MTH2302D Probabilités et statistiques

Travail de session #2 sur

La production et la vente d'électricité et la température

Présenté à :
Kwassi-Joseph Dzahini

Soumis par :
Guillaume Proulx -1899371
Marc-André Primeau Breton - 1856799

Section : 03

Le 2 décembre 2019

Table des matières

Contexte général des données

Les données présentées dans le fichier Excel joint à ce document proviennent principalement de « Statistique Canada ». Les données recueillies sont :

– La quantité d'électricité produite au Canada mensuellement
– La température moyenne des mois
– Le prix de vente de l'électricité
– La date

Premièrement, les données sur la quantité d'électricité produite au Canada mensuellement correspondent à la quantité d'énergie produite mensuellement par tous les moyens de production électrique, c'est-à-dire l'électricité produite par les turbines à hydraulique, les turbines à vapeur nucléaire, par la combustion, etc.

Deuxièmement, les données sur la température moyenne proviennent de la station météo Bonsecours. Cette station a pris en note la température quotidiennement pour, finalement, en faire un rapport météorologique sur la température moyenne pour le mois.

Troisièmement, les données sur le prix de vente de l'électricité correspondent à la moyenne nationale mensuelle du prix de vente provenant de chacune des provinces du Canada.

Finalement, la date correspond au mois et à l'année pour les données de chacune des catégories.

L'analyse de ces données est pertinente, car on suppose qu'elle permettra de prévoir le prix de vente d'une certaine quantité d'électricité quotidiennement ou mensuellement selon la température du mois et la quantité d'électricité produite au Canada durant ce même mois.

Provenance des données

Comme annoncé ci-dessus, les données proviennent principalement du site de « Statistique Canada ». Voici le lien pour chacune des catégories :

- La quantité d'électricité produite au Canada mensuellement : https://www150.statcan.gc.ca/t1/tbl1/fr/cv.action?pid=2510001501

Les données ont été recueillies à l'aide de l'Enquête mensuelle sur l'approvisionnement et l'écoulement de l'électricité.

- La température moyenne des mois : http://climat.meteo.gc.ca/climate_data/monthly_data_f.html?hlyRange=%7C&dlyRange=1967-01-01%7C2019-04-30&mlyRange=1967-01-01%7C2018-0201&StationID=5322&Prov=QC&urlExtension=_f.html&searchType=stnProv&optLimit=yearRange&StartYear=2008&EndYear=2019&selRowPerPage=25&Line=21&lstProvince=QC&timeframe=3&Month=9&Day=18&Year=2018 #

Les données ont été recueillies dans les rapports de données mensuelles de la station.

- Le prix de vente de l'électricité : https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=1810020401

Les données ont été recueillies à l'aide de l'Indice des prix de vente de l'énergie électrique pour les acheteurs non résidentiels.

Le format des données

Pour la quantité d'électricité produite, on a recueilli 138 observations différentes sur la production électrique au Canada et celles-ci sont sous la forme de millions de mégawattheures. Les dates des observations proviennent du 1er janvier 2008 jusqu'au 1er juin 2019. La quantité d'électricité produite est une variable discrète, car c'est une quantité moyenne nette produite durant le mois partout au Canada.

Du côté de la température moyenne, on a recueilli 122 observations différentes sur la température moyenne d'un mois. La température est en degré Celsius et elle a été recueillie du 1er janvier 2008 jusqu'au 1er février 2018. Lorsqu'un « x » apparait dans la case, cela signifie qu'aucune donnée n'était disponible pour ce mois. La température moyenne est considérée comme une variable continue, car la température peut être mesurée précisément dans les dixièmes ou les centièmes de décimales.

Finalement, du côté du prix de vente moyen de l'électricité, on a observé 138 données distinctes sur la période du 1er janvier 2008 au 1er juin 2019. Le prix est affiché en dollar canadien pour 5000 KW. Le prix moyen est considéré comme une variable continue, car le prix de vente peut varier dans les décimales sur le marché d'un ou deux cents.

Observation des données

En observant les données, on peut remarquer certaines relations entre les données. Par exemple, lorsqu'on arrive en hiver, on remarque une augmentation assez conséquente de la production d'électricité moyenne quand la température passe en dessous de la barre des 0 °C.

Donc, est-ce que la production électrique nationale moyenne est accrue de façon exponentielle selon la température moyenne du mois ?

D'un autre côté, on peut remarquer une fluctuation du prix de vente selon la production électrique nationale. Donc, existe-t-il une relation entre le prix de vente de l'électricité et la production électrique moyenne ? De ce fait même, existe-t-il une relation entre le prix de vente de l'électricité, la production électrique moyenne et la température moyenne ?

Analyse des données

L'analyse des données recueillies permettra de voir s'il existe des relations entre la production électrique, la température et le prix de vente de l'électricité. Afin de minimiser les facteurs externes, nous ne prendrons en compte que les années 2016-2017. Cela permettra de minimiser par exemple l'impact de l'inflation et les éventuels évènements tels que l'ouverture d'une nouvelle centrale électrique qui influencerait les données. Nous garderons cependant la totalité des données dans la partie statistique descriptive afin de mieux illustrer la tendance du prix et les cycles de la température et de la production.

Statistique descriptive

Avant tout, il est possible de décrire les différentes données à l'aide de différents diagrammes pour mieux comprendre leur répartition. Dans cette section, les différentes données seront décrites sous forme d'histogramme, de « plot-box », d'un graphique quantile-quantile et d'un graphique à bande. De plus, les données seront accompagnées de leur moyenne, l'écart-type, la variance et la taille de l'échantillon.

L'avantage de l'histogramme est qu'il permet de rapidement voir la distribution et de formuler des hypothèses sur celle-ci. Cependant, il sera également difficile de donner des valeurs précises sur l'entrée statistique. Du côté du graphique « quantile-quantile », ce diagramme permet de rapidement comparer un ensemble de données à un modèle théorique de loi normale. Finalement, le diagramme à bandes permet de visualiser nos données selon le temps, notamment lorsque combiné à un diagramme en boite, il permet d'illustrer le cycle que subissent certaines données comme la température et la production électrique.

