Beleg

Visualisierung von Daten mittels D3

Programmierung von Benutzeroberflächen

Ali Abdin, S. Tejan Sandi-Gahun, Helge Thiessen

Betreuender Hochschullehrer: Prof. Georg Freitag
(fertig) eingereicht am 03.02.2019

1. Inhaltsverzeichnis

2. Vorbemerkung

Insofern Sie gespannt sind, was wir (unbeabsichtigt) herausgefunden haben, oder Sie nur wenig Zeit haben und die Ausführungen zur Darstellung und den Datenkonstrukten eher redundant sind, so überspringen Sie diese bitte und werfen Sie gleich einen Blick in Abschnitt **6. Befund**.

3. Aufgabenbeschreibung

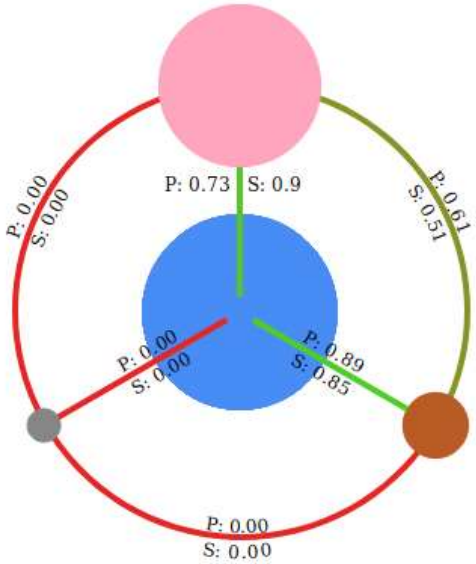
Die Aufgabe bestand in der Visualisierung eines von der Gesellschaft für Technische Visualistik mbH bereitgestellten Datensatzes. Dieses Datensatz wiederum soll von der WHO stammen. Die Einzeldaten beziehen sich jeweils auf ein Land und ein Jahr.
Unser Team sollte Daten zeitstrahlbasiert darstellen. Von den zugelosten 6 Datenwerten wurden dafür von uns die folgenden 4 ausgewählt:

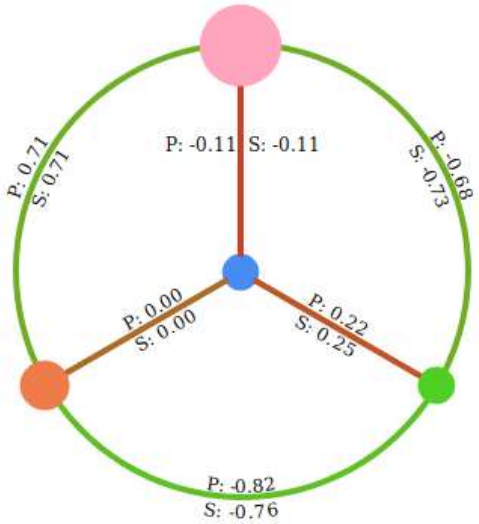
- Bruttoinlandsprodukt pro Einwohner
- Durchschnittliche Proteinaufnahme pro Einwohner und Tag
- Anteil der Bevölkerung mit Unterernährung
- Anteil der Bevölkerung mit Fettleibigkeit

4. Darstellung

Wir entschieden uns für eine Darstellung, die zum einen den Vergleich zwischen 2 Ländern erlaubt, es zum anderen aber auch gestattet, Datensätze eines Landes untereinander und mit denen des Vergleichslandes zu korrelieren.

Hierzu werden in einen ersten Schritt die beiden zu vergleichenden Länder per Dropdown ausgewählt. Als Ergebnis gibt es eine Netzwerkdarstellung der Daten eines konkreten Jahres und der internen Korrelation.





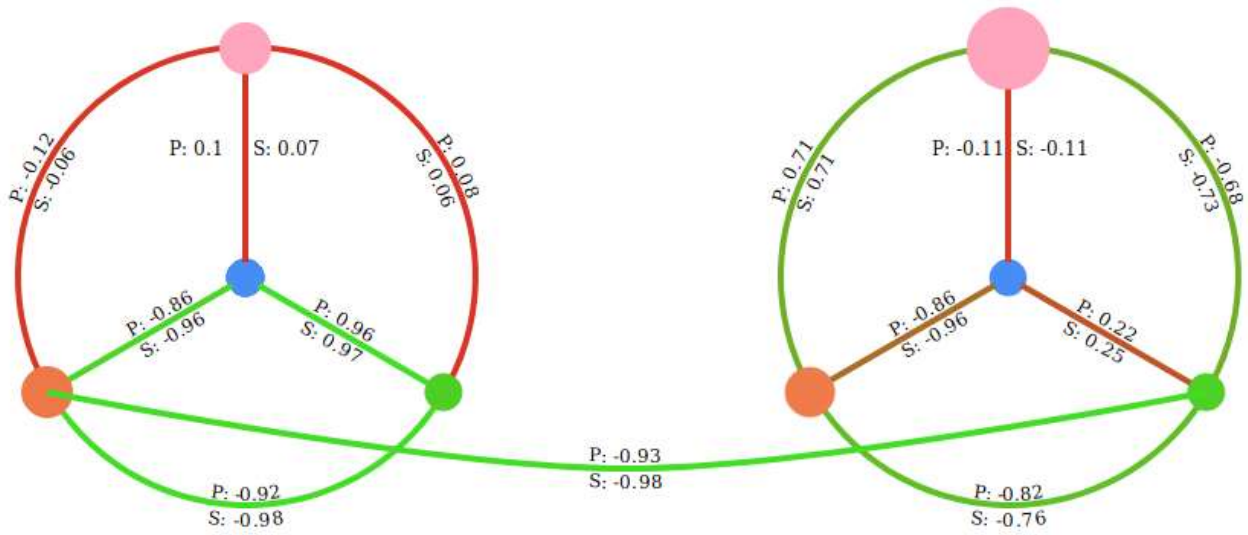
Der Kreis im Norden stellt die Proteinaufnahme dar. Seine Farbe ist unverändert fleischrosa.
Der zentrale Kreis in der Mitte das Bruttoinlandsprodukt. Seine Farbe ist unverändert blau.
Links unten findet man die Unterernährung. Die Farbe verändert sich hier von grün bei einer Unterernährung von 2,5% hin zu rot bei einer Unterernährung von 40%.
Rechts unten im Diagramm findet man die Fettleibigkeit. Ihre Farbe rangiert von grün bei einer Fettleibigkeit von 2,5% hin zu rot bei einer Fettleibigkeit von 40%.

Die Größe der Kreise skaliert mit den Einzelwerten. Konkrete Einzelwerte werden im Mouseover angezeigt.
Die Verbindungslinien stellen über Farbigkeit und die angegebenen Werte die Korrelation dar. Die Farbigkeit rangiert hier von rot bei einer Korrelation von 0 bis hin zu grün bei einer Korrelation von 1. Bei den Werten wird hinter dem P der Pearson-Korrelationskoeffizient angegeben, hinter dem S steht der Spearman-Korrelationkoeffizient.

Insofern bei einem Datensatz keine Werte hinterlegt sind, so werden die betroffenen Kreise grau ausgefüllt dargestellt. Beim Mouseover wird dann bei der Anzeige des Wertes „Value not provided“ ausgegeben.

4.1 Korrelation von Werten verschiedener Länder

Wählt man mittels Mouseklick einen Wertekreis aus Diagramm 1 und einen aus Diagramm 2 so werden diese miteinander korreliert. Dies sieht dann beispielsweise so aus:



Farbigkeit und Darstellung sind analog zur Darstellung der Korrelation im Einzeldiagramm.
Klickt man einen der verbundenen Datensätze erneut an, wird die Korrelationslinie deaktiviert.
Alternativ kann man auch auf einen der noch nicht ausgewählten Datenkreise klicken und erhält dann die zu diesem gehörende Korrelation.

