Regard sur les données de l'épidémie de Covid-19

adjectif.net/spip/spip.php

PDF

Pour citer cet article :

Khaneboubi, Mehdi (2020). Regard sur les données de l'épidémie de Covid-19. *Revue Adjectif, 2020 T1.* Mis en ligne mardi 24 mars 2020 [En ligne] http://www.adjectif.net/spip/spip.php?article524

Résumé :

Cet article à caractère méthodologique présente un cas concret d'analyse de données publiques relativement à la pandémie de coronavirus.

Mots clés :

Statistiques, Analyse de données

Grâce aux données publiées sur le web, il est possible d'analyser le nombre de malades et de victimes du coronavirus dans le monde. C'est une analyse simple, mais qui permet prendre une mesure du phénomène. C'est aussi l'occasion de comprendre comment les statistiques et l'analyse de données permettent de prédire et de prévoir.

Nous allons nous fonder sur deux types de données disponibles en ligne. La première est celle du Centre européen de prévention et contrôle des maladies1 qui est une agence de l'Union européenne. La seconde base de données que nous allons utiliser, est celle du John Hopkins Institute qui est une université américaine située dans la ville de Baltimore. Cette institution a produit la carte qui circule beaucoup2 dans la presse et sur les réseaux sociaux. Les données ne sont pas identiques, ni mises à jour de la même façon. C'est tout à fait normal, celles du *John Hopkins institute* sont produites à partir de plusieurs sources dont l'Union européenne et l'Organisation mondiale de la santé. En revanche, si les valeurs ne sont pas exactement les mêmes, les ordres de grandeur sont identiques.

Vue générale

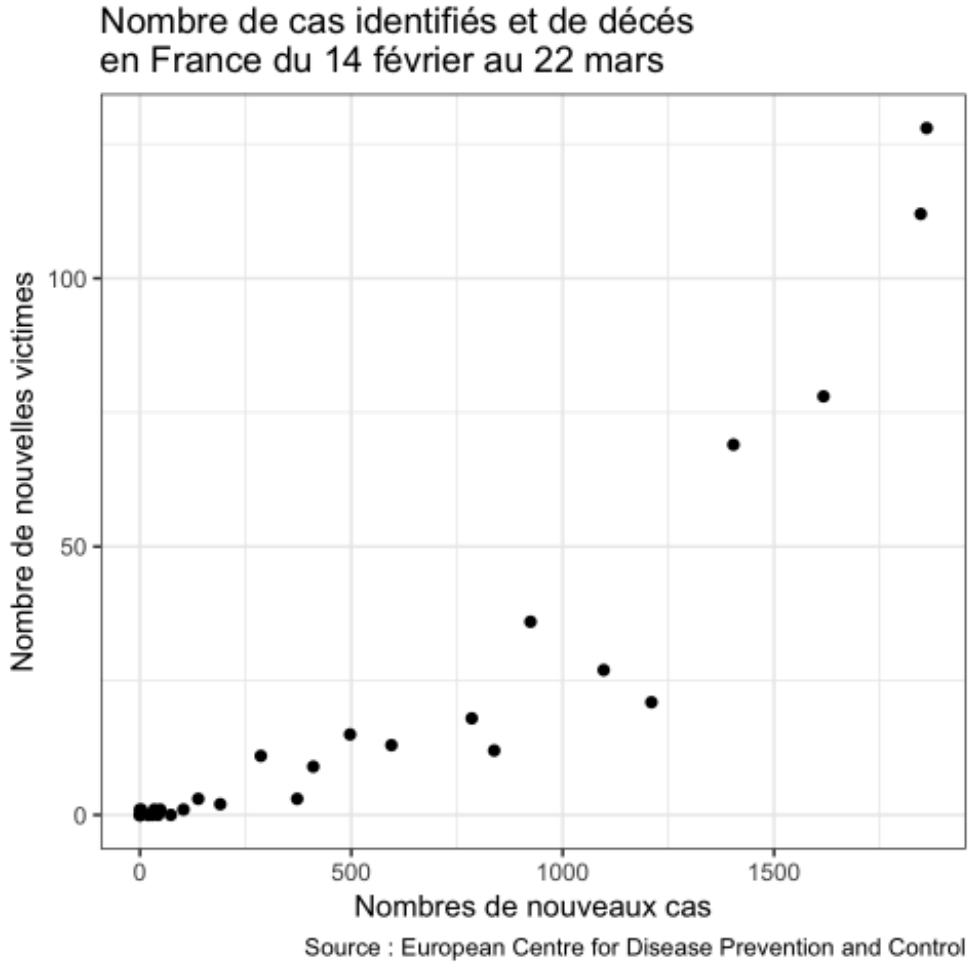
Dans le grand tableau mis à jour quotidiennement que l'on peut télécharger sur le site [3] du Centre européen de prévention et contrôle des maladies, quatre colonnes nous intéressent et se présentent ainsi :

DateRep	Cases	Deaths	Countries and territories
2020-02-14	0	0	Afghanistan
2020-02-14	0	0	Algeria
2020-02-14	0	0	Armenia
2020-02-14	1	0	Australia
2020-02-14	0	0	Austria
2020-02-14	0	0	Azerbaijan