La production électrique

Tout d'abord, il y a 138 différentes entrées discrètes pour la production d'électricité moyenne par mois. La moyenne est de 50 662 293 MW/h avec une médiane à 48 800 364 MW/h. L'écart-type entre les données est de 5 825 505 MW/h. Finalement, le coefficient de variance est de 0.114987 avec une variance de 3.393651e+13. D'un autre côté, les graphiques permettent de mieux visualiser la répartition des données.

Nous avons choisi de représenter ces données sous trois diagrammes, soit un diagramme en bande (Figure 7), un histogramme (Figure 1) et un diagramme en boite (Figure 12). Le diagramme en bande permet de représenter les valeurs de la production au fil des mois. L'histogramme permet de voir la distribution des valeurs de la production électrique afin d'orienter le choix des tests et modèles plausibles pour ladite distribution. Le diagramme à boite est conçu de manière à illustrer la périodicité de la production mensuelle, il est possible de voir une répétition des valeurs, ce qui suggère un cycle annuel. Il oriente donc fortement le choix de modèle vers les modèles de Time series.

La température

Comme pour la production électrique, il y a 138 différentes entrées continues pour la température moyenne par mois. La moyenne est de 5.607 Celsius avec une médiane à 6.300 Celsius. L'écart-type entre les données est de 10.44092. Finalement, le coefficient de variance est de 1.862281.

Tout comme la production, la température est représentée par trois diagrammes, soit un diagramme en bande (Figure 9), un histogramme (Figure 3) et un diagramme en boite (Figure 12). Nous avons calqué l'analyse de la température sur celle de la production, en effet, les deux s'avèrent fortement corrélées (voir analyse de la production électrique).

Le prix

Tout d'abord, il y a, comme pour la température et la production, 138 différentes entrées continues pour le prix moyen de vente de 5000 kW par mois. La moyenne est de 96.49 $ avec une médiane à 98.35 $. L'écart-type entre les données est de 11.08646. Finalement, le coefficient de variance est de 0.1148916.

L'analyse du prix est difficile, nous avons usé d'un diagramme à bandes (Figure 8) dans l'objectif de voir son évolution. La distribution du prix est représentée par un histogramme (Figure 2) afin de pouvoir guider l'analyse future. Pour une illustration du cycle potentiel du prix, afin de faire une comparaison avec la production et le temps, un diagramme en boite (Figure 12) a aussi été inclus. Ici, aucun cycle ne semble visible contrairement aux autres données.

Modèle pour les variables

Dans cette section, on proposera un modèle de lois de probabilité pour les différentes variables étudiées.

Ici, selon les diagrammes en boites vus dans la section de statistique descriptive, deux de nos ensembles de données semblent avoir une cyclicité annuelle. Un ensemble de données qui est mesurée à intervalle régulier se nomme une time series, ou série temporelle. Nous basons notre analyse sur l'approche décrite par Tavish Srivastava dans (R. A 2019) (visité en novembre 2019). L'explication des times-series pouvant être à elle seule un autre devoir, nous n'expliquerons que

sommairement l'ensemble des concepts laissant au lecteur la possibilité d'aller voir les méthodes via nos sources.

Afin de pouvoir modéliser ce genre de données, il faut considérer trois critères, soit :

- la moyenne des cycles de la série doit être une constante ;
- la variance ne doit pas être une fonction du temps ;
- et la covariance entre un terme et son suivant dans le cycle ne doit pas être une fonction du temps.

Lorsque les données suivent les trois critères, on dit qu'elles sont une série temporelle stabilisée. Afin de vérifier si la production et la température suivent les critères, il est possible de faire le test de Dickers-Fuller.

La température

Ici on modifie les données de la température à l'aide de boxcox() qui permet de stabiliser la série de température selon la méthode expliquée dans (Gaur, 2019) (visité en novembre 2019). On applique ensuite la fonction diff () afin de stabiliser la moyenne entre les cycles.

