TP

Becnhmarking

Note

Les sujets sont issus du MOOC Recherche reproductible : principes méthodologiques pour une science transparente disponible sur le plateforme `https://fun-mooc.fr`.

1 Concentration de CO2 dans l'atmosphère depuis 1958

Prérequis : traitement de suites chronologiques

En 1958, Charles David Keeling a initié une mesure de la concentration de CO2 dans l'atmosphère à l'observatoire de Mauna Loa, Hawaii, États-Unis qui continue jusqu'à aujourd'hui. L'objectif initial était d'étudier la variation saisonnière, mais l'intérêt s'est déplacé plus tard vers l'étude de la tendance croissante dans le contexte du changement climatique. En honneur à Keeling, ce jeu de données est souvent appelé "Keeling Curve" (voir https://en.wikipedia.org/wiki/Keeling_Curve pour l'histoire et l'importance de ces données).

Les données sont disponibles sur le site Web de l'institut Scripps. Utilisez le fichier avec les observations hebdomadaires. Attention, ce fichier est mis à jour régulièrement avec de nouvelles observations. Notez donc bien la date du téléchargement, et gardez une copie locale de la version précise que vous analysez. Faites aussi attention aux données manquantes.

Votre mission si vous l'acceptez :

1. Réalisez un graphique qui vous montrera une oscillation périodique superposée à une évolution systématique plus lente.
2. Séparez ces deux phénomènes. Caractérisez l'oscillation périodique. Proposez un modèle simple de la contribution lente, estimez ses paramètres et tentez une extrapolation jusqu'à 2025 (dans le but de pouvoir valider le modèle par des observations futures).

2 Le pouvoir d'achat des ouvriers anglais du XVIe au XIXe siècle

Prérequis : techniques de présentation graphique

William Playfair était un des pionniers de la présentation graphique des données. Il est notamment considéré comme l'inventeur de l'histogramme. Un de ses graphes célèbres, tiré de son livre "A Letter on Our Agricultural Distresses, Their Causes and Remedies", montre l'évolution du prix du blé et du salaire moyen entre 1565 et 1821. Playfair n'a pas publié les données numériques brutes qu'il a utilisées, car à son époque la réplicabilité n'était pas encore considérée comme essentielle. Des valeurs obtenues par numérisation du graphe sont aujourd'hui téléchargeables, la version en format CSV étant la plus pratique.

Quelques remarques pour la compréhension des données :
— Jusqu'en 1971, la livre sterling était divisée en 20 shillings, et un shilling en 12 pences.
— Le prix du blé est donné en shillings pour un quart de boisseau de blé. Un quart de boisseau équivaut 15 livres britanniques ou 6,8 kg.
— Les salaires sont donnés en shillings par semaine.

Votre mission si vous l'acceptez :

1. Votre première tâche est de reproduire le graphe de Playfair à partir des données numériques. Représentez, comme Playfair, le prix du blé par des barres et les salaires par une surface bleue délimitée par une courbe rouge. Superposez les deux de la même façon dans un seul graphique. Le style de votre graphique pourra rester différent par rapport à l'original, mais l'impression globale devrait être la même.
2. Par la suite, améliorez la présentation de ces données. Pour commencer, Playfair a combiné les deux quantités dans un même graphique en simplifiant les unités "shillings par quart de boisseau de blé" et "shillings par semaine" à un simple "shillings", ce qui aujourd'hui n'est plus admissible. Utilisez deux ordonnées différentes, une à gauche et une à droite, et indiquez les unités correctes. À cette occasion, n'hésitez pas à proposer d'autres représentations que des barres et des surface/courbes pour les deux jeux de données si ceci vous paraît judicieux.
3. L'objectif de Playfair était de montrer que le pouvoir d'achat des ouvriers avait augmenté au cours du temps. Essayez de mieux faire ressortir cette information. Pour cela, faites une représentation graphique du pouvoir d'achat au cours du temps, définie comme la quantité de blé qu'un ouvrier peut acheter avec son salaire hebdomadaire. Dans un autre graphique, montrez les deux quantités (prix du blé, salaire) sur deux axes différents, sans l'axe du temps. Trouvez une autre façon d'indiquer la progression du temps dans ce graphique. Quelle représentation des données vous paraît la plus claire ? N'hésitez pas à en proposer d'autres.

3 L'épidémie de choléra à Londres en 1854

Prérequis : représentation de données géographiques

En 1854, le quartier de Soho à Londres a vécu une des pires épidémies de choléra du Royaume-Uni, avec 616 morts. Cette épidémie est devenue célèbre à cause de l'analyse détaillée de ses causes réalisée par le médecin John Snow. Ce dernier a notamment montré que le choléra est transmis par l'eau plutôt que par l'air, ce qui était la théorie dominante de l'époque.

Un élément clé de cette analyse était une carte sur laquelle John Snow avait marqué les lieux des décès et les endroits où se trouvaient les pompes à eau publiques. Ces données sont aujourd'hui disponibles sous forme numérique. Nous vous proposons de les utiliser pour recréer la carte de John Snow dans un document computationnel réplicable.

Votre mission si vous l'acceptez :

1. Londres a bien sûr évolué depuis 1854, mais une carte d'aujourd'hui est tout à fait utilisable comme support pour ces données historiques. À partir des données numériques, réalisez une carte dans l'esprit de celle de John Snow. Montrez les lieux de décès avec des symboles dont la taille indique le nombre de décès. Indiquez sur la même carte la localisation des pompes en utilisant une autre couleur et/ou un autre symbole. 2. Par la suite, essayez de trouver d'autres façons pour montrer que la pompe de Broad Street est au centre de l'épidémie. Vous pouvez par exemple calculer la densité des décès dans le quartier et l'afficher sur la carte, mais n'hésitez pas à expérimenter avec d'autres approches.

Conseils techniques pour l'affichage de cartes

En R, nous vous suggérons l'utilisation de la bibliothèque ggmap. Evitez Google Maps comme source de cartes car ce service est devenu payant en 2018 (et même si vous restez dans le quota gratuit, vous devez déposer vos coordonnées bancaires pour vous inscrire). OpenStreetMaps est une bonne alternative que ggmap propose également (source="osm").

En Python dans l'environnement Jupyter, la bibliothèque folium produit des belles cartes qui en plus sont interactives. Le seul inconvénient est que les cartes stockées dans les notebooks ne sont pas affichées dans GitLab. Pour que vos évaluateurs puissent les voir plus facilement, pensez à exporter votre notebook au format HTML et déposer cette version aussi dans votre dépôt.

4 Estimation de la latence et de la capacité d'une connexion à partir de mesures asymétriques

Prérequis : régression linéaire

Un modèle simple et fréquemment utilisé pour décrire la performance d'une connexion de réseau consiste à supposer que le temps d'envoi T pour un message dépend principalement de sa taille S (nombre d'octets) et de deux grandeurs propres à la connexion : la latence L (en secondes) et la capacité C (en octets/seconde). La relation entre ces quatre quantités est $T(S) = L + S/C$. Ce modèle néglige un grand nombre de détails. D'une part, L et C dépendent bien sûr du protocole de communication choisi mais aussi dans une certaine mesure de S. D'autre part, la mesure de $T(S)$ comporte en général une forte composante aléatoire. Nous nous intéressons ici au temps moyen qu'il faut pour envoyer un message d'une taille donnée.

Votre tâche est d'estimer L et C à partir d'une série d'observations de T pour des valeurs différentes de S Préparez votre analyse sous forme d'un document computationnel réplicable qui commence avec la lectures des données brutes, disponibles pour deux connexions différentes, qui ont été obtenues avec l'outil ping :

Le premier jeu de données examine une connexion courte à l'intérieur d'un campus : `libglab2.log.gz` Le deuxième jeu de données mesure la performance d'une connexion vers un site Web éloigné assez populaire et donc chargé : `stackoverflow.log.gz`

Les deux fichiers contiennent la sortie brute de l'outil ping qui a été exécuté dans une boucle en variant de façon aléatoire la taille du message. Chaque ligne a la forme suivante:

```
[1421761682.052172] 665 bytes from lig-publig.imag.fr (129.88.11.7): icmp_seq=1 ttl=60 time=22.5 ms
```

Au début, entre crochet, vous trouvez la date à laquelle la mesure a été prise, exprimée en secondes depuis le 1er janvier 1970. La taille du message en octets est donnée juste après, suivie par le nom de la machine cible et son adresse IP, qui sont normalement identiques pour toutes les lignes à l'intérieur d'un jeu de données. À la fin de la ligne, nous trouvons le temps d'envoi (aller-retour) en millisecondes. Les autres indications, `icmp_seq` et `ttl`, n'ont pas d'importance pour notre analyse.

> **Attention**
> Il peut arriver qu'une ligne soit incomplète et il faut donc vérifier chaque ligne avant d'en extraire des informations !

Votre mission si vous l'acceptez :

1. Commencez par travailler avec le premier jeu de données (liglab2). Représentez graphiquement l'évolution du temps de transmission au cours du temps (éventuellement à différents instants et différentes échelles de temps) pour évaluer la stabilité temporelle du phénomène. Ces variations peuvent-elles être expliquées seulement par la taille des messages ?
2. Représentez le temps de transmission en fonction de la taille des messages. Vous devriez observer une "rupture", une taille à partir de laquelle la nature de la variabilité change. Vous estimerez (graphiquement) cette taille afin de traiter les deux classes de tailles de message séparément.
3. Effectuez une régression linéaire pour chacune des deux classes et évaluez les valeurs de L et de C correspondantes. Vous superposerez le résultat de cette régression linéaire au graphe précédent.
4. (Optionnel) La variabilité est tellement forte et asymétrique que la régression du temps moyen peut être considérée comme peu pertinente. On peut vouloir s'intéresser à caractériser plutôt le plus petit temps de transmission. Une approche possible consiste donc à filtrer le plus petit temps de transmission pour chaque taille de message et à effectuer la régression sur ce sous-ensemble de données. Cela peut également être l'occasion pour ceux qui le souhaitent de se familiariser avec la `régression de quantiles` (implémentée en R dans la bibliothèque quantreg et en Python dans la bibliothèque statsmodels).
5. Répétez les étapes précédentes avec le second jeu de données (stackoverflow).

5 Analyse des dialogues dans l'Avare de Molière

Prérequis : analyse de texte, éventuellement techniques de présentation graphique

L'Observatoire de la vie littéraire (OBVIL) promeut une approche de l'analyse des textes littéraires fondée sur le numérique. Dans le cadre du Projet Molière, des pièces de cet auteur ont été numérisées et sont accessibles librement dans différents formats utilisables par un programme informatique. Pour l'Avare de Molière, voici les formats disponibles : TEI, epub, kindle, markdown, Texte iramuteq, Texte dit/Paroles, TXM, html complet avec table des matières, fragment html.

Grâce à ces numérisations, il est possible d'écrire des programmes pour réaliser des analyses syntaxiques et sémantiques. Nous vous proposons dans ce sujet de reproduire une étude réalisée par l'OBVIL sur les dialogues de l'Avare de Molière.

Nous vous conseillons de réaliser cet exercice en Python plutôt qu'en R, surtout pour la partie qui traite le texte.

Votre mission si vous l'acceptez :

1. Classez les personnages selon la quantité de parole grâce à une analyse syntaxique du texte (scènes / répliques / mots). En particulier, quel est celui qui parle le plus ? Quel est celui qui ne parle pas du tout ? Attention, les noms des personnages ne sont pas forcément homogènes (casse et accents par exemple).
2. Réalisez un graphique qui montrera le nombre de mots que chaque acteur prononce dans chaque scène. Pour cela, vous pouvez vous inspirer de l'étude de l'Avare de Molière réalisée par l'OBVIL (graphe de gauche). Dans ce graphique, les lignes sont de longueur égale et la hauteur représente le nombre de mots prononcés au total dans la scène. La largeur de chaque rectangle indique le pourcentage de la scène qu'un acteur occupe.
3. Facultatif : Construisez un graphe d'interlocution permettant de visualiser les échanges entre les personnages. Pour cela, vous pouvez vous inspirer de l'étude de l'Avare de Molière réalisée par l'OBVIL (graphe de droite).

6 Autour du Paradoxe de Simpson

Prérequis : calcul de moyennes et de ratios, techniques de présentations graphiques simples, éventuellement régression logistique

En 1972-1974, à Whickham, une ville du nord-est de l'Angleterre, située à environ 6,5 kilomètres au sud-ouest de Newcastle upon Tyne, un sondage d'un sixième des électeurs a été effectué afin d'éclairer des travaux sur les maladies thyroïdiennes et cardiaques (Tunbridge et al. 1977). Une suite de cette étude a été menée vingt ans plus tard (Vanderpump et al. 1995). Certains des résultats avaient trait au tabagisme et cherchaient à savoir si les individus étaient toujours en vie lors de la seconde étude. Par simplicité, nous nous restreindrons aux femmes et parmi celles-ci aux 1314 qui ont été catégorisées comme "fumant actuellement" ou "n'ayant jamais fumé". Il y avait relativement peu de femmes dans le sondage initial ayant fumé et ayant arrêté depuis (162) et très peu pour lesquelles l'information n'était pas disponible (18). La survie à 20 ans a été déterminée pour l'ensemble des femmes du premier sondage.

Les données sont disponibles dans ce `fichier CSV`. Vous trouverez sur chaque ligne si la personne fume ou non, si elle est vivante ou décédée au moment de la seconde étude, et son âge lors du premier sondage.

Cet exercice peut être réalisé indifféremment en R ou en Python.

Votre mission si vous l'acceptez :

1. Représentez dans un tableau le nombre total de femmes vivantes et décédées sur la période en fonction de leur habitude de tabagisme. Calculez dans chaque groupe (fumeuses / non fumeuses) le taux de mortalité (le rapport entre le nombre de femmes décédées dans un groupe et le nombre total de femmes dans ce groupe). Vous pourrez proposer une représentation graphique de ces données et calculer des intervalles de confiance si vous le souhaitez. En quoi ce résultat est-il surprenant ?
2. Reprenez la question 1 (effectifs et taux de mortalité) en rajoutant une nouvelle catégorie liée à la classe d'âge. On considérera par exemple les classes suivantes : 18-34 ans, 34-54 ans, 55-64 ans, plus de 65 ans. En quoi ce résultat est-il surprenant ? Arrivez-vous à expliquer ce paradoxe ? De même, vous pourrez proposer une représentation graphique de ces données pour étayer vos explications.
3. Afin d'éviter un biais induit par des regroupements en tranches d'âges arbitraires et non régulières, il est envisageable d'essayer de réaliser une régression logistique. Si on introduit une variable `Death` valant 1 ou 0 pour indiquer si l'individu est décédé durant la période de 20 ans, on peut étudier le modèle `Death ~ Age` pour étudier la probabilité de décès en fonction de l'âge selon que l'on considère le groupe des fumeuses ou des non fumeuses. Ces régressions vous permettent-elles de conclure sur la nocivité du tabagisme ? Vous pourrez proposer une représentation graphique de ces régressions (en n'omettant pas les régions de confiance).

7 Autour du SARS-CoV-2 (Covid-19)

Prérequis : Techniques de présentation graphique. Cet exercice peut être réalisé indifféremment en R ou en Python.

Le but est ici de reproduire des graphes semblables à ceux du South China Morning Post (SCMP), sur la page The Coronavirus Pandemic et qui montrent pour différents pays le nombre cumulé (c'est-à-dire le nombre total de cas depuis le début de l'épidémie) de personnes atteintes de la maladie à coronavirus 2019.

Les données que nous utiliserons dans un premier temps sont compilées par le Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) et sont mises à disposition sur GitHub. C'est plus particulièrement sur les données time_series_covid19_confirmed_global.csv (des suites chronologiques au format csv) disponibles à l'adresse : https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv, que nous allons nous concentrer.

Vous commencerez par télécharger les données pour créer un graphe montrant l'évolution du nombre de cas cumulé au cours du temps pour les pays suivants : la Belgique (Belgium), la Chine - toutes les provinces sauf Hong-Kong (China), Hong Kong (China, Hong-Kong), la France métropolitaine (France), l'Allemagne (Germany), l'Iran (Iran), l'Italie (Italy), le Japon (Japan), la Corée du Sud (Korea, South), la Hollande sans les colonies (Netherlands), le Portugal (Portugal), l'Espagne (Spain), le Royaume-Unis sans les colonies (United Kingdom), les États-Unis (US).

Le nom entre parenthèses est le nom du « pays » tel qu'il apparaît dans le fichier time_series_covid19_confirmed_global.csv. Les données de la Chine apparaissent par province et nous avons séparé Hong-Kong, non pour prendre parti dans les différences entre cette province et l'état chinois, mais parce que c'est ainsi qu'apparaissent les données sur le site du SCMP. Les données pour la France, la Hollande et le Royaume-Uni excluent les territoires d'outre-mer.

Ensuite vous ferez un graphe avec la date en abscisse et le nombre cumulé de cas à cette date en ordonnée. Nous vous proposons de faire deux versions de ce graphe, une avec une échelle linéaire et une avec une échelle logarithmique.

Question subsidiaire

Vous pourrez également utiliser les données de décès (time_series_covid19_deaths_global.csv) et refaire les courbes, mais là encore, faites attention lors de l'interprétation. Ces courbes, même si elles paraissent effrayantes, doivent être comparées à la mortalité « normale ». Pour la France des données sont disponibles sur le site de l'INSEE : https://www.insee.fr/fr/information/4470857, ainsi que dans les « Points hebdomadaires » de surveillance de la mortalité diffusés par Santé publique France, comme celui de la semaine 12 (le site étant très mal conçu pour quiconque souhaite une information spécifique, le plus simple est de passer par un moteur de recherche généraliste…).

Pour atténuer les effets dus aux méthodes de comptage, etc., vous pourrez, une fois l'épidémie terminée, prendre les données du nombre total de décès et les normaliser pour 1000 habitants du pays concerné. Vous irez ensuite chercher les données sur le nombre de lits d'hôpital pour 1000 habitants sur le site de l'OCDE et vous pourrez corréler les deux (c'est-à-dire, faire un graphe avec le nombre de lits en abscisse et le nombre de décès en ordonnée)…