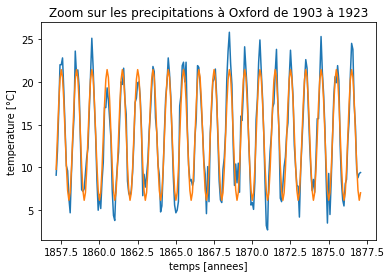


**Projet d’estimation robuste**

Moindres Carrés et RANSAC

Sur les températures maximales

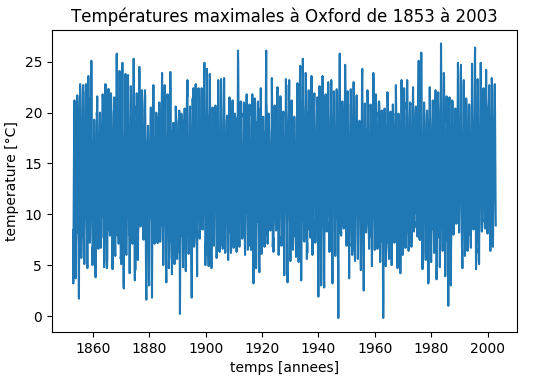
à Oxford entre 1853 et 2003



30 mars 2018

Mannaïg L’Haridon et Iris de Gelis

1. Présentation des données
2. Les données

Au tout départ nous avions pensé travailler sur des données pluviométriques. En recherchant sur internet, nous sommes tombés sur des données libres d’accès et gratuites fournies par le gouvernement britannique sur le site suivant : <https://data.gov.uk/dataset/historic-monthly-meteorological-station-data>. Le choix de la ville de Oxford a été arbitraire : une autre des villes présente sur le site aurait pu être choisie.

Cependant, lorsque nous avons regardé le graphe pluviométrique de Oxford, nous nous sommes rendues compte qu’il serait quasiment impossible d’en tirer un modèle étant donné que le taux de précipitation en Angleterre varie très peu voir pas entre les saisons. Le fichier proposant d’autres observations météorologiques, nous avons donc décidé de conserver ce fichier répertoriant les données météorologiques de la ville d’Oxford de 1853 à 2003, mais d’étudier les températures maximales.

Figure : Données sur les températures maximales à Oxford entre 1853 et 2003, soit sur 150 ans à raison d'un point par mois

L’intérêt principal de ce jeu de données est qu’il propose un total de 1800 valeurs de températures maximales réparties sur 150 années, soit une valeur par mois.

1. Le modèle

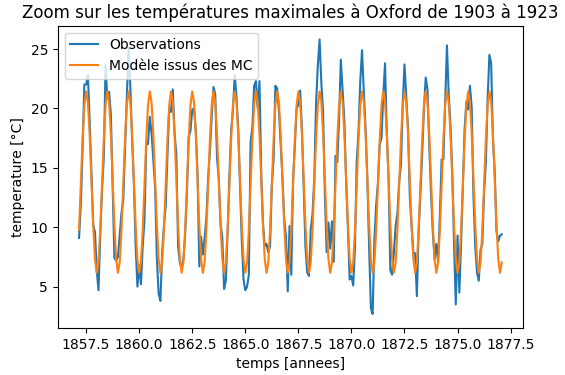
En sélectionnant juste 20 ans de données de façon arbitraire, on remarque que les observations sont périodiques – de période 12 mois soit 1 an – et adoptent un mouvement quasi-sinusoidal.

Figure 2 : Observations et modèle sur 20 années

Le modèle des observations adopté est donc celui d’un signal sinusoidal, soit :

Avec l’amplitude, la pulsation, la phase, une constante et la variation du temps.

1. Conditions initiales

Pour le modèle choisi les conditions initiales ont été choisies en regardant les tracés des observations. En effet, pour une simple sinusoïde, il est facile d’approximer l’amplitude, la période, et la constante. En regardant, les données sur 10 ans, on peut avoir un a priori sur les valeurs de l’amplitude, de la période et de la constante. D’après la figure ci-dessous, on a pris :

* **Amplitude** : 2\*A0 = 15 => A0 = 7.5 par lecture
* **Omega** : La période des oscillations à priori est d’un an, donc T = 1 an. Or ω = 2\*π/T. Ainsi on a pris ω = 2\* π
* **Constante** : 14 par lecture graphique

