Les Cahiers du CEDIMES

2024, Volume 19, numéro 3, https://doi.org/10.69611/cahiers19-3-16

ISSN 2110-6045

Sur les séries temporelles économiques ou démographiques très corrélées.

Philippe RYCKELYNCK

Maître de Conférences, Agrégé et Docteur, Université du Littoral-Côte d'Opale, France,
philippe.ryckelynck@univ-littoral.fr

Résumé. Nous examinons un ensemble de séries temporelles venant de la démographie et de l'économie, pour lesquelles l'ajustement linéaire ou l'ajustement quadratique se réalise avec des corrélations très élevées. Pour ces séries chronologiques, nous examinons également les autocorrélations. Nous donnons aussi quelques exemples très résistants à la prévision, quelque soit le modèle linéaire.

Mots clés : corrélation, ajustement linéaire, ajustement parabolique, autocorrélations, moyennes mobiles.

Abstract. This paper is devoted to the examination of a set of chronological series either economical or demographical, for which linear fit or parabolic fit may be done with very high correlations. We study also internal correlations. We give lastly some examples which fail linear models.

Keywords : Correlation, linear fit, parabolic fit, autocorrelation, moving averages, least squares method.

Classification JEL : C22, C25, C35, C52, C82, J11.

1. Introduction et méthodologie.

Cet article est consacré à l'examen de certaines suites chronologiques d'origine démographique ou économique, présentant un phénomène de corrélation excessivement grand ou, à l'inverse, une irrégularité rendant caduques différentes méthodes de prévision. Quoique le sujet ne soit nullement nouveau d'un point de vue purement mathématique, ce qui doit être mis en exergue ici est *le caractère authentique des données, leur collection, et la quantité de phénomènes desquels il faut extraire ces données pour isoler de telles circonstances relatives à la prédiction de l'évolution temporelle*. Rassembler de tels résultats a demandé de nombreux tests souvent décevants ou infructueux, en conservant les pépites statistiques que le hasard pouvait apporter. Mais pour organiser ces analyses, nous avons abordé le sujet à travers des exemples rassemblés par sujets, en premier chef la démographie, en second, l'économie des pays, et en reportant à un travail ultérieur les données financières et les modèles plus complexes.

Pour commencer, nous donnons quelques indications méthodologiques sur les modèles employés et leur traitement informatique. Toutes les séries ci-après sont temporelles, scalaires, et viennent des institutions les plus soigneuses pour collecter des données, l'Insee, le Sénat, l'Assemblée Nationale, et aussi la Banque Mondiale, etc. Quand cela a été possible, nous avons comparé les données d'un Institut avec celles d'un autre, et conservé constamment les données des sites ci-dessus signalés (dans l'ordre où ils sont cités).

Rappelons en premier chef les quelques formules qui seront employées. Une suite temporelle (ou chronologique) est une suite brute $Y = (y_t)$ indicée par le temps $X = (x_t)$. Lorsque le nuage de points $(x_1, y_1), \dots, (x_n, y_n)$ d'une série statistique bivariée (X, Y) présente une forme « allongée », il est naturel d'approcher le nuage par une droite de la forme $y = ax + b$ où a et b sont des paramètres à déterminer. En pratique, les variables X et Y ne sont pas directement liées par une droite : pour chaque donnée i, il existe une erreur e_i entre la réalité et l'approximation « idéale » par la droite. On note cette erreur : $e_i = y_i - (ax_i + b_i)$. Le plus souvent, la recherche de a et de b s'entend au sens des moindres carrés : on choisit ces paramètres de telle sorte qu'ils rendent minimale l'erreur : $\Delta = \sum_{i=1}^{n} e_i^2$. On sait qu'il existe un unique couple (a, b) rendant minimale l'erreur au sens des moindres carrés (pour la régression linéaire) et celui-ci est donné par : $a = Cov(X, Y)/Var(X)$ et $b = \overline{y} - a\overline{x}$, avec les notations traditionnelles. Systématiquement, la série X sera celle des dates, tandis que Y sera la série intéressante. L'interprétation de a est importante, et a correspond à l'augmentation moyenne de la quantité Y par période, au moins lorsque les dates composant la suite $X = (x_t)$ sont équiréparties. Cette dernière condition n'est pas garantie systématiquement lorsqu'on travaille avec des données démographiques issues de recensements à des dates irrégulières, de données financières indexées par les jours ouvrés, etc. L'écart-type $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ et son carré, la variance, permettent de gagner en précision dans l'ajustement affine, si l'on « écarte » les mesures dans le temps, mais cela sera au détriment du nombre de points collectés. L'équation en considération est donc :

$$y - \overline{y} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}(x - \overline{x}).$$

Ensuite, en adoptant des notations vectorielles, on trouve l'erreur commise sur le vecteur empirique $\mathbf{y}$ par la régression $a\mathbf{x} - b\mathbf{z}$ (où $\mathbf{z}$ est le vecteur unité) sous la forme classique

$$\|\mathbf{y} - a\mathbf{x} - b\mathbf{z}\|^2 = n\mathrm{Var}(y)\left(1 - \frac{\mathrm{Cov}^2(x, y)}{\mathrm{Var}(x)\mathrm{Var}(y)}\right) = n\mathrm{Var}(y)(1 - \rho^2),$$

où $\rho \in [-1,1]$ est le coefficient de corrélation linéaire, de sorte que ρ^2 n'est autre que le fameux coefficient R2 des tableurs. Les calculs faits ci-dessous l'ont été dans le logiciel de calcul formel Maple, qui permet d'introduire des modèles linéaires de toute sorte, et aussi des paramètres variables dans les séries comme la suite l'indiquera. La figure ci-dessous donne un petit programme pour les logiciels Maple, Mupad ou Xcas pour la régression linéaire par les moindres carrés. Outre sa concision, on observera d'une part sa simplicité par rapport à l'emploi d'un tableur et d'autre part la possibilité de créer des simulations de façon répétée et sûre[1].

```
RLMCO := proc(X, Y)
local n, k, a, b, r, m1, m2, m3, m4;
    n := nops(X);
    m1 := sum(X[k], k = 1 .. n) / n;
    m2 := sum(X[k]^2, k = 1 .. n) / n;
    m3 := sum(Y[k], k = 1 .. n) / n;
    m4 := sum(X[k]*Y[k], k = 1 .. n) / n;
    a := (m1*m3 - m4) / (m1^2 - m2);
    b := (m1*m4 - m2*m3) / (m1^2 - m2);
    [a, b]
end proc
```

Ajoutons une remarque sur la corrélation et la causalité. Le fait que deux variables soient fortement corrélées provient, a priori, du fait que les variables sont liées. En revanche, une forte

[1] Les expérimentations avec un tableur occasionnent des erreurs dont la plupart viennent de l'adressage des plages de données. Par ailleurs, les tableurs ne permettent pas de traiter des séries avec paramètre indéterminé, et ne gèrent pas les modèles à plusieurs variables. Le logiciel R est une excellente alternative, et mieux encore Maple.

corrélation ne suffit pas pour établir une causalité entre ces deux variables : d'autres facteurs peuvent entrer en ligne de compte. C'est tout particulièrement le cas lorsque X est une suite de dates, parce que si on trouve deux phénomènes $Y = (y_t)$ et $Z = (z_t)$, d'ordres très dissemblables (par exemple Y est une donnée liée au PIB et Z est une donnée sur la vaccination des individus), pour lesquels chacun des deux modèles de régression $Y = a_1X + b_1 + \varepsilon_1$ et $Z = a_2X + b_2 + \varepsilon_2$ est très concluant, alors on pourra inférer que les relations précédentes sont causales et, par transitivité, qu'on a également une explication du PIB par la vaccination des citoyens, ce qui ne saurait manquer d'être à juste titre contesté. Nous donnerons aussi ci-dessous les courbes fournissant les coefficients d'autocorrélation de différents ordres, autrement dit les coefficients de corrélation linéaire entre la série originelle brute **x** et la série obtenue par décalage d'indices, ce décalage variant entre 1 et la longueur n-1 des observations réalisées. Il arrive que ces auto-corrélations puissent encore affiner la prévision de la série, et nous en donnerons quelques exemples.

Lorsqu'on aborde l'adéquation parabolique au sens des moindres carrés, la matrice des moments non-centrés ci-après joue un rôle fondamental pour trouver la loi régressive :

$$moment := \begin{bmatrix} 1 & \frac{\sum_{k=1}^{n} x_k}{n} & \frac{\sum_{k=1}^{n} x_k^2}{n} \\ \frac{\sum_{k=1}^{n} x_k}{n} & \frac{\sum_{k=1}^{n} x_k^2}{n} & \frac{\sum_{k=1}^{n} x_k^3}{n} \\ \frac{\sum_{k=1}^{n} x_k^2}{n} & \frac{\sum_{k=1}^{n} x_k^3}{n} & \frac{\sum_{k=1}^{n} x_k^4}{n} \end{bmatrix}$$

Après quelques calculs formels, on peut donner, dans la figure ci-dessous, un programme dans le langage Maple pour la régression parabolique par les moindres carrés et l'obtention des coefficients de la parabole par degré ascendant :

```
moments := proc(X, Y)
local a, b, c, n, res;
    n := nops(X);
    a := (sum(X[k]^2, k = 1 .. n)*sum(X[k]^4, k = 1 .. n) - sum(X[k]^3, k = 1 .. n)^2)*sum(Y[k], k = 1 .. n) / n^3
        - (sum(X[k], k = 1 .. n)*sum(X[k]^4, k = 1 .. n) - sum(X[k]^2, k = 1 .. n)*sum(X[k]^3, k = 1 .. n))*sum(Y[k]*X[k], k = 1 .. n) / n^3
        + (sum(X[k], k = 1 .. n)*sum(X[k]^3, k = 1 .. n) - sum(X[k]^2, k = 1 .. n)^2)*sum(Y[k]*X[k]^2, k = 1 .. n) / n^3;
    b := -(sum(X[k], k = 1 .. n)*sum(X[k]^4, k = 1 .. n) - sum(X[k]^2, k = 1 .. n)*sum(X[k]^3, k = 1 .. n))*sum(Y[k], k = 1 .. n) / n^3
        + (sum(X[k]^4, k = 1 .. n)*n - sum(X[k]^2, k = 1 .. n)^2)*sum(Y[k]*X[k], k = 1 .. n) / n^3
        + (sum(X[k], k = 1 .. n)*sum(X[k]^2, k = 1 .. n) - sum(X[k]^3, k = 1 .. n)*n)*sum(Y[k]*X[k]^2, k = 1 .. n) / n^3;
    c := (sum(X[k], k = 1 .. n)*sum(X[k]^3, k = 1 .. n) - sum(X[k]^2, k = 1 .. n)^2)*sum(Y[k], k = 1 .. n) / n^3
        + (sum(X[k], k = 1 .. n)*sum(X[k]^2, k = 1 .. n) - sum(X[k]^3, k = 1 .. n)*n)*sum(Y[k]*X[k], k = 1 .. n) / n^3
        - (sum(X[k], k = 1 .. n)^2 - sum(X[k]^2, k = 1 .. n)*n)*sum(Y[k]*X[k]^2, k = 1 .. n) / n^3;
    res := [a, b, c];
    RETURN(res)
end proc
```

Pour finir, lorsqu'on emploie la série des moyennes mobiles (y_t) d'une variable on emploiera toujours la version la plus usitée[1], à savoir :

