Etude de données météorologiques à Bâle, Suisse

Charles Medous
Aristote Koen

Février 2020

1 Introduction

1.1 Description de la source

Les données météorologiques proviennent de la plate-forme Meteoblue, et sont disponibles librement. Issue de la recherche académique à l'Université de Bâle, en Suisse, Meteoblue a été fondée en tant que société en 2006 et propose en plus des données brutes, des modèles prédicatifs. L'accès aux données plus spécifiques (enneigement, rayonnement solaire,...) étant payant, nous nous sommes concentré sur les données de température par heure sur 20 ans.

1.2 Mise en forme des données

Les données brutes sont récupérées au format .csv et présentent d'autres informations que la température, telles que la vitesse du vent ou la pression. Les données sont récupérées toutes les heures depuis le 9 Février 2000.

	A	B	C	D	E	F	G	H	I	J
1	Year	Month	Day	Hour	Minute	Temperature	Mean Sea Level Pressure	Total Precipitation	Wind Speed	Wind Direction
2	2000	2	9	0	0	7.33	1020.60	NaN	38.53	274.79
3	2000	2	9	1	0	6.48	1021.60	NaN	37.21	275.32
4	2000	2	9	2	0	5.88	1022.70	NaN	36.76	272.82
5	2000	2	9	3	0	5.52	1023.40	NaN	35.71	269.13
6	2000	2	9	4	0	5.26	1024.10	NaN	35.79	266.12
7	2000	2	9	5	0	5.02	1024.70	NaN	35.63	263.07
8	2000	2	9	6	0	4.84	1025.00	NaN	35.58	260.00
9	2000	2	9	7	0	4.79	1025.80	NaN	36.37	259.09
10	2000	2	9	8	0	4.70	1026.30	NaN	37.40	259.03
11	2000	2	9	9	0	4.86	1027.20	NaN	38.72	259.41
12	2000	2	9	10	0	5.04	1028.00	NaN	40.47	261.89
13	2000	2	9	11	0	5.10	1028.40	NaN	41.61	264.39
14	2000	2	9	12	0	5.50	1028.60	NaN	41.86	265.72
15	2000	2	9	13	0	5.87	1028.70	NaN	40.46	266.90
16	2000	2	9	14	0	6.05	1029.10	NaN	39.42	267.85
17	2000	2	9	15	0	6.02	1029.30	NaN	38.09	267.42
18	2000	2	9	16	0	5.92	1029.50	NaN	35.81	265.75
19	2000	2	9	17	0	5.73	1030.00	NaN	32.91	263.32
20	2000	2	9	18	0	4.87	1030.70	NaN	29.98	261.76
21	2000	2	9	19	0	5.16	1030.70	NaN	28.87	276.87
22	2000	2	9	20	0	5.56	1030.80	NaN	25.66	271.94
23	2000	2	9	21	0	5.59	1031.30	NaN	23.34	266.94
24	2000	2	9	22	0	5.58	1031.60	NaN	22.85	262.14
25	2000	2	9	23	0	5.50	1032.00	NaN	19.05	257.68

FIGURE 1 – Données brutes

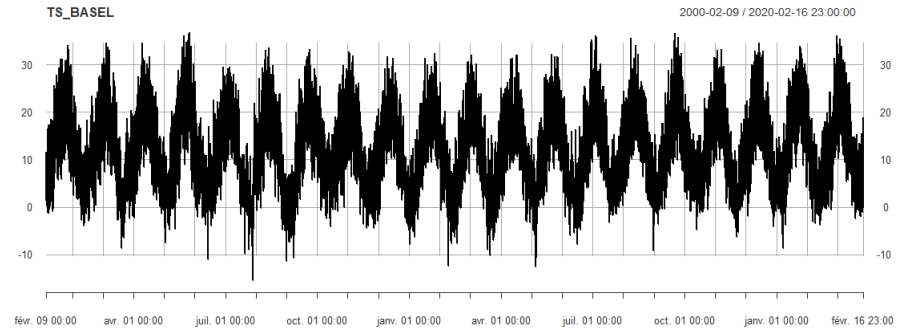
Pour récupérer la date sous un format utilisable on crée une variable `donnees$paste`. On peut ensuite convertir au format `xts` nos observations de températures. on utilise le code suivant :

```
# Importation des données
donnees =read.csv("basel.csv",sep =";")

# Mise en forme des données et gestion des dates
donnees$paste = paste(donnees$Year,donnees$Month,donnees$Day,donnees$Hour,
                      donnees$Minute,sep = '-')
TEMP_BASEL = donnees$Temperature...2.m.above.gnd.

# Conversion en serie temporelle au format xts
DATE_BASEL = as.POSIXct(strptime(donnees$paste,format ='%Y-%m-%d-%H-%M',tz ='GMT'))
TS_BASEL = xts(TEMP_BASEL,order.by = DATE_BASEL)
```

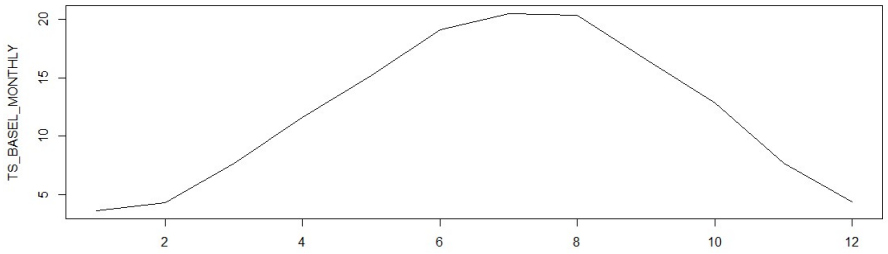
FIGURE 2 – Code R pour la mise en forme des données

2 Analyse descriptive des données

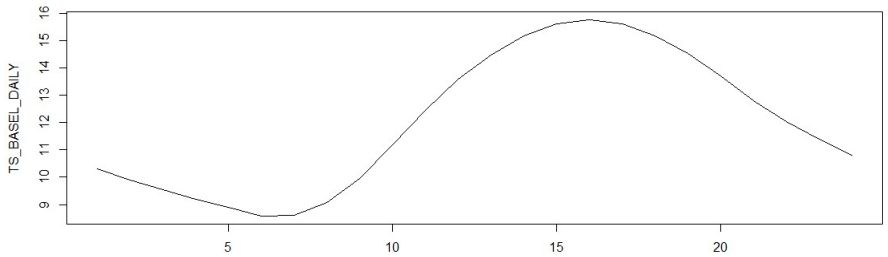
La figure suivante présente l'évolution de la température sur les 20 dernières années. On remarque une périodicité annuelle d'amplitude 35°C, ainsi qu'une journalière. Ces périodes correspondent simplement aux révolutions de la terre autour du soleil et sur elle-même. On peut donc moyenner sur la journée ainsi que sur l'année.



Température (en °C) à bâle sur les 20 dernières années

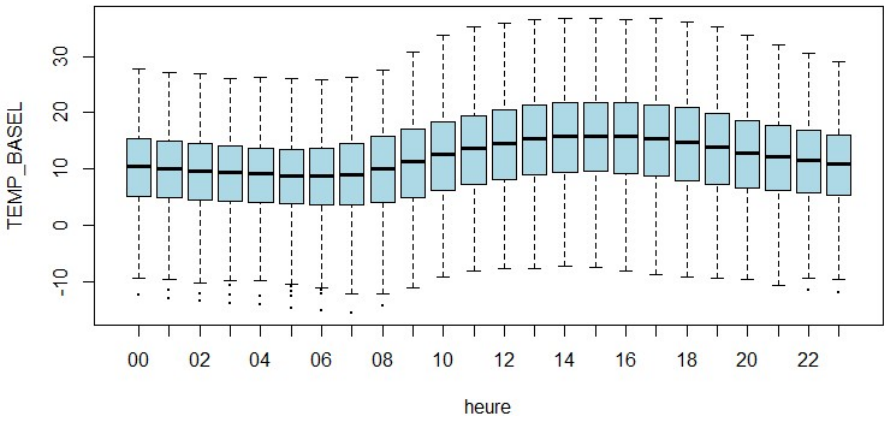


Températures dans l'année moyennées sur 20 ans

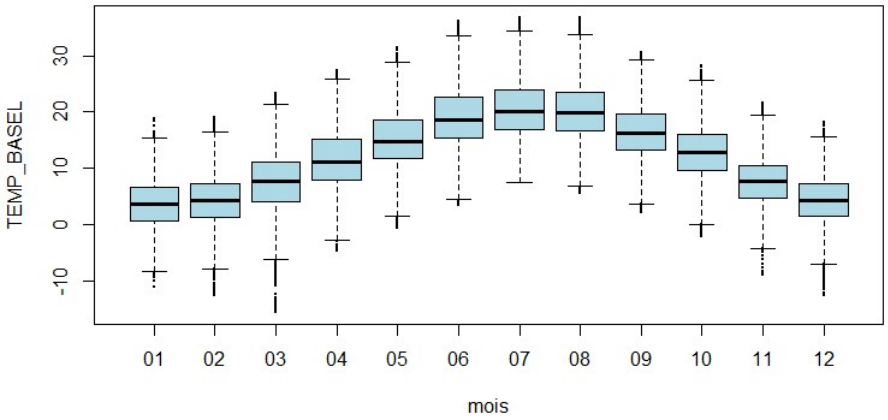


Températures dans la journée moyennées sur 20 ans

Cependant il faut noter que la moyenne journalière est effectuée sur tous les mois et aura donc une variance très importante car il fait plus chaud en été et plus froid en hiver. Pour les températures mensuelles les variations proviennent des années plus ou moins chaudes observées. Nous considérons par la suite une série temporelle à saisonnalité "simple", i.e à une seule période, et nous pourrons donc considérer la série moyennée sur la journée pour n'avoir qu'un point de mesure par jour et donc réduire la série de 175512 à 7313 points.



Boxplot des températures moyennes sur une journée



Boxplot des températures moyennes sur un mois

Nous cherchons à évaluer le réchauffement climatique terrestre sur ces 20 dernières années et à prévoir son évolution à l'aide de modèles simples. En effet près de la surface terrestre, le réchauffement s'est accentué ; "depuis le milieu des années 1970, il a atteint une moyenne de 0,17 C par décennie" (source : météo france). L'étude de la ville de Bâle nous a semblé pertinent (comparé à des villes comme Hong-Kong où les variations sont bien plus spectaculaires) car la démographie ainsi que l'expansion industrielle de la ville est restée très stable ces 30 dernières années (source : wikipédia).

3 Modélisation des données

Afin de modéliser convenablement la série temporelle des températures et de prévoir ses futures réalisations, nous décomposons cette série en une tendance linéaire, une saisonnalité et un bruit stationnaire. Le modèle considéré est additif.

3.1 Tendance

Nous avons comparé différents procédés d'analyse de la tendance ; la moyenne mobile, la régression linéaire et l'estimation à noyau gaussien. Les résultats sont présentés à la figure suivante. L'estimation par noyau gaussien a été faite sur la série moyennée quotidiennement, le temps de calcul étant trop long sur la série à 175512 point. Par la suite tous les calculs seront effectués sur la série moyennée.

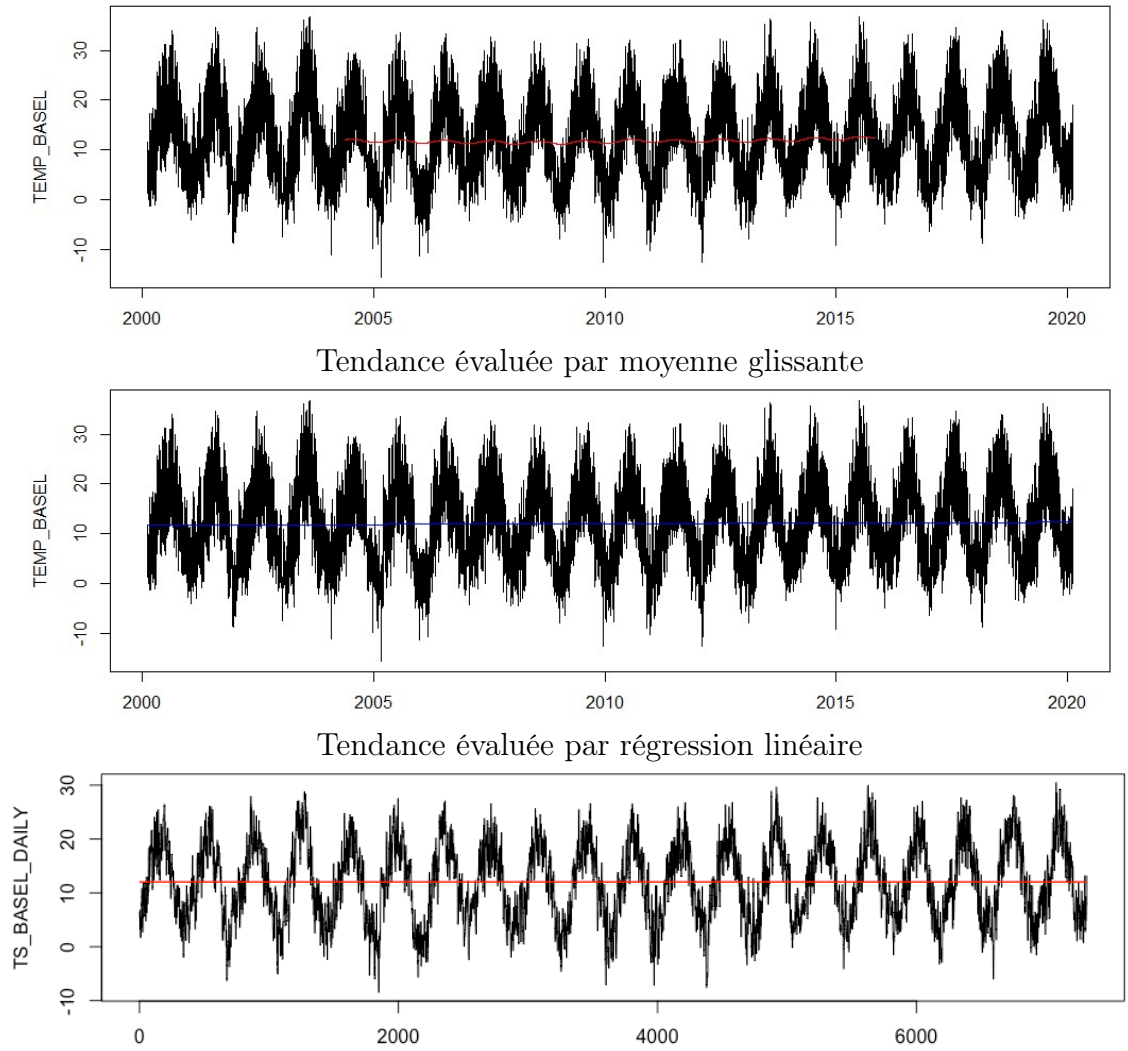
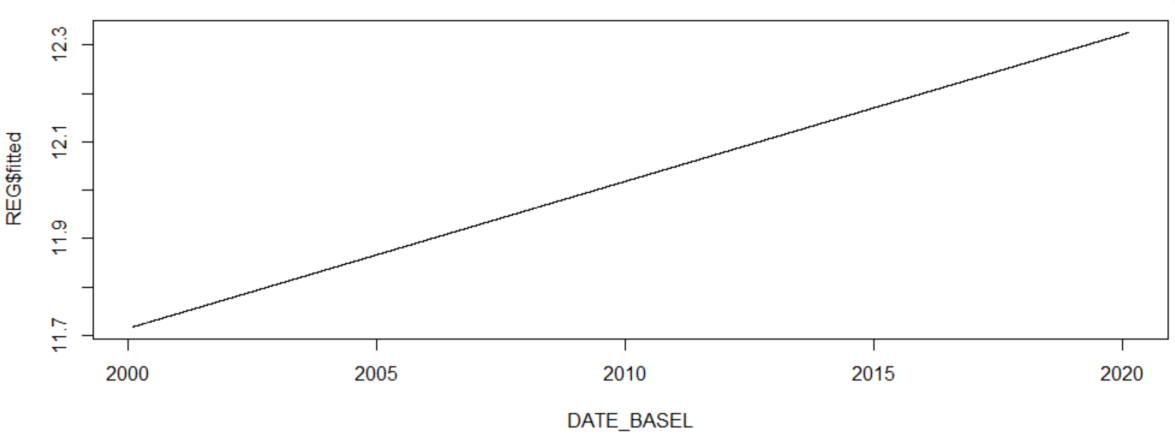


FIGURE 3 – Différentes méthodes de modélisation de la tendance

La tendance étant très faible comparée à l'amplitude de la saisonnalité, il ressort une apparente stationnarité de la tendance (i.e une constante égale à 11.5 °C). L'estimation par moyenne glissante ne donne pas de résultats satisfaisants, mais les estimations par regression

linéaire et noyau Gaussien semblent convenir
Au vu de ces résultats, nous avons décidé de travailler sur l'estimation par régression linéaire de la tendance, dont le détail est donné ci-dessous. On remarque que l'ordonnée à l'origine se situe autour de 11.5 °C, et que la pente de la droite de régression est de 0.3 °C par décennie. Ces estimations sont significatives au regard de la p-value ($< 2.2e^{-16}$), cependant le coefficient de détermination est très faible ($5 * 10^{-4}$) car l'amplitude de la saisonnalité est très importante devant la pente de la droite.



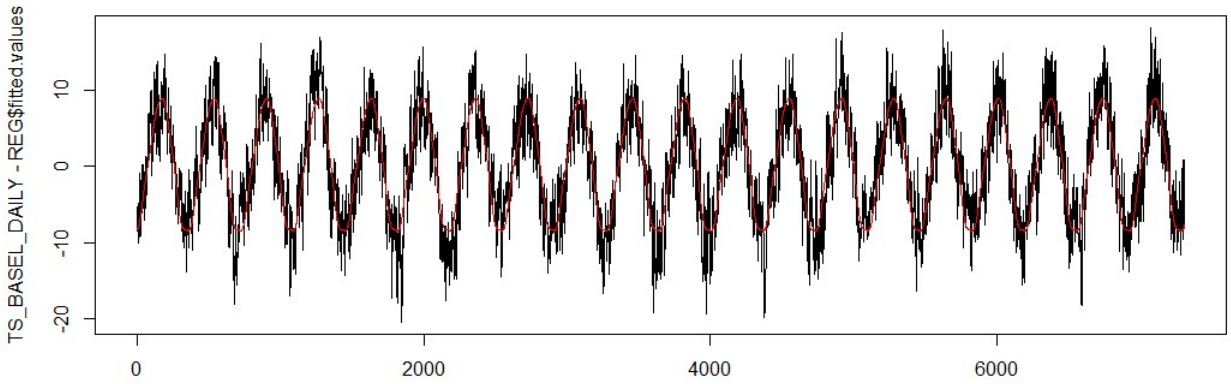

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.080e+01  1.301e-01  83.053   <2e-16 ***
DATE_BASEL  9.615e-10  1.017e-10   9.453   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.772 on 175510 degrees of freedom
Multiple R-squared:  0.0005089, Adjusted R-squared:  0.0005032
F-statistic: 89.36 on 1 and 175510 DF,  p-value: < 2.2e-16
```

FIGURE 4 – coefficients de la régression

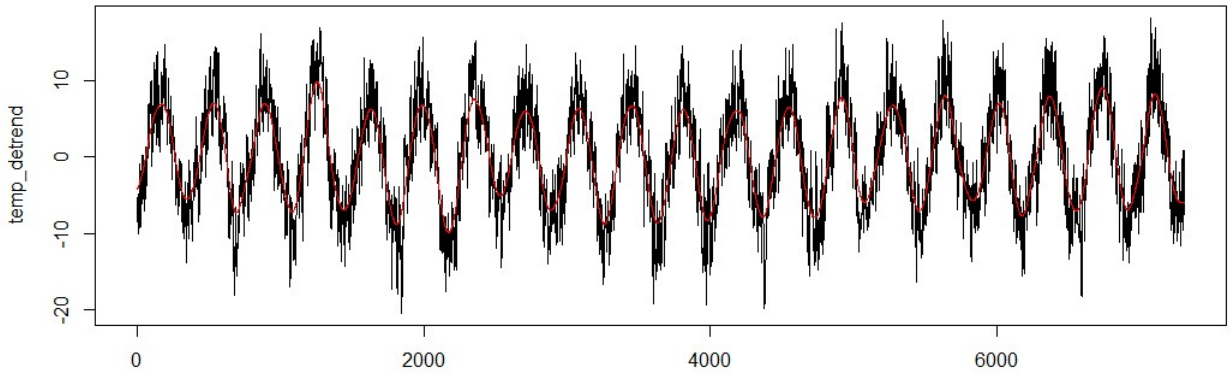
3.2 Saisonnalité

On travaille maintenant sur la série corrigée de sa tendance. On considère un modèle additif, ce qui est cohérent avec l'observation de la série, mais qui pourrait être questionné d'un point de vue météorologique. En effet la tendance au réchauffement peut aussi avoir pour effet d'augmenter l'amplitude des variations de températures annuelles. On travaillera de plus sur la série moyennée sur la journée (1 point par jour), ce qui permettra d'obtenir les résultats des calculs en temps raisonnable.
La modélisation de la saisonnalité permettra d'accéder à la partie aléatoire de la série. Notre comparaison des différentes méthodes prend donc en compte l'estimation de l'erreur, même si ces résultats sont présentés à la partie suivante.
La figure suivante présente les différentes modélisation testées. Pour la modélisation par régression sur série de Fourier, nous avons limité le nombre d'harmoniques à 5 pour éviter le sur-apprentissage. En effet, le cycle observé étant de 20 ans, si l'on garde les périodes proches de 20 ans dans notre modélisation alors les évènements exceptionnels (une année trop chaude par exemple) seront pris en compte pour nos estimations futures comme étant périodiques et déterministes, ce qui n'est pas le cas. En ce qui concerne les autres méthodes, le réglage des paramètres est empirique.

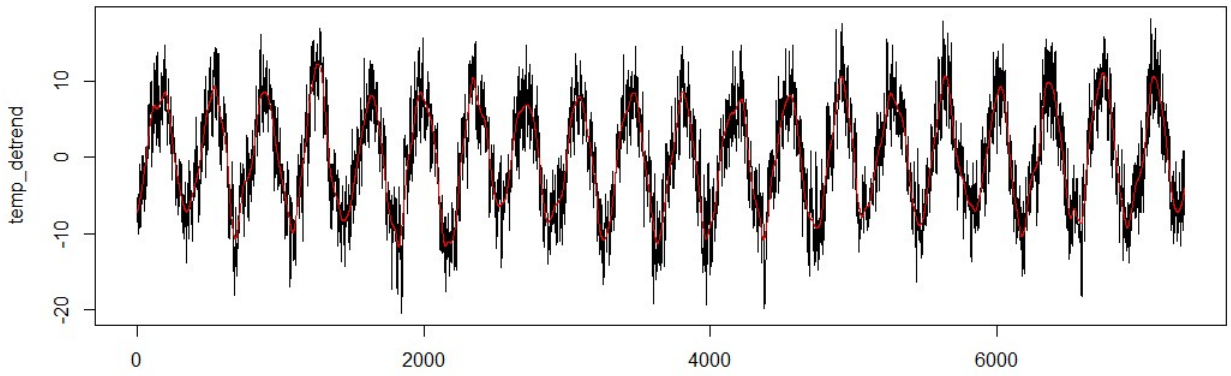
Chacune de ces méthodes semble donner des résultats similaires et acceptables, cependant en y regardant de plus près (figure 6) on note quelques différences de comportement :



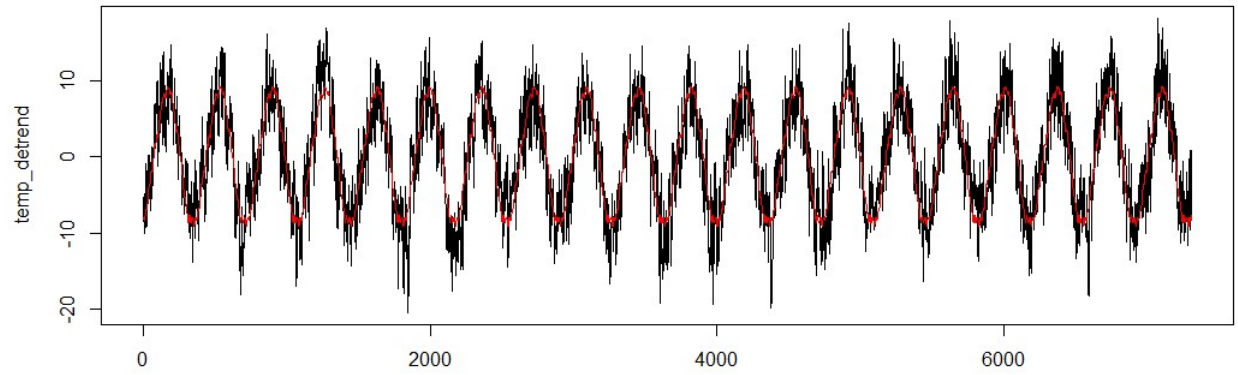
Saisonnalité évaluée par régression sur série de Fourier



Saisonnalité évaluée par régression à noyau Gaussien



Saisonnalité évaluée par polynômes locaux



Saisonnalité évaluée par régression sur une base de splines cycliques

FIGURE 5 – Modélisation de la saisonnalité

1. La modélisation par polynômes locaux, contrairement aux autres méthodes suit les irrégularités de la série et n'est donc pas périodique. Il y a dans ce cas sur-apprentissage
2. La régression sur splines cycliques présente des irrégularité de hautes fréquences qui ne sont pas désirables pour une modélisation de la partie déterministe.
3. La méthode à noyau Gaussien présente un bon comportement mais le réglage des coefficients amène soit à accepter des irrégularités non désirables, soit à accepter une diminution de l'amplitude des variations annuelles de températures.

On choisira donc la modélisation par régression sur série de Fourier.

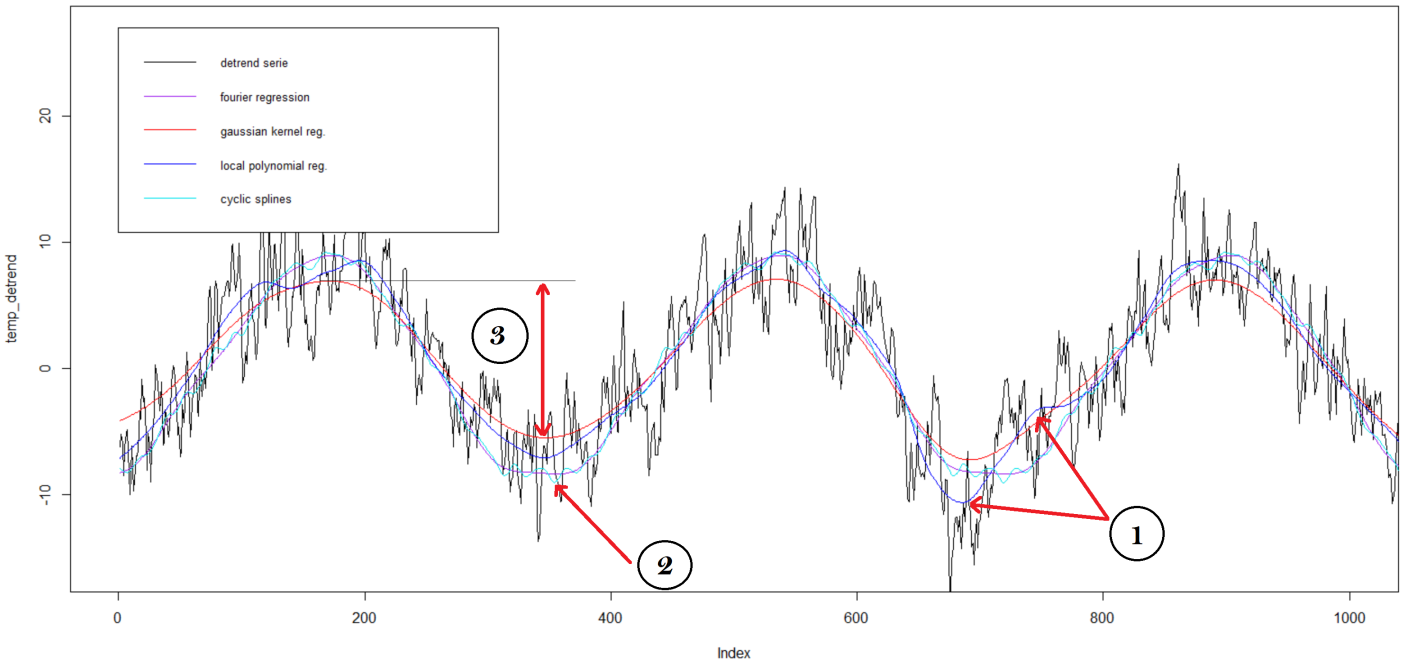
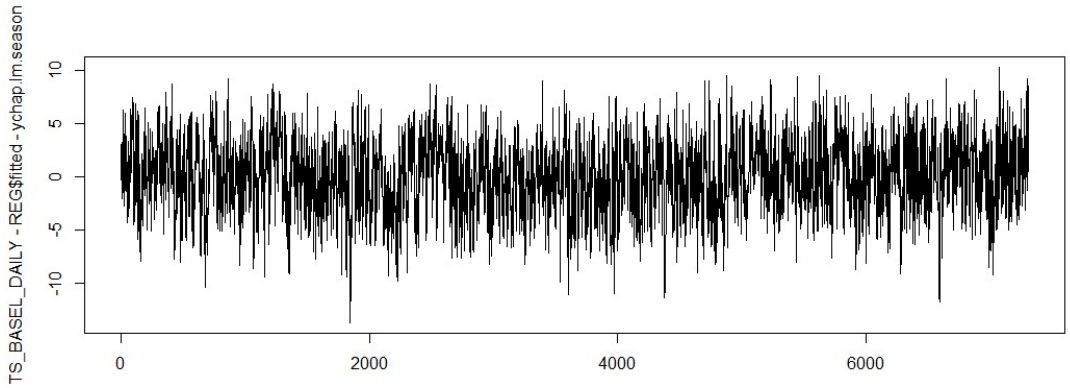


FIGURE 6 – Comparaison des différentes méthodes sur 3 périodes

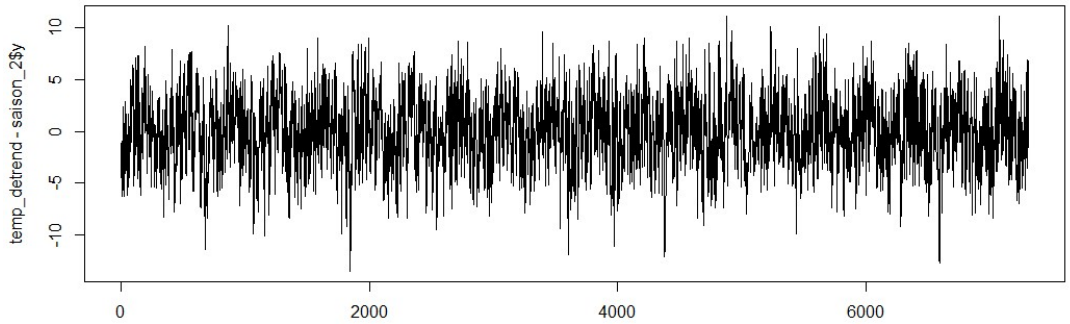
3.3 Erreur

En considérant un modèle additif, l'estimation de l'erreur se fait en corrigeant la série de l'estimation de sa tendance et de sa saisonnalité. Cette erreur est la partie aléatoire de la série et l'on aimerait obtenir idéalement un processus stationnaire, centré. Le tableau suivant présente les moyennes et variances des différentes erreurs obtenues.

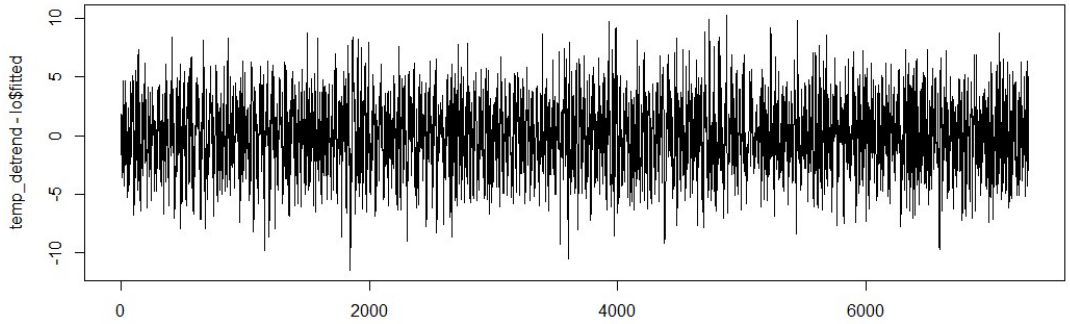
	mean	var
Fourier	$3,4810^{-17}$	11,93
Gauss	$-4,1210^{-3}$	11,76
Loess	$-4,8,4810^{-4}$	9,61
Splines	$4,4410^{-16}$	11,73



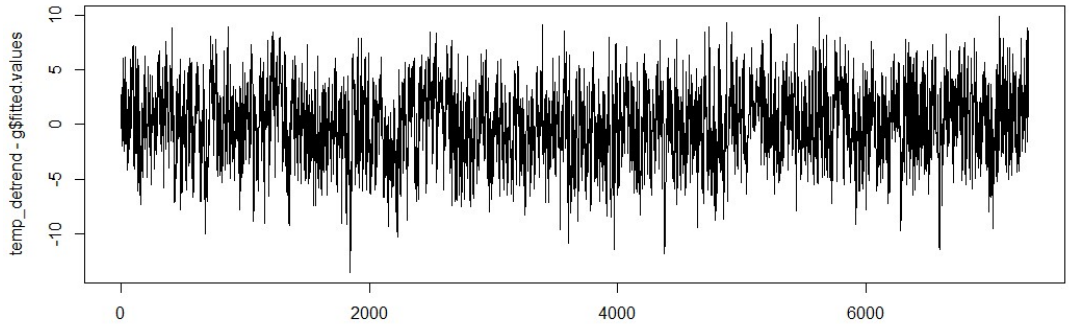
Erreur évaluée par régression sur série de Fourier



Erreur évaluée par régression à noyau Gaussien



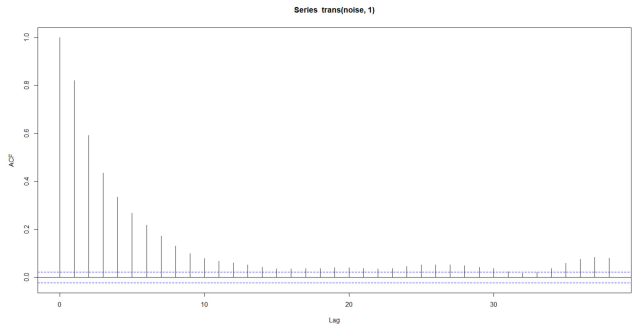
Erreur évaluée par polynômes locaux



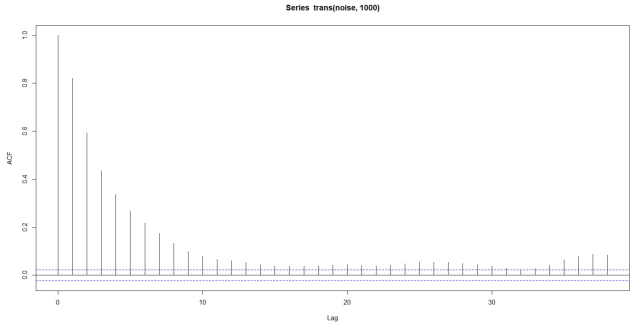
Erreur évaluée par régression sur une base de splines cycliques

FIGURE 7 – Modélisation de l'erreur

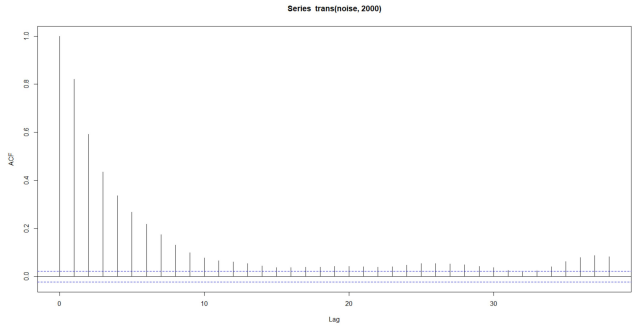
On vérifie que ces processus ont bien une moyenne ainsi qu'une variance constante dans le temps, mais pour conclure sur la stationnarité (faible) il nous faudra regarder plus précisément les corrélogrammes décalés en temps et vérifier qu'ils sont bien égaux. On vérifie pour quelques valeurs, que l'erreur modélisée par série de Fourier semble bien stationnaire faible.



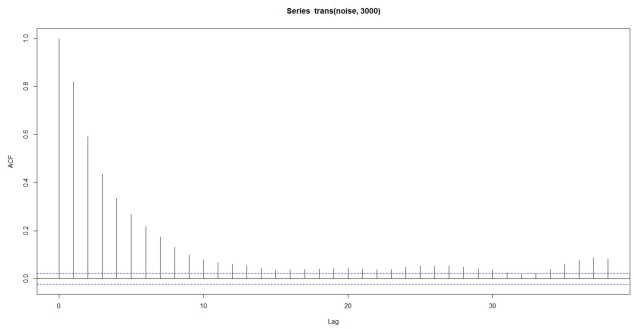
Autocorrélation non décalée dans le temps



Autocorrélation décalée de 1000 jours



Autocorrélation décalée de 2000 jours



Autocorrélation décalée de 3000 jours

Cependant il nous est impossible de conclure à ce niveau car d'une part nous n'avons pas affiché tous les pas et d'autre part cette vérification est visuelle. Il nous faut donc un test pour conclure sur la stationnarité, on va utiliser le test de Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Ce test permet de déterminer si la série n'est pas stationnaire. En effet, l'hypothèse nulle correspond à la stationnarité de la série autour d'une tendance, donc une p-value faible indiquera que l'on rejette cette hypothèse et donc que la série n'est pas stationnaire. Le tableau suivant récapitule les résultats.

	p-value	conclusion
Fourier	< 0.01	non-stationnaire
Gauss	> 0.1	stationnaire
Loess	> 0.1	stationnaire
Splines	< 0.01	non-stationnaire
Série complète	> 0.1	stationnaire

On remarque donc que la série complète, i.e la série avec sa tendance et sa stationnarité est stationnaire au regard de ce test (ce qui veut dire que la partie aléatoire de cette série l'est). On s'attend donc à ce qu'une bonne modélisation de l'erreur conserve ce caractère stationnaire. On ne peut donc pas conserver la modélisation par série de Fourier car celle-ci perd ce caractère. Il nous reste la modélisation par noyau Gaussien ou par polynômes locaux, mais l'on a vu précédemment que les polynômes locaux donnent une modélisation non périodique. On conserve donc la modélisation par noyau Gaussien.

4 Simulation de prévision sur un échantillon test

4.1 Lissage exponentiel simple

Le lissage exponentiel simple pour la prédiction sur notre série temporelle n'est pas très intéressant car il estime le futur de la série par une constante à partir de la dernière valeur du lissage. Cela n'est évidemment pas convenable car la tendance et la saisonnalité sont alors effacés mais nous présentons toutefois le résultat.

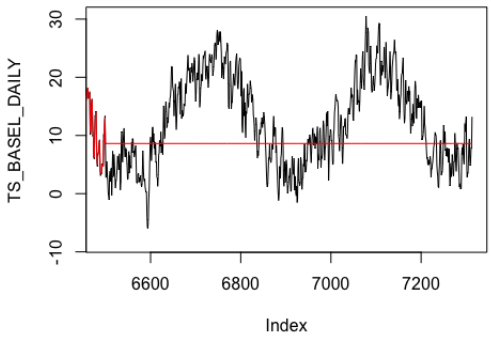


FIGURE 8 – Prédiction par lissage exponentiel simple

4.2 Lissage exponentiel double

Comme pour le lissage exponentiel simple, le lissage exponentiel double n'est pas très intéressant pour la prédiction sur notre série temporelle car il estime le futur de la série par une droite. Cela n'est toujours pas convenable car la tendance et la saisonnalité sont encore effacés. De plus cette estimation est très sensible au bruit car, comme on le voit ici, la pente de l'estimation est fortement influencée par la dernière pente calculée. Pour remédier à cela il faut baisser α mais cette correction ne suffit pas à donner une bonne prédiction.

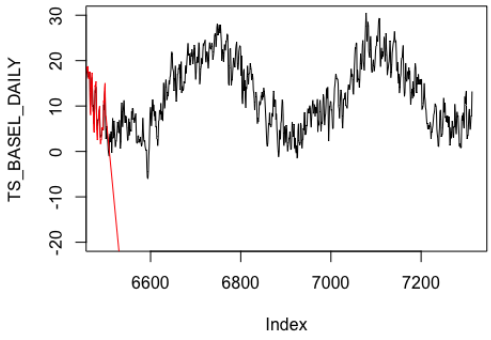


FIGURE 9 – Prédiction par lissage exponentiel double

4.3 Lissage exponentiel de Holt-Winters

Le lissage de Holt-Winters permet de modéliser la saisonnalité et une tendance linéaire locale selon un modèle additif ou multiplicatif. Il est aussi possible de ne pas prendre en compte la saisonnalité, les résultats dans ce cas sont très similaires aux performances d'un lissage exponentiel double.
La modélisation obtenue, présentée figure 10, donne xhat (la série estimée par le lissage), level (le niveau), trend (la pente locale) et season (la saisonnalité). On remarque que la pente locale est constante donc la modélisation nous donne une tendance du type $ax + b$ avec $a = \texttt{trend} = 5.76 * 10^{-05}$ et $b = \texttt{level}$, soit une augmentation de $0.21°C$ par décade. Cette méthode nous extrait aussi la saisonnalité.
