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 suggeré 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 temperature 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 calcule du R0, 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 avions 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.

À chaque point est associée une classe selon trois critères : le climat principal, les précipitations et la température. Voici la répartition des classes de Köppen-Geiger sur la surface du globe. (fig. 1)

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.

Fig. 1. Classification climatique de Köppen-Geiger

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

A. Calcul de la propagation du virus

Nous avons à notre disposition des séries temporelles de cas confirmés de COVID-19 pour chaque pays, il nous a fallu en extraire un indicateur de propagation du virus sous la forme d'un unique nombre par pays.

1) Troncature des courbes: En observant les séries temporelles et en les comparant aux manières dont a été gérée la pandémie dans les différents pays, on constate que les mesures prises par les gouvernements (port du masque, confinement, campagnes de vaccination, etc.) ont eu un impact sur la propagation du virus.

Afin d'extraire un indicateur de propagation comparable entre les pays, nous avons essayé de tronquer les courbes de cas confirmés pour ne considérer que les périodes où le virus se propageait de manière naturelle, c'est-à-dire avant que les mesures gouvernementales n'aient un impact significatif.

Pour ce faire nous avons d'abord lissé les courbes en utilisant une moyenne mobile car les données sont très bruitées et que pour cette partie de l'analyse, nous ne cherchons pas à être précis mais plutôt à capturer les grandes tendances. (fig. 2)

Fig. 2. Cas confirmés et dérivée, sans et avec moyenne mobile

Nous avons ensuite essayé de tronquer les courbes avec la bibliothèque Python `ruptures` qui permet de détecter des points singuliers d'une courbe mais nous ne sommes pas parvenus à trouver des paramètres qui nous donnaient des résultats satisfaisants sur l'ensemble des pays.

Pour détecter les changements de tendance dans la propagation du virus, nous avons calculé la dérivée des courbes de cas confirmés. La dérivée d'une série temporelle représente la variation entre chaque point de données et le point précédent. Nous avons repéré le point où la dérivée était maximale, ce qui correspond au pic de propagation du virus. Ensuite, en remontant dans le temps à partir de ce point, nous avons tronqué la courbe aux endroits où la dérivée passait en dessous d'un premier seuil fixé à 5%. Nous avons procédé de la même manière en avançant dans le temps à partir du pic, en tronquant la courbe lorsque la dérivée descendait sous un second seuil de 70%. (fig. 3)

Les seuils ont été choisis de manière empirique

en observant les résultats produits sur un échantillon de pays avec des courbes variées. (fig. 8)

Fig. 3. Exemple de troncature, cas de la France

2) Calcul du facteur de propagation: Il a ensuite fallu extraire des portions de courbes conservées un indicateur de propagation du virus pour chaque pays. Nous avons testé quatre méthodes différentes :

- Régression linéaire
- Régression exponentielle
- Temps nécessaire pour doubler le nombre de cas
- Ratio de reproduction quotidien

Régression linéaire :

Nous avons ajusté une droite de la forme $y = ax+b$ sur les données de cas confirmés en utilisant la bibliothèque Python `scikit-learn` et avons pris le coefficient a comme indicateur de propagation.

Fig. 4. Régression linéaire

Comme on peut le voir sur la figure 4, la régression linéaire ne semble pas être un bon modèle lorsqu'il s'agit de modéliser la propagation d'un virus. Dans le cas de certains pays, la droite est proche de la courbe, mais pour la pluaprt ce n'est pas le cas et le coefficient directeur est sensible à la population totale du pays ce qui n'est pas souhaitable.

Régression exponentielle :

Nous avons ajusté une courbe de la forme $y = e^{ax+b}$ sur les données de cas confirmés en utilisant la bibliothèque Python `scipy` et avons pris le coefficient a comme indicateur de propagation.

Nous nous sommes assurés de translater les courbes tronquées pour que le premier point de la série aient une ordonnée de 0 avant de faire la régression, et nous avons ensuite translaté la courbe résultant de la régression pour qu'elle corresponde à la courbe originale.