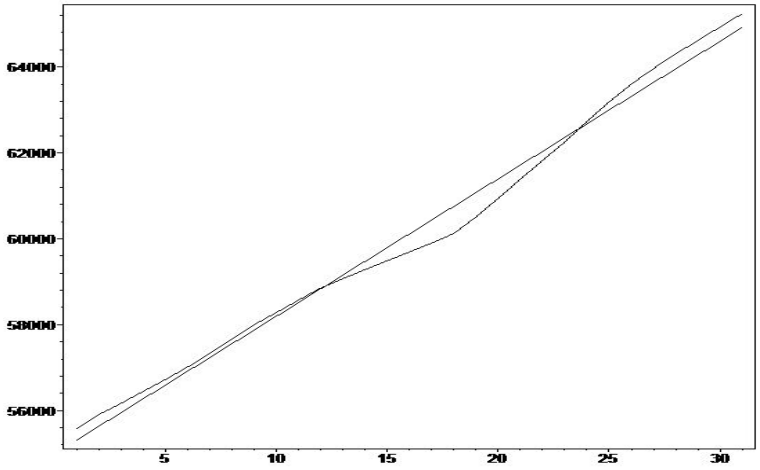
[1] On consultera avec profit le livre de Montfort et Gourieroux, devenu le livre de chevet des statisticiens-économistes en France. On note que l'algorithme des moyennes mobiles, dûment complété par les valeurs idoines au bord, c'est-à-dire pour les valeurs $\widetilde{y_1}$, $\widetilde{y_2}$ et $\widetilde{y_{T-1}}, \widetilde{y_T}$, donne un endomorphisme sur l'espace des suites, qui est contractant autour de la projection sur la moyenne. Ainsi, il ne faut pas réitérer cet algorithme plusieurs fois au risque de trouver une série statistique constante.

$$\tilde{y_t} = \frac{1}{4}\left(\frac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}y_{t+2}\right).$$

2. Les résultats des recensements de la France métropolitaine.

L'Insee publie régulièrement les chiffres de la population totale de la France métropolitaine (en milliers d'habitants). Le tableau ci-dessous les rassemble dans la période de 31 ans allant de 1982 à 2012.

Année	Pop.	Année	Pop.	Année	Pop.	Année	Pop.
1982	55573	1990	57996	1998	59899	2006	63186
1983	55905	1991	58280	1999	60123	2007	63601
1984	56166	1992	58571	2000	60508	2008	63962
1985	56445	1993	58852	2001	60941	2009	64305
1986	56720	1994	59070	2002	61385	2010	64613
1987	57012	1995	59281	2003	61824	2011	64933
1988	57325	1996	59487	2004	62251	2012	65241
1989	57660	1997	59691	2005	62731		



La population totale de la France au fil des recensements.

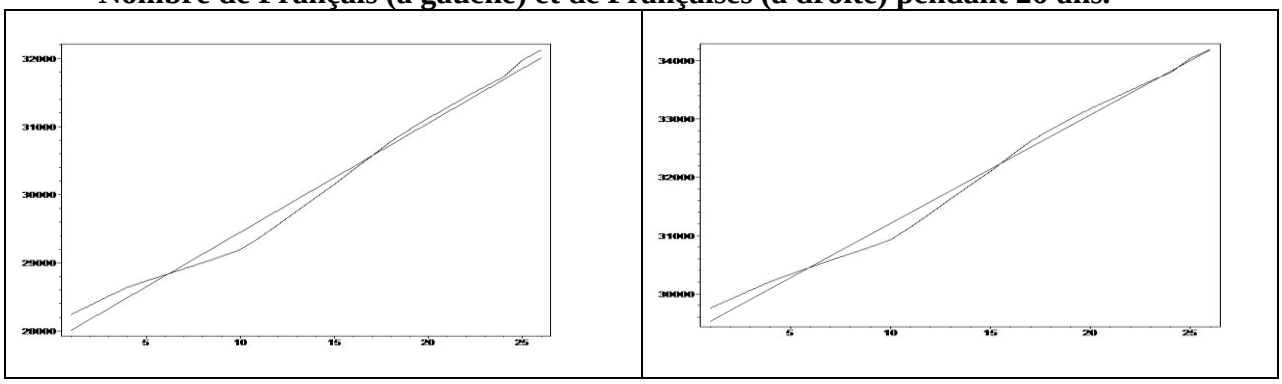
Notons X la date (comptée de 1 à 31, soit $X = 0$ en 1980), Y (resp. H, F) le nombre de milliers d'habitants (resp. de français, de françaises). On trouve ainsi les deux valeurs a=320.2810484 et b=54989.6 de sorte que $Y = 320.28 \cdot X + 54989.6 + \varepsilon_Y$, et le coefficient ρ_Y est exceptionnel, ρ_Y =99,5%. On donne maintenant les chiffres sur la période de 26 ans entre 1990 et 2015 les nombres de Français et de Françaises en milliers, respectivement.

Année	Hommes	Femmes	Année	Hommes	Femmes
1990	28241	29755	2003	29957	31867

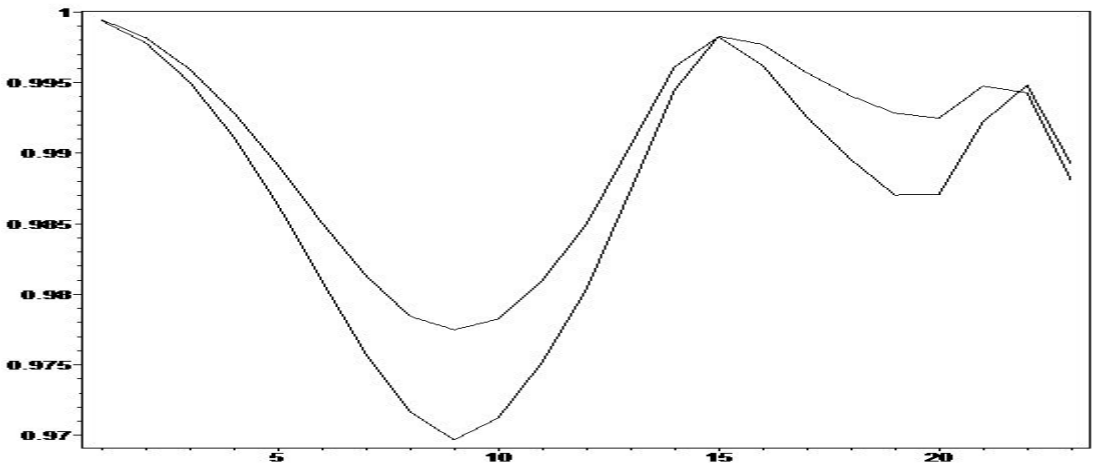
Année	Hommes	Femmes	Année	Hommes	Femmes
1991	28375	29905	2004	30152	32099
1992	28511	30060	2005	30366	32365
1993	28641	30211	2006	30570	32616
1994	28735	30335	2007	30782	32819
1995	28824	30456	2008	30958	33004
1996	28912	30575	2009	31126	33179
1997	29003	30688	2010	31280	33333
1998	29095	30804	2011	31441	33492
1999	29196	30926	2012	31588	33653
2000	29366	31142	2013	31732	33793
2001	29560	31381	2014	31978	34043
2002	29759	31626	2015	32126	34192

Ici, la régression est exceptionnellement bonne. Voici les nuages de points pour ces recensements suivant le genre.

Nombre de Français (à gauche) et de Françaises (à droite) pendant 26 ans.



La courbe suivante donne, sur un même graphique, les deux séries des auto-corrélations des courbes de populations suivant le genre et révèlent une constance du caractère linéaire dans le temps de ces deux variables.



On obtient trois modèles particulièrement dignes de confiance
$H = 160.1818803 \cdot X + 27848.08308 + \varepsilon_H$, avec ρ_H =99,34%,
$F = 186.1596581 \cdot X + 29345.26769 + \varepsilon_F$, avec ρ_F =99,53%,
$F = 1.158903137 \cdot H - 2920.884099 + \varepsilon_{FH}$, avec ρ_{FH} =99,91%.
Le quatrième modèle est ainsi celui qui fournit l'un des coefficients de détermination les plus remarquables trouvés empiriquement.

3. Recensements historiques de la France métropolitaine.

Le premier recensement exhaustif[1], appuyé par des méthodes scientifiques de récolte de données, en France, date de 1851. Avant la seconde guerre mondiale, 17 recensements ont eu lieu. Le site de l'Insee couvre aussi les 33 premiers recensements scientifiques de la France, répartis sur une période de 165 ans et fournit les chiffres de 87 années non consécutives. Lors des années intermédiaires assez récentes, les chiffres sont interpolés par l'Insee et sont respectés, bien sûr, dans le tableau officiel ci-dessous[2]. Le tableau ne conserve pas les données des 87 années fournies par l'Insee, mais seulement 19 d'entre elles. Le traitement a porté bien entendu sur les 87 années.

Année	Ensemble	Hommes	Femmes	Année	Ensemble	Hommes	Femmes
1851	36452	18128	18324	1972	51486	25179	26307
1856	36696	18196	18500	1973	51916	25407	26509
1861	37386	18645	18741	1974	52321	25630	26691
1962	46422	22552	23870	2005	60963	29519	31444
1963	47573	23149	24425	2006	61400	29714	31685
1964	48059	23389	24670	2007	61795	29918	31878
1969	50108	24418	25690	2012	63376	30699	32677
1970	50528	24656	25873	2013	63652	30840	32812
1971	51016	24929	26087	2014	63920	30976	32944
…	…	…	…	2015	64204	31119	33085

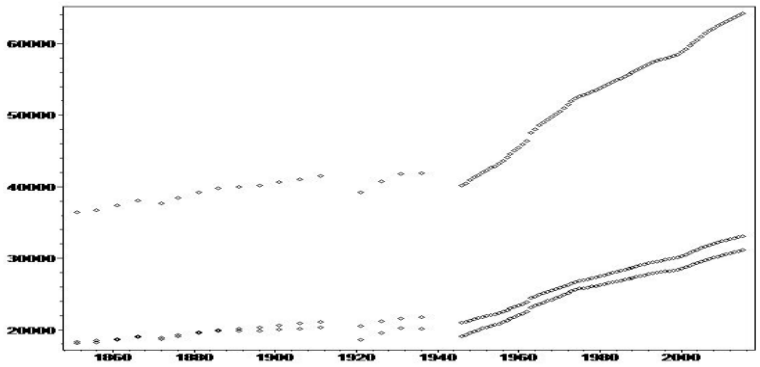
Notons X la date (comptée de 1 à 87, soit $X = 0$ en 1850), Y (resp. H, F) le nombre de milliers d'habitants (resp. de français, de françaises). Par rapport au paragraphe précédent, une triple modification surgit : les dates sont légèrement irrégulières, et la période est considérablement plus longue, enfin les données empiriques manquantes à certaines dates sont remplacées par des estimations obtenues par des méthodes de maximisation de vraisemblance. Cela étant on trouve quatre modèles régressifs :

$Y = 183.1211740 \cdot X - 309140.4009 + \varepsilon_Y$, avec $\rho_F = 90{,}8\%$;
$H = 86.3419188 \cdot X - 144998.2616 + \varepsilon_H$, avec ρ_H =89% ;
$F = 96.7781933 \cdot X - 164140.0662 + \varepsilon_F$, avec ρ_f =92% ;
$F = 1.077412772 \cdot H - 550.3592224 + \varepsilon_{FH}$, avec $\rho_{FH} = 99{,}55\%$.

[1] Voir https://www.insee.fr/fr/statistiques/6683031?sommaire=6683037

[2] Différentes méthodes d'interpolation entre les recensements ont été employées et sont possibles ; la méthode d'interpolation linéaire n'est pas la meilleure. La méthode de maximisation de la vraisemblance est plus pertinente mais elle suppose qu'on dispose de la loi de probabilités à laquelle obéit la population et que cette loi ait peu de paramètres.

