**Dokumentation**   
Big Data Praktikum Projekt:   
Is the Globe really warming?

Philipp Lang, Pit Müller

Inhalt

[Inhalt 2](#_Toc495506535)

[Einführung 3](#_Toc495506536)

[Datenquellen 3](#_Toc495506537)

[Vorgehen und Funktionen 3](#_Toc495506538)

[Data Extraction, Cleaning and Conversion 3](#_Toc495506539)

[generate\_weather\_data.r 4](#_Toc495506540)

[weather\_data.r 4](#_Toc495506541)

[station.r 4](#_Toc495506542)

[generate\_baselines.r 4](#_Toc495506543)

[generalization.r 5](#_Toc495506544)

[initial\_tiles.r 5](#_Toc495506545)

[emission.r 6](#_Toc495506546)

[convert\_polygon\_data.r 6](#_Toc495506547)

[Data Presentation 6](#_Toc495506548)

[setup.r 6](#_Toc495506549)

[main.R 6](#_Toc495506550)

[init.r 6](#_Toc495506551)

[config.r 6](#_Toc495506552)

[lang.r 6](#_Toc495506553)

[ui.r 6](#_Toc495506554)

[server.r 7](#_Toc495506555)

[style.css 7](#_Toc495506556)

[Initiales Setup 7](#_Toc495506557)

[Bedienung Frontend 7](#_Toc495506558)

[Weltkarte 8](#_Toc495506559)

[Globale Plots 8](#_Toc495506560)

[Weitere Informationen 8](#_Toc495506561)

Einführung

Obwohl es allgemein bekannt zu sein scheint und der Großteil der Wissenschaftler dies auch bewiesen hat, wird der Klimawandel von einigen Individuen noch heute hinterfragt.

Dazu haben wir uns im Rahmen einer Big Data Projektarbeit mit dem Thema „Is the Globe really warming?“ auseinandergesetzt. Darin haben wir eine Weboberfläche entwickelt die grafisch den Verlauf einiger Wettergrößen darstellt, um die Klimaerwärmung aufzuzeigen.

Datenquellen

Der Global Historical Climatology Network (GHCN) Daily Datensatz dient dazu als Grundlage für die Wetterdaten der Anwendung. Herausgegeben werden diese Daten von den National Centers for Environmental Information (NCEI) der National Oceanic and Atmospheric Administration (NOAA) des U. S. Department of Commerce (DOC).

Für die Emissionsdaten gelten die Daten zu „Global CO2 Emissions from Fossil-Fuel Burning, cement Manufacture, and Gas Flaring: 1751-2014“ vom Carbon Dioxide Information Analysis Center (CDIAC) des U. S. Department of Energy (DOE) als Grundlage.

Vorgehen und Funktionen

Für die visuelle Darstellung der Daten über den Zeitraum in dem diese aufgenommen wurden, sollen eine interaktive Weltkarte und Plots dienen.

Die interaktive Weltkarte wird dazu in ein Raster unterteilt, das aus Kacheln der Größe 10x10 (Längengrad x Breitengrad) besteht. Auf diese Kacheln sollen die Wetterdaten der darin liegenden Stationen aggregiert werden. Als anzuzeigende Wetterdaten auf der Weltkarte wurde nur die Temperatur (TMAX), der Niederschlag (PRCP) und die Anzahl der Unwettervorkommen (WT\*\*, engl. Weather Types) ausgewählt, da für die sonstigen Wetterdaten nicht ausreichend Messungen vorhanden waren.

Die Plots werden für die globale Darstellung der Entwicklung über den gesamten Zeitraum verwendet. Hier wird der Verlauf der globalen Temperatur, Niederschlags, CO2 Emissionen und der Anzahl Unwettervorkommen dargestellt.

Data Extraction, Cleaning and Conversion

Die Rohdaten aus den oben genannten Quellen gilt es nun in ein passendes Format umzuwandeln, welches dann für die Weltkarte und die Plots verwendet werden kann.