Fig. 5. Régression exponentielle

Comme on peut le voir sur la figure 5, la régression exponentielle semble donner des résultats satisfaisants sur différents types de courbes.

Nous avons appliqué un min-max scaling sur les coefficients obtenus pour chaque pays afin de les ramener entre 0 et 1 et de pouvoir les comparer.

Temps moyen requis pour doubler le nombre de cas :

Une autre manière de calculer le taux de propagation est de calculer le temps qu'il faut pour que le nombre de cas confirmés double. Cette méthode est robuste à la forme de la courbe et facile à calculer mais puisque nous avons tronqué les séries temporelles à un certain point, l'initialisation ne sera pas la même pour chaque pays même après translation. Certains pays pourraient déjà avoir gagné un certain élan au début de la série temporelle tronquée et cela ferait que le temps pour doubler les cas serait plus petit qu'il ne devrait l'être.

Pour éviter cela, nous avons décidé de ne prendre en compte le temps pour doubler les cas qu'à partir d'une certaine portion de la série temporelle en espérant que le taux de propagation se soit stabilisé.

Ainsi, nous avons choisi de ne considérer que les derniers 75% de la série temporelle tronquée.

Fig. 6. Temps pour doubler les cas

Comme on peut le voir sur la figure 6, notre manipulation a permis de stabiliser le temps pour doubler les cas pour chaque pays et de minimiser l'impact de l'initialisation.

Pour ramener cet indicateur entre 0 et 1, nous avons utilisé la formule suivante :

$$v' = \frac{v_{max} - v}{v_{max}}$$

où v_{max} est le temps pour doubler les cas le plus grand parmi tous les pays.

Ratio de reproduction quotidien :

Cette méthode consiste à calculer l'augmentation relative moyenne du nombre de cas confirmés par jour sur la période considérée. C'est une méthode simple et intuitive qui semble fonctionner correctement. Nous avons, comme pour l'indicateur précédent, décidé de ne considérer que les derniers 75% de la série temporelle tronquée.

La formule utilisée est la suivante :

$$v = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{c_{i+1} - c_i}{c_i}$$

où c_i est le nombre de cas confirmés au jour i et n est le nombre de jours considérés.

Voici les valeurs que l'on obtient pour quelques pays :

Pays	Ratio de reproduction quotidien
France	0.052
United Kingdom	0.042
United States	0.015
China	0.098
Russia	0.031
Ethiopia	0.035
Afghanistan	0.052

À nouveau, un min-max scaling a été appliqué pour ramener les valeurs entre 0 et 1.

Comparaison des méthodes :

Afin de comparer les différentes méthodes, nous avons affiché les positions de quelques pays pour chacune d'entre elles. (fig. 7)

On constate notamment que le temps moyen pour doubler les cas présente un classement assez différent des deux autres méthodes.

En calculant les corrélations deux à deux entre les différents indicateurs (Régression exponentielle (exp), Temps pour doubler les cas (T2C) et Ratio de reproduction quotidien (repr)), on obtient les valeurs

Fig. 7. Comparaison des méthodes de calcul de la propagation

corr	exp	T2D	repr
exp	1.0	0.43	0.78
T2C	0.43	1.0	0.61
repr	0.78	0.61	1.0

suivantes qui corroborent les observations faites sur les graphiques :

Après avoir comparé les résultats de la régression exponentielle et du ratio de reproduction quotidien entre eux et avec les courbes, nous avons décidé de poursuivre notre étude en utilisant le coefficient de le ratio de reproduction quotidien comme indicateur de propagation car il semble être moins sensibles à la forme des courbes et donne des résultats plus cohérents dans l'ensemble.

B. Résultats et interprétation

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

C. Applications concrètes

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

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.

IV. ANNEXES

Fig. 8. Troncature des courbes de cas confirmés