**Einflussfaktoren auf die Ausbreitung von Covid-19**

|  |  |  |
| --- | --- | --- |
| Niklas Wagner  7283479 | Lukas Gerspers  6984850 | Robin Hammer  7224832 |

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 Modell CRISP-DM Modell verwendet.

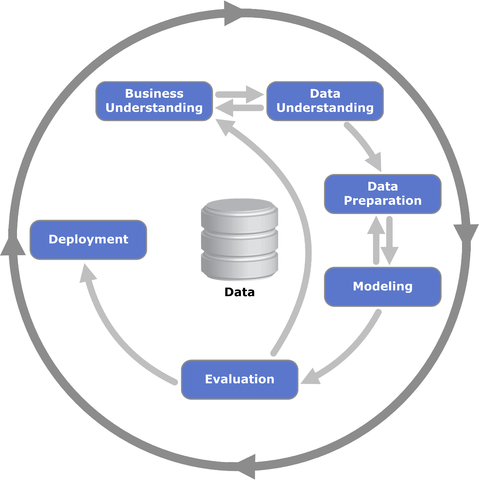


Abb. 1: CRISP-DM

1. 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]](#footnote-2)

1. 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, bspw. die Anzahl an Tagen zwischen der 100. und 1000. Infektion 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.

Anders als logischerweise erwartbar, steigt mit dem BIP allerdings die Anzahl der Fälle und Tode:

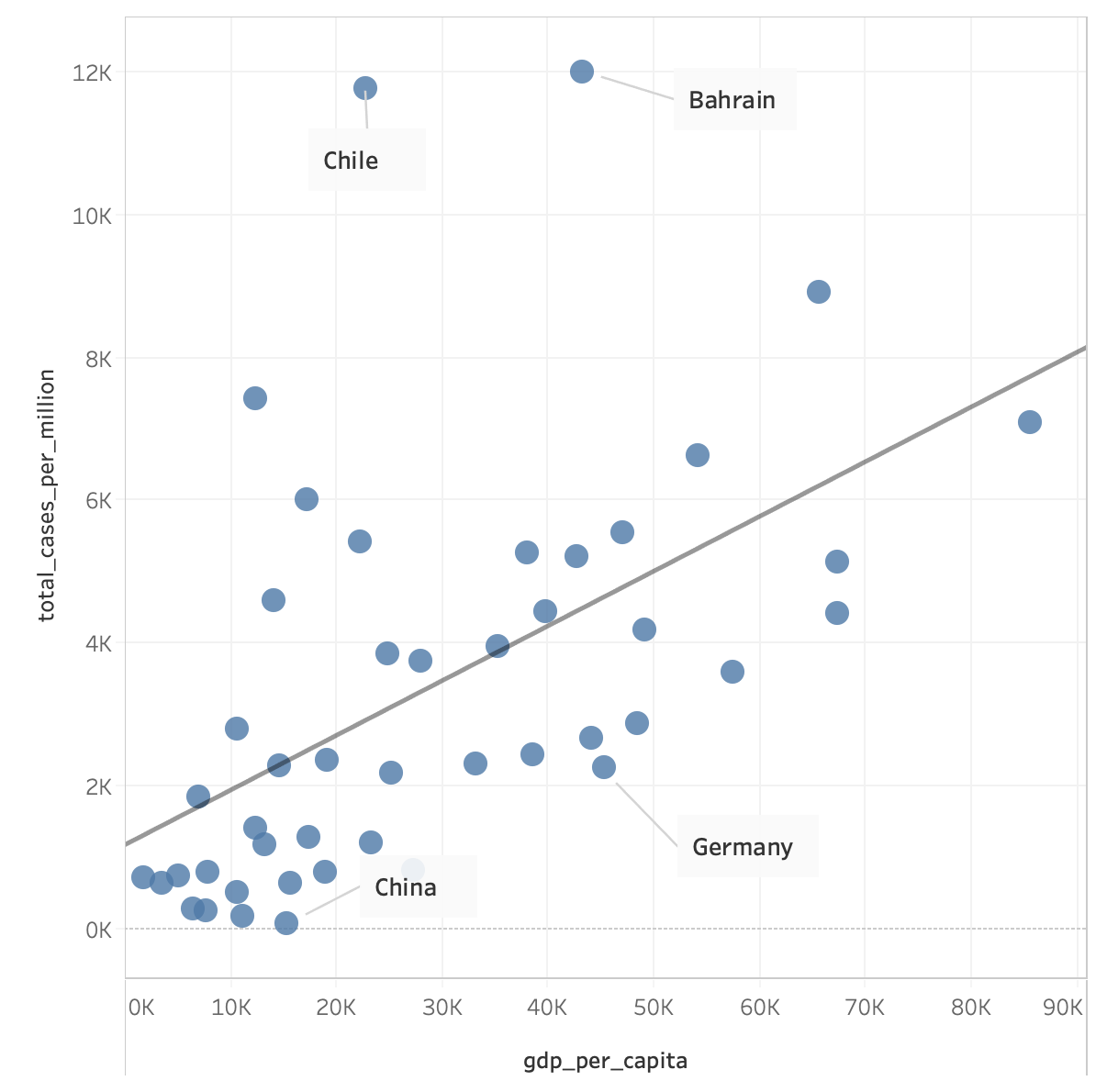


Abb. 2: Einfluss des BIP pro Kopf auf die Anzahl der Fälle pro Mio. Einwohner

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.

1. Clustering

Innerhalb des Projektes wurden drei unterschiedliche Arten von Clustering-Algorithmen verwendet - partitionierend, dichtebasiert und hierarchisch. Der Datensatz setzt sich aus dem BIP pro Kopf, Kennzahlen zur Ausbreitungsgeschwindigkeit und aus Infektionen & Toden pro Millionen Einwohner zum Stichtag 20.06.2020 zusammen. [[2]](#footnote-3) Des Weiteren wurden die Daten um Ausreißer bereinigt und zur besseren Vergleichbarkeit normalisiert. Ebenfalls wurden nur Länder mit einer Fallanzahl größer 20.000 betrachtet.

## 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 und der Wert 25 gewählt.

Die Elbow-Knick Methode () 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:

Ein Bild, das Text, Karte enthält.

Automatisch generierte Beschreibung

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

## Dichtebasierter Algorithmus - DBSCAN

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

Ergebnis der Analyse war, dass viele Länder als noise-Points bei zu geringem keinem Cluster zugeordnet wurden und bei größer werdendem 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 nicht weiter berücksichtigt.

## Hierarchischer Algorithmus

Auf Basis der normalisierten Daten wurde eine Distanzmatrix mit “euklidischer Distanz” erstellt. Anschließend wurde der agglomerative Ansatz des hierarchischen Clusterings mit der Ward.D2 Methode gewählt. Analog zum k-Means Algorithmus resultierten aus Ausreißern wie Qatar, Singapore, Kuwait und Bahrain separate Cluster. Aus dem Clustering ohne Ausreißer ergab sich folgendes Dendrogram:

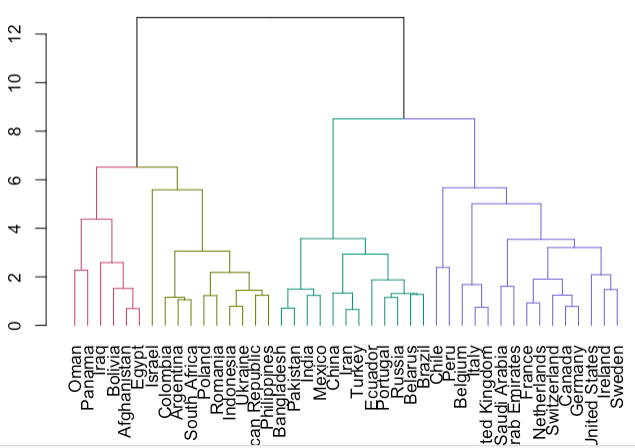
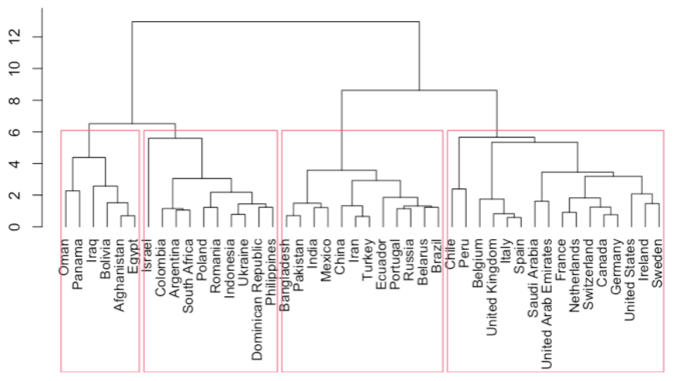
 

Abb. 4: Ergebnisse des hierarchischen Clusterings

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

## Beschreibung der Cluster

Interessanterweise liefern die beiden Clusterverfahren Kmeans und Hierarchisches Clustern 4 nahezu identische Cluster. (Ausnahme: Panama, Peru und Deutschland)

Ein Bild, das Screenshot enthält.

Automatisch generierte Beschreibung

Abb. 5: Differenzierte Betrachtung der 4 Cluster

1. 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.

