

IFT870 - Projet de session

Louis-Vincent CAPELLI (CAPL1101)

Tom SARTORI (SART0701)

Alexandre THEISSE (THEA1804)

I. INTRODUCTION

A. Titre du projet

Analyse de la propagation du COVID-19 en fonction de la position géographique, de la part de population urbaine et du climat.

B. Motivations

La pandémie de COVID-19 a eu un impact majeur sur le monde entier. Les gouvernements ont dû prendre des mesures drastiques pour limiter la propagation du virus. Cependant, la propagation du virus a été très inégale d'un pays à l'autre. L'objectif de ce projet est d'analyser les facteurs qui ont influencé la propagation du virus.

Une étude [4] a déjà été réalisée sur la relation entre la propagation du virus et la latitude mais de nombreuses critiques ont mentionné qu'elle ne prenait pas assez de facteurs en compte.

Notre objectif est de réaliser une étude plus complète en prenant en compte en plus de la position, l'urbanisation du pays et son climat comme cela a été suggéré dans les commentaires de l'étude précédente.

II. DONNÉES

A. Description du projet

Le projet consiste à mettre en perspective les données de propagation du COVID-19 avec des données géographiques, climatiques et démographiques. Le but est dans un premier temps de retrouver les résultats de l'étude précédente [4] et dans un second temps de comparer ces résultats en prenant en compte des facteurs supplémentaires pour voir si la position est bien un facteur déterminant dans la propagation du virus ou si elle est simplement corrélée à d'autres facteurs plus importants.

B. Contexte général

Le COVID-19 est une maladie infectieuse causée par le coronavirus SARS-CoV-2. Elle a été découverte en Chine en décembre 2019 et s'est rapidement propagée dans le monde entier. Le virus se transmet par les gouttelettes respiratoires et les surfaces contaminées. Les symptômes les plus courants sont la fièvre, la toux et la fatigue. La maladie peut être grave et entraîner la mort, en particulier chez les personnes âgées ou souffrant de problèmes de santé sous-jacents.

On sait que la mortalité d'un virus est fortement liée aux conditions climatiques, la température par exemple pouvant être un facteur aggravant chez les personnes les plus fragiles notamment. [7]

De plus, la densité de population est un facteur important dans la propagation d'une maladie infectieuse puisqu'elle augmente les contacts entre les individus et donc la probabilité de transmission du virus. C'est d'ailleurs un des facteurs à prendre en compte dans le calcul du R_0 , le nombre de reproduction de base du virus. [11] [8]

Ces observations nous amènent à nous poser la question de savoir si la position géographique est vraiment une variable qui permettrait de prédire la propagation du virus ou si elle n'est qu'un agrégat de variables plus importantes comme la densité de population ou le climat.

C. Source des données

Pour réaliser notre étude, une longue phase de collecte de données a été nécessaire. Voici les différentes sources de données que nous avons utilisées :

- **Population urbaine** : données de la Banque Mondiale [3]
- **Population totale** : données de la Banque Mondiale [2]

- **Position géographique** : données de Tadas Talaikis [9]
- **Température moyenne** : données de Wikipedia [13]
- **Cas confirmés de COVID-19** : données de Kaggle [10]
- **Classification climatique de Köppen-Geiger** : données de l'Université de Vienne [6]
- **Taux de mortalité** : données de Kaggle [5]

D. Traitement des données

1) *Sélection des pays étudiés*: Nous avons commencé par sélectionner un sous-ensemble de pays pour lesquels nous avons toutes les données nécessaires, ce qui nous a donné un total de 159 pays.

Les pays ayant plusieurs dénominations dans les différentes bases de données, nous avons utilisé la librairie `difflib` avec un seuil de similarité de 0.8 pour matcher les noms des pays et ainsi pouvoir lier les données des différentes bases.

2) *Position géographique*: La position géographique de chaque pays étant disponible dans plusieurs bases de données, nous avons conservé les données de la base de Tadas Talaikis [9] qui représentent la position moyenne des pays en latitude et longitude.

3) *Climat*: Pour refléter le climat global d'un pays, nous avons conservé deux indicateurs : la température moyenne et la classification climatique de Köppen-Geiger.

La température moyenne a été extraite de Wikipedia [13]. La classification climatique de Köppen-Geiger est fournie par l'Université de Vienne [6] pour un peu plus de 90000 points répartis sur la surface du globe. Nous avons calculé la distance entre chaque pays et les points de la base de données grâce à la formule de Haversine [12] et avons sélectionné le point le plus proche pour chaque pays.

4) *Population*: Nous avons utilisé les données de la Banque Mondiale de population totale [2] et de population urbaine [3] de chaque pays pour être en mesure de calculer la part de population urbaine. Ceci servira à estimer l'urbanisation de chaque pays.

Nous avons conservé les données de 2020 seulement pour être en phase avec les données de COVID-19.

