Einflussfaktoren auf die Ausbreitung von Covid-19

Niklas Wagner	Lukas Gerspers	Robin Hammer
7283479	6984850	7224832

GitHub: https://github.com/LuGee94/TeamDB

1. Ziel der Analyse & Vorgehen

Die Covid-19 Pandemie prägt die Welt wie kaum ein anderes Ereignis im 21. Jahrhundert. Vor diesem Hintergrund ist das Ziel des Projekts, Einflussfaktoren auf die Ausbreitungsgeschwindigkeit und Mortalitätsrate zu identifizieren. Für die Analyse wurde das CRISP-DM Modell verwendet.

Abb. 1: CRISP-DM Modell

2. Datenvorbereitung

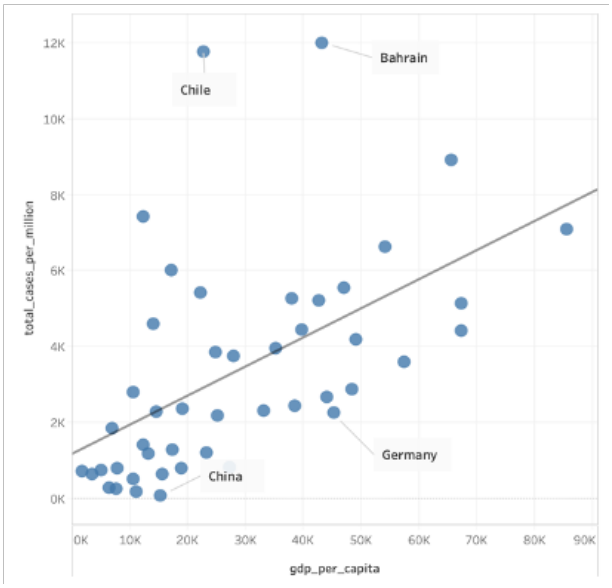
Bei der Datenbeschaffung standen aktuelle Länderdaten, sowohl geographischer als auch demographischer Natur im Vordergrund. Ebenfalls wurden Daten zur Ausbreitungs-, Mortalitäts-, Gesundheits- und Vorsorgesituation in den Ländern inkludiert.
Besondere Herausforderungen waren Null-Werte, uneinheitliche Länderbezeichnungen und der Umgang mit Zeitreihen in Verbindung mit der Gewinnung von Kennzahlen, welche den Stand der Infektionen zu einem vergleichbaren Stichtag abbilden. Es ergab sich ein vereinheitlichter Datensatz aus 173 Ländern und 68 Variablen.[1]

3. Regression

Die ursprüngliche Idee des Projekts war es, nur mit verschiedenen Regressions-Algorithmen Einflussfaktoren zu ermitteln. Konkret wurde versucht, sowohl die Anzahl der Fälle als auch der Tode pro Millionen Einwohner "vorherzusagen". Als Stichtag wurden verschiedene Variablen vorbereitet, darunter die Anzahl an Tagen zwischen der 100. und 1000. Infektion

[1] R-Skript (current_coronadata)

des jeweiligen Landes. Jedoch wurde schnell klar, dass die Ergebnisse der Regressionen je nach Betrachtung stark variieren und sich insbesondere nicht als Kausalität interpretieren lassen. Besonders gut illustriert dies das Beispiel des BIPs pro Kopf, denn nur diese Kennzahl wies einen signifikanten Einfluss bei dem Großteil der Analysen auf. Entgegen der Erwartungen steigt mit dem BIP allerdings die Anzahl der Fälle und Tode:



Abb, 2: Einfluss des BIP pro Kopf auf die Anzahl der Fälle pro Mio. Einwohner

Die Analyse ergibt zwar eine klare Korrelation, es handelt es sich hierbei jedoch wahrscheinlich um eine Scheinkausalität. Vielmehr lässt sich die These aufstellen, dass weitere unbekannte Faktoren, wie z.B. Dunkelziffer und mangelnde Testkapazitäten eine Rolle spielen. Der starke Einfluss des BIPs und die damit verbundene Erkenntnis, dass Infektionsdaten von reichen Ländern kaum vergleichbar mit Zahlen von ärmeren Ländern sind, führte zur Entscheidung, das BIP zusammen mit Corona-Entwicklungskennzahlen zu clustern, um besser vergleichbare Ländergruppen zu erhalten.

4. Clustering

Innerhalb des Projektes wurden drei Arten von Clustering-Algorithmen verwendet - partitionierend, dichtebasiert und hierarchisch. Der Datensatz setzt sich aus dem BIP pro Kopf, Kennzahlen zur Ausbreitungsgeschwindigkeit und aus Infektionen und Toden pro Millionen Einwohner zum Stichtag 20.06.2020 zusammen.[2] Des Weiteren wurden die Daten um Ausreißer bereinigt und zur besseren Vergleichbarkeit normalisiert. Ebenfalls wurden nur Länder betrachtet welche eine Infektionsanzahl größer als 20.000 aufwiesen.

