Projet Série Temporelle : Les recherches internet des aliments

CHAUVET Hugo - LEUCHI Ilias

Contents

Introduction

Objectif

Nous voulons comprendre les recherches internet de certains aliments au cours d'une année. L'objectif va être de trouver la meilleure façon de prédire les recherches internet de ces aliments au cours des prochaines années.

Présentation des données

Pour cela, nous avons récupéré sur kaggle un jeu de données comprenant le nombre de recherches internet de 200 aliments pour chaque semaine entre 2004 et 2016.

Table 1: Extrait du jeu de donnée

id	week_id	value
frozen-yogurt	2004-01	20
frozen-yogurt	2004-02	16
frozen-yogurt	2004-03	7
frozen-yogurt	2004-04	7
frozen-yogurt	2004-05	13
frozen-yogurt	2004-06	9

Voici un extrait de ces données. On retrouve le type d'aliment, la semaine, ainsi que la valeur du nombre de recherche pour cette semaine. Cette valeur correspond exactement aux proportions de recherches portant sur un aliment, ou le 100% correspond au taux de recherche le plus élevé de cet aliment entre 2004 et 2016.

Table 2: Extrait des aliments de nos données

anise	elderberry	okra
apple	empanada	old-fashioned
apple-au	endive	onion
apple-cider	energy-drink	orange
apple-pie	feijoa	parsnip
apple-ru	fennel	pasta-salad

Voici un extrait de quelques aliments présents dans nos données. Pour notre analyse, nous avons décidé de conserver deux aliments, il s'agit du chocolat et pizza. Nous commencerons par analyser la série, puis nous estimerons les différents paramètres, et pour finir, nous chercherons la meilleure manière de réaliser des prédictions sur la série. Nous réaliserons ce travail pour nos deux aliments. Commençons d'abord avec le chocolat.

Analyse des recherches internet du chocolat

Voyons ensemble comment on évolué les recherches internet du chocolat entre 2004 et 2016, puis regardons comment cela évoluera dans les années suivantes.

Caractéristique de la série

Nous allons analyser notre série en regardant si elle présente une tendance et une saisonnalité. Pour rappel, il s'agit de données hebdomadaires donc la période est de 52, la fonction d'auto-corrélations n'est donc pas nécessaire. La création de la série, ainsi que l'analyse descriptive se fait à l'aide de fonctions crée au préalable, qui s'adapte à l'aliment entré en paramètre. Ici, on utilise donc ces fonctions pour le chocolat.

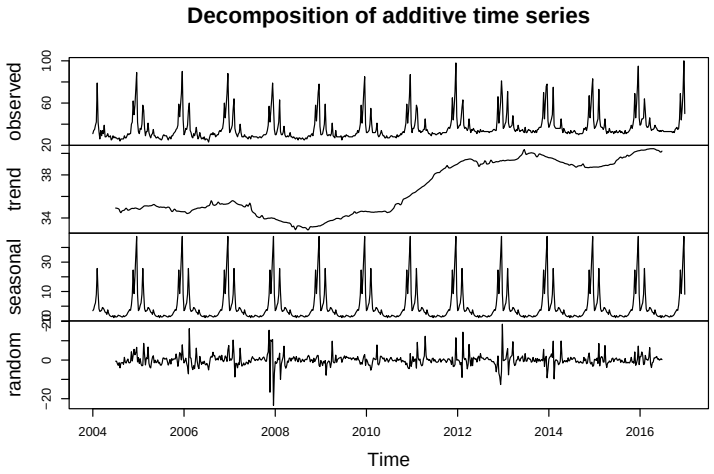

Figure 1: Décomposition de la série chocolat

Quand on regarde la tendance générale des recherches internet sur le chocolat entre 2004 et 2016, on remarque qu'il y a une légère augmentation qui se dégage. Quand on isole la tendance de la série, on voit qu'il y a eu une tendance négative entre 2007 et 2009, puis la tendance c'est inversé et une augmentation a eu lieu entre 2009 et 2014 et c'est stabilisé jusqu'en 2016.

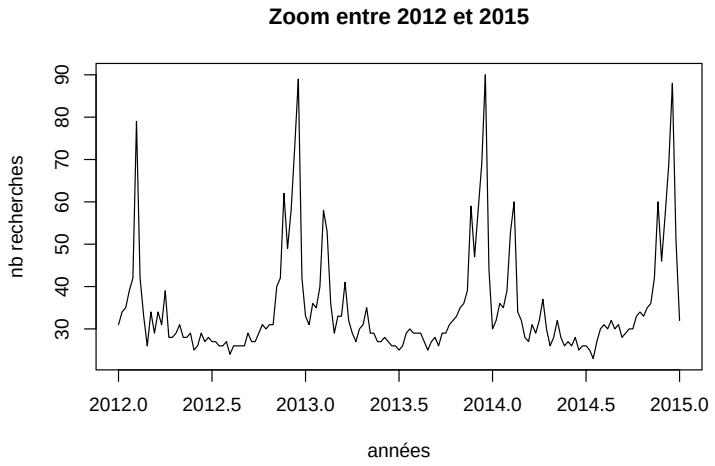

Figure 2: Zoom entre 2012 et 2015

Dans cette série, on remarque également qu'il y a une saisonnalité, en effet, on voit qu'une courbure se répète entre les années. Quand on zoome entre trois-quatre années, on constate qu'entre les fins d'années et début d'années suivantes, il y a une forte augmentation des recherches internet du chocolat. On peut imaginer que celle-ci est due aux différentes fêtes de fin d'année.