4.2 Darstellung der Einzelwerte

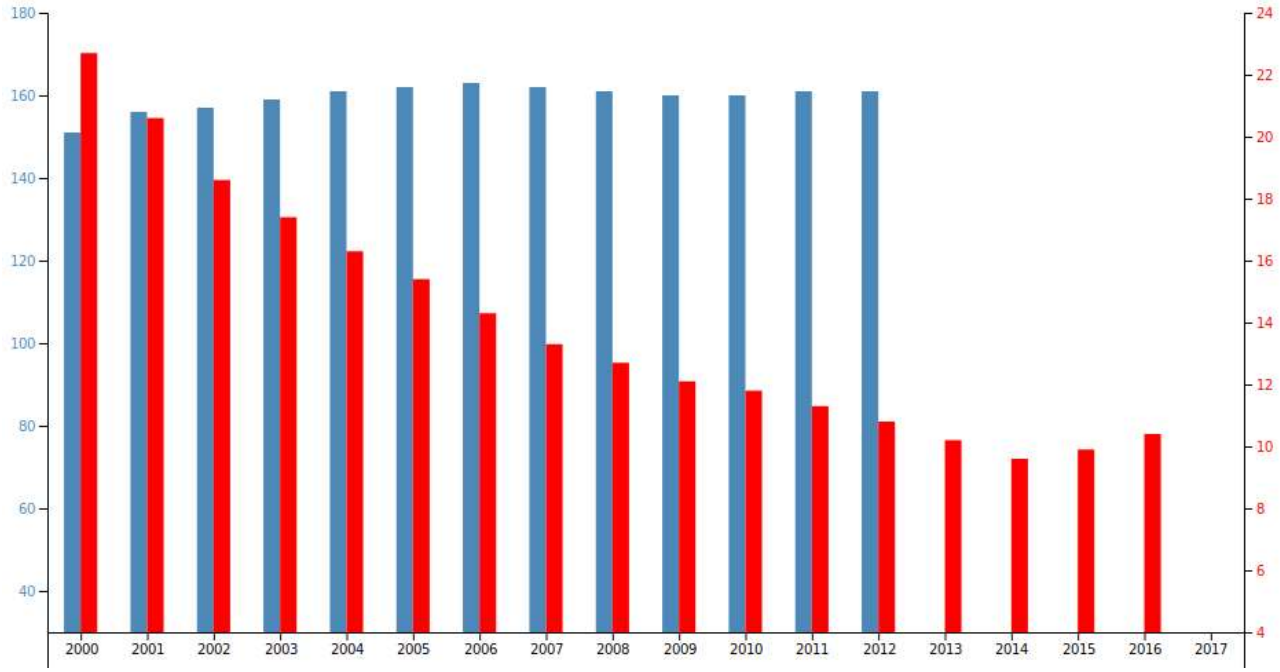
Da die Kreisdarstellung es zwar erlaubt die Korrelationswerte für einen kompletten Datensatz anzuzeigen, aber eben nur die Einzelwerte für ein einzelnes Jahr, war es notwendig eine Variante zu finden, die auch die Darstellung aller anderen Einzelwerte erlaubt. Hierfür gibt es 2 Varianten.

4.2.1 Darstellung der Einzelwerte per Zeitstrahl

Unter den Kreisdiagrammen befindet sich ein Zeitstrahl. Über ihn kann ein konkretes Jahr ausgewählt werden, dessen Einzelwerte dann angezeigt werden. Auf die Korrelationswerte, die ja immer auf allen Werten eines Datensatzes beruhen, hat dies natürlich keinen Einfluß.

4.2.2 Anzeige von Datensätzen als Zeitreihe

Genauso wie Datensätze von unterschiedlichen Ländern zur Korrelation ausgewählt werden (per Klick auf einen Datenkreis aus Diagramm 1 und einen aus Diagramm 2) werden auch Datensätze zur Darstellung als Zeitreihe ausgewählt. Dabei wird dann ein Diagramm wie das folgende dargestellt:



Die blauen Werte inklusive der dazugehörenden linken Skala gehören zum Datensatz aus Diagramm 1. Die roten inklusive der rechten Skala zum Datensatz aus Diagramm 2. Die Korrelation kann man entweder wie unter 4.1 beschrieben ablesen. Es gibt aber auch eine Extrafunktion, die sie noch einmal als Gaugeansicht darstellt.

5. Verwendete Datenkonstrukten und Hilfsfunktionen

Das Modul Dataprovider wurde um folgende Funktionen erweitert:

- getDataForCountrySelection (callback)
  - Hier werden nur die Namen der Regionen und Länder zurückgegeben. Sie werden für die Dropdownauswahl verwendet.
- getDataForTimeline (country, code, callback)
  - Eingabe sind das gewünschte Land und der Code der benötigten Datensatzes. Zurückgegeben wird ein Array mit Jahreszahlen und den dazugehörigen Werten.
- getDataForCountryAndYear (country, year, callback)
  - Eingabe sind das gewünschte Land und das Jahr. Rückgabe ist ein JSON-Objekt mit den Werten für die 4 von uns ausgewählten Datensätze für das gewählte Land und Jahr.

