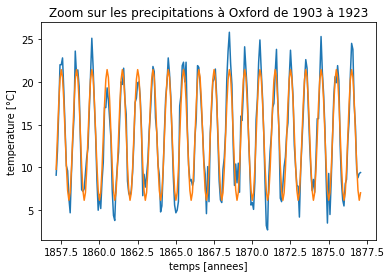


**Projet d’estimation robuste**

Moindres Carrés et RANSAC

Sur les températures maximales

à Oxford entre 1853 et 2003



30 mars 2018

Mannaïg L’Haridon et Iris de Gelis

Table des matières

[**I.** **Présentation des données** 3](#_Toc510203374)

[**1.** **Les données** 3](#_Toc510203375)

[**2.** **Le modèle** 3](#_Toc510203376)

[**3.** **Conditions initiales** 4](#_Toc510203377)

[**II.** **Moindres Carrés** 4](#_Toc510203378)

[**1.** **Estimation simple** 4](#_Toc510203379)

[**2.** **Estimation avec élimination des points faux** 5](#_Toc510203380)

[**3.** **Comparaison des méthodes** 7](#_Toc510203381)

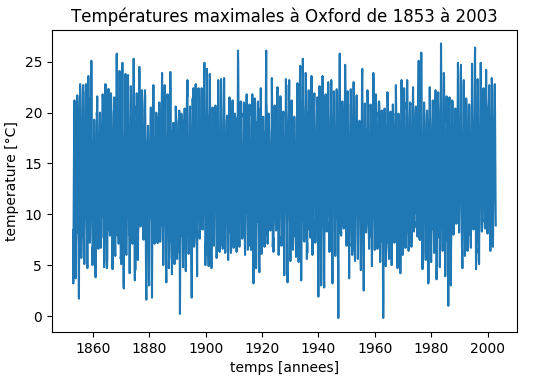
[**III.** **RANSAC** 7](#_Toc510203382)

[**1.** **La méthode RANSAC et ses paramètres** 7](#_Toc510203383)

[**2.** **Résultats** 8](#_Toc510203384)

[**IV.** **Conclusion** 9](#_Toc510203385)

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 d’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.

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

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.

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. Les températures sont exprimées en degrés Celsius [°C].

1. **Le modèle**

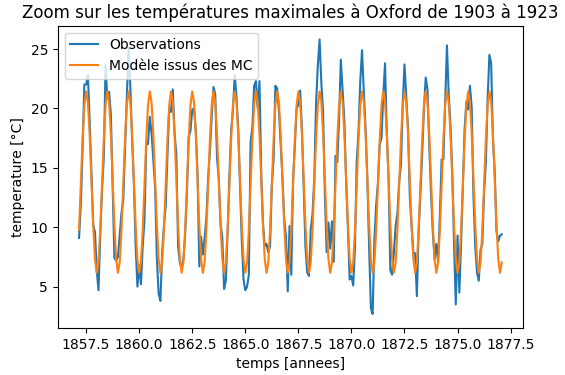
En sélectionnant juste 20 ans de données de façon arbitraire, on remarque que les observations sont périodiques – de période 12 mois soit 1 an – et adoptent un mouvement quasi-sinusoidal.

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**

Les conditions initiales du modèle choisi ont été optées 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 graphique.
* **Période** : La période des oscillations a priori est d’un an, donc T = 1 an. Or ω = 2\*π/T. Ainsi, nous avons choisi ω = 2\* π.
* **Constante** : 14 par lecture graphique.
* **Phase**: fixée à 0