La population de la France au fil des recensements, suivant le genre.



Le quatrième modèle, liant F, H, est le seul qui nous contente pleinement, avec un coefficient remarquable de 99,55 %, montre trois conséquences inattendues. D'abord, l'évolution de la population d'un genre ou de l'autre n'est pas linéaire, pas davantage que la population totale. Le graphique montre clairement les deux décrochements lors des deux guerres mondiales du XXème siècle, qui ont créé la décorrélation entre les dates et le nombre d'habitants de chacun des deux sexes. Ensuite, chaque année, en moyenne, le solde naturel de la population féminine croît de 7 % en faveur des femmes, et bien entendu, cet effet n'est pas dû aux naissances mais à l'espérance de vie nettement supérieure pour les femmes par rapport aux hommes. Une troisième observation, d'ordre méthodologique, tient en ceci : si on dispose de trois variables X, Y, Z et qu'on ait expérimenté que X et Z sont peu corrélées (autrement dit pas assez), et qu'il en va de même de Y et Z, alors il peut se faire que la corrélation de X et Z soit toutefois de bonne qualité.

4. L'extraordinaire prévisibilité de la population bulgare.

Pour finir cette famille d'exemples démographiques, prenons maintenant le cas de la population bulgare pendant la décade de 2006 à 2015

2006	**2007**	**2008**	**2009**	**2010**
7688573	7629371	7572673	7518002	7467119
2011	**2012**	**2013**	**2014**	**2015**
7421766	7369431	7327224	7284552	7245677

La prévision est excellente avec un taux exceptionnel de $\rho = 99{,}8$ % et donne lieu à la droite de régression $Y = -49200{,}65455X + 106370354{,}8 + \varepsilon_B$, en excellent accord pendant les huit ans qui ont suivi. Ici, avec une grande stabilité politique et sociale, la perte de 49 mille personnes annuellement doit être interprétée principalement comme le départ définitif de personnes étudiant ou travaillant à l'étranger. Il est remarquable que si l'on procède à une régression quadratique $y = ax^2 + bx + c + \varepsilon$ pour cet exemple, le coefficient de détermination ρ atteint un « Everest » de précision et est égal à 99,999999 % (il y a bien huit chiffres 9). Cependant les trois coefficients a, b, c sont plutôt compliqués, hormis $a = 1195{,}5$, et le gain de précision est entâché par le caractère « artificiel » que prend la parabole régressive.

5. Taux d'accroissement naturel de huit pays.

Le dernier exemple démographique que nous traitons concerne les taux annuels de progression

des populations des pays de l'Afrique entre 1960 et 2015, soit 56 années[1]. Rappelons que la croissance de la population (% annuel) correspond au taux exponentiel de croissance de la population au milieu de l'année n-1 à n, exprimée en pourcentage. Ainsi, le bon moyen d'évaluer la croissance de la population sur une longue période consiste à calculer la moyenne géométrique des taux augmentés de 1. Nous donnons ci-dessous les cas de seulement huit pays africains, et les taux des quatre années qui enferment la période. La cinquième colonne contient le taux moyen, obtenu à partir des moyennes géométriques indiquées auparavant, et la sixième colonne ces moyennes géométriques elles-mêmes.

	1960	**1962**	**2014**	**2015**	**taux moy**	**Mult.**
Gambie	3,025	1,786	3,232	3,201	1,031	5,428
Ghana	3,161	3,138	2,350	2,300	1,026	4,173
Guinée	1,571	1,561	2,698	2,677	1,023	3,514
Guinée-Bissau	1,246	1,047	2,439	2,404	1,020	2,996
Guinée-Equat.	1,281	1,107	2,943	2,902	1,021	3,274
Tanzanie	2,895	2,949	3,154	3,130	1,032	5,765
Ouganda	3,038	3,279	3,254	3,253	1,030	5,327
Madagascar	2,373	2,441	2,784	2,777	1,028	4,761
Monde	NA	1,353	1,180	1,182	0,0162	2,4206

Les études d'ajustement de tous ces taux donnent des conclusions très décevantes, le meilleur coefficient de corrélation linéaire ρ de l'ensemble concerne la Guinée-Bissau, mais n'est que de 85 %, vient ensuite celui de la série de Madagascar égal à 60 %, les autres étant inférieurs à 46 %. Il en va de même de l'analyse des corrélations entre deux séries de taux annuels de progression pour deux pays différents, indépendamment de l'année. Il est surprenant que ces séries soient aussi irrégulières et elles résistent aussi à la régression parabolique. En revanche, l'ajustement de la série relative au Monde en sa totalité donne un coefficient nettement plus favorable, de 90 %, qui traduit une certaine régularité des données globales sans laisser de place à la prévision, par cette méthode élémentaire, du moins.

6. Composition de la population mondiale future.

Dans ce paragraphe nous étudions des prévisions sur la composition de la population mondiale future, répartie en différentes tranches d'âges à 17 époques lointaines[2]. Les chiffres indiqués sont les pourcentages de la population mondiale dans les tranches d'âge indiquées. Bien entendu, à partir de 2025, il s'agit d'estimations.

Année	**0-4**	**5-14**	**15-24**	**25-49**	**50+**	**Année**	**0-4**	**5-14**	**15-24**	**25-49**	**50+**
2020	8,7	16,8	15,5	34,9	24,2	**2065**	9,9	17,2	14,6	29,1	29,2
2025	8,7	16,2	15,2	34,2	25,8	**2070**	10,2	17,6	14,8	28,9	28,5
2030	8,6	15,9	15,0	33,5	27,0	**2075**	10,5	18,1	15,0	28,7	27,7
2035	8,7	15,9	14,6	32,7	28,2	**2080**	10,9	18,5	15,3	28,5	26,8
2040	8,7	15,9	14,3	31,7	29,4	**2085**	11,2	19,0	15,5	28,4	26,0
2045	8,8	16,0	14,3	30,8	30,0	**2090**	11,6	19,4	15,6	28,3	25,1
2050	9,0	16,2	14,3	30,3	30,1	**2095**	11,9	19,9	15,8	28,1	24,2

[1] Voir http://donnees.banquemondiale.org/indicateur/SP.POP.GROW?view=chart

[2] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, Online Edition. Rev. 1.