4.1. Partitionierender Algorithmus - k-Means

Um das Ergebnis der k-Means-Clustering-Analyse stabil zu halten und ein möglichst geringes "residual sum of squares" (RSS) zu erzielen, wurde für $nstart$ und $iter.max$ der Wert 25 gewählt.

[2] R-Skript (cluster_20000_bip)

Die Elbow-Knick Methode ($k \in [2:9]$) auf Basis des RSS ergab eine optimale Clusterzahl von 3. Um Vergleichbarkeit mit der hierarchischen Clusteranalyse zu gewährleisten, wurden jedoch 4 Cluster gewählt. Die folgende Abbildung stellt das Ergebnis dar:

Abb. 3: K-*means Algorithmus mit k=4*

4.2. Dichtebasierter Algorithmus - DBSCAN

Die Vorgehensweise des DBSCAN bestand in der Anwendung des Algorithmus mit den folgenden Parametern:

$$\epsilon \in [1:1.5], \qquad minPoints \in [2:5], \qquad borderPoints = enabled, \qquad weigths = NULL$$

Ergebnis der Analyse war, dass viele Länder als „Noise-Points“ bei zu geringem ϵ keinem Cluster zugeordnet wurden und bei größer werdendem ϵ ein Cluster dominierte. Einen wesentlichen Einfluss auf die mangelhaften Ergebnisse hat die geringe Größe der Stichprobe, gepaart mit einer hohen Varianz innerhalb der Daten. Deshalb wurde der DB-Scan zur weiteren Analyse nicht weiterverwendet.

4.3. Hierarchischer Algorithmus - Ward

Auf Basis der normalisierten Daten wurde eine Distanzmatrix mittels “euklidischer Distanz” berechnet. Anschließend wurde der agglomerative Ansatz des hierarchischen Clusterings mit der Ward-Methode angewendet. Analog zum k-Means Algorithmus resultierten aus den Ausreißern Katar, Singapur, Kuwait und Bahrain eigene Cluster. Daher wurden sie ebenfalls herausgefiltert. Aus dem Clustering ohne Ausreißer ergab sich folgendes Dendrogramm:

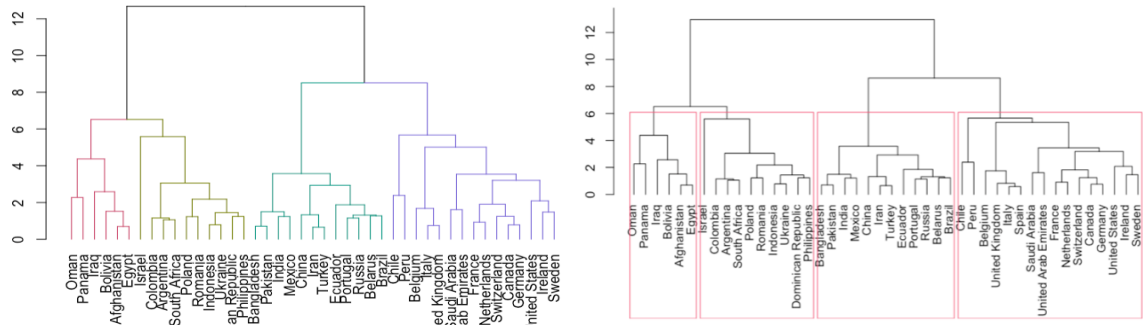


Abb. 4: Ergebnisse des hierarchischen Clusterings

Somit bildeten sich die 4 Cluster, welche als Grundlage für den Vergleich mit den Ergebnissen des k-Means Algorithmus dienten.

4.4. Beschreibung der Cluster

Interessanterweise liefern beiden Clusterverfahren 4 nahezu identische Cluster. (Ausnahme: Panama, Peru und Deutschland). Diese Cluster sehen folgendermaßen aus:

	Clusterzuordnung			
	1	2	3	4
Median days_10000_to_20000_infections	9	10	24	15
Median days_1000_to_10000_infections	14	16	38	38
Median days_100_to_1000_infections	10	10	13	23
Median Death rate	8,3%	3,7%	2,8%	3,2%
Median total_cases_per_million	4.424	2.297	1.183	707
Median total_deaths_per_million	355	82	35	21
Median gdp_per_capita	46.949	16.238	14.601	10.550

Cluster 1
Farbe hclust: lila
Farbe kmeans: lila

Beschreibung:

Die reichsten mit den meisten Toden pro Kopf, überdurchschnittlich vielen Fällen und überdurchschnittlicher Ausbreitungsgeschwindigkeit

Cluster 2
Farbe hclust: gelb
Farbe kmeans: rot

Beschreibung:

Eher ärmere Länder mit eher wenigen Fällen, gemischt vielen Toden, aber besonders hoher Ausbreitungsgeschwindigkeit