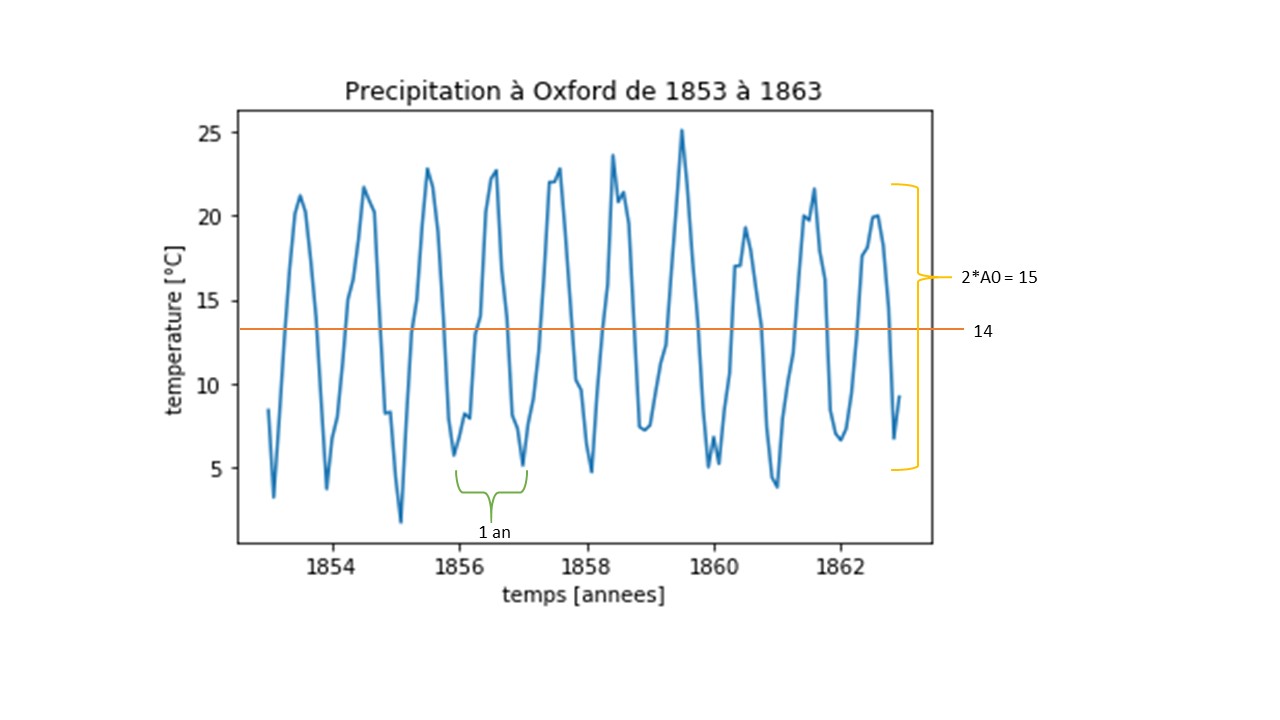
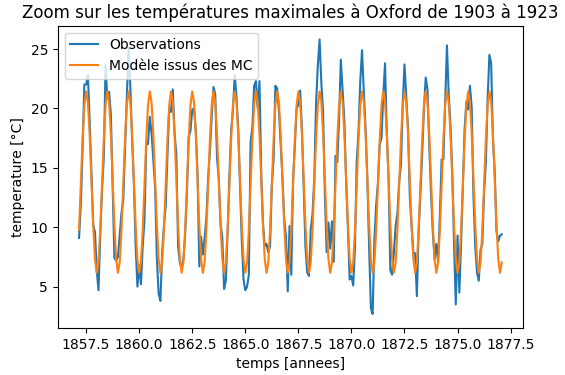


Figure 3 : Choix des paramètres

1. **Moindres Carrés**
2. **Estimation simple**

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

**Paramètres estimés** :

* A = -7.638
* ω = 6.283
* ϕ = 0.857
* cste = 13.798

Figure 4: Zoom sur 20 ans pour mieux voir le modèle issus des moindres carrés basiques et les observations

Tout d’abord, on peut constater que les paramètres estimés sont assez proches des conditions initiales estimées visuellement. La figure 4 montre un zoom sur les années 1903 à 1923 des observations (en bleu) et du modèle issus des moindres carrés basique (en orange). On peut remarquer que le modèle semble assez bien coller aux oscillations des observations. L’amplitude effective semble correcte car ne semble pas prendre les mesures « fausses ».

Nous obtenons une variance de 3,108 au bout de 4 itérations où le critère d’arrêt est la stabilisation de la variance à 10-6.

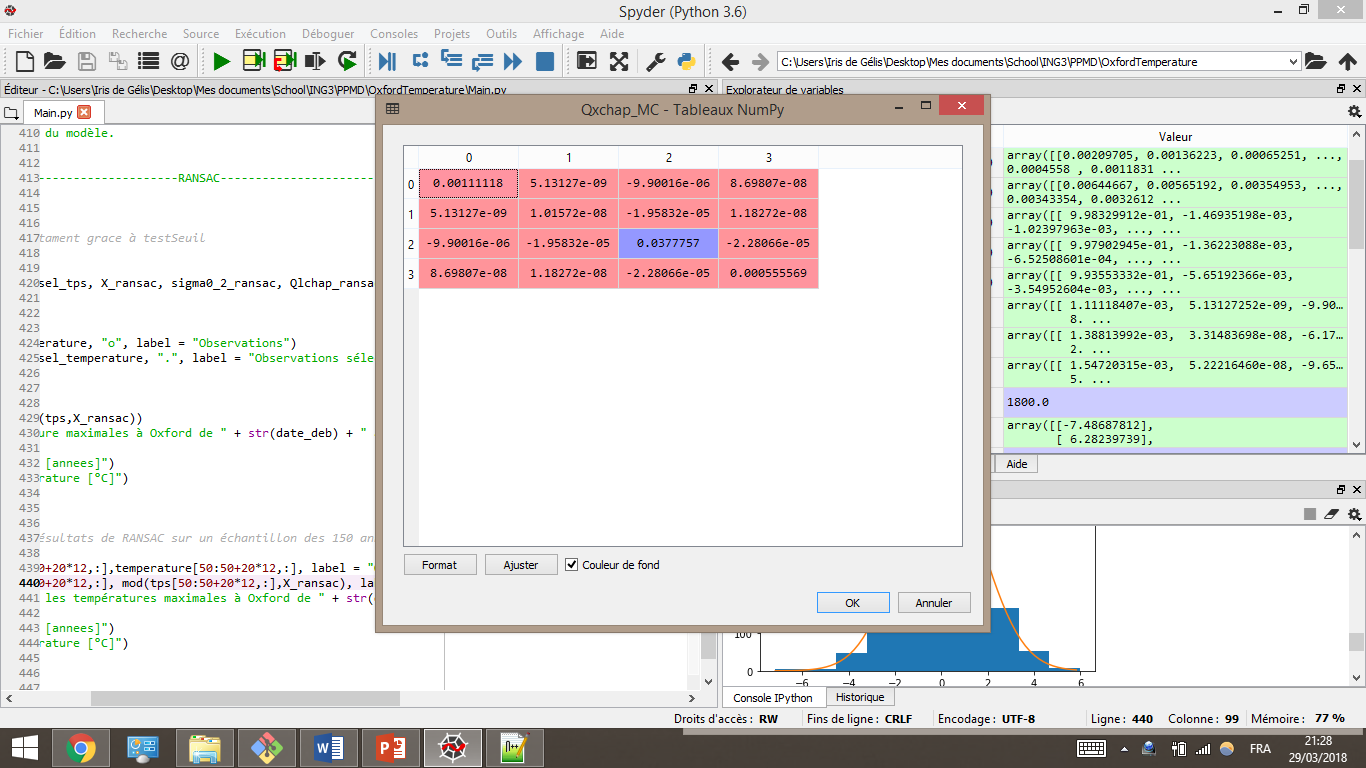
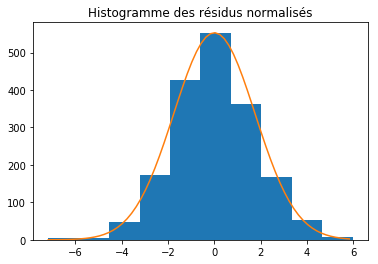
Matrice de variance-covariance des paramètres: on peut voir grâce à cette matrice que les quatre paramètres du modèle sont bien décorrélés entre eux, ils ont donc tous une réelle importance. Le modèle semble cohérent.

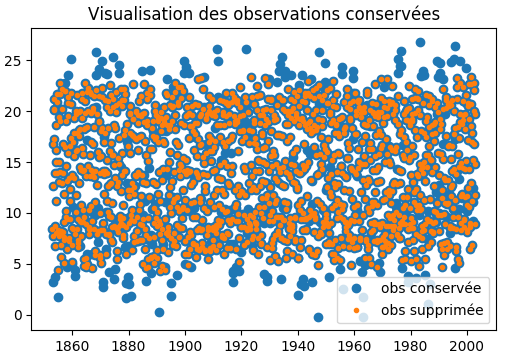
Figure 5 : Matrice de variances-covariances

Afin de valider le modèle, l’histogramme des résidus normalisés a été tracé. Comme on peut le voir ci-contre, l’histogramme montre grâce à la distribution gaussienne qui lui est superposée, que les résidus suivent bien une loi Normale réduite et centrée en 0.

Figure 6 : Histogramme des résidus normalisés pour des moindres carrés simples

Un test du χ² a ensuite été effectué. Pour être validé, le facteur unitaire de variance doit être compris dans l’intervalle [0.936,1.066], or , par conséquent le test du χ² ne peut pas être validé. Cela peut s’expliquer par le fait que les observations n’ont pas été corrigées des points faux.

1. **Estimation avec élimination des points faux**

Cette seconde estimation des moindres carrés a été effectuée en éliminant les points faux des observations afin d’obtenir de meilleurs résultats qu’auparavant. Il a été choisi de conserver les observations dont la valeur absolue des résidus est inférieure au critère .

La figure 6 représente l’ensemble des observations qui ont été conservées en orange ainsi que celles qui ont été supprimées en bleu au bout de 10 itérations où le critère d’arrêt est l’absence de détection de nouveaux points faux. On peut observer que les points conservés sont bien situés dans une zone bien définie et que certains points dans cette zone rectangulaire ont été tout de même supprimés.

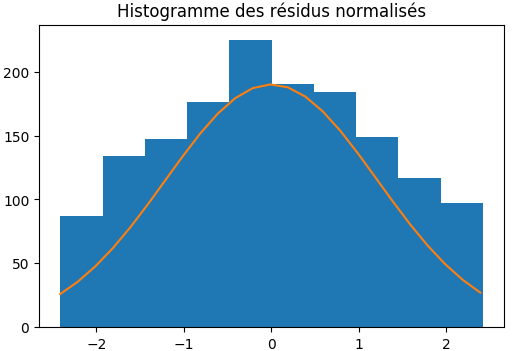
Figure 7 : Observations conservées ou non en fonction du critère à .

Les paramètres estimés ci-dessous montrent que les paramètres dépendent bien des valeurs fausses puisqu’elles varient, surtout la phase, selon si des points faux sont enlevés ou non.

**Paramètres estimés** :

* A = -7.365
* ω = 6.282
* ϕ = 1.234
* cste = 13.786

Nous obtenons une variance de 1,463 au bout de 4 itérations où le critère d’arrêt est la stabilisation de la variance à 10-6 et de 10 étapes de suppression des points faux jusqu’à absence de détection de nouveaux points faux. On remarque alors que 16,33% des observations sont alors considérées comme étant fausse alors que le critère des est respecté. Cependant la règle des trois sigmas n’est alors plus respectée étant donné que pour ce critère on est censé conserver environ 95% des observations, et non 84,77%.