Année	0-4	5-14	15-24	25-49	50+	Année	0-4	5-14	15-24	25-49	50+
2055	9,3	16,4	14,4	29,9	30,0	**2100**	12,2	20,3	16,1	28,0	23,4
2060	9,6	16,8	14,4	29,5	29,7						

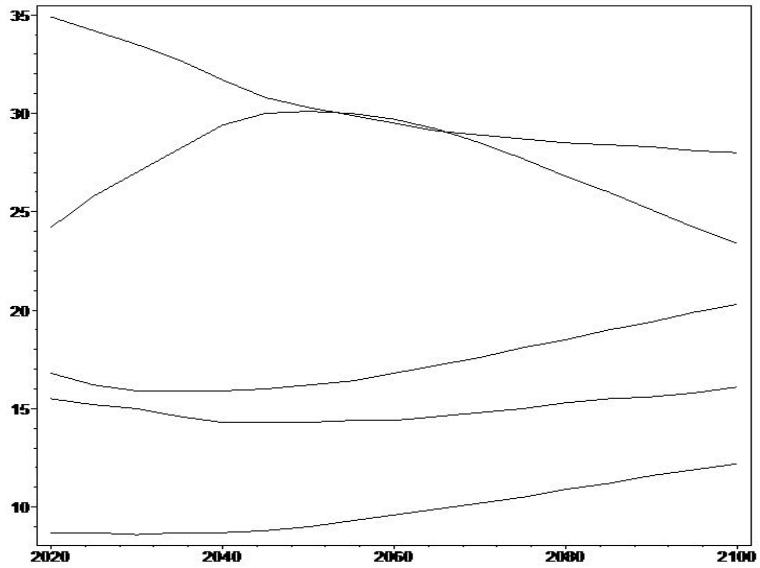
Les cinq scénarii de régression linéaire correspondant aux cinq tranches d'âges proposées offrent des résultats mitigés. Pour les deux tranches d'âge 15-24 ans et 50 ans et +, les coefficients ρ valent respectivement 52 % et 33 %. On ne peut guère augurer des suites futures de l'importance de ces tranches d'âge dans plus de 80 ans. Les trois autres tranches, dans l'ordre du tableau, offrent des coefficients ρ de 96 %, 91 %, 95 %. On peut donc retenir que les tendances suggérées dans le tableau peuvent être extrapolées sous les formes :

"Tranche 0-4ans" = 0,048 X -88,85 et "Tranche 25-49ans" = -0,0085 X + 205,32.

Ici X est la date en années, autrement dit le millésime. Ces conclusions ne manquent pas de nous laisser rêveur, car la tranche d'âge de jeunes adultes serait amenée à perdre de l'importance dans les décades lointaines (au rythme moyen de 0,85 %), tandis que, dans ce temps même, la tranche des tout-petits croîtrait au rythme très important de 5 %. Ces deux modèles ne doivent donc être employés que dans les décennies précédant 2030, en dépit de leur bonne qualité attestée par ρ.

Le graphique donne les proportions de la population mondiale par tranches d'âges à 17 époques lointaines

Les tranches d'âge futures de la population mondiale.



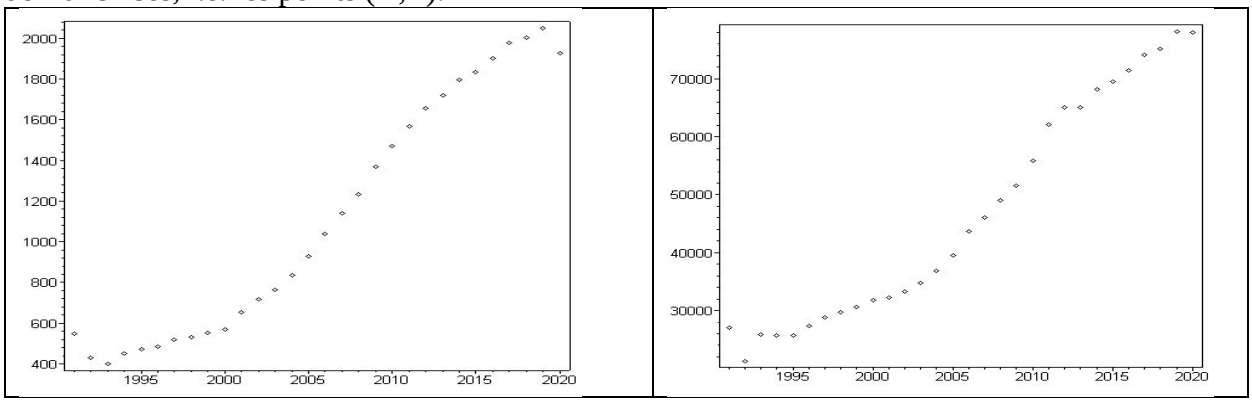
7. L'essor du secteur franchisé en France pendant 30 années.

Les sections qui suivent quittent la démographie pour les séries temporelles économiques. On dispose des chiffres du secteur de la franchise en France pendant 30 ans. On note X la suite des années, X=0 en 1990, Y la série des nombres de franchiseurs, Z la série des nombres de franchisés, W enfin la série donnant les cumuls annuels des chiffres d'affaires de la franchise en milliards d'euros de l'année chaînée (avec conversion en cette unité pour les années antérieures à l'an 2000).