Cluster 3
Farbe hclust: rot
Farbe kmeans: türkis

Beschreibung:

Die langsamste Ausbreitungsgeschwindigkeit bei eher wenigen Fällen und Toden und eher ärmeren Ländern

Cluster 4
Farbe hclust: türkis
Farbe kmeans: grün

Beschreibung:

Enthält nur 6 Länder mit Werten, die für einen gemäßigten Coronaverlauf sprechen. Alleinstellungsmerkmal dieses Clusters ist eine besonders lange Dauer von der 100. bis zur 1000. Infektion

Abb. 5: Differenzierte Betrachtung der 4 Cluster

5. Analyse der Cluster

Ausgehend von den Ergebnissen der Regression folgt die Analyse der Cluster. Eine Erkenntnis stellt der oftmals inverse Verlauf mehrerer Variablen dar. Beispielsweise sinkt die Anzahl der Fälle im reichen Cluster bei Ländern mit höherer Lebenserwartung, während in anderen Clustern die Fallzahl steigt. Dieses Verhalten scheint die These aus Kapitel 3 zu bestätigen, dass in wohlhabenden Ländern die Fallzahl so hoch sind, weil die technischen Möglichkeiten eine Diagnose erleichtern. Im Folgenden wird eines der Erkenntnisse genauer erläutert – der Einfluss von Krankenhausbetten auf die Anzahl der Tode.

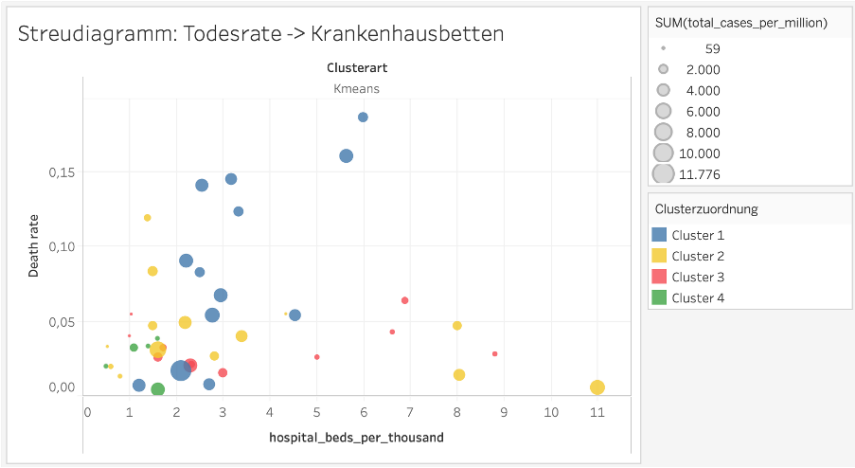


Abb. 6: Streudiagramm

Die Länder mit den meisten Krankenhausbetten wurden nicht dem "reichen" Cluster zugeordnet (Abb. 6). Gleichzeitig haben die Länder mit den meisten Betten wenige Tote pro Millionen Einwohner - auch bei Ländern mit hoher Fallzahl wie bspw. Deutschland. Somit konnten starke Indizien gefunden werden, dass die Anzahl der Krankenhausbetten einen Einfluss auf die Anzahl der Tode sowie die Todesrate hat. Zwischen einem direkten kausalen Effekt (Betten verhindern Tode) und möglichen Begleiteffekten („Gesunde" Länder haben ohnehin mehr Betten) lässt sich ohne individuelle Recherche nicht trennen. Für Begleiteffekte sprechen eine starke Korrelation zwischen Krankenhausbetten und der von Anzahl von Ärzten und Pflegern sowie den Gesundheitsausgaben.

6. Zusammenfassung, Limitierungen und Ausblick

Ausgehend von der Regression konnten verschiedene Einflussfaktoren auf die Corona-Entwicklung gefunden werden. Die Clustering-Analyse ermöglichte anschließend eine bessere Vergleichbarkeit der Ländergruppen in Bezug auf signifikante Einflussfaktoren. Klare Kausalitäten ließen sich aufgrund der komplexen Datenlage nicht identifizieren. Insbesondere Dunkelziffern und unterschlagene (Todes-)Fälle erschweren die Erkenntnis, ob ein höherer Fallstand tatsächlich eine stärkere Verbreitung bedeutet, oder lediglich eine bessere Testabdeckung. Somit sind die Ergebnisse dieser Arbeit ambivalent zu betrachten und bedürfen im Einzelfall genauere Analysen. Die Untersuchung könnte durch eine weitreichende, individuelle Recherche in einzelnen Ländern verbessert werden, um die Validität der Zusammenhänge besser bewerten zu können.

Abschließend bleibt zu sagen, dass globale Vergleiche nur mit einer hohen Datenqualität zu aussagekräftigen Ergebnissen führen. Die genaue Betrachtung regionaler Studien kann daher wertvoller sein, um den Effekt einzelner Maßnahmen genauer analysieren zu können.