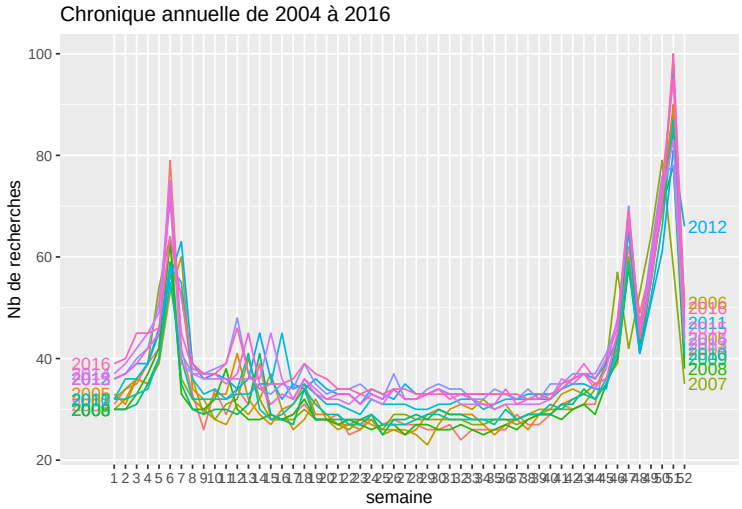

Figure 3: Chronique annuelle de 2004 à 2016

Avec la chronique annuelle, on remarque également des pics entre les 10e et 17e semaines de l'année, là aussi une fête est sûrement à l'origine de cette hausse. Les recherches se stabilisent pour le reste de l'année.

Estimation de la tendance, saisonnalité et résidus

On va maintenant chercher à estimer la tendance, la saisonnalité et les résidus, qui serviront à faire des prévisions sur nos données. Ici aussi, nous avons développé plusieurs fonctions pour pouvoir adapter les résultats selon l'aliment souhaité. Nous les appliquons donc au chocolat.

Partie saisonnière

On commence avec la partie saisonnière. L'estimation des coefficients saisonniers est obtenue à l'aide de la fonction décompose appliquée à notre série.

Table 3: Coefficients saisonniers de la série du chocolat

-3.08	-2.54	-7.33	-5.52
-2.09	-4.05	-7.34	-4.28
0.48	-4.39	-7.02	-3.37
2.22	-5.57	-6.12	-2.30
8.05	-2.49	-6.46	-2.81
25.70	-4.83	-6.53	0.84
9.27	-6.16	-6.94	7.58
-1.74	-6.75	-7.29	24.65
-3.58	-7.00	-7.39	8.56
-3.26	-7.16	-6.48	20.38
-2.77	-6.41	-7.16	34.37
-0.86	-7.83	-6.68	47.53
-1.37	-6.43	-6.36	8.10

Notre série à une période de 52, nous obtenons donc 52 coefficients saisonniers présents dans le tableau ci-dessus.

Estimation de la tendance

Le prochain objectif va être d'estimer la tendance. Pour cela, on commence par retirer la partie saisonnière de notre série.

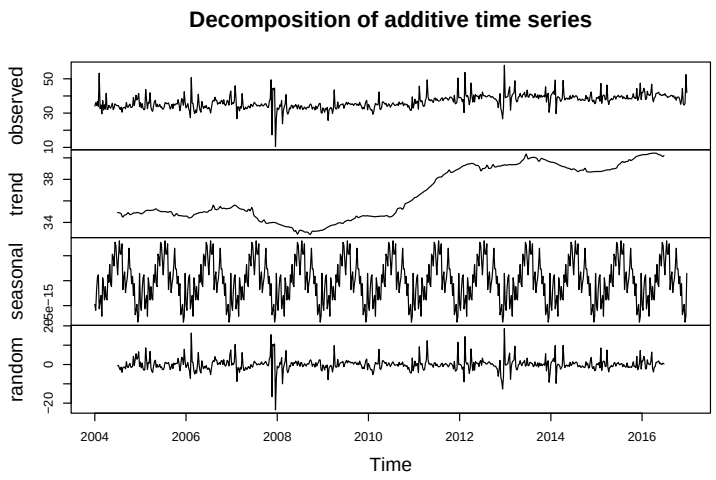

Figure 4: Décomposition de la série corrigée des variations saisonnières

On voit que la série a bien été désaisonnalisé, car la partie saisonnière varie entre de très petites valeurs. La série comporte donc uniquement la tendance ainsi que les résidus. Pour estimer la tendance, on va ajuster plusieurs polynômes avec différents degrés sur cette série désaisonnalisé, afin de trouver à partir de combien de degrés l'estimation de la tendance est la meilleure. Grâce à une fonction, nous testons jusqu'à 50 degrés de polynôme et affichons les R^2 ajustés au fur et à mesure.

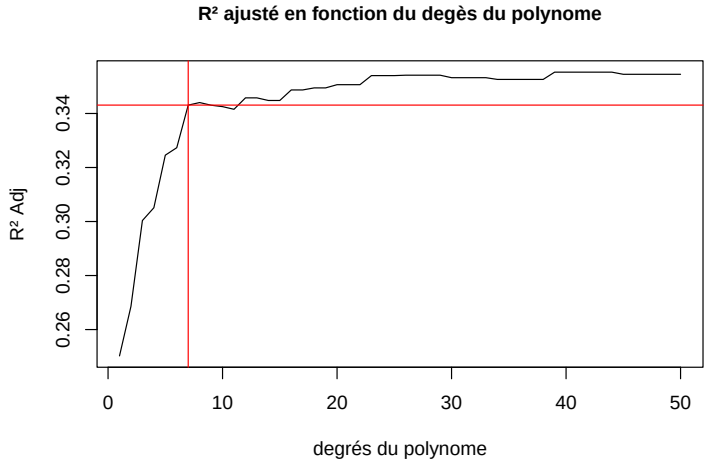

Figure 5: R^2 ajusté en fonction du degès du polynome

Au vu des R^2 ajustés et de la courbe suivante, on va considérer 7 degrés de polynômes pour la série sur le chocolat, car après cette valeur, il n y a pas d'augmentation significative des R^2 ajustés, il y a même une légère chute. On obtient donc un pourcentage de variance expliqué de 34,31%. Cette valeur reste assez faible la tendance n'aura donc pas une estimation consistante.

Quand on regarde les pvaleurs des paramètres de notre polynôme à 7 degrés, on remarque qu'elles sont toutes très faibles. Nous allons donc conserver tous les paramètres de notre modèle.

Estimation des résidus

Maintenant que la saisonnalité et la tendance sont estimées, il nous reste plus qu'à déterminé un modèle qui estimera les résidus de la série du chocolat. On commence par récupérer la série des résidus en enlevant à la série désaisonnalisé la tendance que nous venons d'estimer.

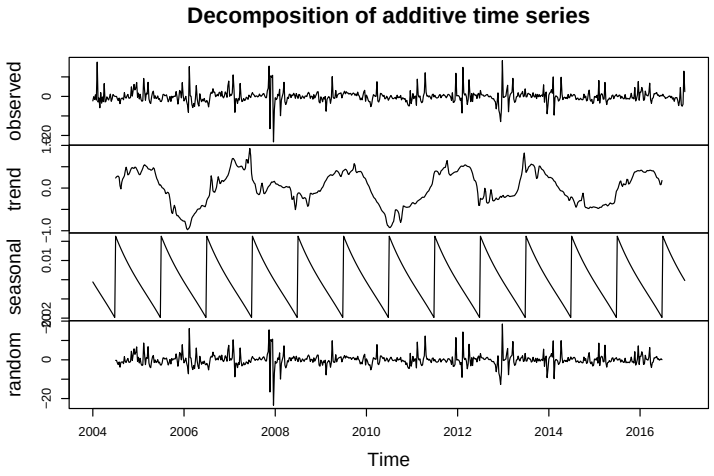

Figure 6: Décomposition de la série résiduelle du chocolat

Sur la figure ci-dessus, il s'agit bien de la série des résidus. La saisonnalité varie entre de petites valeurs et il n y a pas de tendance visible. La série des résidus semble donc stationnaire.

Nous allons voir à l'aide d'un Box.test s'il reste de l'information à extraire dans nos résidus. On obtient une pvaleur de 0.428, cette valeur n'est pas suffisamment proche de 1, on peut encore extraire de l'information dans nos résidus. Nous allons donc extraire l'information restante à l'aide d'un processus ARMA.

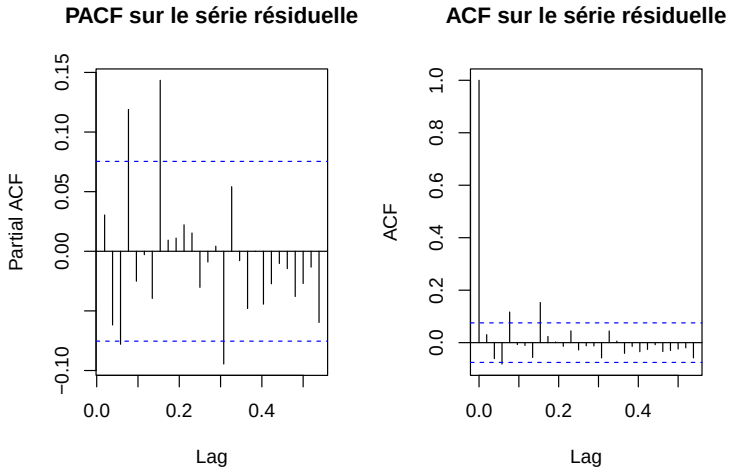

Figure 7: PACF et ACF sur la série résiduelle du chocolat

Nous commençons par choisir la dimension p et q de modèles AR(p) et MA(q) à l'aide d'un PACF et d'un ACF sur la série résiduelle. Au vu des graphiques, on choisit les valeurs p = 4 et q = 4. Nous allons maintenant investiguer différents modèles ARMA tels que p+q = 4.

Pour chacun des modèles tel que p+q=4 on obtient les AIC suivants :

- AIC ARMA(2,2) : 3540.18
- AIC ARMA(3,1) : 3536.79
- AIC ARMA(1,3) : 3537.26

Ici, parmi les ARMA ou p+q = 4, on retiendra ARMA(3,1) qui a l'AIC le plus bas. On utilisera donc un modèle ARMA(3,1) pour estimer nos résidus.

On refait ensuite un Box.test sur les résidus de notre estimation. Cette fois-ci la pvaleur est de 0.919, ce qui est proche de 1. On peut donc considère qu'il n y a plus d'information à extraire.

Prédiction

Passons à la partie prédiction. Pour cela, nous allons voir plusieurs méthodes de prédiction et les comparer pour savoir laquelle est la meilleure.

Prédiction manuelle

Commençons avec les différentes estimations du modèle trouvé. Nous allons réaliser une prévision à horizon 1, c'est-à-dire pour la première semaine de 2017. Après réalisation des calculs, on obtient une valeur de 36.014. Nous allons maintenant réaliser des prévisions pour toute l'année 2017.

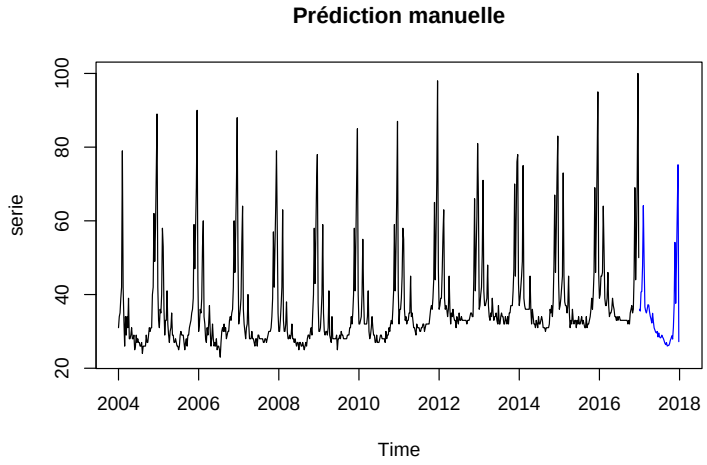

Figure 8: Prédiction manuelle de la série chocolat en 2017

Voici le graphique des prévisions manuelle réaliser pour l'année 2017. Ces prévisions semble juste, cependant, on a l'impression qu'elles sous-évaluent légèrement le nombre de recherches internet. Nous allons par la suite comparer avec les autres méthodes de prévision pour en tirer de meilleures conclusions.

Lissage exponentiel

Passons au lissage exponentiel. Parmi les différents lissages possibles, nous allons chercher lequel est le meilleur à l'aide d'une fonction qui compare les RMSE.

Table 4: RMSE selon les lissages sur la série chocolat

lissage	RMSE
LES	14.419
LED	12.679
HWNS	18.507
HWSA	2.566
HWSM	2.479

On remarque que le lissage de HoltWinters saisonnier multiplicatif (HWSM) obtient le plus faible RMSE. Nous utiliserons donc ce type de lissage pour faire des prédictions sur nos données chocolat. D'abord à horizon 1.

Pour la première semaine de 2017, on obtient une prévision de 37.764 avec un intervalle de confiance à 95% de [35.325 , 40.203]. Maintenant, réalisons les prévisions pour l'année 2017.

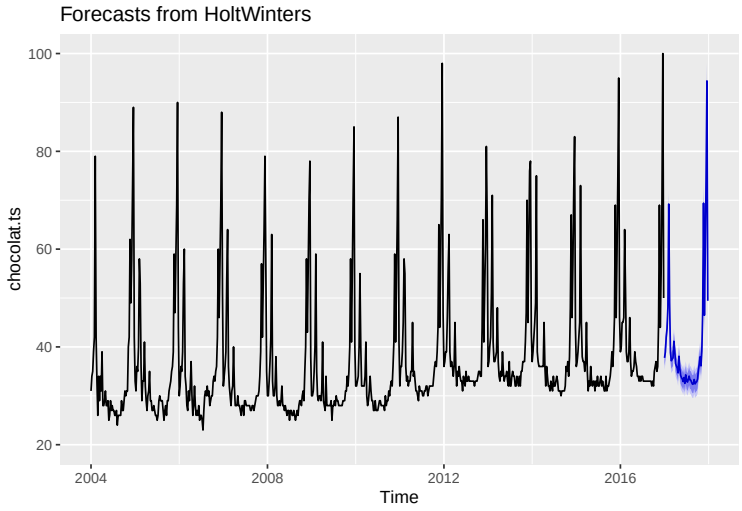

Figure 9: Prévision de HWSM pour de la série chocolat en 2017

Avec la figure suivant on remarque que les prévisions avec un HWSM semblent plus cohérentes par rapport aux données. De plus, l'intervalle de confiance à un étendu très peu élevé pour l'ensemble des prévisions de 2017. Nous verrons par la suite s'il s'agit de la meilleure manière de prédire nos données sur la série chocolat.

Processus automatique

Pour finir, nous allons utiliser la méthode SARIMA à l'aide de la fonction auto.arima sur notre série chocolat. Le modèle choisi par la fonction est : ARIMA(0,1,1)(2,1,0). Faisons les prévisions à horizon 1 et en 2017.

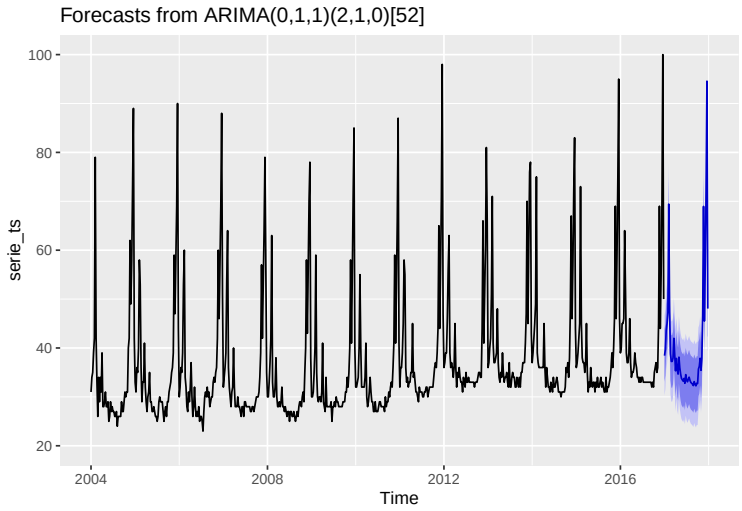

Figure 10: Prévision avec la méthode SARIMA de la série chocolat en 2017

Pour la première semaine de 2017, on obtient grâce à l'auto arima une prévision de 38.433 avec un intervalle de confiance de [30.301 , 46.566]. Le graphique nous montre que la méthode SARIMA produit des prévisions cohérent, cependant l'intervalle des valeurs est plus étendu que celui des prévisions avec le HWSM. Comparons toutes ces méthodes pour n'en garder qu'une seule.

Meilleure prédiction

On va maintenant voir quelle est la meilleure méthode pour la prédiction de nos données sur le chocolat. Pour cela, on va comparer les RMSE des différentes méthodes.

Table 5: RMSE selon les méthodes de prévision sur la série chocolat

méthode	RMSE
HWSM	2.479
SARIMA	3.566
manuelle	18.396

On remarque que le lissage de HoltWinters saisonnier multiplicatif (HWSM) nous donne le plus petit RMSE. Avec l'ensemble des résultats que nous avons obtenus, on peut en déduire que le HWSM est la meilleure façon de modéliser notre série sur les recherches internet sur le chocolat.

Analyse des recherches internet de la pizza

On va maintenant s'intéresser aux recherches internet de la pizza entre 2004 et 2016.

Caractéristique de la série

On commence par l'analyse de la série pizza.

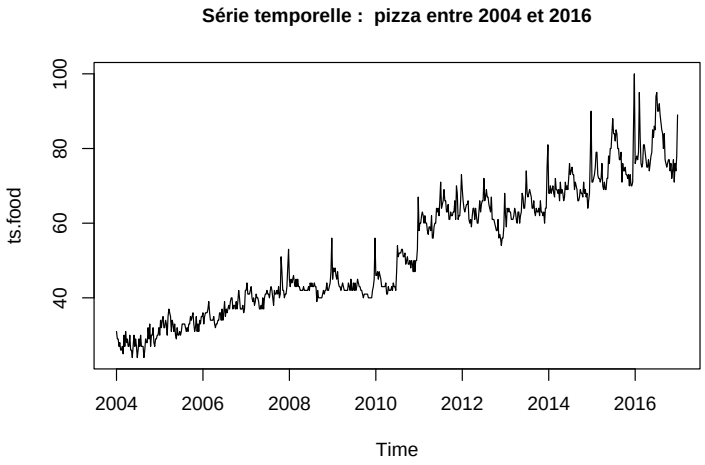

Figure 11: Série pizza

Quand on observe la série sur la pizza, on remarque une forte tendance positive croissante. Une saisonnalité est moins visible sur ce graphique. Utilisons la décomposition de la série et la chronique annuelle pour une meilleure visualisation de la saisonnalité.

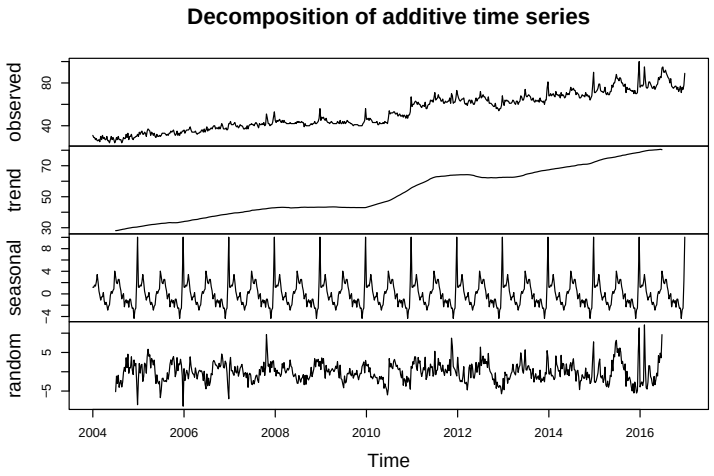

Figure 12: Décomposition de série pizza

Avec la décomposition de la série, on voit qu'il y a une saisonnalité, cependant elle semble plutôt légère, regardons la plus en détail avec la chronique annuelle.

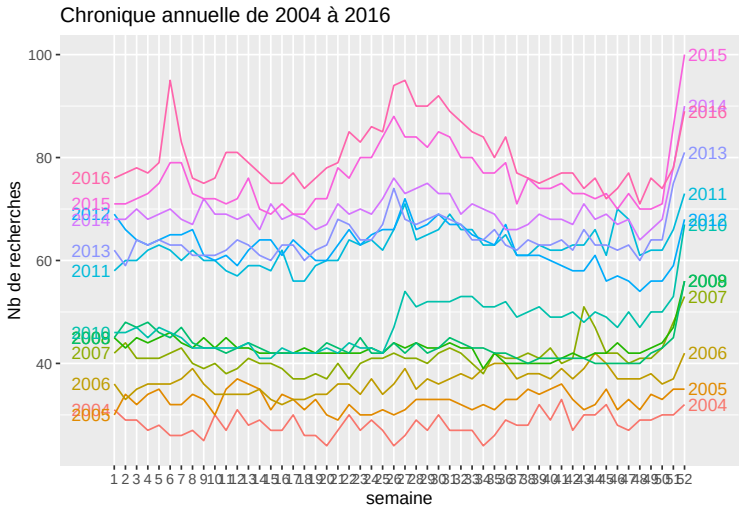

Figure 13: Chronique annuelle de la série pizza

Ici, on remarque que les courbes entre les différentes années ne sont pas superposé comme avec la série du chocolat, ceci est dû à la forte tendance. Concernant la saisonnalité, on observe très peu de variations saisonnières entre les années.

Estimation de la tendance, saisonnalité et résidus

L'analyse de la série étant réalisée, passons à l'estimation des différents paramètres de notre série des recherches internet de la pizza.

Estimation de la saisonnalité

Commençons avec la saisonnalité. Pour cela, nous récupérons les coefficients saisonniers grâce à la décomposition de la série.

Table 6: Coefficients saisonniers de la série de la pizza

1.18	-1.25	3.04	-1.61
1.16	-1.90	1.97	-1.19
1.56	-1.48	1.81	-2.35
1.37	-1.97	2.56	-0.93
2.11	-2.88	2.56	-0.92
3.43	-2.19	1.30	-1.92
2.08	-1.93	0.53	-2.58
0.49	-0.26	-0.88	-2.50
-0.27	0.42	-0.30	-4.34
-1.11	0.10	0.04	-2.84
-0.52	0.61	-2.29	-2.52
-0.52	0.87	-1.37	2.31
0.24	4.03	-1.03	10.04

Nous obtenons les coefficients suivants.

Estimation de la tendance

Passons maintenant à l'estimation de la tendance. Pour cela, on commence par retirer les variations saisonnières de notre série pizza.

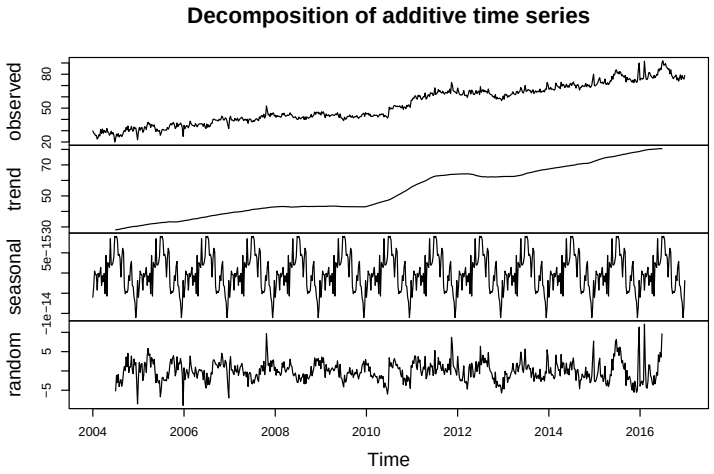

Figure 14: Décomposition de la série corrigée des variations saisonnières

La partie saisonnière varie entre de très faibles valeurs, il s'agit bien de la série corrigée des variations saisonnières. Nous pouvons maintenant estimer le nombre de degrés à prendre en compte pour l'estimation de notre tendance.

Voici la sortie R que nous obtenons avec la fonction que nous avons développée :

- R^2 ajusté du polynôme de degré 1 : 0.9407
- R^2 ajusté du polynôme de degré 2 : 0.9444
- R^2 ajusté du polynôme de degré 3 : 0.9448
- R^2 ajusté du polynôme de degré 4 : 0.9457
- R^2 ajusté du polynôme de degré 5 : 0.9472

Au vu des R^2 ajusté, nous pouvons considérer 1 degré de polynôme. Notre tendance est donc estimée par une fonction affine. Nous obtenons 94.07% de variance expliqué. Au vu des pvaleurs nous gardons tous les coefficients.

Estimation des résidus

La tendance étant estimée, nous pouvons passer à l'estimation des résidus.

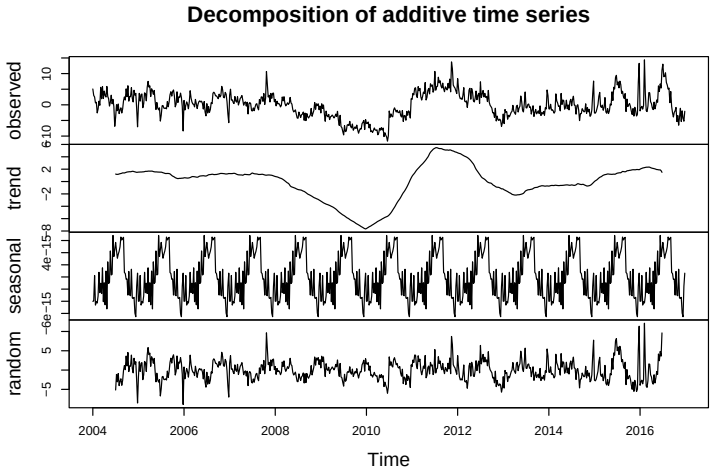

Figure 15: Décomposition de la série résiduelle de la pizza

Nous commençons par enlever la tendance estimée, à la série corrigée des variations saisonnières. La figure ci-dessus montre que nous obtenons bien la série des résidus. Nous réalisons ensuite un Box.test de notre série des résidus, nous obtenons une pvaleur de 0, il reste donc énormément d'information à extraire.

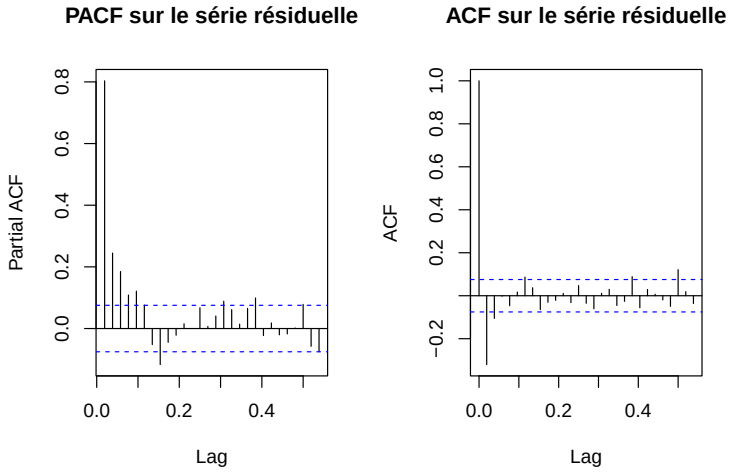

Figure 16: PACF et ACF sur la série résiduelle de la pizza

Cherchons les paramètres p et q pour nos résidus estimés. Grace au PACF nous trouvons une valeur de p=3. Avec l'ACF, nous obtenons une valeur de q=2, de plus nous avons utilisé la différence donc nous aurons d=1.

Nous avons ensuite testé plusieurs modèles ARMA avec une différence, telle que p et q ne dépassent pas leur valeur initiale. Nous choisissons donc un ARMA(3,2) qui a l'AIC le plus bas avec une valeur de 3046.26 .

Nous réalisons un Box.test sur les résidus de cette estimation. Nous obtenons une pvaleur de 0.916, ce qui est suffisant. Il n y a plus d'information à extraire.

Prédiction

Tous nos paramètres étant estimés, nous pouvons effectuer les différentes prévisions. Comme pour la série chocolat, nous ferons pour chaque méthode une prédiction à horizon 1 et une prédiction pour l'année 2017. Puis nous comparerons les différentes méthodes.

Prédiction manuelle

Commençons avec la prédiction qui utilise les différents paramètres que nous venons d'estimer.

Avec la méthode manuelle, nous obtenons une prédiction de 79.469 à horizon 1.

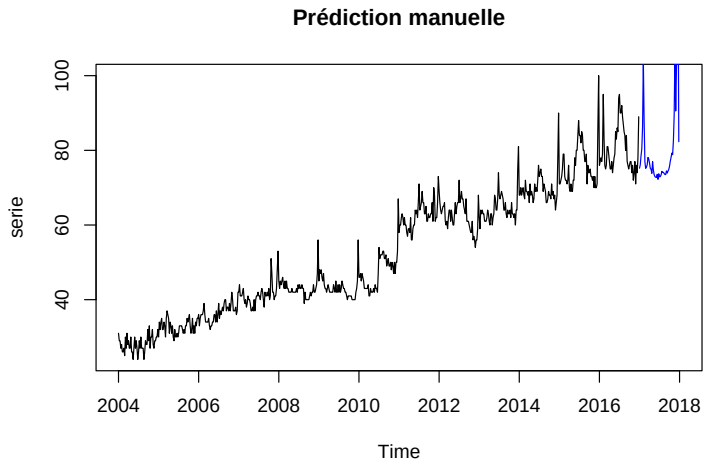

Figure 17: Prédiction manuelle de la série pizza en 2017

Pour l'année 2017, les prévisions ne semblent pas aberrantes, la tendance semble être correctement représentée. Par la suite, nous calculerons le RMSE pour connaître la qualité des estimations.

Lissage exponentiel

Passons maintenant aux méthodes plus automatiques, avec d'abord une comparaison des lissages exponentiels.