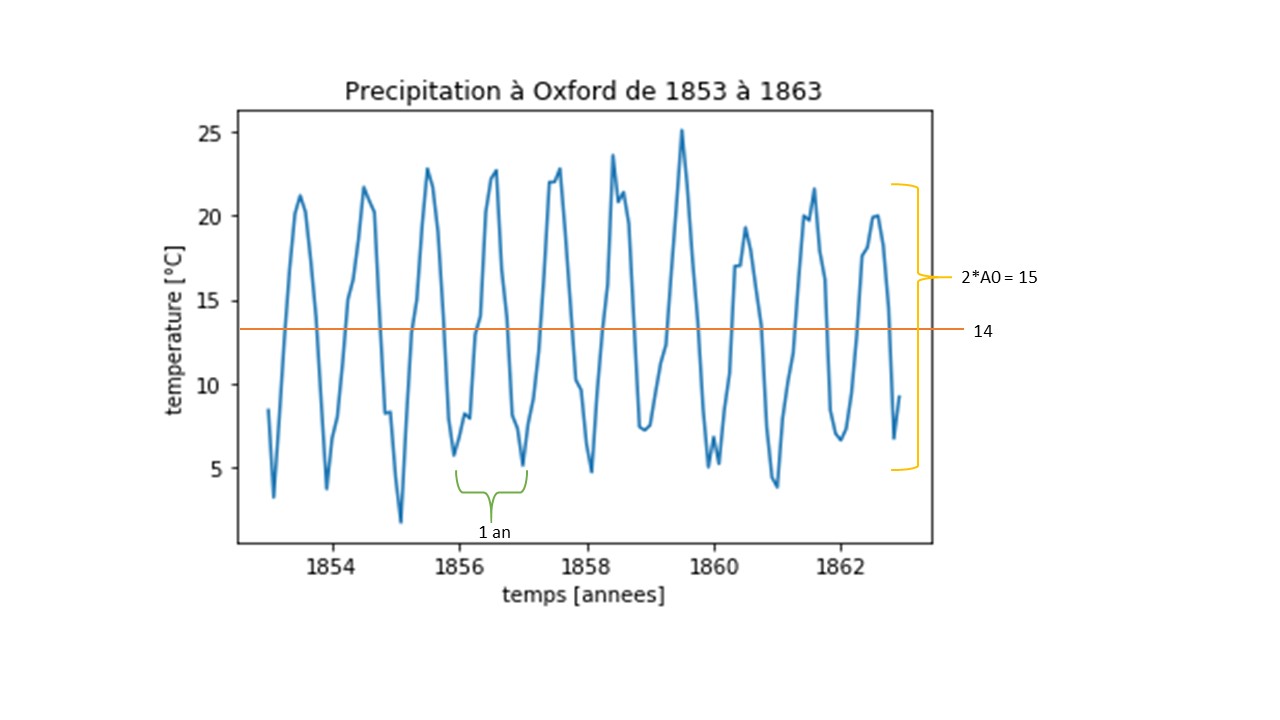


Figure : Choix des paramètres

Pour la phase, nous avons testé avec une condition initiale de 0.

1. Moindres Carrés
2. Basique

Une première estimation par moindres carrés a été faite. Le paragraphe ci-dessous présente les résultats.

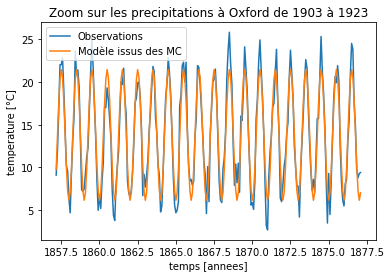
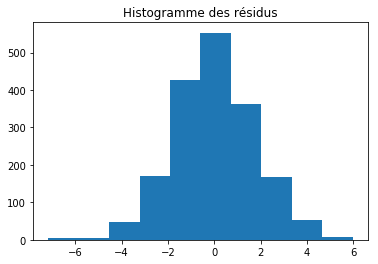


Figure : Zoom sur 20 ans pour mieux voir le modèle issus des moindres carrés basiques et les observations



Sigma0\_2

Test chi2

Var covar paramètre

1. Elimination des points

Explication.

Paramètres obtenus

IMAGE MODELE

HISTOGRAMME

Sigma0\_2

Test chi2

Var covar paramètre

Amelioration par rapport MC normal

1. RANSAC
2. La méthode RANSAC et ses paramètres

La méthode d’optimisation Random sample consensus (RANSAC) est une méthode d’optimisation se basant sur la méthode des moindres carrés appliqués à un ensemble de points tiré aléatoirement. Ensuite, on évalue ce modèle en fonction du nombre de points de l’ensemble total correspondant à ce modèle à un seuil près. Le modèle correspondant au plus de points possibles est alors sélectionné.

Paramètres :

* **Nombre de points de l’échantillon test :** Le nombre de points dans l’échantillon test est choisi par rapport au théorème de Shannon. En effet, le théorème de Shannon, précise que la représentation discrète d'un signal exige des échantillons régulièrement espacés à une fréquence d'échantillonnage supérieure au double de la fréquence maximale présente dans ce signal. Or ici fréquence est d’environ 1/12, si on compte la période en mois. Donc la fréquence d’échantillonnage doit être supérieur à 2/12, soit 1/6. Ainsi au minimum, il faut prendre un point tous les 6 mois. Sachant qu’il n’y a pas de données manquantes, le théorème de Shannon nous indique de prendre 150\*12/6 = 300 points.
* **Nombre d’itération K :** Pour avoir le plus de chance de tomber sur le meilleur modèle, nous avons décidé d’itérer la méthode de RANSAC sur un assez grand nombre de fois. K = 100
* **Seuil de sélection d’un point considéré comme valide au modèle t :** C’est au regard des résidus, que le seuil t a été choisi. Le seuil t a été fixé à 2°C car cela signifie qu’environ 90% des observations sont sélectionnées. De plus par logique c’est normal d’avoir des fluctuations de quelques degrés d’une année sur l’autre.
* **Seuil T :** Ce seuil est fixé à 98% du nombre donné pour validé automatique un tirage qui serait très bon pour générer un modèle le plus proches des données.

1. Résultats

IMAGE MODELE

HISTOGRAMME

Sigma0\_2

Test chi2

Var covar paramètre

1. Conclusion