Zuerst wurden die Datensätze aus dem GHCN-Daily Datensatz entfernt, die bei dem Flag zur Qualitätssicherung (QFLAG) einen Wert gesetzt haben, was bedeutet, dass die Messung eine Überprüfung nicht bestanden hat. Darauf wurden alle Datensätze entfernt, die wir nicht in unseren Darstellungen verwenden. Dazu wurde für die Temperatur TMAX (max. Temperatur) verwendet, da dieser Wert deutlich mehr Datensätze hat als die durchschnittliche Temperatur (TAVG). Für den Niederschlag wurden die dazugehörigen Werte für PRCP verwendet. Bei den Wettertypen wurde genauer definiert, welche Typen zu Unwetter gehören und welche nur übliche Wetterereignisse sind. Folgende Typen werden in der Anwendung als Unwetter behandelt (WT\*\*, mit \*\* gleich folgenden):

* 02 = Heavy fog or heaving freezing fog
* 03 = Thunder
* 04 = Ice pellets, sleet, snow pellets, or small hail
* 05 = Hail (may include small hail)
* 07 = Dust, volcanic ash, blowing dust, blowing sand, or blowing obstruction
* 10 = tornado, waterspout, or funnel cloud
* 11 = High or damaging winds
* 16 = Rain (my include freezing rain, drizzle, and freezing drizzle)
* 17 = Freezing rain
* 18 = Snow, snow pellets, snow grains, or ice crystals

Die letzte Veränderung am GHCN-Daily Datensatz ist die Beschränkung der berücksichtigten Daten ab 1917, da erst ab diesem Zeitraum ausreichend globale Abdeckung in den Werten besteht und aus den Anfangsjahren des angepassten Datensatzes eine sogenannte Baseline erstellt werden soll, die für spätere Jahre als Basisvergleich gelten soll.

Für jede Kachel, wie oben beschrieben, wurde ein Baseline-Jahr erstellt, das für die späteren Jahre als Vergleich gelten soll. Dazu wurde der Durchschnitt des jeweiligen Tageskachelwertes der ersten 30 Jahre des angepassten Datensatzes (1917-1946) berechnet und daraus ein Basisjahr (Baseline) generiert. Ein Tageskachelwert ist der Durchschnittswert an einem Tag einer Wettergröße aller Stationen, die in der bestimmten Kachel liegen.

Die Funktionen dazu sind in den folgenden Dateien enthalten.

generate\_weather\_data.r

Diese Datei ist allgemein zuständig für die Generierung der Wetterdaten im benötigten Format auf Basis der GHCND Daten. Es wird eine Datei für die Stationen erstellt, die alle Stationen die Kachelinformation zuweist (Einbindung station.r). Zusätzlich werden für die Kacheln die entsprechenden Baselines generiert (Einbindung generate\_baselines.r).

Zuletzt werden die Wetterdaten aller betrachteten Jahre durch eine Schleife, die alle Jahre einliest, in einer gemeinsamen Datei erstellt (Einbindung weather\_data.r).

weather\_data.r

Diese Datei bietet Funktionen zum gefilterten Auslesen der GHCND Wetterdaten und übernimmt damit das oben beschriebene Cleaning der GHCND Daten.

read\_weather\_baseline: Hilfsfunktion zum Einlesen und Zurückgeben einer temporären Baseline Datei als Spark Table.

read\_weather\_data\_org\_with\_tile\_id: Liest eine Wetterdaten Datei nach GHCND Format ein und filtert nur die nötigen Daten für diese Anwendung heraus (TMAX, PRCP, WT\*\*), sowie das Entfernen der Elemente mit gesetzten QFLAGs und gibt die gefilterten Daten zurück.

station.r

Durch diese Datei werden Funktionen zum Auslesen der Stationsdaten der GHCND Daten zur Verfügung gestellt.

read\_stations\_org: Liest die Stationsdatei und gibt die Daten gefiltert in einer Spark Table zurück (ID, Latitude und Longitude).

generate\_tiled\_stations\_table: Nutzt die eingelesene Datei und fügt jeder Station Informationen zur zugehörigen Kachel hinzu. Neben der Tile\_Id werden noch der Längengrad und Breitengrad des Kachelzentrums angegeben, welche später für die Darstellung der Kreise auf der Weltkarte wichtig sind. Zurückgegeben wird die angepasste Stationstabelle.

generate\_baselines.r

Diese Datei ist verantwortlich für die Generierung der Baselines.

generate\_tiled\_weather\_baseline: Generiert für alle initial\_tiles eine Baseline über die ersten year\_span\_baseline Jahre des Datensatzes und gibt die Baseline in einer Spark Table zurück und schreibt eine Baseline CSV Datei.

