Air pollution Lucknow

LIAO Puwei
RAOELISON Finaritra

February 2021

1 Introduction

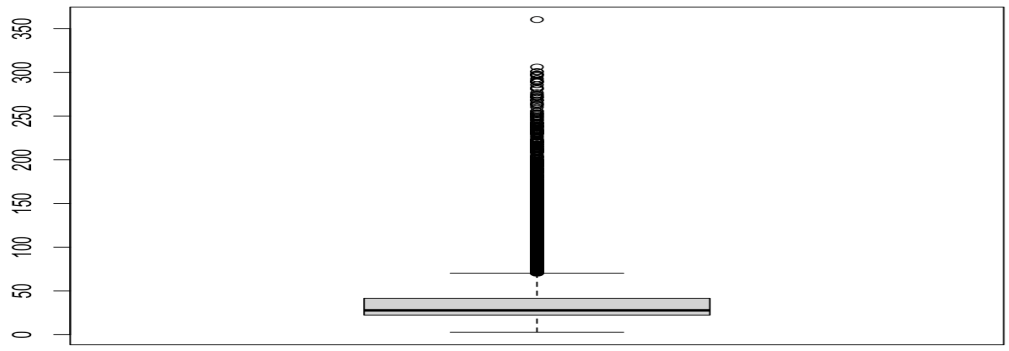
L'air est ce qui maintient les humains en vie. Le surveiller et comprendre sa qualité est d'une importance capitale pour notre bien-être, mais aussi un indicateur sur le niveau de pollution d'une ville. Il existe plusieurs centre de données qui récupèrent ces informations pour différentes villes dans le monde. Pour notre part, le jeu de données que nous avons choisis contient des données sur la qualité de l'air dans plusieurs villes indiennes. La qualité de l'air y ai déterminé en fonction de la quantité de certains gaz. Nous pouvons citer: le dioxyde d'azote, dioxyde de carbone ou encore l'oxyde d'azote. C'est ce dernier (NOx) que nous allons étudier plus en détail. L'oxyde d'azote désigne un ensemble de gaz dont le point commun est leur contenance en azote et dioxyde, mais en différentes quantité. Il est intéressant d'observer les variations de ces gaz car il regroupe les principaux polluants atmosphériques qui sont le monoxyde d'azote (NO) et le dioxyde d'azote(NO2). Nous avons donc choisis d'observer les données pour la ville de Lucknow qui est une des villes les plus polluées d'Inde. Les principales sources de pollutions dans cette ville sont les suivantes: la proxmité avec plusieurs usines, les gaz dus aux véhicules, les feus déclenchés pour se débarrasser des déchets ou encore l'expansion urbaine de la ville. Nous allons tout d'abord étudier cette série temporelle de NOx par heure à Lucknow sur une période qui s'étend de 2017 à 2018,et ensuite on cherchera une modélisation pour faire la prédiction de l'année 2019 pour, qu'on pourra alors commparer avec les valeurs réelles. Une fois que ce modèle aura été réussi,on est capable de l'appliquer à la gestion de la qualité de l'air de Lucknow.

2 Construction de Série temporelle

Pour étudier la série temporelle sur la quantité de NOx par heure à Lucknow , on récupère un jeu de donnée - trouvé à l'adresse suivante (https://www.kaggle.com/rohanrao/air-quality-data-in-india). Les données vont de Janvier 2015 à Avril 2020. On extrait les données de la ville du fichier city_hour.csv, et on construit deux data set; un que nous allons manipuler et l'autre qui nous servira de comparaison. Tout d'abord le sous ensemble qui s'étend du 26/09/2017 au 31/12/2018 qui s'appelle data1718, puis le sous ensemble de toute 2019 qui s'appelle data19 pour vérifier notre modèle série temporelle.

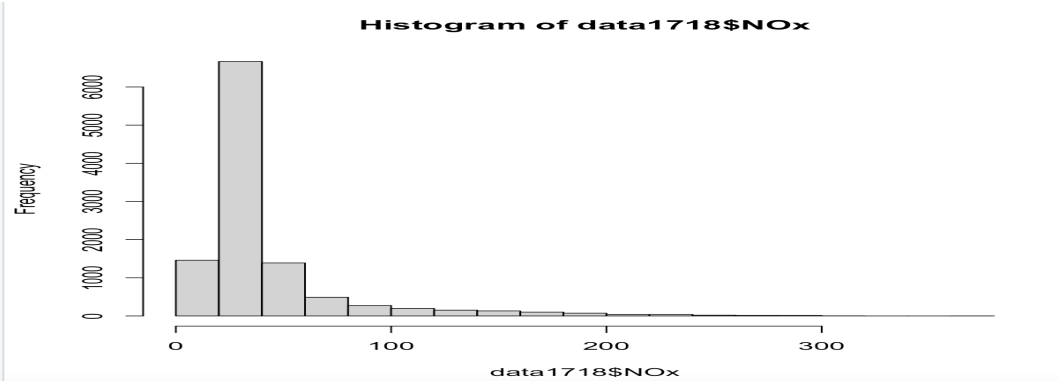
3 Analyse descriptive

Nous allons observer la quantité horaire de NOx à Lucknow entre 2017 et 2018, à travers différents graphiques. Sur cette période nous possédons 11070 observations horaires, et la quantité de NOx est quantifié en ppb (Partie par milliard). Nous allons effectuer une analyse descriptive de nos données à l'aide de quelques graphiques, qui nous permettront de voir la répartition de nos données ainsi qu'une idée des paramètres qui influent dessus.
Pour observer la répartition de la quantité de NOx nous allons construire la boîte à moustache de nos données.



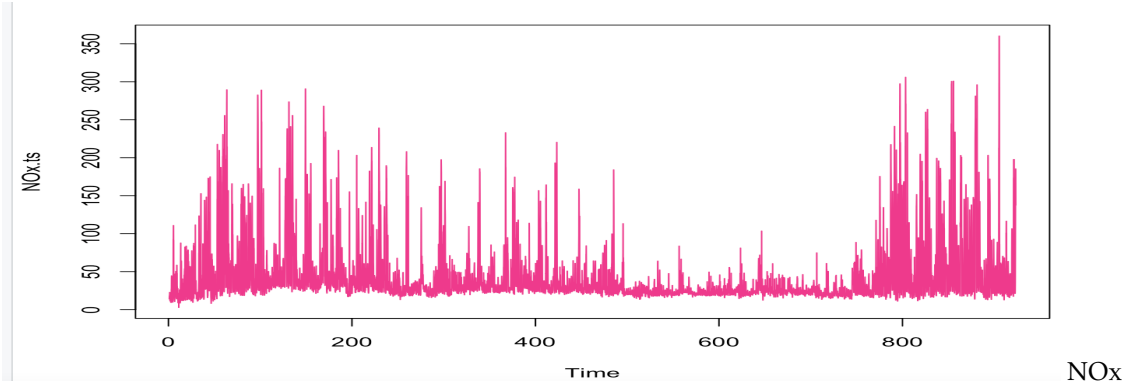
Box-plot de NOx entre 09/2017 et 12/2018

Nous remarquons qu'on a des données légèrement asymétriques par rapport à la médiane mais surtout plusieurs valeurs aberrantes dont nous essaierons de trouver la cause.



Histogramme de NOx entre 09/2017 et 12/2018

Nous remarquons ici que la quantité de NOx varie surtout entre 20 et 40ppb.
Visualisons maintenant l'ensemble de notre série temporelle. On affiche sur le graphique suivant les données à une fréquence de 12h.



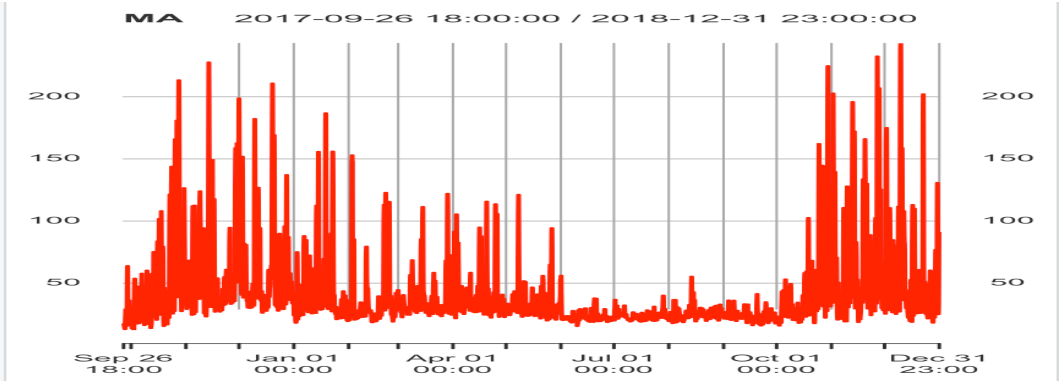
NOx dans l'air pour une fréq=12heures

On remarque que la quantité de NOx a connu une baisse bien conséquente durant plusieurs mois, entre juin et septembre 2018, avant d'atteindre des pics. Cela peut s'expliquer par la saison des moussons, qui se produit à cette période-ci dans cette région de l'Inde. Il semble que ce soit un facteur influant de la qualité de l'air.

4 Tendance

Dans cette partie, pour visualiser plus clairement notre série temporelle,on cherche à étudier sa tendance avec 4 méthodes très typiques: estimation de tendance par moyenne mobile, estimation de tendance par régression linéaire, estimation de tendance par noyau, estimation de tendance par polynômes locaux.

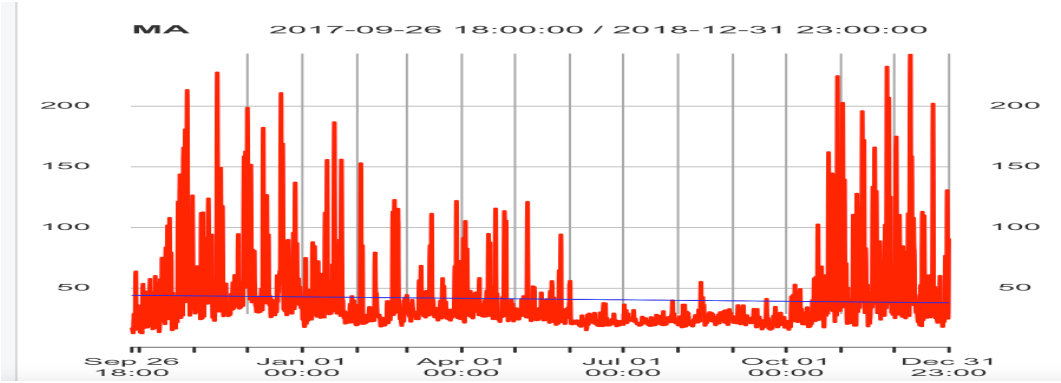
4.1 Tendance par moyenne mobile



Estimation de tendance par moyenne mobile

4.2 Tendance par régression linéaire

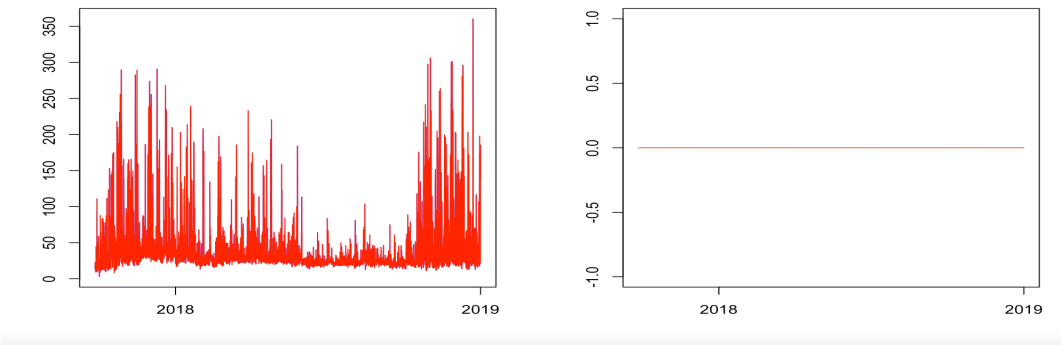
On utilise une méthode par régression linéaire pour estimer la tendance ce qui est le segment bleu sur ce dessin.



Estimation de la tendance de la série par régression linéaire

4.3 Tendance par noyau gaussien

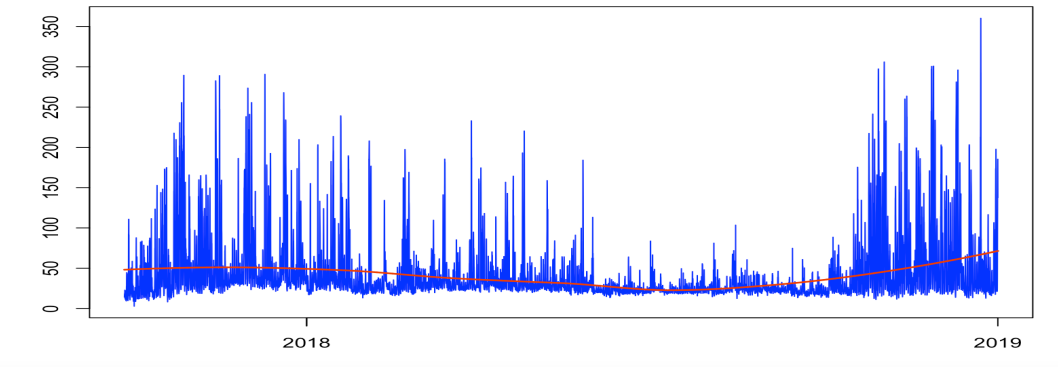
D'après la méthode de noyau gaussien, on voit clairement que la partie bleu(série temporelle) est couvert entièrement par la partie bleu(estimation de noyau),et leurs différence s'annule partout.



Estimation de la tendance de la série par noyau gaussien

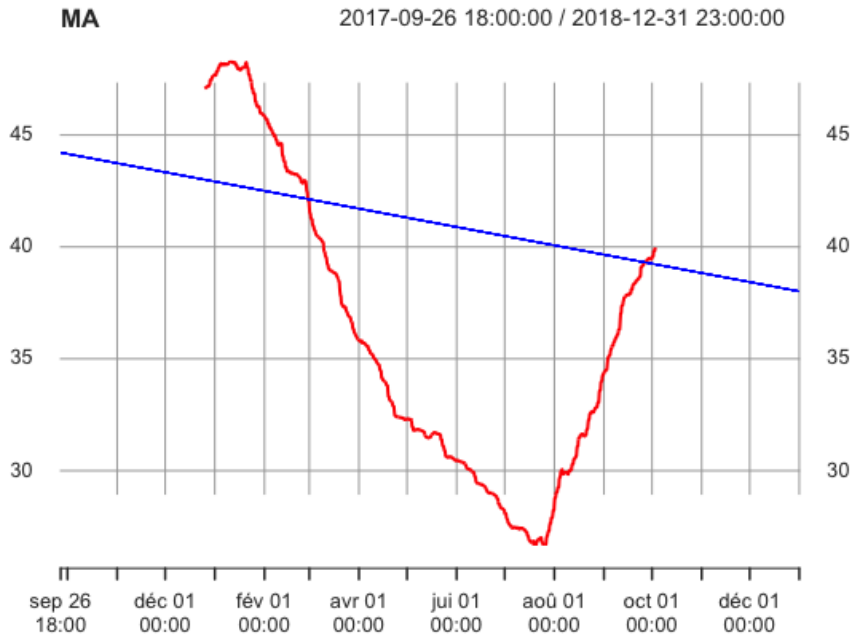
4.4 Polynômes locaux

Utilisons une méthode par polynômes locaux, nous voyons bien que la partie rouge c'est l'estimation de tendance par cette méthode et la partie bleue c'est notre série temporelle.



Estimation de la tendance de la série par polynômes locaux

Regardons plus précisémment l'étude de la tendance pour les méthodes de moyenne mobile et régression linéaire,pour une période d'un an.

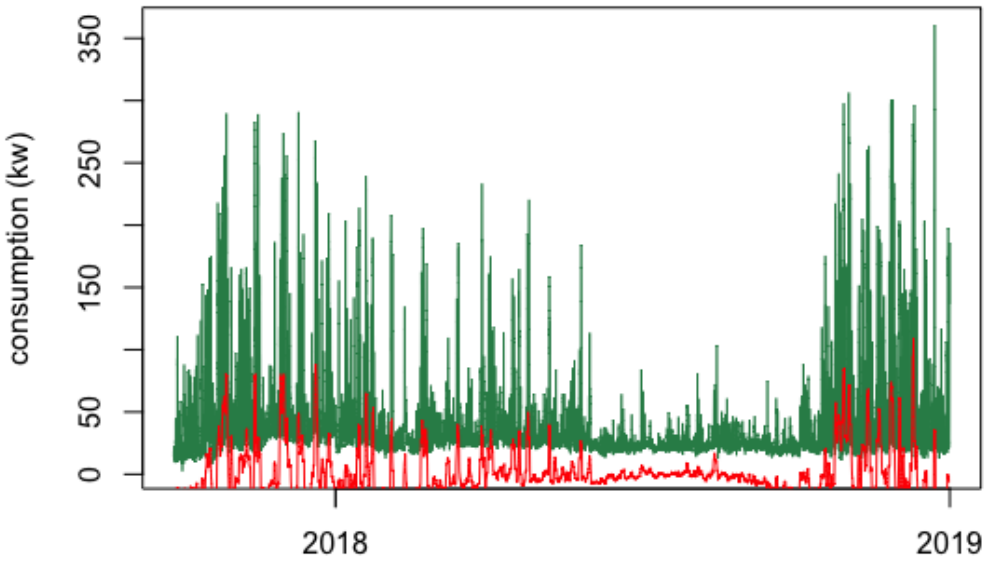


Estimation de la tendance par MA et regression linéaire pour periode d'un an

On remarque bien ici, que la quantité de NOx dans l'air a tendance a diminuer jusqu'au mois de septembre avant de remonter.

5 La Saisonnalité

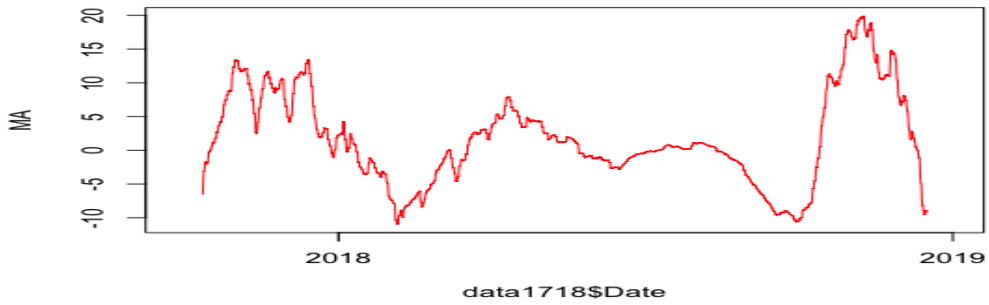
On s'intéresse alors la saisonnalité, ici nous estimons la saisonnalité par méthode de moyenne mobile.



On a en rouge notre série temporelle à laquelle on a retiré la tendance qu'on a obtenu par la méthode des polynômes locaux. C'est donc sur ces données que nous allons étudier la saisonnalité.

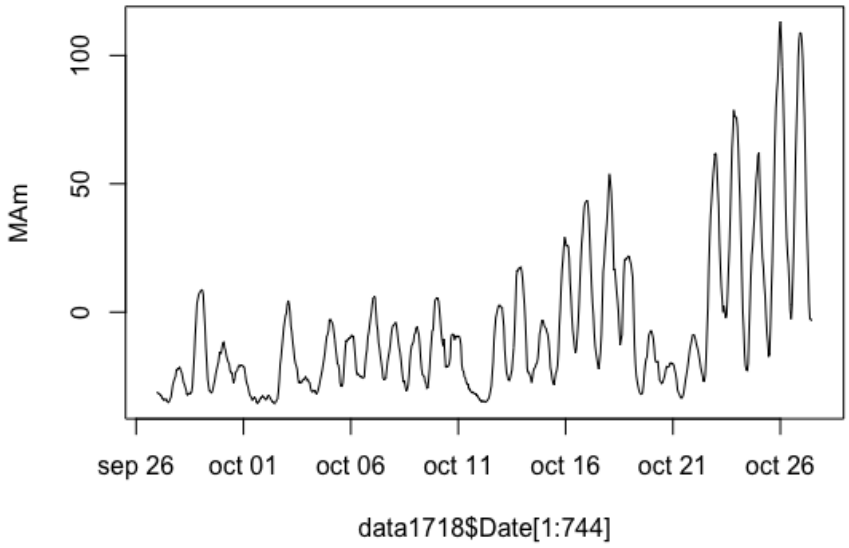
5.1 mobile moyenne

Comme pour la tendance on peut observer la saisonnalité de données par la méthode de la moyenne mobile. Nos données d'étendant sur un peu plus d'un an, on choisis une période de 744h ce qui correspond à un mois.

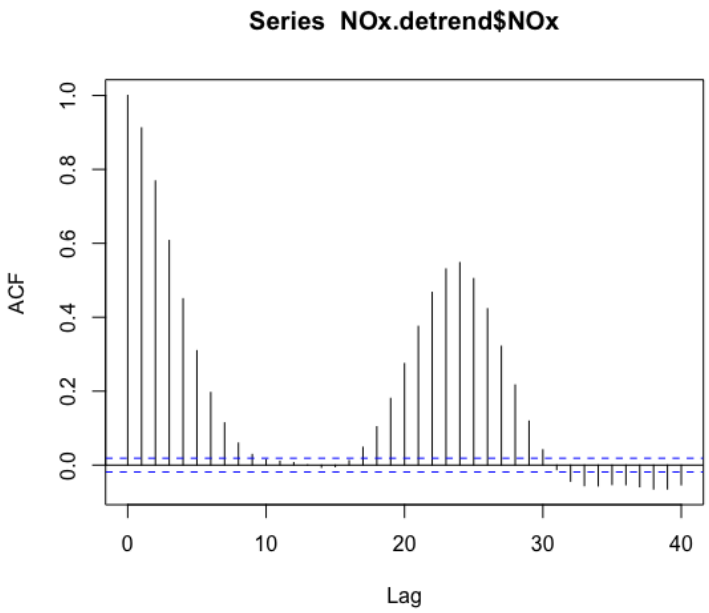


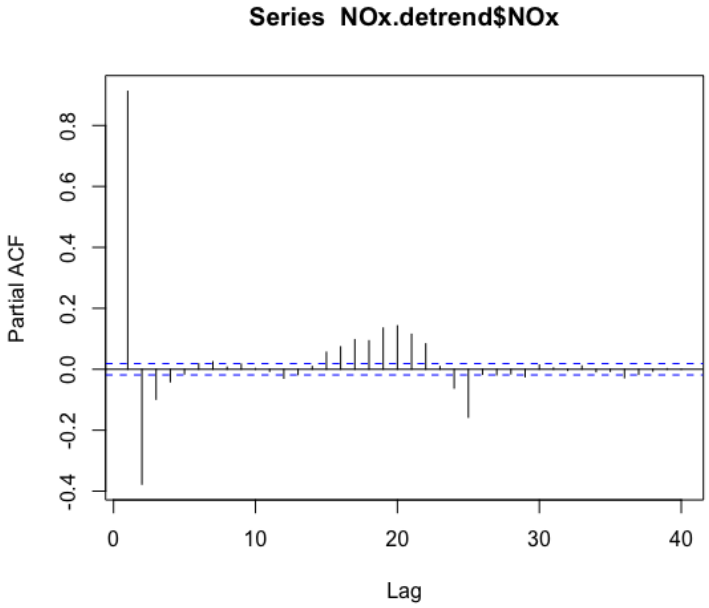
Estimation de la saisonnalité de la série par moyenne mobile

On ne voit pas clairement un schéma qui se répète mais on retrouve la baisse qui correspond à la saison des moussons. Regardons d'un peu plus près ce qu'il se passe au niveau d'une semaine. Toujours par la même méthode, on récupère les données pour le premier mois. On obtient la figure suivante.



On a choisis ici une période d'une demi-journée. Le 26 septembre 2017 était un mardi, on remarque que sur le reste du mois on observe une augmentation de NOx au milieu de semaine,avant d'en voir une diminution en fin de semaine.
Regardons la stationnarité de nos données sans tendance.





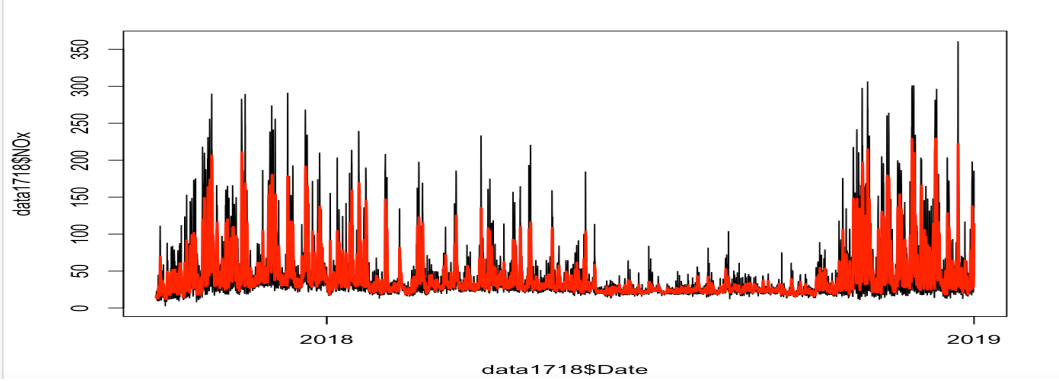
On remarque sur les auto-corrélogramme qu'on semble avoir une auto-corrélation de nos données pour un déclage de 1 à 4 heures. Ce qui est logique puisqu'on a plutôt d'après ce qu'on a vu précédemment des différences plus significatifs du point de vu des journées entières.

6 Lissages exponentielles

Jusqu'à ici, on voudrait pour ce moment faire une prévison de 2019 en comparant avec les vrais données de 2019 pour justifier notre modèle série temporelle,pour faire cela, on propose les méthodes de lissages exponentielles,ces méthodes consistent à ajuster à une chronique de série temporelle une estimation locale de ce que va être sa valeur future

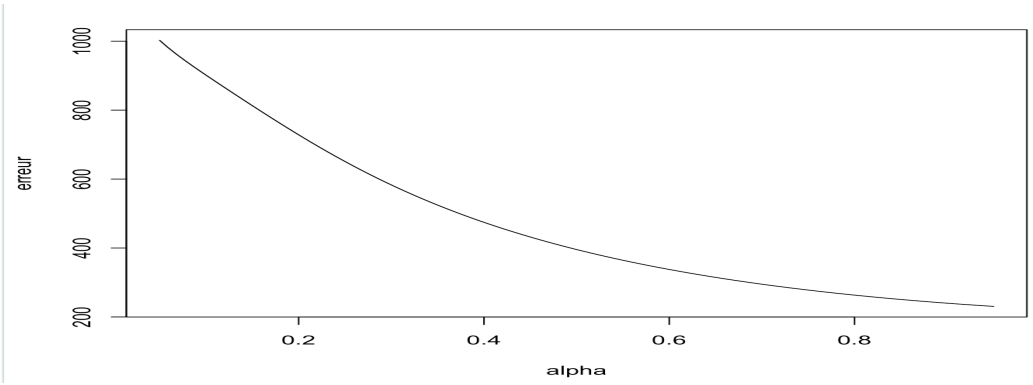
6.1 Lissage exponentiel simple

Voici le lissage simple avec alpha=0.2:



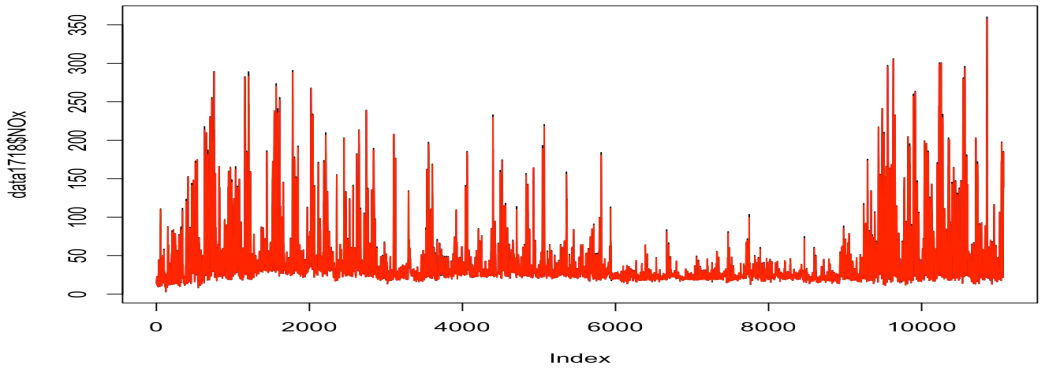
exponentiel simple

Si, pour ce moment, on choisit alpha de 0.05 à 0.95, on va voir optimiser alpha, c'est à dire le alpha qui renvoie l'erreur plus petit:



erreur

En mettant le alpha optimal, on obtient le dessin suivant

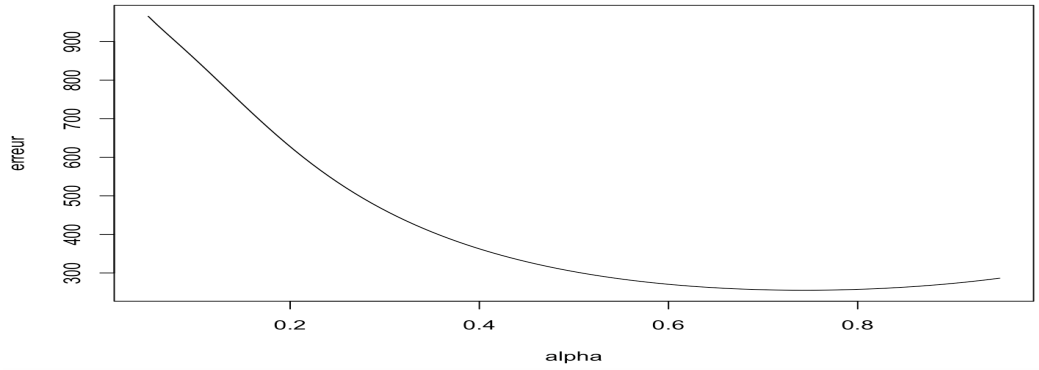


lissage simple avec alpha optimal

Mais dans ce cas, notre prévision va être très local.

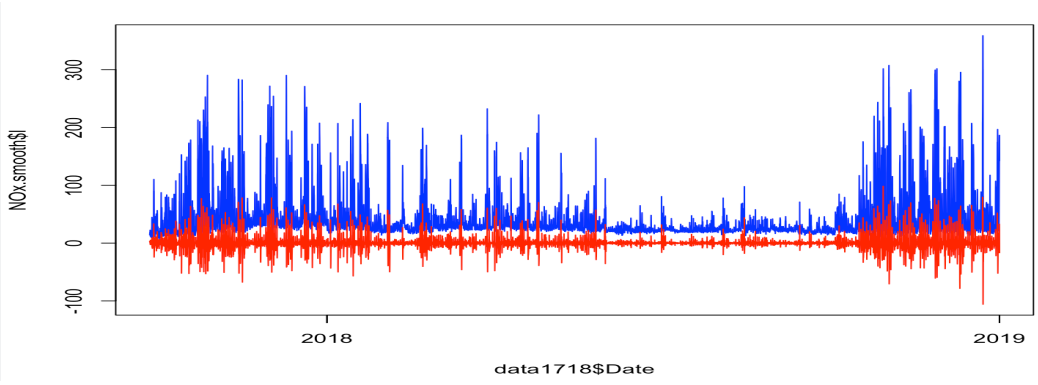
6.2 lissage exponentiel double

Passons à une autre lissage qui s'appelle lissage exponentiel double, l'idée est d'ajuster une droite au lieu d'une constante dans l'approximation locale de la série. Tout d'abord on cherche le alpha optimal entre 0.05 et 0.95:



erreur de alpha

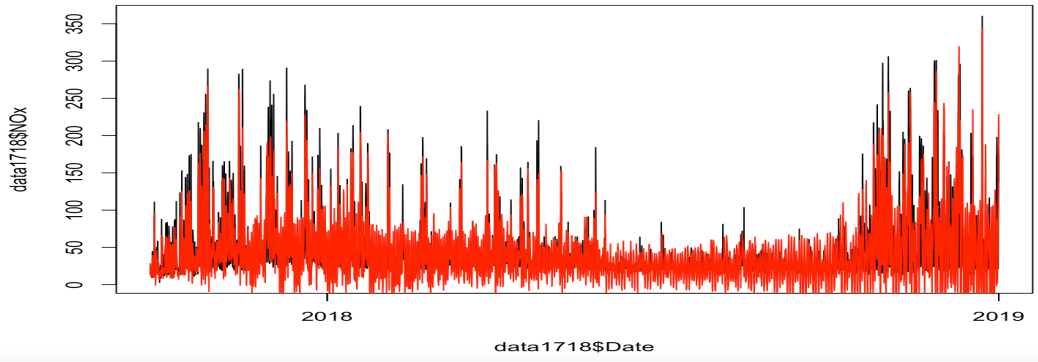
En prenant ce alpha optimal, affichons ce lissage double avec notre série temporelle:



lissage double avec alpha optimal

6.3 Lissage exponentiel de Holt-Winters

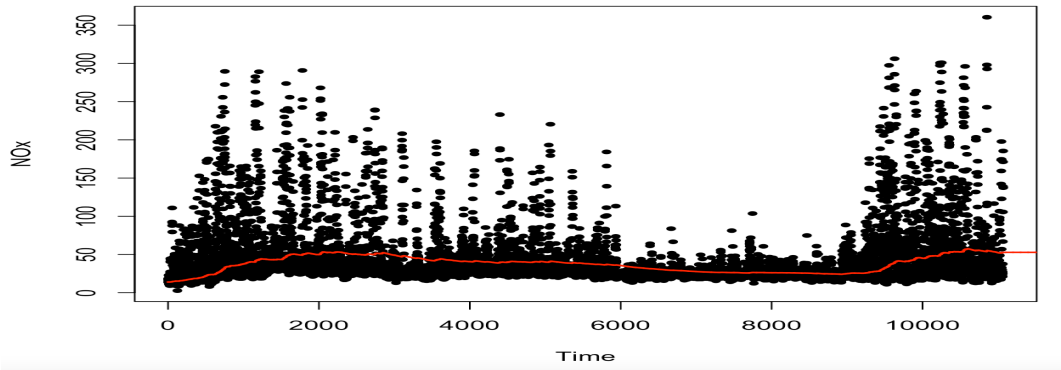
Ici on utilise holt-winters ce qui est une généralisation du lissage double, qui permet entre autre de proposer les modèles suivants: tendance linéaire locale ,tendance linéaire locale + saisonnalité et tendance linéaire locale * saisonnalité , et ici alpha= 0.2, beta=0.2, delta=0.2, et la période T=168h ce qui est une semaine:



lissage holt winters

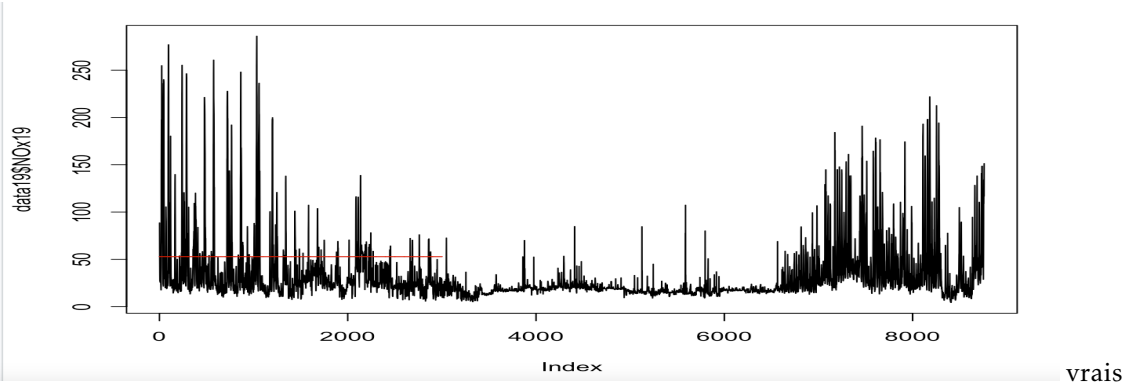
7 prévision

Dans cette partie on essaye de faire une prédiction à 2019(3000 heures au début de 2019),dans notre cas, on considère ce modèle comme une constante avec un bruit , donc ici on utilise plutôt lissage simple pour faire notre prédiction en choissisant un alpha très petit qui coïncide bien avec les vrais données de 2019:



prévision avec lissage simple

On voit bien que avec ce alpha petit ce modèle fonctionne bien pour la prévision en débarrasser le bruit:



vrais données de 2019 avec la prévision

8 Conclusion

Nous avons pu modéliser nos données de quantité de NOx dans l'air au sein de la ville de Lucknow en Inde. On a effectué une analyse despcriptive de données, on a pu voir qu'il y avait une tendance puisque la quantité de NOx diminue avant de remonter jusqu'à un certain pique. L'étude de la saisonnalité nous a montré que les variations dépendaient de la période de l'année, entre autre la saison des moussons où là la baisse est significative. Mais aussi lorqu'on observe les données à la semaine on a des quanttés moins élevées lors des week-end. En vue d'effectuer une précvision pour l'année 2019 et comparer avec les valeurs réelles, nous avons utiliser différentes méthodes de lissage de nos données. Malheureusement nous avons rencontré des difficultés à ce niveau là et la prévision que nous avons n'est pas celle que nous attendions.