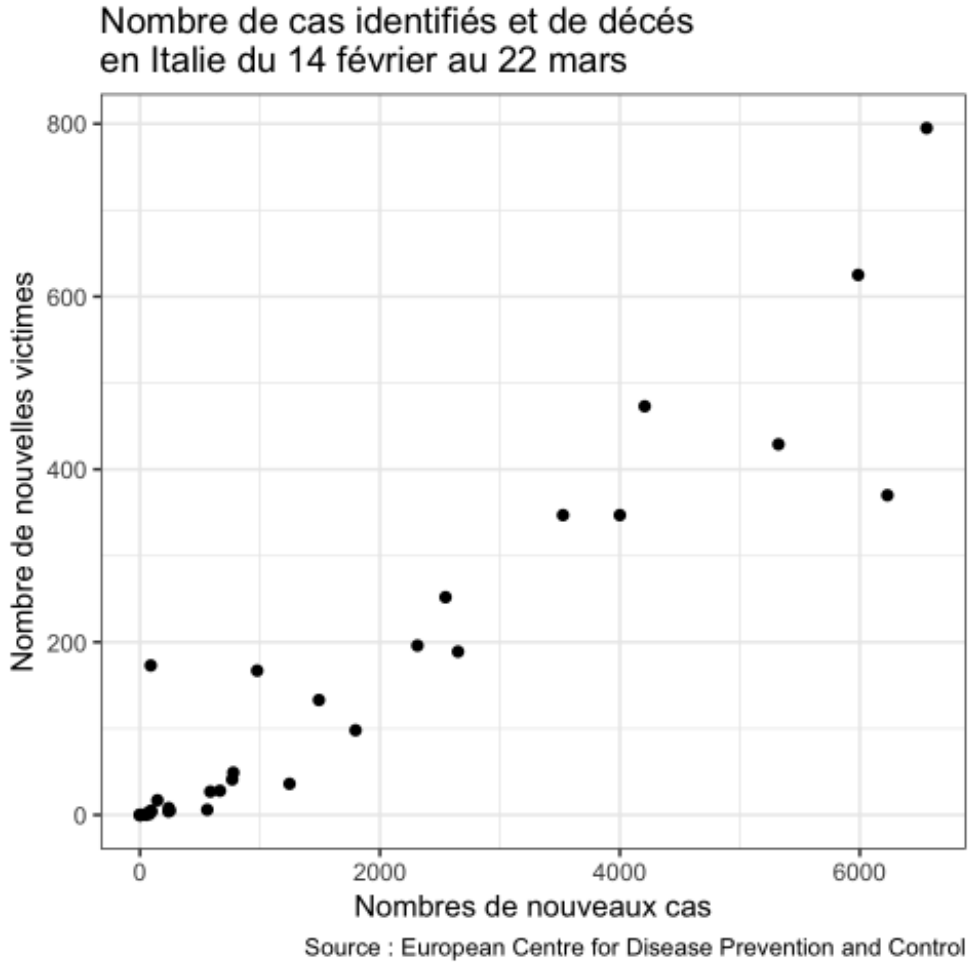
Chaque ligne du tableau correspond à un jour et à un pays. En colonne figure : la date du relevé (DateRep), le nombre de cas de malades ce jour-là (Cases), le nombre de décès comptés ce jour-là (Deaths) et le pays auquel correspondent ces informations (Countries and territories).

On peut donc lire sur la première ligne du tableau que le 14 février 2020 il n'y avait aucuns cas de malades identifiés ni aucun décès en Afghanistan du Coronavirus. Mêmes informations pour la même date en Algérie sur la ligne suivante, puis pour l'Arménie sur la troisième ligne, etc. Le tableau continue ainsi pour tous les pays du monde et pour toutes les dates jusqu'au 22 mars.

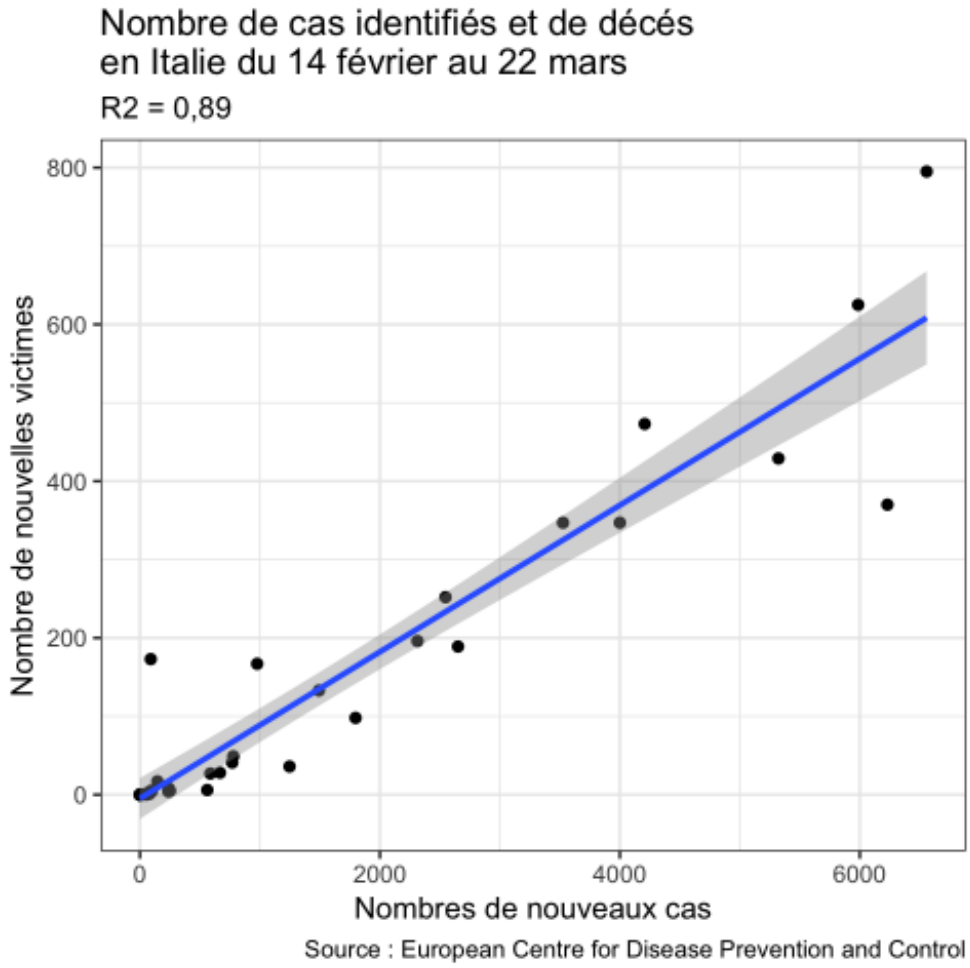
La première chose que l'on peut présenter graphiquement c'est le nombre de cas au regard du nombre de décès. Cela permet de vérifier qu'il y a un lien statistique fort entre les deux. La figure suivante présente le cas de la France.