write\_filtered\_data: In dieser Funktion wird durch die Jahre für die Baseline iteriert und die Ergebnisse in einer temporären Datei abgespeichert. Dabei wird die Filterung entsprechend der initial\_tiles durchgeführt und das Aggregieren der Stationswerte zu den Kachelwerten vorgenommen.

generalization.r

Diese Datei ist umfasst mehrere Funktion zur Generalisierung der Wetterdaten von Stations- auf Kachelebene sowie von Tages- auf Jahresebene.

generalize\_by\_time: Das Aggregieren der Werte von Tagen, die in einem Jahr liegen, findet über diese Funktion statt. Dabei werden für die verschiedenen Werte die Durchschnitte gebildet.

generalize\_to\_list\_of\_time\_segments: Die Wetterdaten werden in drei Zeiträume/Tabellen unterteilt (ganzes Jahr, sommerliches Halbjahr, winterliches Halbjahr). Auf alle drei Zeiträume wird dir Funktion generalize\_by\_time() angewendet. Die drei Tabellen werden als Liste zurückgegeben.

generalize\_from\_stations\_to\_tiles: Das Aggregieren der Werte von Stationen, die in einer Kachel liegen, findet über diese Funktion statt. Dabei werden für TMAX und PRCP die Durchschnitte der Werte gebildet. Bei den Unwettern gibt es keine Werte, sondern die Anzahl der Unwetter wird durch jeden Datensatz dargestellt. Diese wurden für jede Kachel aggregiert und durch die gesamte Anzahl der Stationen in der Kachel dividiert, um einen normalisierten Wert zu bilden. Eng aneinander liegende Stationen führen damit nicht zu einem Anstieg der Unwetteranzahl.

generalize\_from\_stations\_to\_tiles\_and\_calc\_baseline\_differences: Diese Funktion ist sehr ähnlich zu generalize\_from\_stations\_to\_tiles(). Sie berechnet außerdem die Differenz der Messwerte zur Baseline. Auf Grund eines Bugs konnten beide Aufgaben nicht auf zwei Funktion aufgeteilt und generalize\_from\_stations\_to\_tiles() wiederverwendet werden.

initial\_tiles.r

Diese Datei ist eine Hilfsdatei für die Generierung der Baselines. Dadurch wird überprüft, ob für ein bestimmtes Gebiet überhaupt ausreichend Daten vorhanden sind, um dafür eine Baseline zu generieren. Hierfür war es wichtig, dass das Startjahr der Betrachtung nicht zu früh ist, da sonst kaum ein Gebiet abgedeckt werden kann.

read\_tiles\_initial: Hilfsfunktion zum Einlesen einer temporären tiles\_initial Datei mit Rückgabe der Daten in einer Spark Table.

get\_initial\_tiles: Gibt die aktiven Kacheln für die Baselines zurück, die seit Beginn der Baseline bestanden und seitdem genügend Daten beinhalten, um eine Visualisierung der Ergebnisse zu ermöglichen (initial\_tiles).

get\_active\_tiles\_of\_year: Gibt alle aktiven Kacheln für ein Jahr zurück.

get\_active\_tiles\_of\_years: Gibt alle Kacheln zurück, die für den gesamten Zeitraum der Baseline-Jahre aktiv waren.

limit\_data\_to\_considered\_tiles: Verwirft Datensätze inaktiver Kacheln.

add\_station\_count\_per\_tile: Fügt stationsbasierten Datensätzen die Anzahl der Stationen hinzu, die sich in der zugehörigen Kachel befinden.

emission.r

Diese Datei bietet die Funktion zum Auslesen der Emissionsdaten. Diese werden reduziert auf nur einen Wert, den Total Wert des Kohlenstoffausstoßes, der dann nach Angabe der Herausgeber mit 3,667 multipliziert werden soll, um den Wert des CO2 Ausstoßes zu erhalten.

read\_co2\_global: Liest CO2 Daten aus CDIAC Daten ein und gibt die Daten in einer Spark Table zurück.

convert\_polygon\_data.r

Diese Datei dient der Erstellung von Polygonen bzw. den Kacheln für die Weltkarte und der Zuordnung der zugehörigen Daten zu der jeweiligen Kachel. Dazu werden die gefilterten Wetterdaten eingelesen und für jedes Jahr eine geojson-Datei erzeugt, die jede relevante Kachel für das Jahr enthält.