Table 7: RMSE selon les lissages sur la série pizza

lissage	RMSE
LES	11.451
LED	6.102
HWNS	12.293
HWSA	5.907
HWSM	6.397

Au vu des RMSE des différents lissages, le HoltWinters saisonnier additif (HWSA) à la valeur la plus faible, c'est donc celui-ci que nous conservons.

Pour la prévision de la première semaine de 2017, nous obtenons une prévision de 71.732 avec un intervalle de confiance de [66.808 , 76.655].

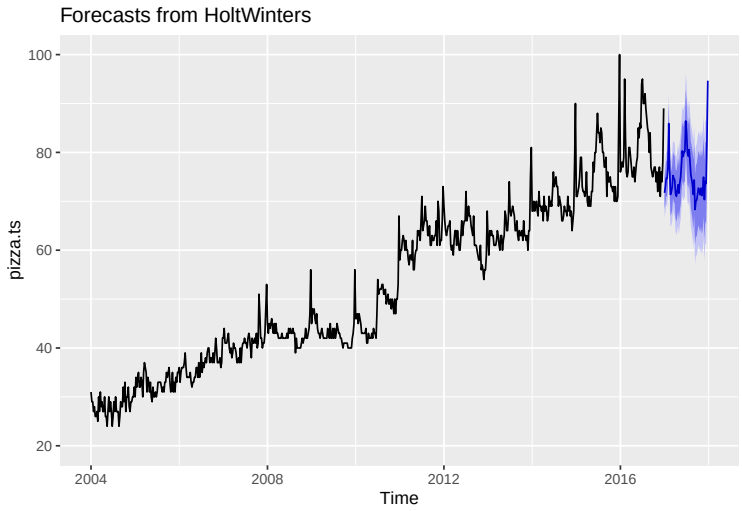

Figure 18: Prévision avec HWSA pour de la série pizza en 2017

Concernant la prédiction pour l'année 2017 avec un HWSA, la tendance est différente qu'avec notre prédiction manuelle, elle semble moins linéaire. Nous vérifierions quelle méthode correspond le plus à la réalité dans la dernière partie.

Prédiction automatique

Pour finir, regardons les prévisions à l'aide de la fonction auto.plot. Le modèle SARIMA obtenu à l'aide la fonction est le suivant : ARIMA(2,0,1)(0,1,1).

À horizon 1, nous obtenons une prévision de 74.248, avec l'intervalle de confiance [69.556 , 78.941]. Cet intervalle est moins étendu que celui avec la méthode HWSA, et englobe de plus grande valeur.

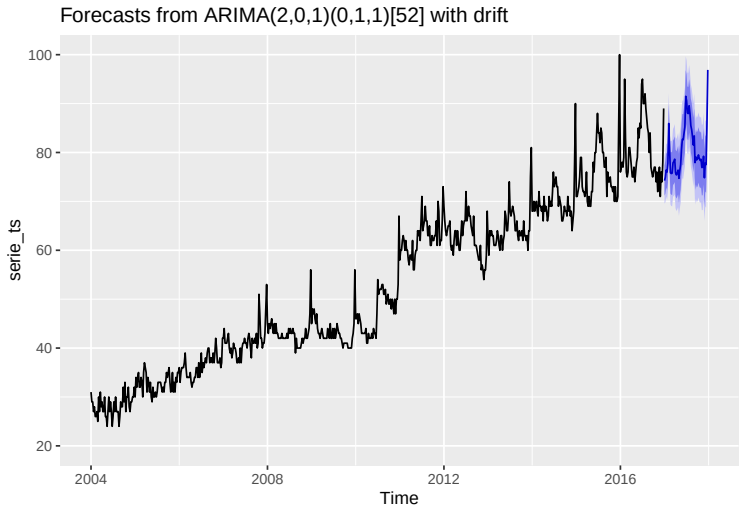

Figure 19: Prévision avec la méthode SARIMA de la série pizza en 2017

La prévision sur l'année 2017 est très similaire à celle réaliser avec le lissage HWSA, les valeurs sont légèrement plus grande.

Meilleure prédiction

On va maintenant voir quelle est la meilleure méthode pour la prédiction de nos données de la pizza. Pour cela, on va comparer les RMSE des différentes méthodes.

Table 8: RMSE selon les méthodes de prévision sur la série pizza

méthode	RMSE
HWSA	5.907
SARIMA	8.718
manuelle	12.348

La méthode manuelle est clairement la plus mauvaise pour prédire nos données, nous obtenons un RMSE de 12.348. Le HWSA est la meilleure méthode pour la prédiction de nos recherches internet sur la pizza, avec un RMSE 5.907. Alors qu'avec le modèle SARIMA, nous obtenons une valeur de 8.718. Nous choisissons donc un lissage HWSA pour la prédiction sur notre série pizza.

Conclusion

Nous avons bien réussi à obtenir des estimations consistantes avec les deux séries que nous avons étudiées. Nous retenons que pour avoir les meilleures estimations sur les séries des recherches internet des aliments, les méthodes de lissage exponentiel semblent retourner les meilleurs résultats. Ceci s'est vu sur nos deux séries étudiées, il faudrait réaliser des analyses avec plus d'aliments pour le confirmer. Le code développé à l'aide de différentes fonctions permet de s'adapter à un aliment particulier, afin de pousser plus loin l'analyse et d'obtenir des conclusions plus solides.