Ce graphique, nommé nuage de point, montre que plus le nombre de cas identifiés est grand, plus le nombre de décès l'est aussi. C'était attendu et cela montre que les données ont une certaine cohérence. On peut réaliser la même figure pour l'Italie avec le graphique ci-dessous.



Il est aussi possible de modéliser cette relation, c’est-à-dire de résumer classiquement le nuage de points par une droite de tendance. C’est ce qui est représenté dans l’illustration suivante.

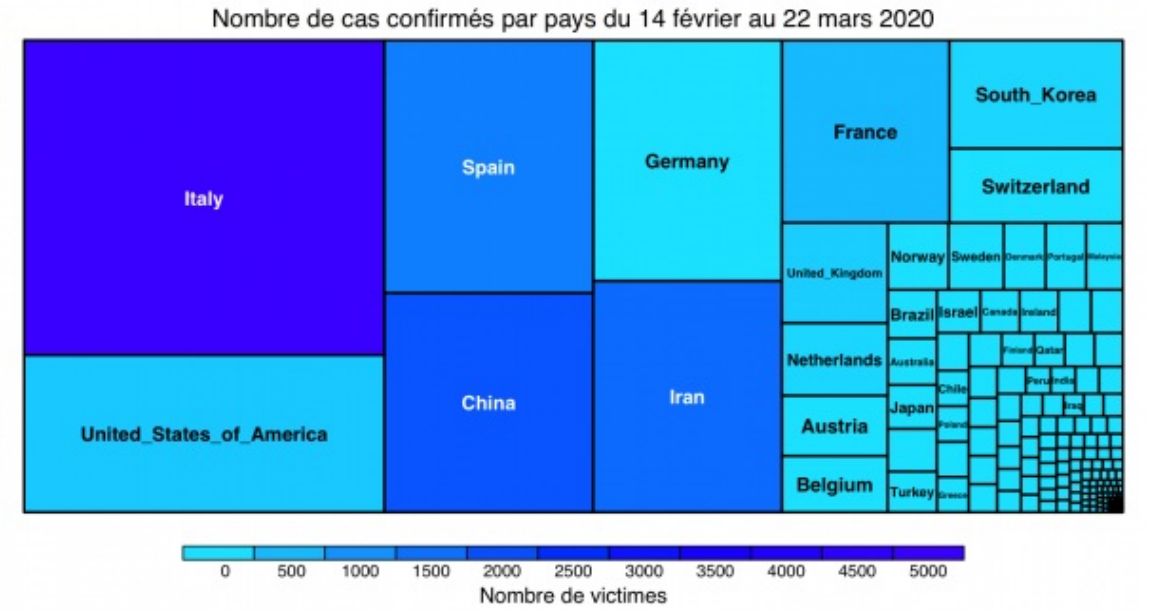


Cette illustration montre qu'il existe une relation linéaire, c'est-à-dire une relation de proportionnalité, entre le nombre de malades identifiés et le nombre de décès Cette relation est modélisable par une droite visible en bleu. Ce n'est pas forcément le cas pour tous les pays, cela dépend du moment et de la réponse de santé publique qui a été apportée par chaque pays.