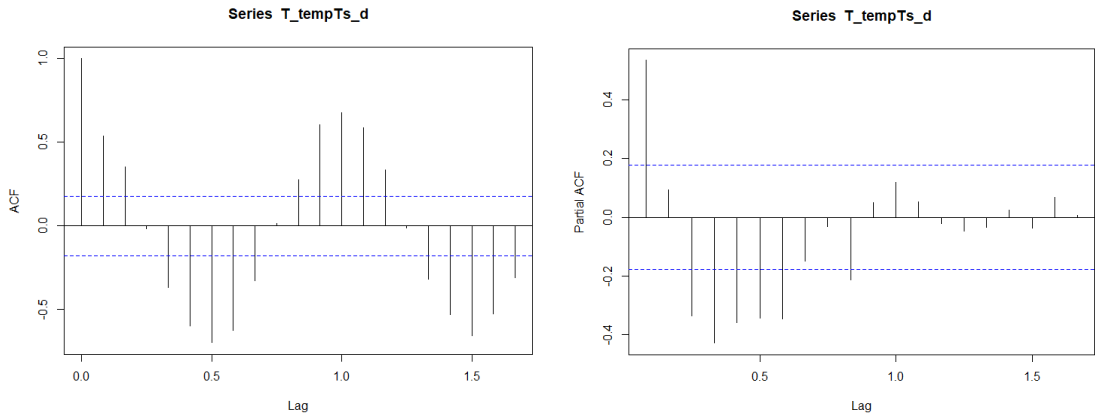
Ensuite, afin de vérifier que la série est bien stable, on use du test de Dickers-Fuller.

```
        Augmented Dickey-Fuller Test

data:  T_tempTs_d
Dickey-Fuller = -5.9034, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

Test de Dickey-Fuller pour la température

Après cela, nous produisons des diagrammes d'autocorrélation de fonction (ACF) et d'autocorrélation partielle (PACF) afin de bien déterminer le type de modèle de série temporelle auquel correspond la température.



Modèle de série temporelle pour la température

Ici deux processus sont possibles, le premier est un modèle AR soit un modèle autorégressif dans lequel un cycle ressemble au dernier cycle avec une erreur. Le deuxième type de cycle est un

modèle MA, où un cycle ressemble au dernier cycle, avec l'ajout des erreurs de celui-ci et d'une nouvelle erreur.

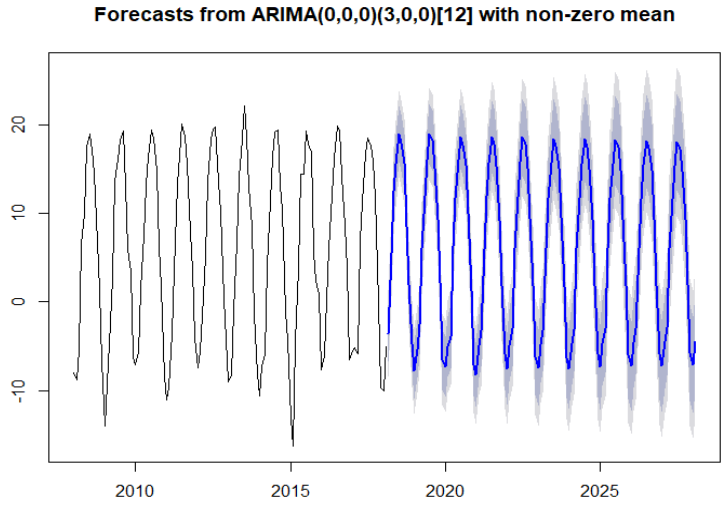
Ici, on remarque à l'aide des diagrammes ACF et PACF que l'autocorrélation partielle du diagramme PACF diminue très rapidement, ce qui identifie un processus autorégressif. Nous avons joué avec les paramètres afin de trouver le modèle qui établit la plus petite erreur quantifiable avec le coefficient AIC, qui permet de jauger l'exactitude du modèle. Après plusieurs itérations nous avons trouvé le modèle qui possède le plus petit AIC et qui permet de modéliser et de prévoir la température au fil du temps. Le paramètre AIC soit *Akaike Information Critera* présente, en une statistique, la corrélation entre le modèle et la théorie ainsi que la simplicité du modèle.

Ici le paramètre 3 correspond au paramètre de l'auto-régression, nous avons donc simplifié le modèle ARIMA qui est un modèle ‘*’Auto Regressive Integrated Moving Average*’’ en modèle autorégressif plus simple qui use simplement des anciennes observations pour prévoir les futures, étant donné que celles-ci sont fortement corrélées selon les diagrammes ACF. Le paramètre 3 signifie que nous prenons en compte les 3 derniers cycles afin de prévoir celui qui vient. Il est à noter que l'incertitude augmente telle que le signifient les barres de confidences du graphique.

```
Call:
arima(x = tempTs, order = c(0, 0, 0), seasonal = list(order = c(3, 0, 0), period = 12))

Coefficients:
        sar1    sar2    sar3  intercept
      0.3537  0.2536  0.3775     5.8183
s.e.  0.0932  0.0975  0.0957     2.8555

sigma^2 estimated as 6.193:  log likelihood = -304.45,  aic = 618.91
```



Graphique ARIMA de la température

La production électrique modélisation ARIMA

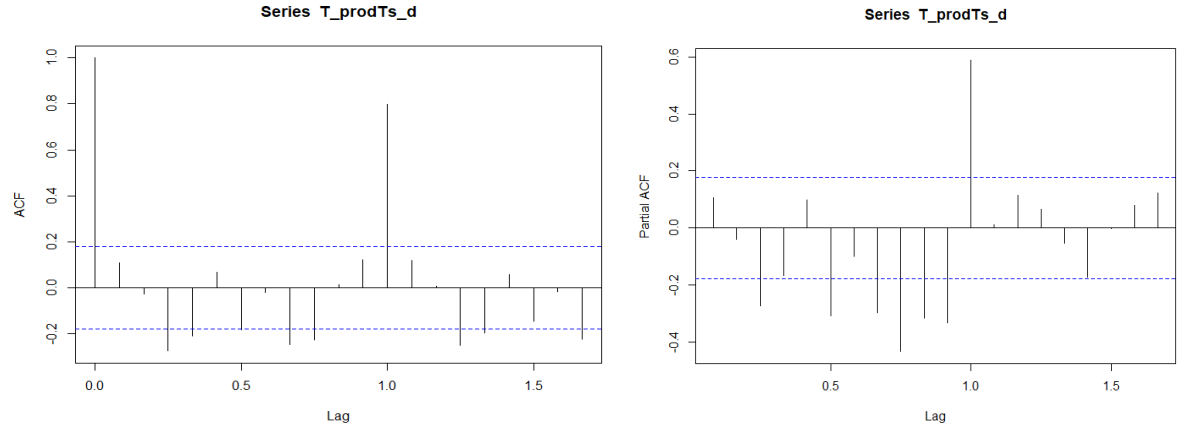
De la même manière que la température, nous effectuons le test de Dickers-Fuller pour nous assurer de la stationnarité de la série après avoir traité nos données avec boxcox().

```
        Augmented Dickey-Fuller Test

data:  T_prodTs_d
Dickey-Fuller = -9.5485, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

Test de Dickey-Fuller pour la production

Ensuite nous générons les diagrammes ACF et PACF afin d'analyser l'autocorrélation entre les cycles et l'autocorrélation partielle qui ne s'explique pas par la similitude entre les cycles.



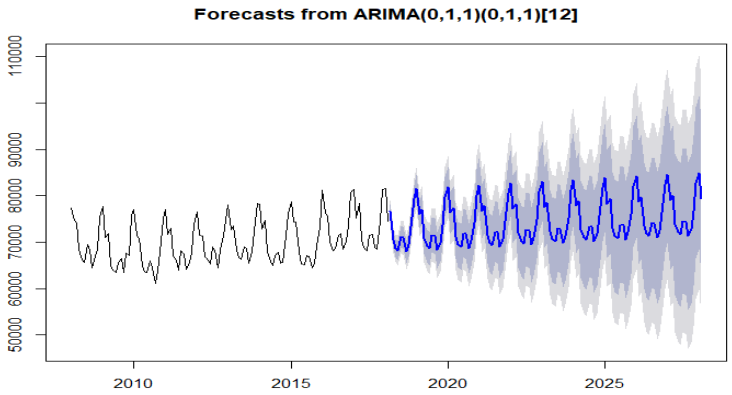
Modèle de série temporelle pour la production

Ici, les diagrammes suggèrent que la fonction est fortement autocorrélée avec ses itérations précédentes soit sur un lag de 0, 1, 2, etc. Par itération sur le coefficient AIC qui est le meilleur nous trouvons donc qu'un modèle ARIMA (0, 1, 1) (01,1) possède la plus forte corrélation avec les données observée.

```
Call:
arima(x = T_prodTs, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1),
    period = 12))

Coefficients:
          ma1     sma1
      -0.5095  -0.8159
s.e.   0.0904   0.1285

sigma^2 estimated as 1546279:  log likelihood = -937.99,  aic = 1881.98
```



Graphique ARIMA de la production

Ici le modèle est beaucoup plus incertain et moins fiable, ce qui se voit dans les barres de confidences et le coefficient AIC beaucoup plus élevé.