Auch das Modul Helper musste erweitert werden. Es enthält nun die folgenden Funktionen.

- getPearsonCoeffizient (timeline1, timeline2, callback)
  - Eingabe sind hier die Rückgabewerte von getDataForTimeline. Zurückgegeben wird der Pearsoncoeffizient für die beiden Datenreihen.
- getSpearmanCoeffizient (timeline1, timeline2, callback)
  - Dito für den Spearmancoeffizienten. Dieser ist allerdings deutlich aufwändiger in der Berechnung.
- getCoeffizientsForRegion (country, callback)
  - Hier werden für ein Land gleich alle 12 möglichen Coeffizienten berechnet und als JSON-Objekt zurückgegeben.
- getDatasetFrom2Timelines (timeline1, timeline2, callback)
  - Dies ist nur eine Datenaufbereitung für die in 4.2.2. vorgestellte Darstellung als Zeitreihe.

Die Rückgabewerte werden grundsätzlich per Callback bereitgestellt.

6. Befund

In der Präsentation hatten wir angekündigt, einige interessante Korrelationen zeigen zu wollen. Dies ließ sich so nicht aufrechterhalten.

Wenn die Werte für die Fettleibigkeit eines beliebigen Landes gegen die Werte für Fettleibigkeit eines anderen Landes korreliert werden, so ergibt sich unter der Voraussetzung, dass Werte hinterlegt sind, immer ein Korrelationskoeffizient von 1. Dies ist kein Berechnungsfehler. Es wurde mehrfach von uns überprüft.

Es ist tatsächlich so, dass die WHO (oder wer auch immer für sie diese Werte bereitstellte) hier keine individuellen Werte pro Land erhoben hat, sondern dass die Werte offenbar aus einem Satz von Basiswerten jeweils über Verwendung einer linearen Transformation entstanden sind.

Dies führt natürlich dazu, dass zwei beliebige Zeitreihen zur Fettleibigkeit unterschiedlicher Länder ebenfalls über eine lineare Funktion gekoppelt sind. Und das wiederum führt zur engen Kopplung der Daten, die durch den Korrelationswert von 1 klar belegt wird.

Es wird klar, dass die Datenwerte zur Fettleibigkeit nicht stimmen können. Es ist schlicht undenkbar, dass alle aufgeführten Länder mit Daten zur Fettleibigkeit die gleiche enge Kopplung aufweisen.

Die Fettleibigkeit in den USA korreliert direkt gegen die von Benin, die von Deutschland, die von Schweden, die der Elfenbeinküste. Etc. Das ist schlicht absurd.

Es wurde gesagt, dass wir davon ausgehen sollen, dass die bereitgestellten Daten stimmen. Das Problem ist, dass wir jetzt sicher wissen, dass dies unmöglich der Fall sein kann.

6.1 Erläuterung

Insofern es interessiert, wie wir diesen völlig unerwarteten Zusammenhang auf die Spur kamen:
Wir hatten beim Testen der Korrelation der Daten unterschiedlicher Länder die Fettleibigkeit der USA und Kanadas gegeneinander korreliert und erhielten 1. Nun gut, so etwas kann passieren und die USA und Kanada sind ja auch recht ähnlich. Wir wählten dann weitere Länder und erhielten bei Fettleibigkeit Land 1 gegen Fettleibigkeit Land 2 weiterhin immer 1. Natürlich dachten wir zuerst an einen Fehler im Code. Vielleicht hatten wir ja in Wahrheit Fettleibigkeit Land 1 gegen Fettleibigkeit Land 1 korreliert. Dann müsste ja 1 rauskommen.
Dem war aber nicht so. Es waren tatsächlich immer unterschiedliche Datensätze. Und auch die Einzelwerte bei der Coeffizientenberechnung (sumCounter, sumDenominator1, sumDenominator2) ergaben immer völlig andere Werte. Das Gesamtergebnis war aber immer wieder 1. Mal abgesehen von ein paar Fällen, wo über Rundungsfehler ein Wert von 0.99 ausgegeben wurde.