Les pays les plus touchés

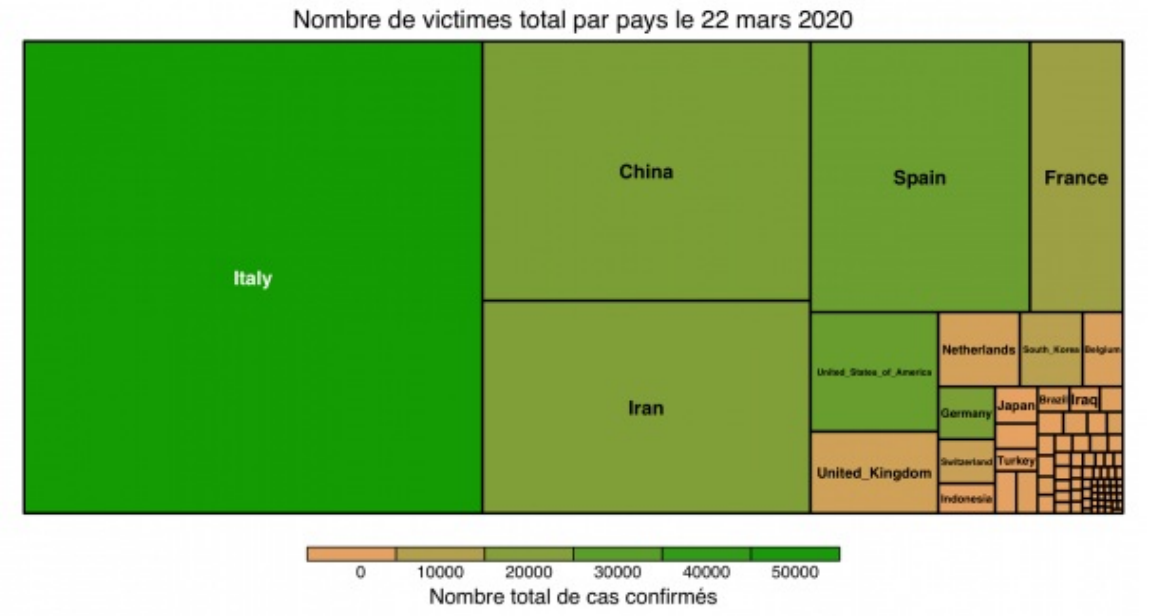
À partir de ces données, on peut classer les pays par leur nombre de cas confirmés. C'est une donnée très approximative puisque les tests de dépistages ne sont pas pratiqués de la même façon dans tous les pays. On dispose donc de grandeurs, mais il ne s'agit pas d'informations exactes.

La figure ci-dessous représente le nombre de cas confirmés : plus la case est grande plus le nombre de personnes malades (ou qui l'ont été) est important, plus la case est foncée plus le nombre de décès est important. Bien que le comptage des victimes soit une donnée plus fiable que le nombre de cas, il ne s'agit pas de données totalement exactes non plus, puisqu'elles ne sont pas produites avec les mêmes critères d'un pays à l'autre. Définir la source d'un décès comme étant le produit du coronavirus n'est pas une tâche simple ni forcément possible à chaque fois.

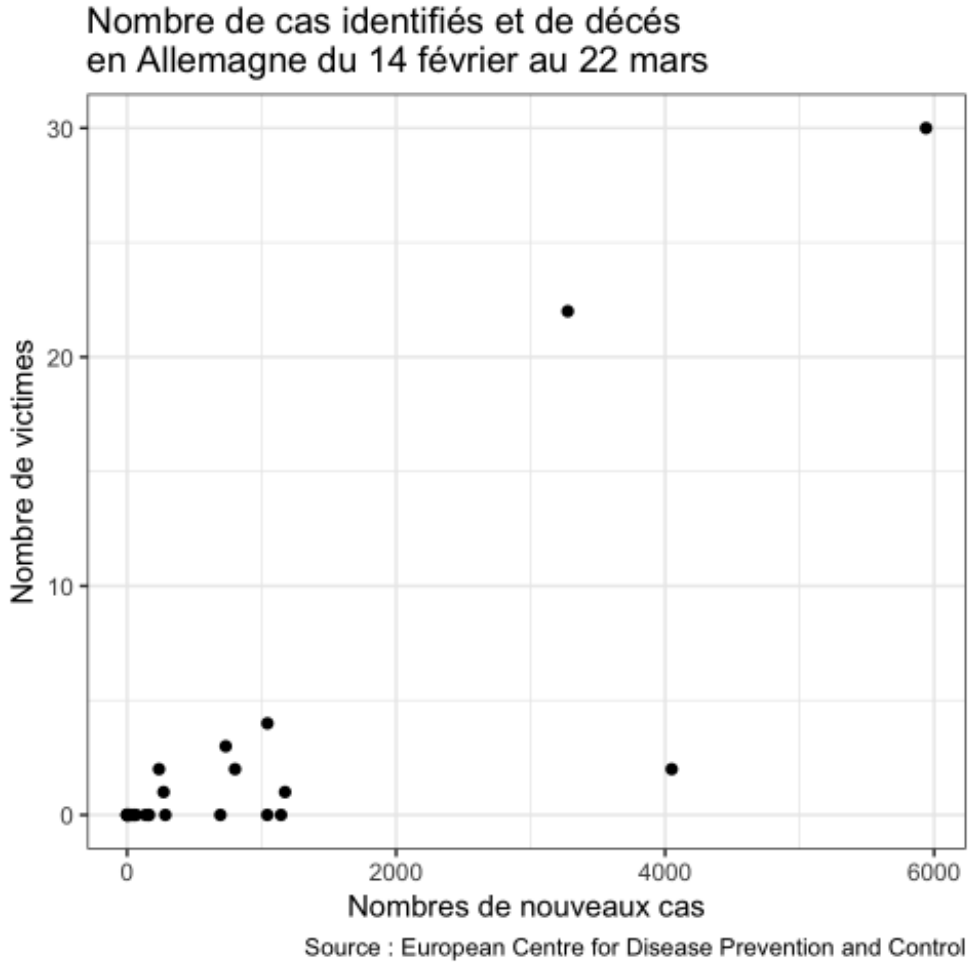


Sur la figure ci-dessus on voit que l'Italie, la Chine, l'Espagne, l'Iran, les USA et l'Allemagne sont les pays qui ont déclaré le plus grand nombre de malades entre le 14 février et le 22 mars. On voit aussi que globalement, plus le nombre de cas confirmés est important plus le nombre de victimes l'est aussi. Sauf pour l'Allemagne et les USA qui ont un nombre de victimes moins important que l'Espagne, l'Iran ou la Chine ou même la France.

Pour le voir plus clairement on peut faire la même figure que la précédente, en représentant le nombre de décès par la surface de la case et le nombre de malades par la couleur : plus la case est grande plus le nombre de victimes est important et plus la case est verte plus le nombre de cas comptés est important.