Le prix

Tout d'abord, afin de faire corréler nos données avec une loi de probabilité, nous avons dû exclure certaines données du calcul, car celles-ci provoquaient des variables inconnues dues à l'inflation du prix au cours du temps, donc les données des années 2008 jusqu'à 2012 sont exclues. Ce qui nous fait un total de 88 données étudiées des années 2013 à 2019.

Avec ce lot de données, nous pouvons proposer comme modèle une loi Normale. Lorsqu'on passe le test de Shapiro-Wilk à nos données, celui-ci nous renvoie une p-value de 0.8132 et donc on peut voir que la normalité est plausible puisque la p-value est élevée.

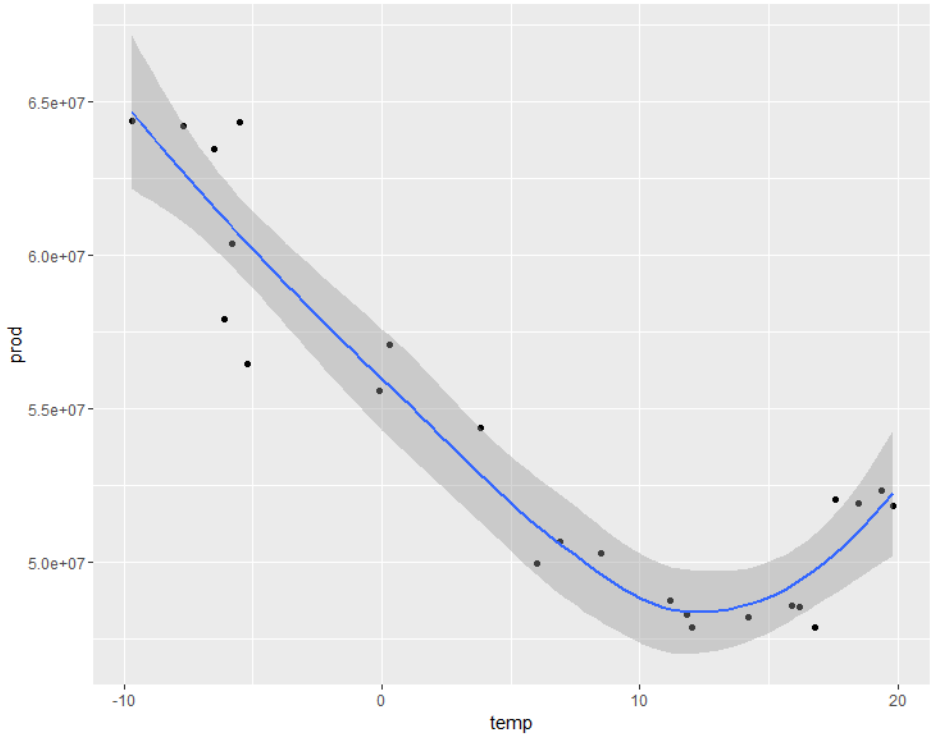
Maintenant que nous avons déterminé que la loi Normale est valide, il nous reste qu'à déterminer la moyenne théorique et la variance théorique. Les deux paramètres ont été calculés à l'aide de l'estimation ponctuelle et de l'intervalle de confiance et cela nous donne 103.4216 et 106.1732 pour la moyenne théorique donc une moyenne de 104.7974. Du côté de la variance, nous avons calculé une variance de 27.79325 et 52.48852.

Finalement, le test d'hypothèse d'une moyenne avec une variance connue t.test nous permet d'accepter ou de rejeter **H0 : μ=Lm** contre **H1 : μ > Lm** ou Lm est la moyenne théorique faible c'est-à-dire 103.4216. Pour donner suite au calcul, nous obtenons un **t = 1.9913 ; df= 77 ; p-value= 0.025**, ce qui nous permet de ne pas rejeter H0, car **t>p-value**. Ceci est donc une conclusion faible, mais valide pour notre modèle.

Régression polynomiale :