Die Länder mit den meisten Krankenhausbetten wurden nicht dem “reichen” Cluster zugeordnet. (siehe Abb. 6) Gleichzeitig haben die Länder mit den meisten Betten wenige Tote pro Millionen Einwohner - auch bei Ländern mit hoher Fallzahl wie Deutschland und Belarus. 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.

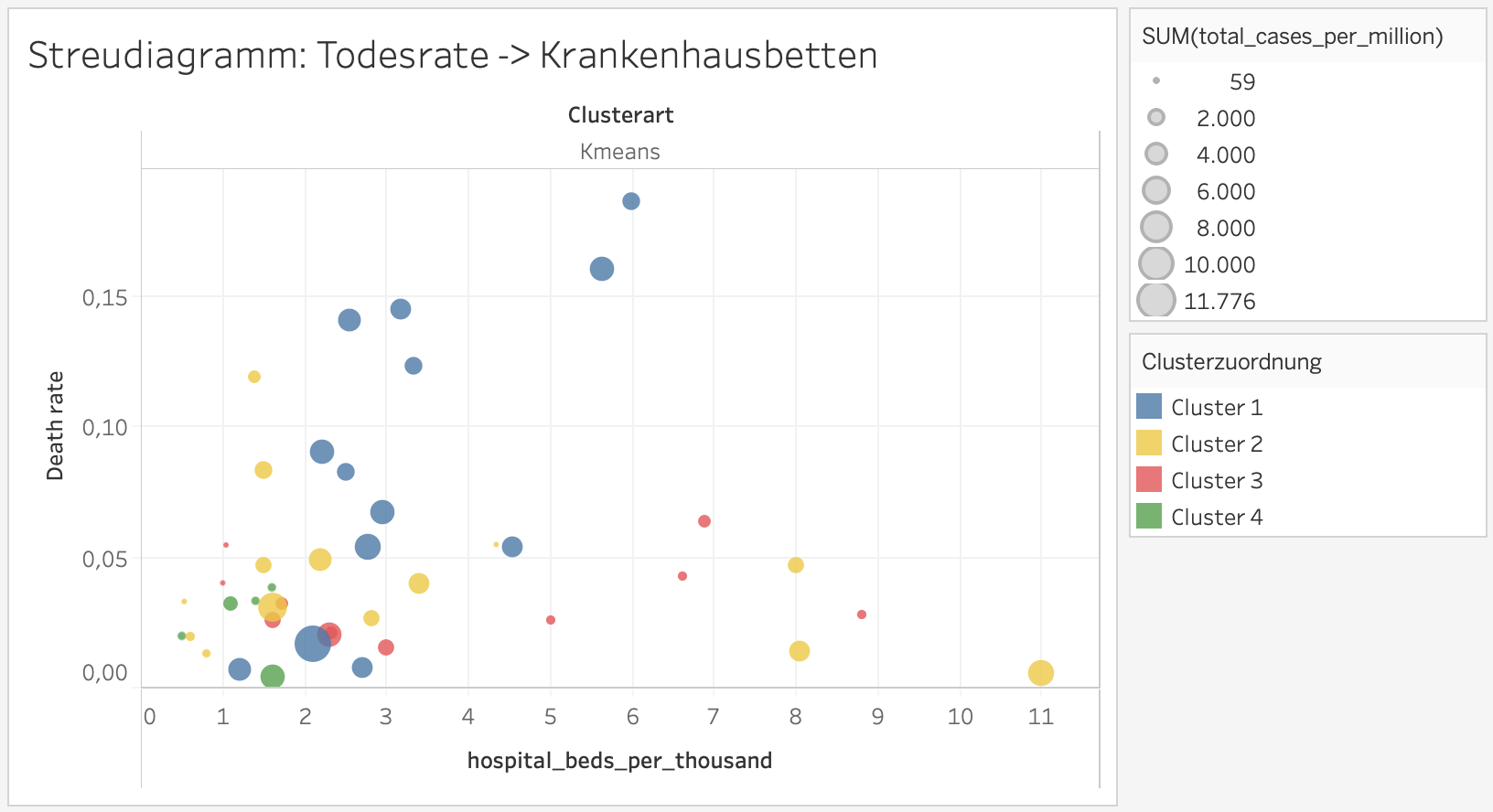


Abb. 6: Streudiagramm

1. Zusammenfassung, Limitierungen und Ausblick

Ausgehend von der Regression konnten verschiedene Einflussfaktoren für 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.

1. siehe R-Skript (current\_coronadata) [↑](#footnote-ref-2)
2. R-Skript: cluster\_20000\_bip [↑](#footnote-ref-3)