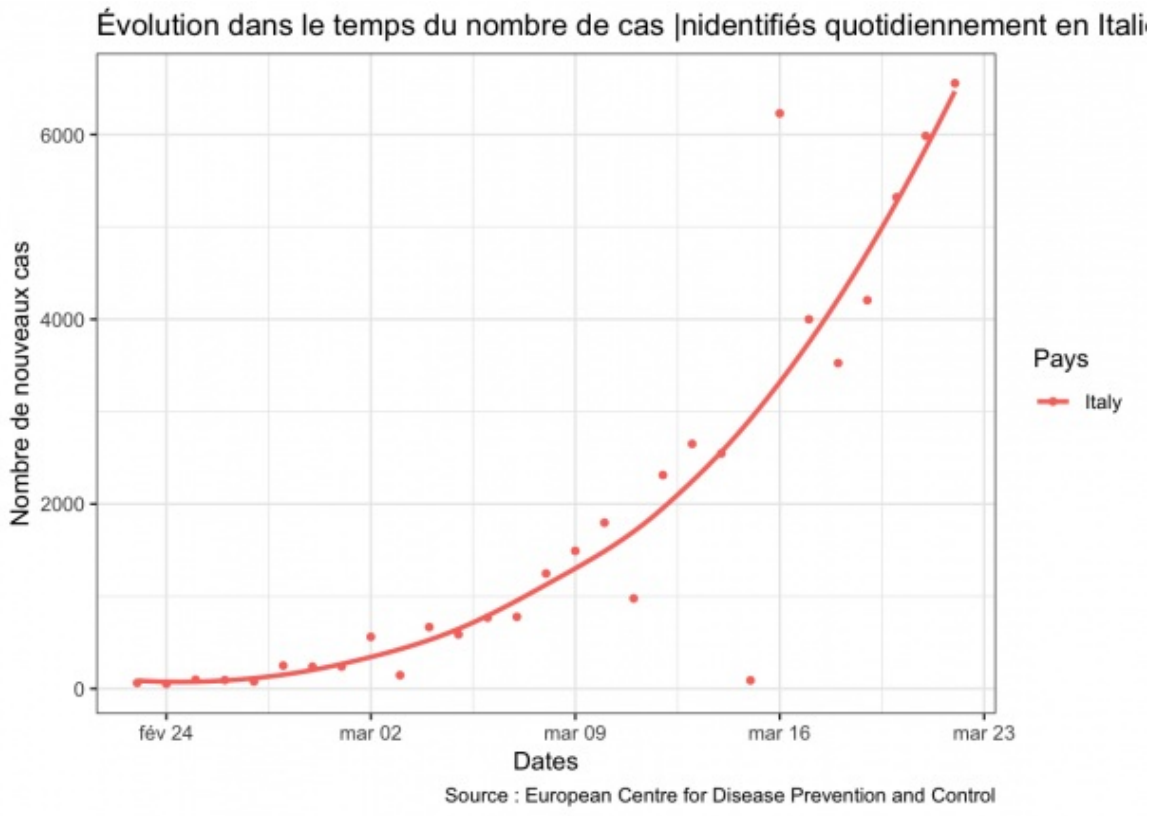
On voit sur cette figure que le nombre de victimes en Allemagne est moindre par rapport aux nombres de cas comptés. On peut regarder la relation entre le nombre de cas et le nombre de victimes comme on l'a fait précédemment pour la France et l'Italie.



Pour l'Allemagne le nombre de cas quotidiens relevés ne dépasse pas 30 alors que la France est aux alentours de 120. Il peut y avoir beaucoup d'explications que nous ne sommes pas en mesure d'identifier : cela peut être dû à la façon de construire les données ou bien à la réponse de santé publique apportée par le pays.

Évolution dans le temps en Italie

Le temps est le critère qui représente le mieux le phénomène épidémique. Comme nous l'avons vu précédemment nous pouvons modéliser la relation entre le temps et le nombre de nouveaux cas comme sur la figure suivante.