La température et la production électrique

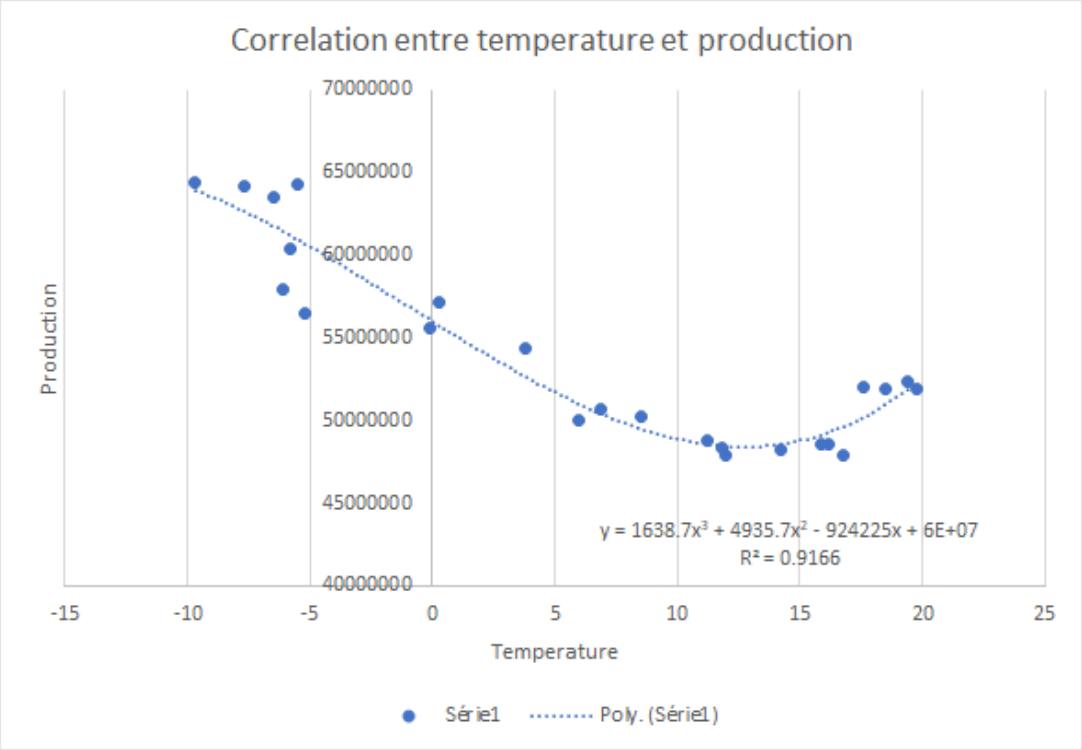
Afin de pouvoir modéliser la relation entre la température et la production, on observe d'abord le graphique de la production en fonction de la température. Afin de limiter l'impact des facteurs externes, seules les années 2016-2017 sont considérées.



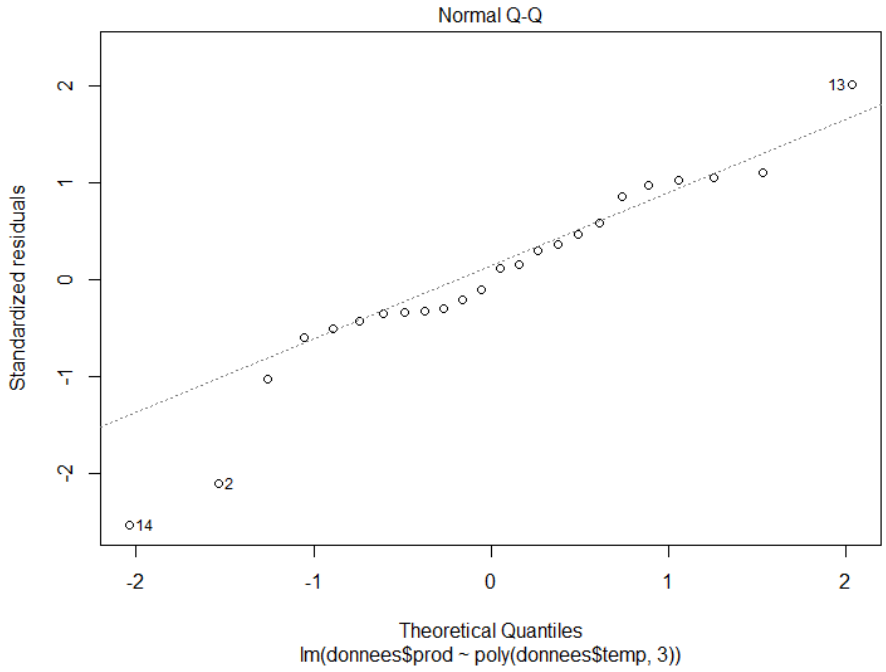
Régression polynomiale de la production selon la température

On remarque que généralement la production d'électricité augmente si la température est basse, diminue lorsque la température augmente et tend vers 10 degrés Celsius puis commence à augmenter de nouveau lorsque la température augmente. Cela correspond intuitivement au fait que nous chauffons à l'électricité en hiver et usons de climatisation en été.

Il est donc immédiatement possible de voir qu'une simple régression linéaire ne suffirait pas à présenter cette relation, c'est pourquoi nous usons d'une régression polynomiale de degré 3 afin d'obtenir une régression linéaire significative.



Régression polynomiale de la production selon la température



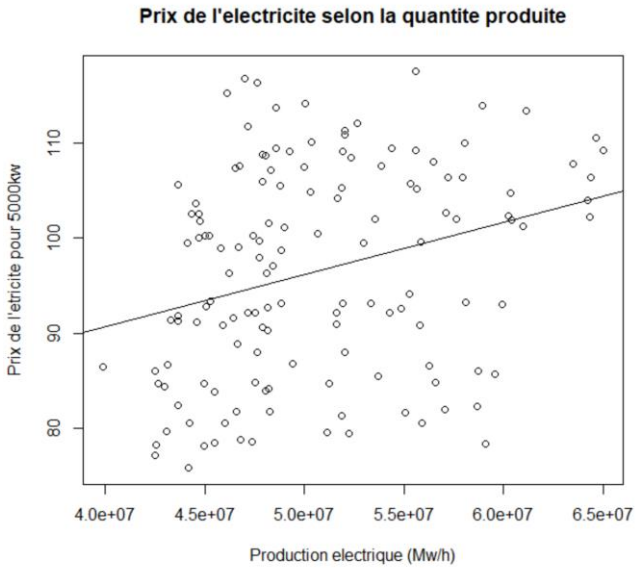
Graphique quantile-quantile de la production selon la température

Le diagramme quartile-quartile permet de voir l'ajustement des données au modèle théorique de la régression linéaire. Outre certaines valeurs aberrantes, les valeurs semblent suivre la régression proposée.