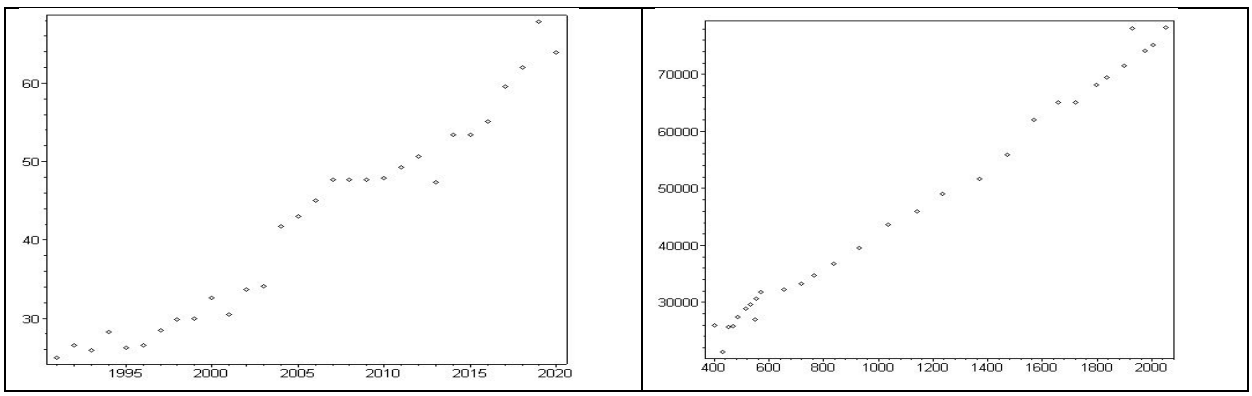
X	Y	Z	W	X	Y	Z	W
1991	550	27000	NC	**2006**	1037	43680	45

X	Y	Z	W	X	Y	Z	W
1992	430	21300	26,6	**2007**	1141	45996	47,7
1993	400	25900	25,9	**2008**	1234	49094	47,7
1994	450	25700	28,31	**2009**	1369	51619	47,72
1995	470	25750	26,22	**2010**	1472	55871	47,88
2001	653	32240	30,49	**2016**	1900	71508	55,1
2002	719	33268	33,71	**2017**	1976	74102	59,55
2003	765	34745	34,12	**2018**	2004	75193	62
2004	835	36773	41,76	**2019**	2049	78218	67,8
2005	929	39510	43	**2020**	1927	78032	63,88

Voici les quatre nuages dignes d'intérêt formés avec les séries X, Y, Z, W. D'abord à gauche, on donne les nombres de franchiseurs, autrement dit les points (X,Y) ; à droite, ce sont les nombres de franchisés, i.e. les points (X,Z).



Puis à gauche, on donne le cumul des chiffres d'affaires de la franchise, autrement dit les points (X,W) ; et enfin, à droite, c'est le nuage des nombres de franchisés, vs les nombres de franchiseurs, i.e. les points (Y,Z) connectés entre eux par le lien chronologique d'une année à la suivante.



Les six différents scénarii d'ajustement linéaires sont tous remarquables de précision sur la période et sont résumés dans le tableau suivant :

Régression	a	b	ρ
X vs. Y	65,5657397	-130374,024	0,973

X vs. Z	2124,38131	-4213657,46	0,975
Y vs. Z	32,1728257	10817,9027	0,996
X vs. W	1,43677855	-2839,42672	0,976
Y vs. W	0,02109098	18,4515484	0,966
Z vs. W	0,0006526	11,4977651	0,966

Il est quand même quelque peu surprenant que ces données s'uniformisent sur une période si longue. Par ailleurs, le chiffre d'affaires étant pris comme fonction d'une des variables Y ou Z, on peut sans doute inférer que le ratio chiffre d'affaires par employé dans les entreprises franchisées semble indépendant du secteur précis où cette personne œuvre. L'étude des auto-corrélations révèle aussi des faits surprenants. La série Y (resp. Z, resp. W) admet des auto-corrélations très prononcées à tout décalage entre 1 et 5 (resp. entre 1 et 6, resp. 1, 2, 3, 13 ans.) Le défaut de corrélation interne à sept ans d'intervalle, par exemple, pose question, de même que l'absence complète de corrélation de la série du nombre de franchiseurs Y après 17 ans d'intervalle, avec des coefficients en ce cas tous <0,70.

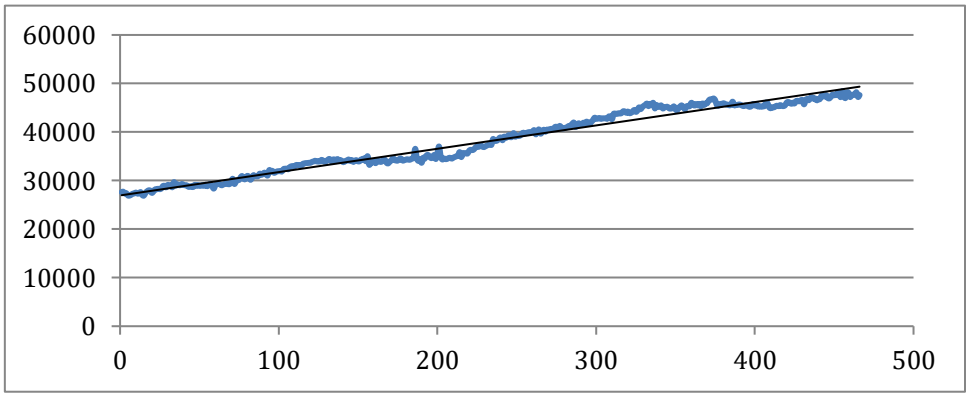
8. La consommation des ménages.