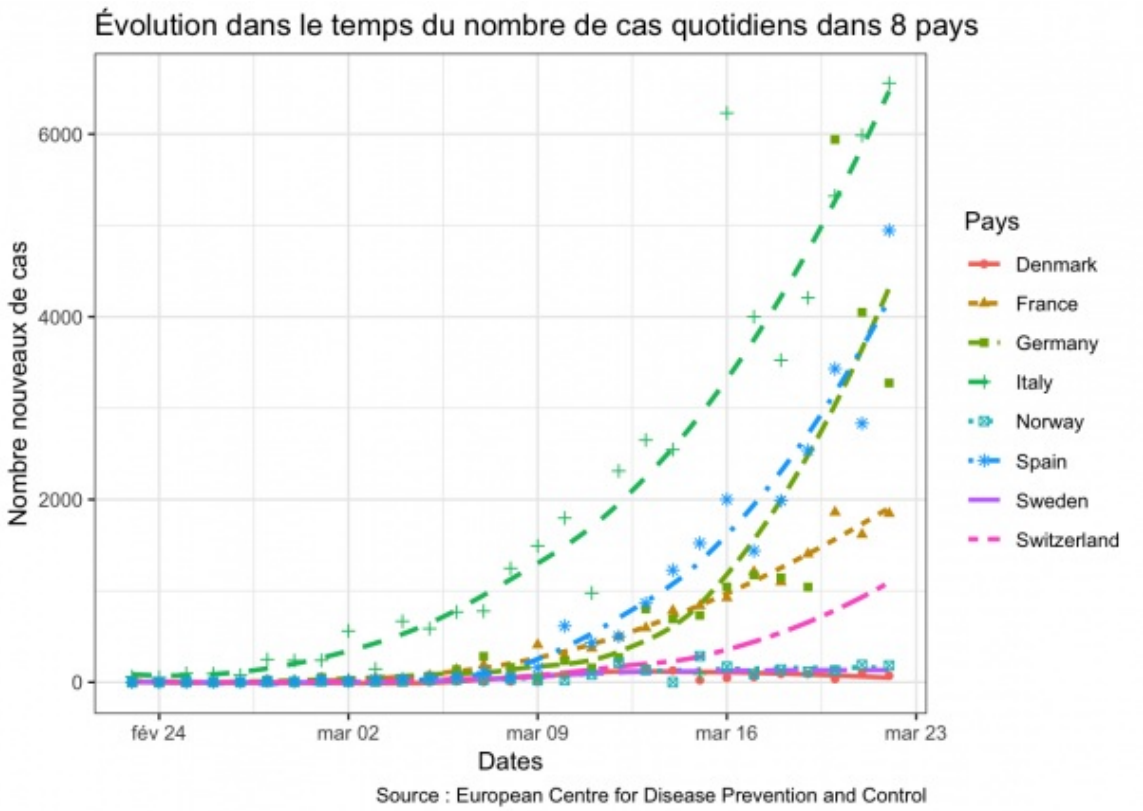


Sur l'axe horizontal (dit axe des abscisses) figure le temps : 24 février, 2 mars, 9 mars… Sur l'axe vertical (dit des ordonnées) est représenté le nombre de nouveaux malades du coronavirus identifiés chaque jour : 2000 malades dénombrés, 4000 et 6000.

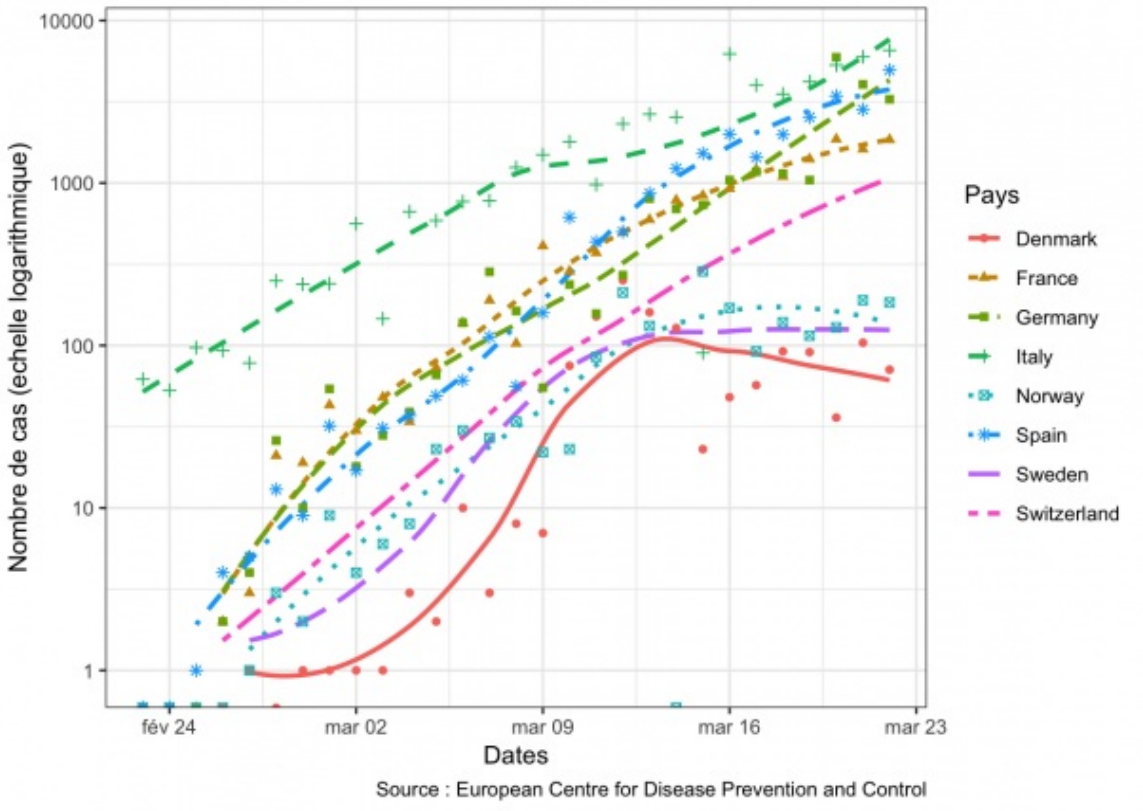
On voit sur ce graphique la forme caractéristique d'une fonction mathématique : la fonction exponentielle. Si on résume les données (représentées par des points sur le graphique) par une courbe (représenté par la ligne rouge sur le graphique), cette courbe pourra être définie en employant la fonction exponentielle (\(e^x\)), dont la caractéristique est de croître de plus en plus. On voit par exemple, qu'en Italie le 25 février, on comptait entre 0 et 5 nouveaux cas quotidiens. Trois semaines plus tard, le 15 mars, c'est aux alentours de 3000 cas quotidiens qui étaient dénombrés. Ensuite le 22 mars, une semaine plus tard, c'est 6000 cas quotidiens qui furent identifiés. La particularité de la fonction exponentielle c'est une augmentation de plus en plus importante et c'est là tout le problème : sur un temps très court, traiter un grand nombre de malades nécessitera une hospitalisation.

