**Analyse signifikanter Einflussfaktoren auf die Ausbreitung der Covid-19 Pandemie zur Mustererkennung auf Basis von Ausbreitungs- und Mortalitätsdaten**

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

# 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 Vorgehen des CRISP-DM Modell verwendet.

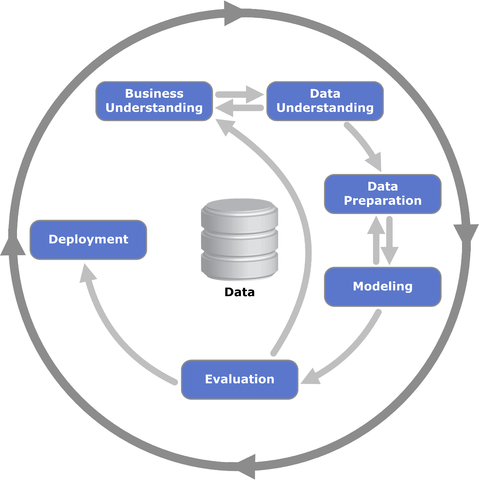


Abbildung 1: CRISP-DM Modell

# Datenvorbereitung

Bei der Datenbeschaffung standen aktuelle Länderdaten, sowohl geographischer als auch demografischer 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 mit 173 Objekten (Ländern) und 68 Variablen.[[1]](#footnote-1)

# Regression

Die ursprüngliche Idee des Projekts war es, nur mit verschiedenen Regressions-Algorithmen Einflussfaktoren zu ermitteln. Getestet wurden verschiedene abhängige Variablen. 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 man logischerweise erwarten würde, steigt mit dem BIP allerdings die Anzahl der Fälle und Tode (siehe Abbildung 2).

A close up of a map

Description automatically generated

Abbildung 2: Einfluss von BIP pro Kopf auf die Anzahl Fälle und Tode

Die Analyse ergibt zwar eine klare Korrelation, aus logischer Sicht 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.

# Clustering

Innerhalb des Projektes wurden drei unterschiedliche Arten von Clustering-Algorithmen verwendet - partitionierend, dichtebasiert und hierarchisch. Resultierend aus der Regressionsanalyse wurden als Datenbasis für alle Clusteranalysen die Corona-Daten der jeweiligen Länder zum Stichtag 20.06.2020 verwendet.[[2]](#footnote-2) Der Datensatz setzt sich aus den pro Kopf Einkommen, Daten zur Ausbreitungsgeschwindigkeit und aus Infektionen & Tode pro Millionen Einwohner zusammen. Des Weiteren wurden die Daten um Ausreißer bereinigt und zur besseren Vergleichbarkeit der einzelnen Merkmale normalisiert. Ebenfalls wurden nur Länder mit einer Fallanzahl größer 20.000 ausgewählt.

## Partionierender Algorithmus - K-Means

Das erste zum Clustering verwendete Verfahren ist der K-Means Algorithmus. Um das Ergebnis der 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 hat ergeben, dass 3 Cluster für die Analyse verwendet werden sollten. Aufgrund der späteren hierarchischen Clusteranalyse wurden jedoch 4 Cluster gewählt, um die Vergleichbarkeit zwischen den Analysen zu ermöglichen. Die folgende Abbildung stellt das Ergebnis der Clusteranalyse mit k means dar.

Ein Bild, das Text, Karte enthält.

Automatisch generierte Beschreibung

Abbildung 3: Ergebnis des k-means Algorithmus

## 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. Daher wurden die Ergebnisse für die weitere Analyse nicht berücksichtigt.

## Hierarchischer Algorithmus -

Auf Basis der normalisierten Daten wurde eine Distanzmatrix mit der Methode der “euklidischen 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. Demnach wurden diese für die globale Betrachtung aus der Analyse ausgeschlossen. Aus der Durchführung des Clusterings ergab sich folgendes Dendrogram:

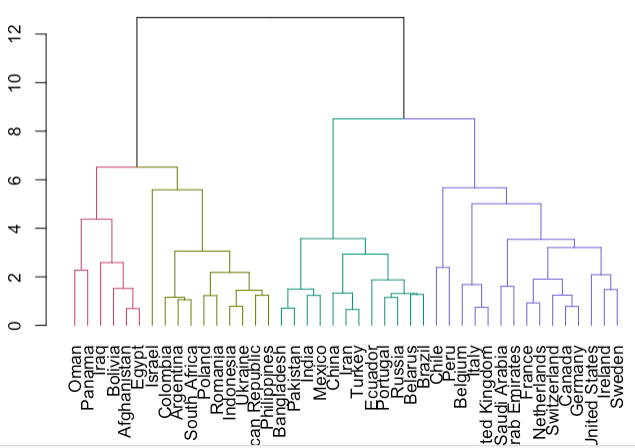
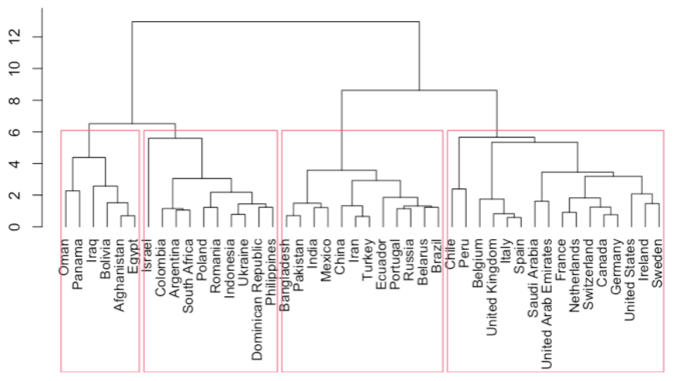
 

Abbildung 4: Ergebnisse des hierarchischen Clusterings

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

Abbildung 5: Differenzierte Betrachtung der 4 Cluster

# Analyse der Cluster

Ausgehend von den Ergebnissen der vorangegangenen Regression folgte 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 besonders in wohlhabenden Ländern die Fallzahl so hoch sind, weil die technischen Möglichkeiten eine Infizierungs- 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 Abbildung 6) Gleichzeitig haben die Länder mit den meisten Betten ziemlich wenige Tote pro Millionen Einwohner. Somit konnten starke Indizien gefunden werden, dass die Anzahl der Krankenhausbetten einen Einfluss auf die Anzahl der Tode sowie die Todesrate hat. Zum einen ist als direkter kausaler Effekt denkbar, dass eine ausreichende Bettenkapazität eine gute Behandlung ermöglicht. Zum anderen suggeriert eine hohe Bettenkapazität ein ausgereiftes Gesundheitssystem mit besseren Möglichkeiten der Behandlung und Versorgung. Beide Effekte lassen sich mit den vorhandenen Daten nicht belegen, da es hierfür einer Recherche auf Länderebene bedarf. Dies verdeutlicht auch die starke Korrelation zwischen Krankenhausbetten und Anzahl von Ärzten, Pflegern sowie die geringere Korrelation zu den Gesundheitsausgaben pro Land.

A close up of a piece of paper

Description automatically generated

Abbildung 6: Einfluss der Anzahl von Krankenhausbetten auf die Fälle & Tode

# Zusammenfassung 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 höhere 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 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 sehr 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-1)
2. siehe R-Skript (cluster\_20000\_bip) [↑](#footnote-ref-2)