Data Presentation

Nachdem alle Daten in den passenden Formaten vorhanden sind gilt es, diese durch die Weboberfläche entsprechend zu visualisieren. Dazu wurde in R das Package Shiny verwendet.

setup.r

Datei für das initiale Setup auf einer neuen Entwicklungsumgebung. Installiert alle nötigen Packages in R und generiert die nötigen Wetterdaten auf Basis der heruntergeladenen Daten von GHCND und CDIAC.

main.R

Diese Datei ruft die Webanwendung auf.

init.r

Diese Datei dient der allgemeinen Einbindung von Packages zum Start des Programms. Diese Packages sind häufig auch schon für die Vorbereitung der Daten relevant gewesen.

init: allgemeiner Aufruf für die Folgefunktionen. install\_packages wird standardmäßig nicht ausgeführt.

install\_packages: Installiert alle nötigen Packages in R für die Anwendung.

load\_packages: Lädt alle nötigen Packages in R für die Ausführung der Anwendung.

start\_spark: Stellt die Spark-Verbindung her.

config.r

Konfigurationsdatei für die Voreinstellung der Pfade zu den Daten für die jeweilige Entwicklungsumgebung (wegen zu großer Menge nicht auf Server abgespeichert). Weitere Voreinstellungen: Startjahr, Endjahr, Startjahr Baseline, Dauer Baseline, Abdeckungsgrenze für aktive Kachel, Größe der Kacheln.

lang.r

Sprachdatei

ui.r

Frontend Datei für das Shiny Plugin in der alle Elemente die für das Frontend relevant sind, ausgezeichnet werden. Dazu wurden drei Tabs erstellt, die alle Funktionen beinhalten sollen:

Interaktive Karte: Hierin wird mit Hilfe des Leaflet Packages die interaktive Weltkarte über den gesamten Arbeitsbereich angezeigt. Daneben ist ein verschiebbares Sidepanel, das dem Nutzer erlaubt die Wettergrößen anzeigen zu lassen, die er sehen möchte und welches Jahr betrachtet werden soll.

Globale Plots: In diesem Tab werden die Plots zu den verschiedenen Wettergrößen über die gesamte Zeitspanne des angepassten Datensatzes dargestellt. Dazu gibt es ein verschiebbares Sidepanel, das den Nutzer auswählen lässt, welche Daten angezeigt werden sollen.

Weitere Informationen: Dieser Tab dient ausschließlich zur Information über die Anwendung. Dadurch kann sich der Nutzer informieren, worauf diese Anwendung basiert und wie die Daten dazu eingesetzt wurden.

server.r

Backend Datei für das Shiny Plugin in der alle Funktionen, die für die Webanwendung relevant sind, enthalten sind. Darin werden die Daten aus den vorbereiteten Dateien ausgelesen und entsprechend den ausgewählten Filtern des Nutzers die auszugebenden Elemente generiert.

Die Temperatur wird aufgezeigt, indem die gesamte Fläche einer Kachel mit einer entsprechenden Farbe ausgefüllt wird, die die Temperaturdifferenz zur Baseline dieser Kachel darstellt.

Unwetter und Niederschlag wird durch 2 unterschiedlich große Kreise aufgezeigt, die im Zentrum einer Kachel liegen. Je nach Differenz zur Baseline der Kachel wird durch Farben die Veränderung dargestellt.

Die Plots wurden mit Hilfe des ggplot Packages erstellt, welche die jährlichen Durchschnittswerte als Punkte darstellen und darüber eine lokal lineare Kernel-Regressions-Kurve legen, um die Entwicklung der Werte zu veranschaulichen.

style.css

Datei mit der Layout-Datei für die Anwendung.

Initiales Setup

Da nun alle Funktionen einzeln erläutert sind, gilt dieses Kapitel nochmal als kurze Übersicht, wie man vorzugehen hat, um die Anwendung auf einem neuen Server lauffähig zu machen.

Als Erstes muss man die Pfade in der config.r-Datei entsprechend des aktuellen Servers anpassen. Darauf folgt das Ausführen der setup.r-Datei, die in RStudio alle nötigen Packages installiert.