Évolution dans le temps en Europe

On va maintenant comparer la situation italienne à celle d'autres pays d'Europe. La figure ci-dessous représente l'évolution du nombre de cas depuis le 24 février. Chaque pays est représenté par une couleur, par la forme d'un point (astérisque, triangle, petit carré…) et par une courbe qui résume ces points.



Sur cette représentation on peut aussi avoir l'impression que les situations de la Suisse et de la France sont différentes. C'est assez difficile à voir en l'état. Pour mieux identifier les tendances et les similarités, il est utile de changer l'échelle de l'axe vertical. Au lieu d'avoir un espacement équivalent pour chaque unité, on va calculer un espacement équivalent pour 10, 100, 1000 et 10 000 cas. C'est ce que l'on appelle une transformation logarithmique qui figure sur la représentation suivante.



Sur cette visualisation, pour l'axe des ordonnées (l'axe vertical) une unité sur le graphique correspond à 10, 100, 1 000 et 10 000 cas, alors que sur le graphique précédent chaque graduation correspondait à 2 000 cas.

Les pays Scandinaves (Danemark, Suède et Norvège) se distinguent des autres pays puisque le nombre de cas stagne voir décroît pour le Danemark. Alors que l'Italie, l'Espagne, la France, l'Allemagne et la Suisse ont des tendances à la hausse et relativement similaires.

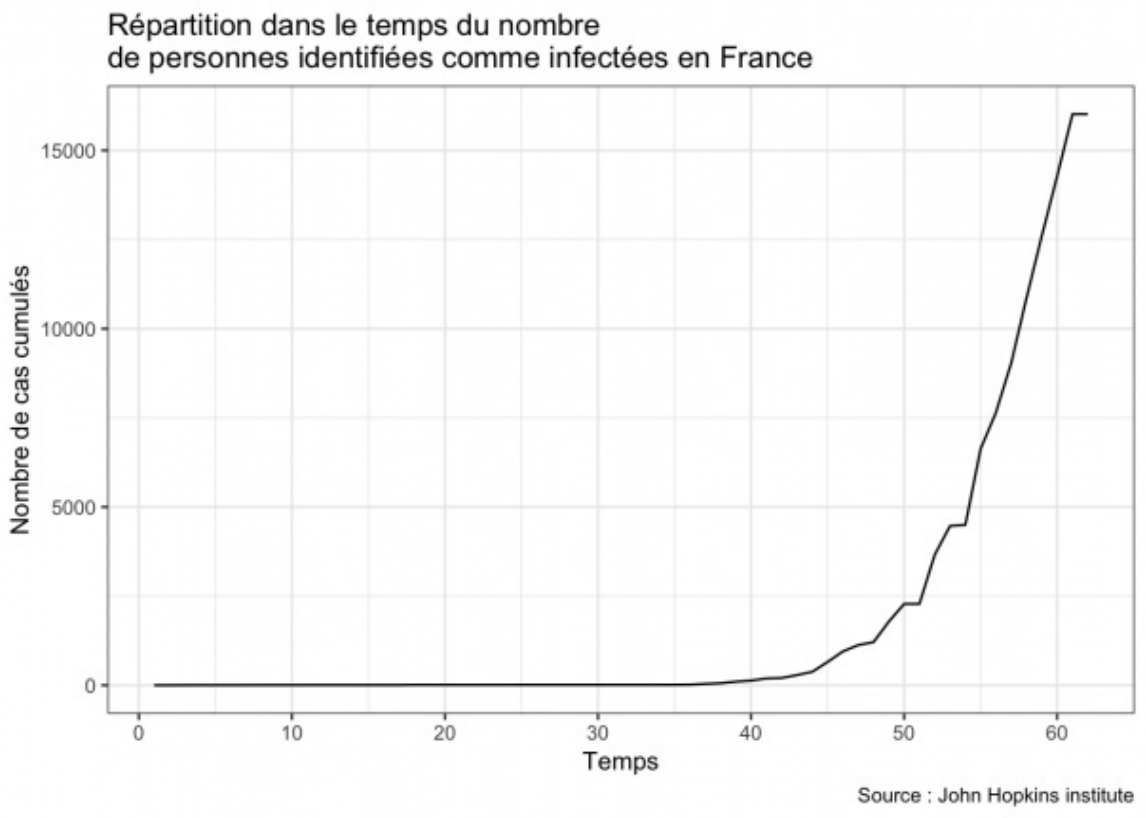
Une autre distinction est à signaler, qui n'est pas visible sur le graphique, c'est de rapporter le nombre de cas à la population totale du pays. Les pays comme la Suisse, le Danemark, La Belgique, la Norvège ou la Suède sont largement moins peuplés (entre 5 et 11 millions d'habitants) que la France, l'Allemagne, l'Italie ou l'Espagne (entre 45 et 80 millions d'habitants).

Établir une prédiction ?