Régression linéaire :

Le prix selon la production

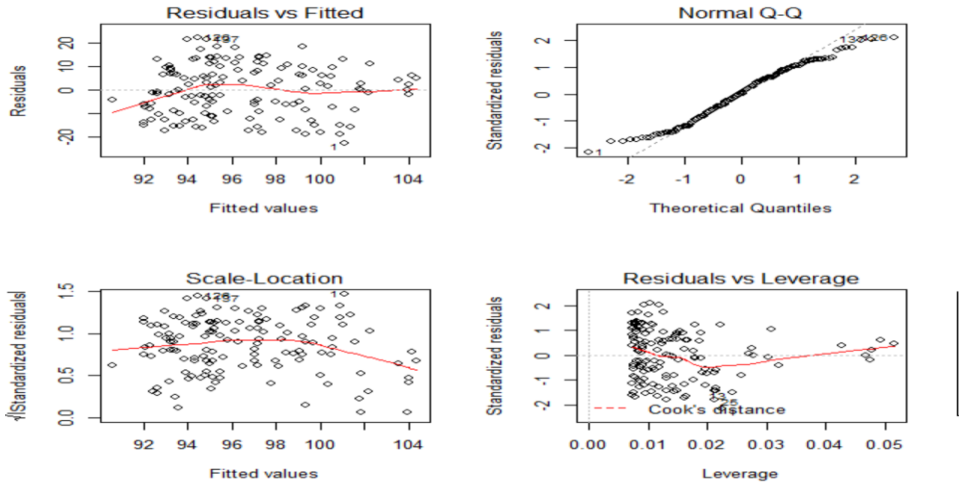
Lorsque nous mettons en relation le prix de l'électricité et et la production électrique dans une régression linéaire, nous obtenons le graphique suivant :



Nuage de points du prix de l'électricité selon la production

Comme on peut le constater, nous avons peu de données qui suivent la droite. De plus, théoriquement, en prenant en compte les lois du marché de base, plus il a de l'offre, plus le prix devrait être bas comme son contraire est valide. Mais ici, nous avons une relation croissante, ce qui ne respecte pas les lois de l'économie de base.

D'un autre côté, nous observons le graphique « quantile-quantile » et les résidus :



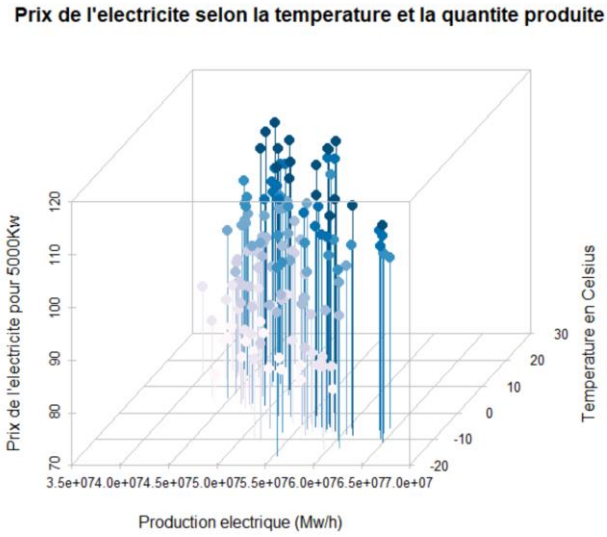
Graphique quantile-quantile et résidus du prix selon la production

Le diagramme quartile-quartile permet de voir permet l'ajustement des données au modèle théorique de la régression linéaire. Outre certaines valeurs aberrantes, les valeurs semblent suivre la régression proposée. Donc, il est tout à fait plausible que nos données suivent cette régression linéaire, mais lorsque nous observons le R^ 2 (0.08363841), nous remarquons que la valeur est beaucoup trop petite pour venir valider notre relation.

Régression multiple :

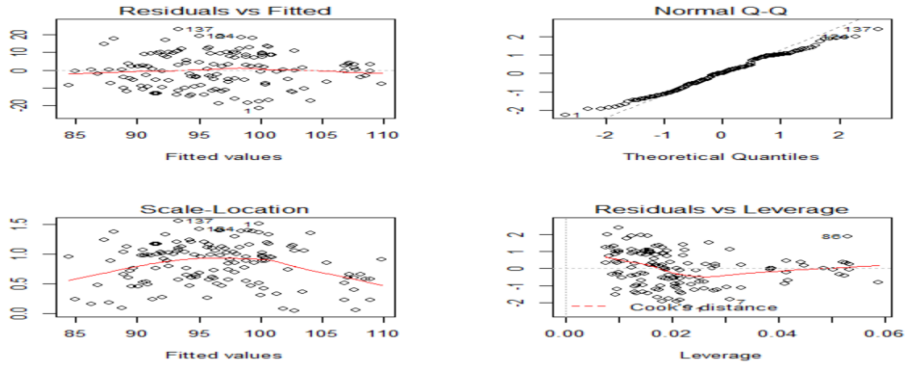
Le prix selon la production et la température

Lorsqu'on met en relation le prix de vente, la production électrique et la température dans une régression multiple, nous obtenons le graphique suivant :



Graphique de la régression multiple du prix selon la température et la production

Comme on peut le constater, nous avons un nuage de points qui ne nous apprend pas grand-chose sur la relation qui unit nos trois variables. Mais, lorsqu'on observe les résidus de cette régression multiple.



Graphique quantile-quantile et résidus de la régression multiple

Le diagramme quartile-quartile permet de voir permet l'ajustement des données au modèle théorique de la régression multiple. Outre certaines valeurs aberrantes, les valeurs semblent suivre la régression proposée. De plus, les données ont l'air d'être réparties uniformément dans le graphique des résidus. Donc, il est tout à fait plausible que nos données suivent cette régression linéaire. Mais lorsque nous observons le R^ 2 (0.2562798), nous remarquons que la valeur est beaucoup trop petite, donc notre modèle est inadéquat. Il n'y a donc pas de lien direct entre la température, le prix et la production.

Conclusion

En conclusion, nous avons pu décrire nos variables, poser des modèles de loi de probabilité à nos différentes variables et les faire corréler entre elles. De plus, nous avons réussi à valider que nos variables de production et de température étaient des times-series, mais, pour le prix, nous avons réussi à apposer le modèle d'une loi normale par une conclusion faible.

Du côté de nos questions ouvertes, il existe une corrélation entre la production électrique et la température. En effet, grâce à la régression polynomiale (Figure 11), on peut voir que la relation entre la température et la production est une relation polynomiale avec un plateau vers 10 degrés Celsius. Donc, nous avions tort sur notre spéculation que la température et la production électrique pourraient être exponentielle. Pour la relation entre le prix de l'électricité et la production électrique, il n'existe aucune corrélation entre les deux qui permettrait de prédire le prix de l'électricité selon la production, donc il n'existe pas de relation entre ses deux variables. De ce fait même, cela vient annuler fait qu'il pourrait avoir une relation multiple entre le prix, la température et la production électrique, donc il n'y pas de modèles valides entre les trois variables. Il serait cependant intéressant d'approfondir la relation trouvée entre la température et la production d'électricité dans une recherche future.

