultisens/ [letzter Zugriff: 23.04.2018]

Rawald, T. (2015): PyRQA Open-Source Paket, verfügbar unter https://pypi.org/project/PyRQA/ [letzter Zugriff: 23.04.2018]

Rawald, T., Sips, M., Marwan, N. (2017): PyRQA - Conducting Recurrence Quantification Analysis on Very Long Time Series Efficiently. - Computers and Geosciences, 104, pp. 101—108. DOI: https://doi.org/10.1016/j.cageo.2016.11.016

Rawald, T., Sips, M., Marwan, N., Dransch, D. (2014): Fast Computation of Recurrences in Long Time Series. - In: Marwan, N., Riley, M., Guiliani, A., Webber, C. (Eds.), Translational Recurrences. From Mathematical Theory to Real-World Applications, (Springer Proceedings in Mathematics and Statistics; 103), pp. 17-29

Scheffler, D., Hollstein, A., Diedrich, H., Segl, K., Hostert, P. (2017): AROSICS: An Automated and Robust Open-Source Image Co-Registration Software for Multi-Sensor Satellite Data. - Remote Sensing, 9, 7, 676. DOI: https://doi.org/10.3390/rs9070676

The Apache Software Foundation (2017): Apache Flink®: Scalable Stream and Batch Data Processing, verfügbar unter https://flink.apache.org [letzter Zugriff: 23.04.2018]

Software-Publikationen