Optional kann man hierüber die Wetterdaten in das notwendige Format konvertieren, falls diese Daten noch nicht vorhanden sind, oder überschrieben werden sollen. Dazu müssen die Rohwetterdaten jedoch auf dem Server vorhanden sein und in der config.r-Datei entsprechenden Struktur abgelegt werden. Desweiteren können die Parameter in config.r bezüglich Zeitraum, Zeitraum für Baseline, Kachelgröße und das Minimum der Datensätze, die nötig sind um eine Kachel in der Baseline zu berücksichtigen, verändert werden.

Zuletzt folgt das Ausführen der Web-Applikation über die main.R-Datei, indem diese komplett ausgeführt wird.

Bedienung Frontend

Für den Standardbenutzer sind die Daten im Backend nur wenig relevant, deshalb geht es hier darum, wie die Anwendung im Frontend bedient werden soll, um die gewünschten Ergebnisse zu erreichen.

Weltkarte

Beim Starten der Anwendung wird direkt der Tab Interaktive Karte angezeigt. Dieser beinhaltet eine Weltkarte, die standardmäßig bereits mit einigen Kacheln überlagert ist, welche die Temperaturdifferenz des letzten abgeschlossenen Jahres gegenüber der Baseline durch verschiedene Farbtöne aufzeigen. Die Werteskala wird unten rechts in einer Legende beschrieben. Geht man mit der Maus über eine Kachel so wird der zugehörige Differenzwert der Kachel angezeigt. Bei grauen Kacheln stehen nicht ausreichend Informationen aus diesem Jahr zur Verfügung. Per Mausklick kann das Zentrum der Karte verschoben werden und durch das Mausrad wird die Zoomfunktion ermöglicht.

Will man nun andere Informationen betrachten so gilt es das Hauptaugenmerk auf die Einstellungsbox (standardmäßig oben rechts) zu werfen. Als Erstes kann man dort das zu betrachtende Jahr über einen Slider auswählen. Verschiebt man diesen Werten die Elemente auf der Karte nach kurzer Berechnungszeit aktualisiert. Als Nächstes stehen drei verschiedene Wettergrößen, die angezeigt werden sollen, zur Auswahl:

* Differenz der maximalen Tagestemperaturen (°C)
* Differenz der normierten Unwetteranzahl
* Niederschlagsdifferenz (mm)

Die Differenz bezieht sich dabei immer auf das Baseline-Jahr, welches ein generiertes Durchschnittsjahr der ersten 30 Jahre der Messungen entspricht.

Wenn man eine oder mehrere Wettergrößen selektiert werden diese entsprechend auf der Weltkarte dargestellt und jeweils eine passende Legende dazu bereitgestellt.

Die letzte Auswahl die man hier treffen kann bezieht sich auf den Zeitraum der Messungen. Standardmäßig werden alle Daten für das selektierte Jahr über die Option „Ganzes Jahr“ berücksichtigt. Alternativ kann hier auch die wärmere Periode des Jahres (auf der Nordhalbkugel, 04.05 – 03.11) oder die kältere Periode (auf der Nordhalbkugel, 03.11. – 04.05.) ausgewählt werden. Dadurch werden nur die Daten für diesen Zeitraum in dem Jahr berücksichtigt.

Globale Plots

Im zweiten Tab werden die globalen Plots dargestellt, die den Verlauf der Differenzen über die gesamte Betrachtungsperiode darstellen.

Ähnlich wie bei der Weltkarte gibt es hier eine Selektion, welche Plots dargestellt werden sollen:

* Differenz der maximalen Tagestemperaturen (°C)
* CO2-Emission
* Niederschlagsdifferenz (mm)
* Differenz der normierten Unwetteranzahl

Dazu gibt es wieder die Selektion des zu betrachtenden Zeitraums analog zur der Selektion für die Weltkarte.

Im Vergleich zur Weltkarte wird hier eine globale und keine regionale Entwicklung aufgezeigt, die teilweise auch auf klare Trends hindeutet. Zusätzlich soll die Betrachtung der verschiedenen Plots dabei helfen, möglich Korrelationen zwischen den verschiedenen Werten ausmachen zu können.

Weitere Informationen

Dieser Tab stellt Informationen für den Nutzer dar, weshalb und wie diese Anwendung erstellt wurde. Dazu werden die Datenquellen und das Vorgehen genauer erläutert.