Sur la Figure 11 nous avons extrait la prédiction de la composante saisonnière obtenue par le lissage de Holt-Winters en l'appliquant à une partie de notre série dépourvue de sa tendance. Nous avons ensuite ajouté la tendance extraite par régression linéaire manuellement à la prédiction de la composante saisonnière. On a ensuite juxtaposé la prédiction sur l'ensemble de la série afin de juger la qualité de cette prédiction et nous observons que la prédiction (en rouge) s'aligne bien avec les valeurs réelles de la série.

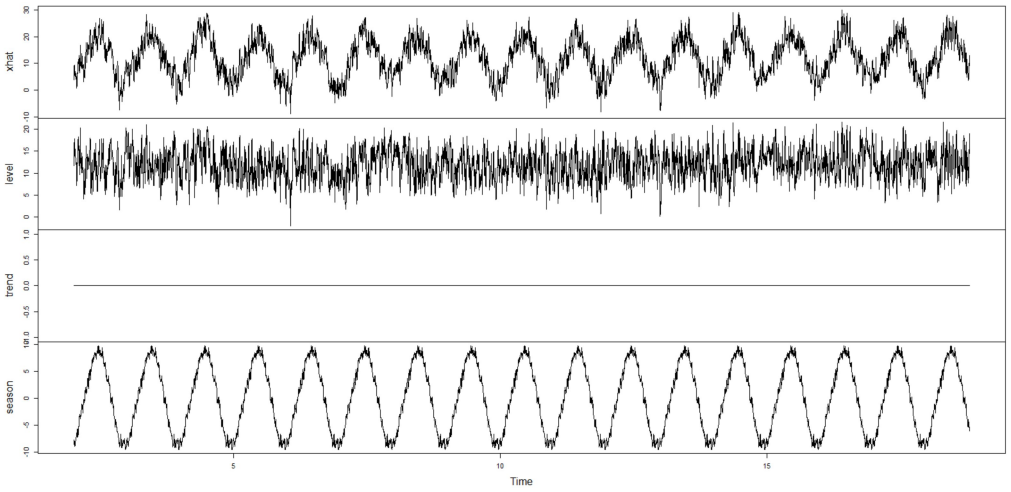


FIGURE 10 – Décomposition de la série modélisée par Holt Winters

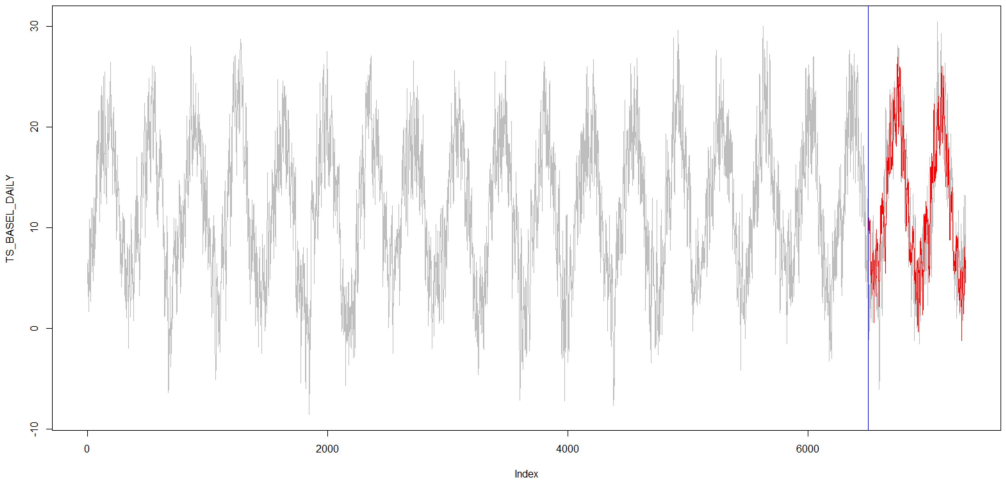


FIGURE 11 – Prévision sur 2 ans par Holt Winters saisonnier + tendance

5 Conclusion

La série chronologique des températures à Bâle depuis 2000 peut être modélisée par un modèle additif du type : $Y_t = T_t \ + \ S_t \ + \ \epsilon_t$.
Avec $T_t = a * t + b$ une tendance linéaire de pente $a \sim 0,25°C$ par décade et d'intersection $b = 11.7°C$.
Une saisonnalité S_t de période 365 jours.
Une erreur aléatoire ϵ_t de moyenne nulle et d'écart type $\sim 3,4°C$
La modélisation de la tendance donnée par régression linéaire est cohérente avec celle du modèle de Holt Winters, avec une pente légèrement plus importante ($a_{lm} = 0.30°C/$dec. , contre $a_{HW} = 0.21°C/$dec.) mais toujours dans les même ordres de grandeurs que celle observée au niveau mondial de 0.17 $°C/$decades.
Nous avons put remarquer que cette tendance était très faible et nos résultats nous incitent à être prudents sur nos conclusions. 20 ans semble une période trop faible pour récupérer une tendance propre, compte tenu des amplitudes des variations observées entre 2 années. Si l'on considère que le réchauffement est de $0.30°C/$dec et que les variations annuelles sont de $\pm 4°C$, il faudrait environ 100 ans pour "sortir du bruit".
De plus l'extrapolation d'une quelconque conclusion sur le climat à Bâle au climat mondial ne peut se faire sans prendre en compte les variations partout sur terre.