Bibliographie

Gaur, P. (2019). 10 Steps to Build a Time Series Model. Retrieved 2 December 2019, from https://techblog.xavient.com/10-steps-to-build-a-time-series-model/

R, A. (2019). A Complete Tutorial on Time Series Analysis and Modelling in R. Retrieved 2 December 2019, from https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

Figures

Figure 1 : Histogramme de la production

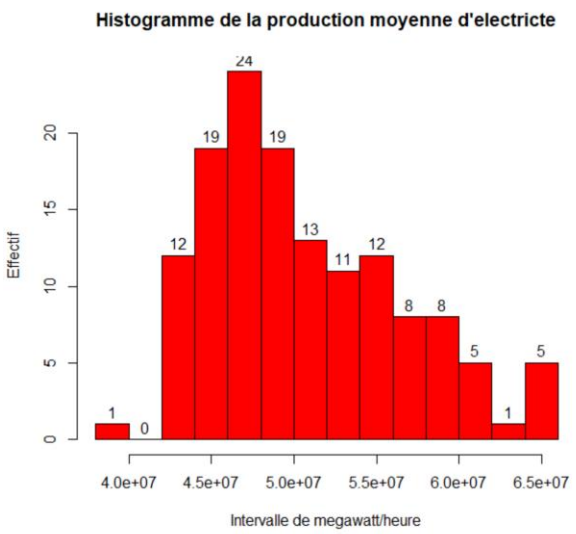


Figure 2 : Histogramme de la vente

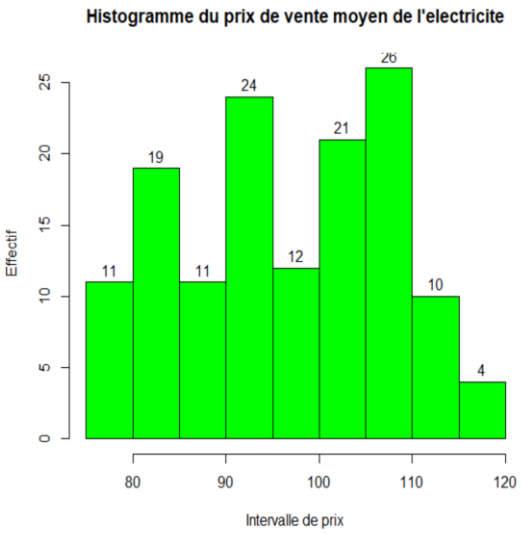


Figure 3 : Histogramme de la température

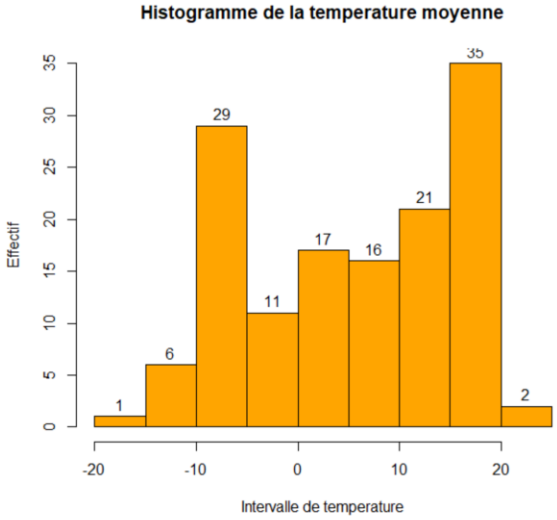


Figure 4 : Diagramme « quantile-quantile » de la production

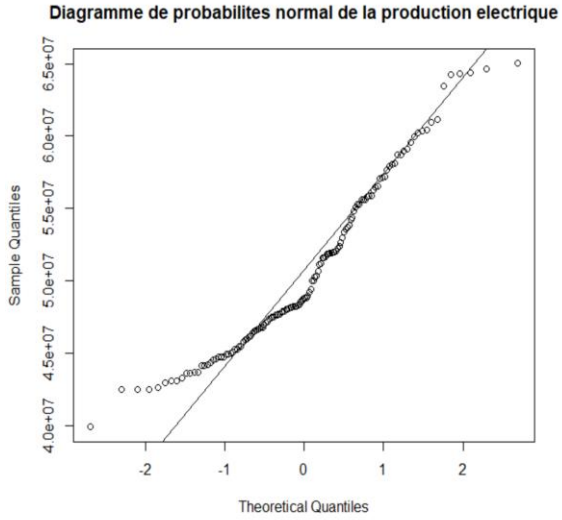


Figure 5 : Diagramme « quantile-quantile » de la vente

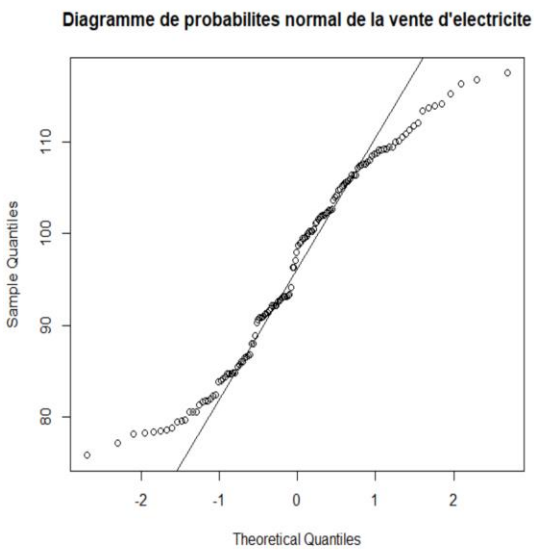


Figure 6 : Diagramme « quantile-quantile » de la température

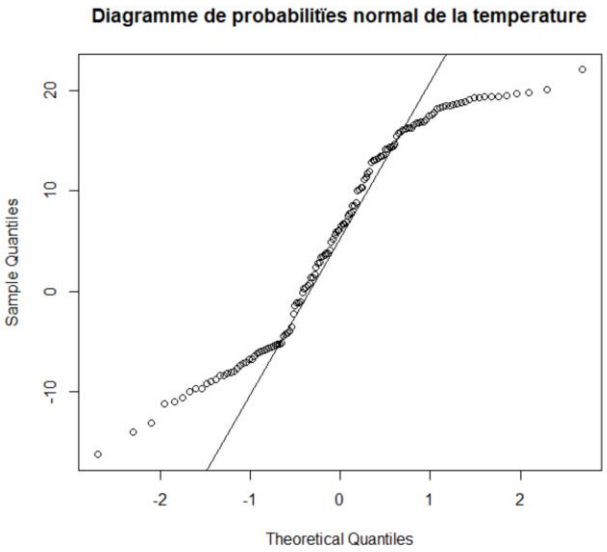


Figure 7 : Dispersion de la production électrique

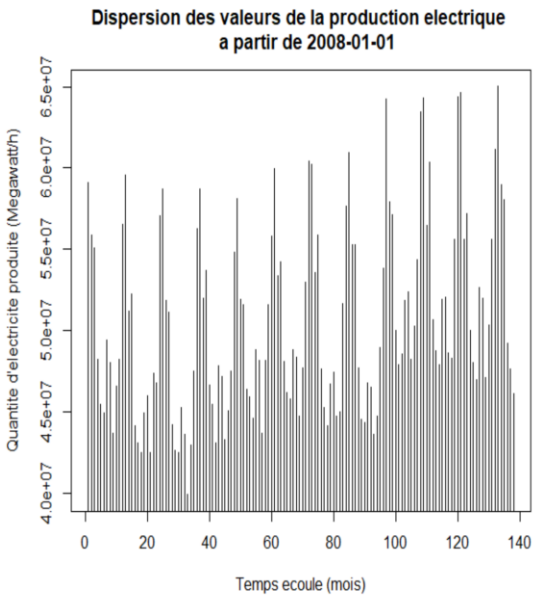


Figure 8 : Dispersion de la vente

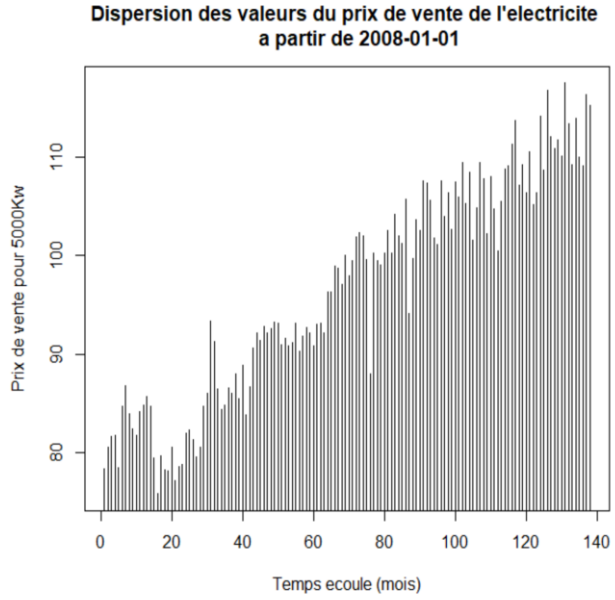


Figure 9 : Dispersion de la température

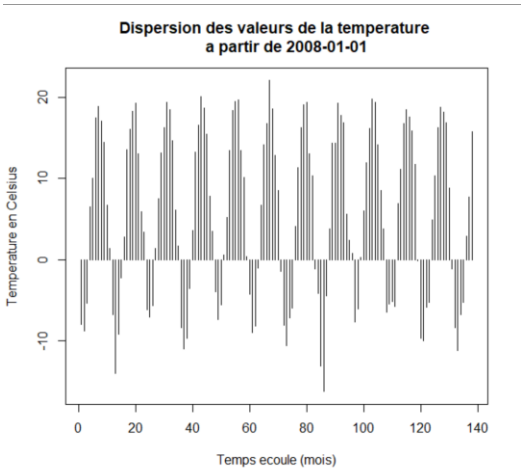


Figure 10 : Diagramme Température-Prix

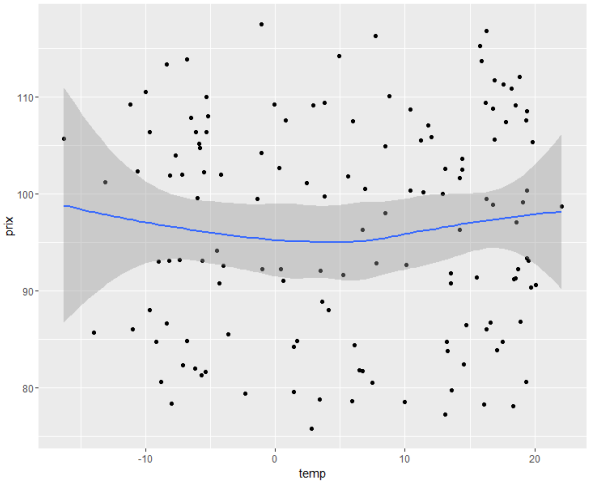


Figure 11 : Diagramme Température-Production

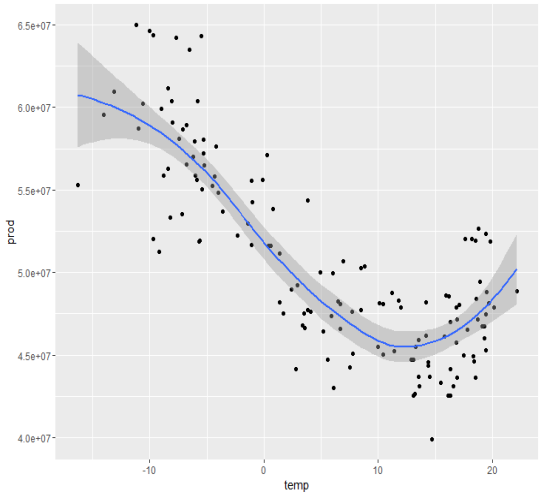


Figure 12 : Diagramme des cycles de la température, de la production et des prix