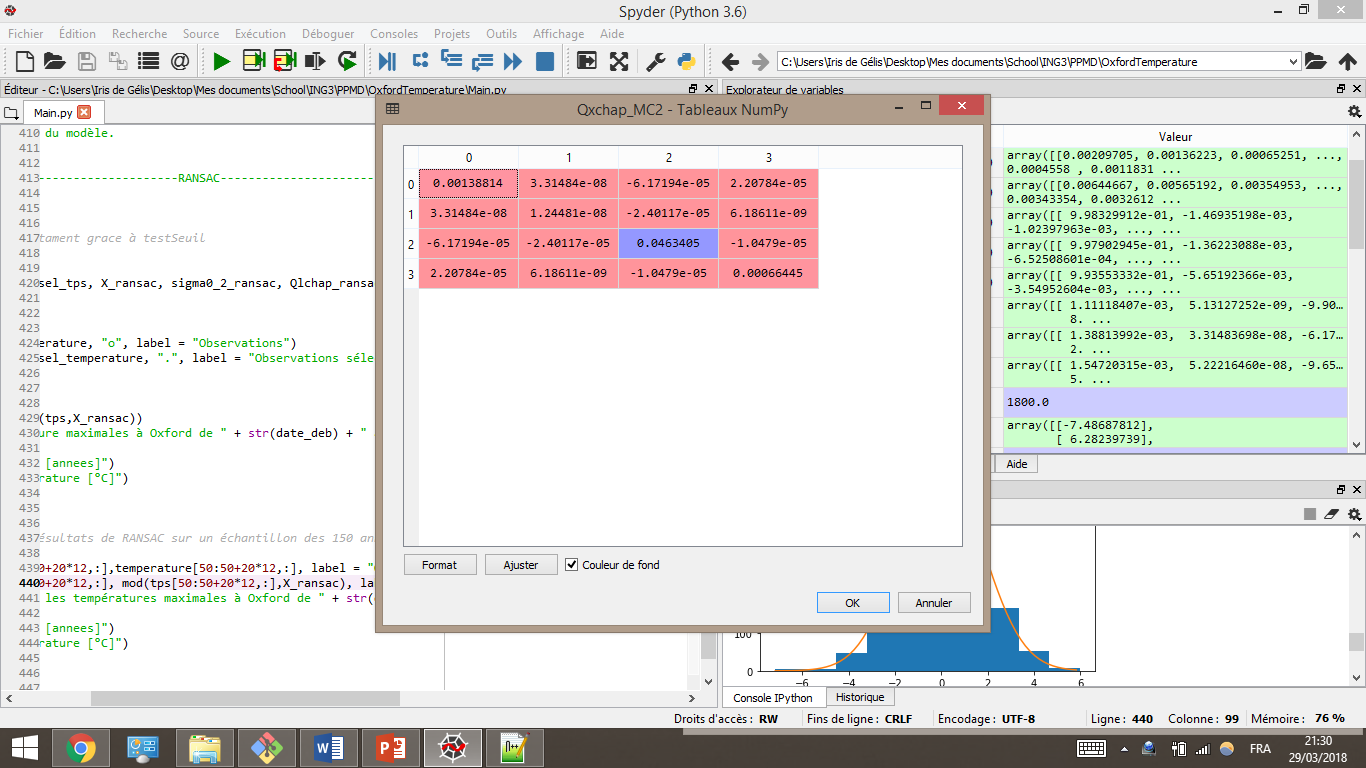


Afin de détecter d’éventuelles erreurs, l’histogramme des résidus normalisés a été tracé. La distribution gaussienne qui est superposée à l’histogramme montre que les résidus suivent bien une loi Normale réduite et centrée en 0. Cependant les résidus dépassent cette la courbe de la gaussienne et il y a une trop forte concentration des résidus aux bords de l’histogramme.

Figure 8 : Histogramme des résidus normalisés pour les moindres carrés avec suppression de points faux

Pour que le test du χ² soit validé sur ces valeurs, le facteur unitaire de variance doit être compris dans l’intervalle [0.843,1.169]. Or , par conséquent le test du χ² ne peut pas être validé mais on s’en rapproche beaucoup plus.

Par ailleurs, on a remarqué que lorsque le critère des est respecté mais que l’on opère une seule opération de détection des points faux, la règle des trois sigmas est cette fois-ci respectée (atteinte de … %). Cependant le facteur de variance obtenu est moins bon : 2,329. De plus, les paramètres estimés sont quasiment similaire à ceux obtenus sans suppression des points faux.

Pour que le test du χ² soit validé sur ces nouvelles valeurs, le facteur unitaire de variance doit être compris dans l’intervalle [0.713, 1.335]. Or , par conséquent le test du χ² ne peut pas être validé.

Var covar paramètre :

1. **Comparaison des méthodes**

On remarque que l’élimination des points faux permet de réduire le facteur de variance avec la méthode des carrés. Les paramètres estimés changent peu d’une estimation à l’autre, la plus grande variation étant de 0.377 pour la phase.

Quelle que soit l’estimation, le test du χ² n’est pas validé. Cela peut-être dû au fait que l’on étudie des températures maximales et non moyennes : les températures maximales dépendant de la saison, mais aussi d’autres facteurs tels que la pollution de l’air qui n’est pas un phénomène périodique, les valeurs ne seront donc pas exactement identiques d’une année. Cela pourrait donc expliquer qu’il y a de nombreux résidus pour et donc la raison pour laquelle l’histogramme des résidus normalisés n’est pas complètement bien étalé.

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 appliquée à un ensemble de points tirés aléatoirement. Ensuite, on évalue le modèle en fonction du nombre de points de l’ensemble total correspondant au modèle à un seuil près. Le modèle correspondant au plus de points possibles est alors sélectionné.

Choix des 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 la 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, il faut prendre au minimum 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 2,5°C car cela signifie qu’environ 90% des observations sont sélectionnées. De plus, il est logique et normal d’avoir des fluctuations de quelques degrés d’une année sur l’autre étant donné qu’il s’agit des températures maximales par mois.
* **Seuil T :** Ce seuil est fixé à 98% du nombre donné pour valider automatique un tirage qui serait très bon afin de générer le modèle le plus proches des données.

1. **Résultats**

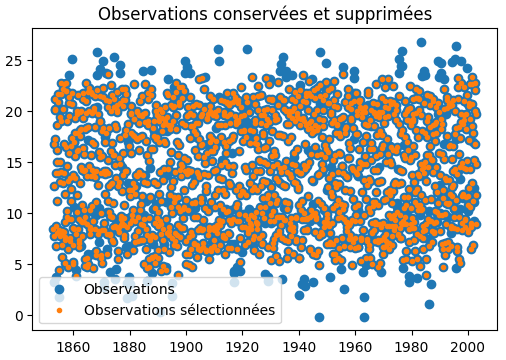
IMAGE MODELE

HISTOGRAMME

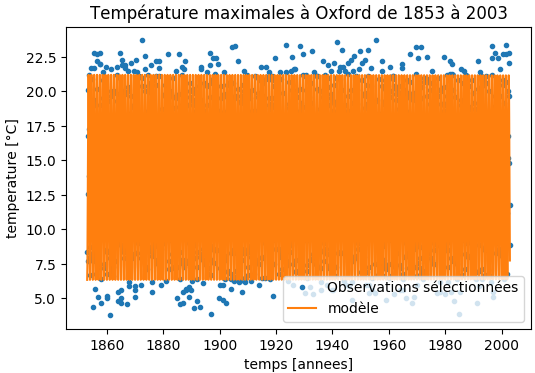
Sigma0\_2

Test chi2

Var covar paramètre



RQ : Ransac semble une bonne éthode mais le choix des paramètres est très arbitraire et influence beaucoup le nombre de pints éliminés.



1. **Conclusion**