5) *COVID-19*: Les données de cas confirmés et de taux de mortalité de COVID-19 ont été extraites de l'étude précédente [4] [5] et de Worldometers [1].

Le nombre de cas confirmés est cumulatif et apparaît sous la forme d'une série temporelle avec un point par jour. Nous avons regroupé les données des USA et du reste du monde et lorsque plusieurs régions étaient disponibles pour un pays, nous les avons sommées.

Le taux de mortalité a été calculé en divisant le nombre de morts par le nombre de cas confirmés et en le multipliant par 100 pour obtenir un pourcentage de mortalité pour chaque pays.

E. Description des données finales

Nous avons donc obtenu un jeu de données avec les caractéristiques suivantes pour chaque pays :

- Country : nom du pays (string)
- Latitude : latitude moyenne du pays (float)
- Longitude : longitude moyenne du pays (float)
- Urban Population : population urbaine du pays (int)
- Total Population : population totale du pays (int)
- Mortality Rate : taux de mortalité du pays (float)
- Mean temperature : température moyenne du pays (float)
- Climate : classification climatique de Köppen-Geiger du pays (string)
- 1/22/20 - 9/23/20 : nombre de cas confirmés de COVID-19 pour chaque jour (int)

F. Informations sur les données

TODO : Ajouter des insights (moyennes, médianes, etc.) sur les données et des graphiques pour les illustrer.

G. Problématique

La propagation du COVID-19 est-elle intrinsèquement liée à la position géographique des pays ou est-elle plutôt liée à d'autres facteurs comme la densité de population ou le climat ?

H. Historique des travaux et développements

Comme évoqué précédemment, une étude [4] a déjà été réalisée sur la relation entre la propagation du virus et la latitude. Cette étude utilisait comme indicateur de la propagation le taux de mortalité. Cependant, de nombreuses critiques ont mentionné que cette étude ne prenait pas assez de facteurs en compte et donc que les résultats obtenus n'étaient pas suffisants pour conclure que la position géographique était un facteur déterminant dans la propagation du virus.

III. ALGORITHMES

TODO : Pas bien compris cette partie, le nettoyage des données a déjà été expliqué il y aura peu d'algorithmes à expliquer ici, plutôt des analyses statistiques.

A. Calcul de la propagation du virus

TODO : Trouver comment calculer le spread factor

B. Applications concrètes

TODO : Meilleure gestion des pandémies, meilleure compréhension des facteurs influant sur la propagation du virus.

C. Résultats et interprétation

TODO : spread/position, spread/urban population, spread/climate et comparaison entre spread et mortality rate

D. Limites

Nous avons été limités par la qualité des données disponibles. En effet, certaines données ne sont disponibles que par pays et d'autres pour des points du globe. Nous avons donc dû faire des approximations pour pouvoir les utiliser. Ce point pose notamment problème pour la classification climatique de Köppen-Geiger qui est disponible pour des points du globe et peut différer d'une région à l'autre d'un même pays. Également, les données de COVID-19 ne sont disponibles qu'à l'échelle des pays ou de grandes régions, ce qui ne nous permet pas de faire des analyses plus fines quant aux différences d'urbanisation au sein des pays.

E. Conclusion

TODO : Répondre à la problématique, donner des pistes pour des études futures.

REFERENCES

- [1] Mortality rate. <https://www.worldometers.info/coronavirus/>.
- [2] World Bank. Total population. <https://data.worldbank.org/indicator/SP.POP.TOTL>.
- [3] World Bank. Urban population. <https://data.worldbank.org/indicator/SP.URB.TOTL>.
- [4] Paul Mooney. Does latitude impact the spread of covid-19? <https://www.kaggle.com/code/paultimothymooney/does-latitude-impact-the-spread-of-covid-19>.
- [5] Paul Timothy Mooney. Mortality rate. <https://www.kaggle.com/datasets/paultimothymooney/coronavirus-covid19-mortality-rate-by-country>.
- [6] University of Vienna. Climate classification. <https://koeppen-geiger.vu-wien.ac.at/data/Koeppen-Geiger-ASCII.zip>.
- [7] Samuel A Sarkodie and Peter A Owusu. Climate and covid-19 pandemic: effect of humidity and temperature on the outbreak in selected nordic countries. *Science of the Total Environment*, 737:140640, 2020.
- [8] Katelyn T L Sy, Laura F White, and Brooke E Nichols. Population density and basic reproductive number of covid-19 across united states counties. *PLoS One*, 16(4):e0249271, 2021.
- [9] Tadas Talaikis. Country position. https://gist.github.com/tadast/8827699#file-countries_codes_and_coordinates-csv.
- [10] Johns Hopkins University. Covid confirmed cases. <https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset/versions/143>.
- [11] Wikipedia. Basic reproduction number. https://en.wikipedia.org/wiki/Basic_reproduction_number.
- [12] Wikipedia. Haversine formula. https://en.wikipedia.org/wiki/Haversine_formula.
- [13] Wikipedia. Mean temperature. https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature.