Hochschule Anhalt
Anhalt University of Applied Sciences

COVID-19: Prognosen auf kommunaler Ebene

Konzept

Niklas Tscheuschner B. Sc. Volkswirtschaftslehre
5025035

Amer Goli B. Sc. Angewandte Informatik
4064312

Fachbereich: Informatik und Sprachen
Dozentin: Prof. Dr Groß
Abgabe: 22.04.2021

Motivation / Analyseideen

Treibend für die Gefahr von Pandemien ist deren hohes Maß an Unsicherheit. Mit Vorhersagemodelle lässt sich Unsicherheit reduzieren sowie die politische Entscheidungsfindung, Ressourcenallokation und Krankheitsprävention unterstützen.

Autokratisch regierte Länder tendieren zu erfolgreicheren Strategien in der Pandemiebekämpfung (vgl. Sorci et. al, 2020). Dessen Natur ist jedoch nicht zwangsläufig auf die Regierungsform zurückzuführen (vgl. Cassan & Van Steenvoort, 2020). Gerade in modernen Demokratien kann eine Chance darin liegen, die Bevölkerung mit in die Pandemiebekämpfung einzubeziehen. Dies setzt jedoch voraus, dass Informationen und Methoden öffentlich zugänglich sind.

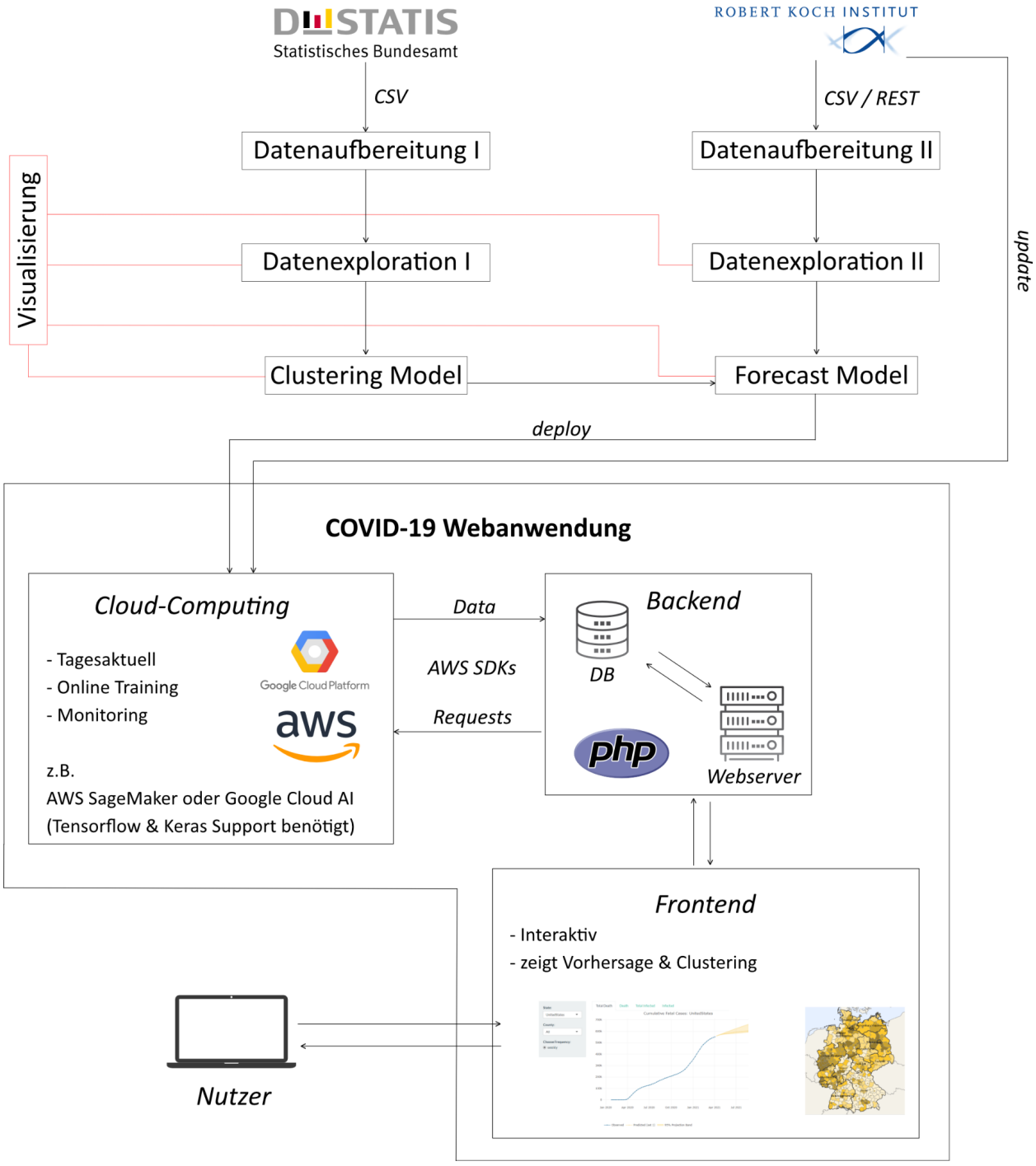
In diesem Projekt und im Rahmen des Moduls „Projekt Data Science “, soll ein Vorhersagemodell für COVID-19 auf kommunaler Ebene entwickelt werden. Wir planen so alle Arbeitsschritte des generischen Data Science Workflows abdecken. Neben den historischen und aktuellen Fallzahlen des *Robert Koch Instituts* sollen Daten des *Statistischen Bundesamtes* zur Landkreisstruktur verwendet werden. Einen ähnlichen Ansatz zur Prognose wählten Wang et. al. (2020) auf Bundestaatenebene in den USA. In Deutschland sind uns keine Ansätze auf regionaler Ebene bekannt. Die Ergebnisse des Projekts sollen als Webanwendung frei zugänglich gemacht werden.

Daten

Das *Robert Koch Institut* veröffentlicht seit März 2020 täglich Statistiken zum Infektionsgeschehen in Deutschland. Diese werden von den kommunalen Gesundheitsämtern übermittelt. Neben den Fallzahlen, werden Daten zur Sterblichkeit/Genesung, zum Alter und zum Geschlecht erfasst. Durch die feine Gliederung (1.695.186 Zeilen x 18 Spalten) ist der Datensatz ca. 250 MB groß. Die Daten stehen unter der ‘Open Data Datenlizenz Deutschland – Namensnennung – Version 2.0’ zur Verfügung.

Im *Regionalatlas* des *Statistischen Bundesamtes* werden demografische und sozioökonomische Statistiken auf kommunaler Ebene gesammelt. Für die Analyse interessante Themenbereiche sind z.B. Bevölkerung, Gebiet und Fläche, Verdienste und Einkommen, Bruttoinlandsprodukt und Bruttowertschöpfung. Momentan sind die meisten Daten für das Jahr 2019 aktualisiert. Gelegentlich muss auf die Daten der Jahre 2018 bzw. 2017 zurückgegriffen werden. Eine Verknüpfung erscheint uns durch die geringe zeitliche Volatilität problemfrei. Die Daten stehen ebenfalls unter der ‘Open Data Datenlizenz Deutschland – Namensnennung – Version 2.0’ zur Verfügung.

Methoden und Implementierung (Teil 1)



Methoden und Implementierung (Teil 2)

Wir planen einen Großteil des Projektes in *Python* umzusetzen. Die Datensätze des *Statistischen Bundesamtes* und des *Robert Koch Instituts* können als CSV-Datei bezogen werden. Letzteres verfügt ebenfalls über eine Programmierschnittstelle, die für die Webanwendung von Bedeutung sein wird.

Pandas als Bibliothek bietet sich generell für die Datenauf- und -nachbereitung an. Hinsichtlich der Größe des Datensatzes muss jedoch überlegt werden, ob die gewünschte Performanz initial mit *Pandas* erreicht werden kann, da diese nur *in-memory* Analysen unterstützt. Gegebenenfalls muss der Datensatz vorher komprimiert werden.

Das Clustering der Landkreise setzt Kenntnisse über die Struktur der Daten / Attribute voraus. Diese sollen in der explorativen Analyse gesammelt werden. So kann die Menge anwendbarer Clusteringmethoden frühzeitig eingegrenzt werden. *Scikit-learn* als Bibliothek zum maschinellen Lernen deckt hier die gängigsten Methoden ab.

Parallel soll die Aufbereitung und Evaluierung des RKI-Datensatzes erfolgen. Hier liegt der Fokus auf der Zeitreihenanalyse. Neben Trends, Saisonalität und Stationarität sollen auch eventuell vorhandene Brüche durch Kontaktbeschränkungen identifiziert werden. Mithilfe der gewonnenen Kenntnisse können potenzielle Vorhersagemodelle gewählt werden. Zum jetzigen Zeitpunkt erscheinen *ARMA*-Modelle und *Machine Learning* Modelle auf Basis Neuronaler Netze (*RNNs, CNNs, Autoencoder, GANs*) als ein plausibler Ansatzpunkt. *Tensorflow* und *Keras* sind für Letztere geeignet. *Statsmodels* für statistische Ansätze. Der Vorhersagezeitraum ist momentan mit einer Woche angedacht.

Geplant ist zudem, den Einfluss der Landkreisstruktur auf die Fallzahlenentwicklung sowie deren Clustering im Vorhersagemodell zu berücksichtigen. Gemäß dem gängigen *Machine Learning* Workflow soll trainiert, validiert und getestet werden. Die Testphase kann zu Teilen an den aktuellen Fallzahlen durchgeführt werden und soll die Modelle benchmarken.

Das fertig trainierte Modell soll dann als Webanwendung öffentlich zugänglich sein. Cloud-Anbieter wie *Amazon Web Services* oder *Google Cloud Platform* machen dies mit minimalem Aufwand möglich. Herausfordernd wird der Aspekt, dass das Modell tagesaktuell Prognosen liefern soll. Die Vorhersagen des Vortages können mit den aktuellen Fallzahlen vom *Robert Koch Institut* verglichen werden. So wird das Modell *Online* ständig weiter trainiert.

Die Nutzer können die Prognosen über eine Webanwendung für einen interaktiv gewählten Landkreis auswählen. Über die *SDK* des Cloudanbieters erhält das Backend die Prognosedaten. *AWS* ermöglicht dies beispielsweise für *PHP*. Die Visualisierung im Frontend soll auf *Wordpress* basieren und über *Godaddy* gehostet werden.

Zeitplan	Wer?	Wann?
Datenaufbereitung I	Amer	April / Mai
Datenexploration I	Amer	April / Mai
Clustering Model	Amer & Niklas	Mai
Datenaufbereitung II	Niklas	April / Mai
Datenexploration II	Niklas	April / Mai
Forecast Model	Niklas	Mai / Juni
Cloud Computing / Backend	Amer & Niklas	Juni / Juli
Frontend	Amer	Mai/ Juni / Juli
Visualisierung	Amer & Niklas	fortlaufend
Präsentation	Amer & Niklas	Juli

Kommunikation

Wir planen uns mindestens einmal wöchentlich für einen Austausch auf Skype zu treffen. Dort besprechen wir unsere Fortschritte und eventuelle Probleme. Prinzipiell soll die Ausarbeitung lokal stattfinden. Um das Projekt aber während des Entwicklung zu sichern, werden wir regelmäßig auf *Gitlab* hochladen.

Quellen

Cassan, Guilhem & Steenvoort, Milan. (2021). Political Regime and COVID 19 death rate: efficient, biasing or simply different autocracies?

Li Wang, Guannan Wang, Lei Gao, Xinyi Li, Shan Yu, Myungjin Kim, Yueying Wang, & Zhiling Gu. (2020). Spatiotemporal Dynamics, Nowcasting and Forecasting of COVID-19 in the United States.

Sorci, Gabriele & Faivre, Bruno & Morand, Serge. (2020). Why Does COVID-19 Case Fatality Rate Vary Among Countries?. SSRN Electronic Journal. 10.2139/ssrn.3576892.

Daten abrufbar unter:

Robert Koch Institut:
https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74

Statistisches Bundesamt:
https://www-genesis.destatis.de/gis/genView?GenMLURL=https://www-genesis.destatis.de/regatlas/AI003-2.xml&CONTEXT=REGATLAS01