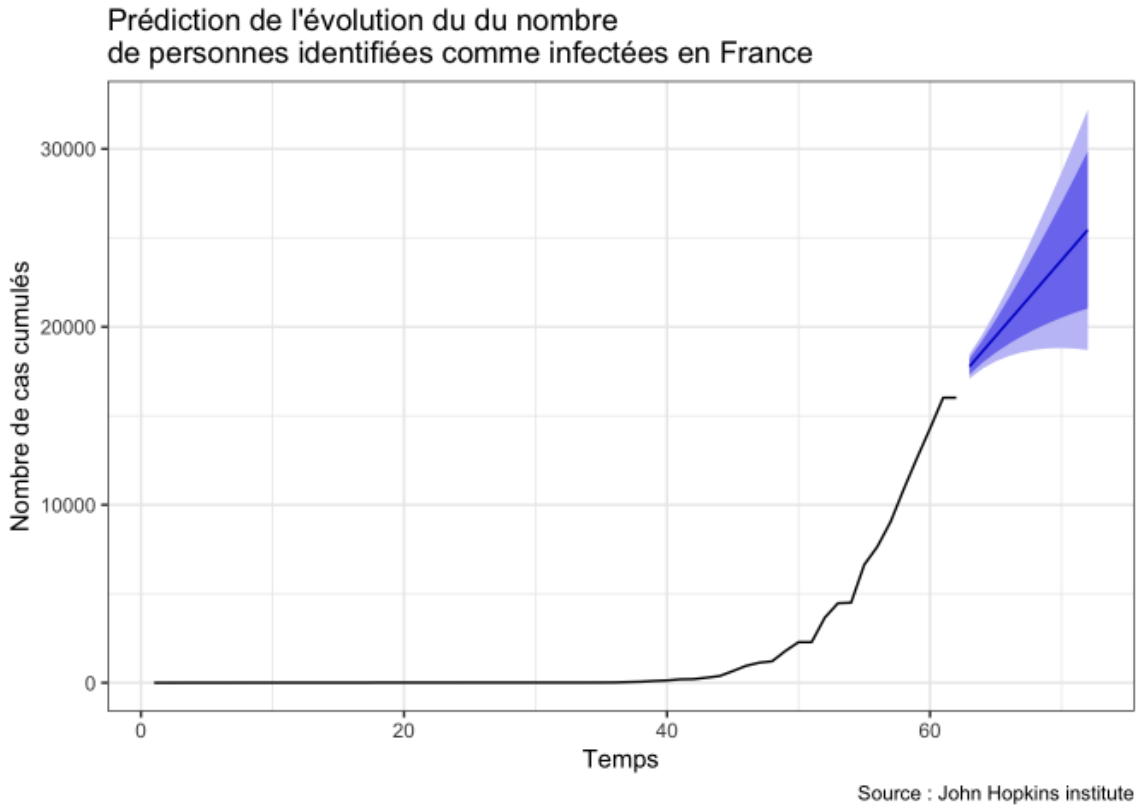
À partir de ces données, il est possible de réaliser des prédictions, mais leur validité est très limitée. Un facteur limitant est en effet qu'on ne connaît pas le nombre de personnes réellement infectées dans la mesure où le dépistage varie beaucoup en fonction des pays. Techniquement, il s'agit de prolonger la courbe et d'établir un champ des possibles purement théorique. Rien dans les données ne permet de savoir ce qui arrivera dans la réalité, il s'agit seulement d'indications approximatives basées sur les informations dont on dispose.

On va utiliser les données du John Hopkins Institute[4]. La différence entre ces données et les précédentes c'est qu'il s'agit de valeurs cumulées, c'est-à-dire qu'elles indiquent le nombre total de malades identifiés : le nombre de nouveaux malades identifiés chaque jour plus ceux des jours précédents.

En représentant les valeurs pour la France du nombre total de personnes identifiées comme infectées, on obtient le graphique suivant sur lequel on a représenté les valeurs par une ligne plutôt que des points.



On remarque toujours cette forme exponentielle que l'on prolonge en réalisant la visualisation suivante.



En bleu figure l'échantillon de confiance, c'est-à-dire l'erreur de la prédiction. On voit que plus on s'éloigne des données plus l'erreur augmente. La prédiction indique que, si aucune mesure n'est prise, le nombre de malades continuera à croître. Il n'est pas nécessaire de la commenter d'avantage.

Conclusion

L'objectif du gouvernement est de limiter le nombre de malades afin que le système de santé ne soit pas submergé et puisse soigner les cas graves les uns après les autres. Toute la question du confinement est mue par cet objectif : étaler dans le temps les besoins en hospitalisation c'est pour cette raison qu'il faut rester confiner et veiller à ne pas propager la maladie en particulier auprès des personnes âgées ou ayant déjà des maladies. Il est fort probable que ces mesures de confinements, parce qu'elles tendent à entraver la transmission, auront des effets visibles sur les chiffres à venir même si ces résultats ne sont pas immédiats. Le problème bien sûr, est celui des conséquences non médicales liées au confinement.

On a déjà mentionné les questions de fiabilité des données. Il serait intéressant de mieux creuser cette question notamment en établissant le ratio entre nombre de victimes et nombre de personnes infectées. Mais là encore on bute sur l'incertitude concernant ce dernier nombre. L'Allemagne se distingue de ses voisins avec relativement peu de victimes. Les pays Scandinaves, relativement peu peuplés, semblent avoir un nombre de personnes infectées qui stagne, voire décroît. Il serait intéressant d'en savoir plus sur la réponse de santé publique qui a été apportée dans ces pays.

Pour aller plus loin

La page web de l'Institut Pasteur sur le coronavirus :
https://www.pasteur.fr/fr/centre-medical/fiches-maladies/coronavirus-wuhan

Un fil Twitter du 11 mars qui explique et présente une analyse de données et des réponses possibles :
https://twitter.com/aktiur/status/1237693562994778112

Un site de concours d'analyse de données qui présente de nombreuses données sur le coronavirus :
https://www.kaggle.com/covid19

Une vidéo qui explique très bien tout ce que l'on vient de présenter :
https://www.youtube.com/watch?v=Kas0tIxDvrg