On étudie dans cette section la consommation des ménages[1] sur une très longue période allant de janvier 1980 à octobre 2018, soit 466 mois. L'Insee offre un tableau particulièrement détaillé de ces dépenses, en indiquant leur total macro-économique et en ventilant celui-ci en différentes catégories, comme les dépenses alimentaires, les dépenses alimentaires hors tabac, etc. Mais l'Insee donne également les valeurs annuelles des productions de biens fabriqués durables, de matériels de transport, d'équipements du logement. Ces tableaux distribués par l'Insee sont une mine d'information relatives aux dépenses de consommation des ménages en biens. Les données sont corrigées des variations saisonnières et sont exprimées en millions d'euros chaînés aux prix de l'année précédente. En voici donc un extrait consacré aux trois séries X, Y, Z formées respectivement par les dépenses alimentaires hors tabac, les dépenses relatives à l'énergie, et les dépenses occasionnées par l'eau et les déchets (que l'INSEE regroupe en une seule classe). La série T désignera ici les numéros des mois de la période (T=0 à Noel 1979 et T=467 à la Toussaint 2018).

Janvier 80	27621	9942	2209	**Août 14**	45554	15501	4347
Février 80	27710	9965	2217	**Septembre 14**	45395	15558	4313
Mars 80	27329	9998	2225	**Octobre 14**	45344	15518	4034
Avril 80	27380	9942	2232	**Novembre 14**	45498	15563	4147
Juin 85	29308	10726	2769	**Décembre 14**	46023	15577	4340
...	...	...	...	...	...	...	...
Novembre 89	33658	12019	3012	**Juillet 18**	47689	15846	4375
Février 10	45447	14775	4780	**Août 18**	48194	15758	4379
Juin 14	45427	15480	4329	**Septembre 18**	47223	15602	4353
Juillet 14	45322	15350	4348	**Octobre 18**	47620	15750	4244

[1] https://www.insee.fr/fr/statistiques/3654025.

Le graphique ci-dessous donne les données brutes et la droite régressive pour le couple de variables (T, X), autrement dit la série des dépenses alimentaires.



Que la régression soit digne de confiance n'a rien de surprenant. Nous donnons dans le tableau ci-dessous les résumés des six modèles :

Input	Output	ρ	Ajustement
Mois	Aliment.	0,985	y=48,07665194x+26938,13825
Mois	Energie	0,995	y=12,98022490x+10108,76985
Mois	Déchets	0,933	y=5,359442075x+2467,104610
Aliment.	Déchets	0,948	y=0,1115575850x-538,9534103
Aliment.	Energie	0,988	y=0,2641388176x+3059,048889
Energie	Déchets	0,939	y=0,4135025230x-1714,745068

Quand bien même la corrélation au temps qui s'écoule en mois des trois postes de dépenses des Français est grande, il est surprenant que les variables de dépenses alimentaires ou en énergie, d'une part, et de dépenses liées à l'eau et aux déchets, d'autre part, soient corrélées sans l'être assez pour qu'on ait confiance dans deux des six modèles proposés ci-dessus. Et d'ailleurs les prévisions faites dans les mois qui suivirent n'étaient pas bonnes pour ces deux couples de variables. Les trois prévisions temporelles annoncées comme bonnes n'ont pas prévalu, non plus, car l'époque qui a suivi était celle du confinement du covid.

9. Variations du PIB au Royaume-Uni et en France pendant 56 années

On étudie les variations du PIB au Royaume-Uni et en France pendant les 56 années de 1960 à 2015 incluses. On note X l'année, Y le PIB du Royaume-Uni et Z le PIB de la France. Nous ne reportons pas les données en provenance de la Banque Mondiale[1]. Voici toutefois les chiffres des PIB des années extrêmes :

- 1960 : Pib UK= 72328047042.16, et Pib Fr = 62651474946.60
- 2015 : Pib Uk = 2858003087965.69 et Pib Fr= 2418835532882.33

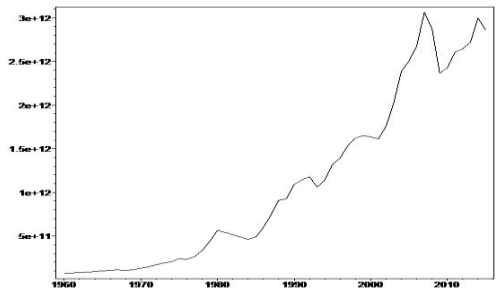
Selon les termes mêmes de la Banque Mondiale, « le PIB au prix des acheteurs est la somme de la valeur ajoutée brute de tous les producteurs résidents d'une économie plus toutes taxes sur les produits et moins les subventions non incluses dans la valeur des produits ». Les données sont en dollars américains courants et les montants sont convertis à partir des devises locales en utilisant les taux de change officiels de l'année. On donne ci-dessous deux graphiques donnant d'une part les PIB

[1] http://donnees.banquemondiale.org/indicateur/NY.GDP.MKTP.CD?view=chart

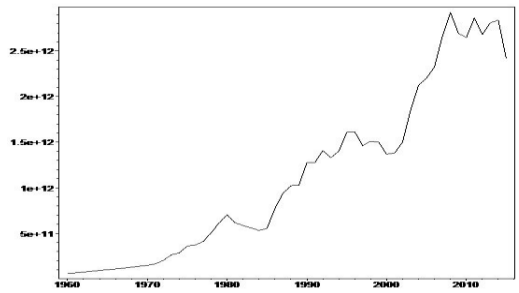
Y et Z des deux pays en fonction du temps X sur un même graphique et d'autre part le nuage de points formé avec les PIB des deux pays. En dépit des apparences qu'ont les courbes des séries brutes, les liaisons sont très significatives. Ainsi les quatre régressions obtenues en prenant séparément les couples de variables (X, Y) ou (X, Z) ou (Y, Z) ou (Z, Y) amènent les modélisations suivantes en dollars :

- $Z = (57{,}605.X - 11338{,}4).10^9 + \varepsilon_{XZ}$ avec $\rho = 95{,}3\%$
- $Y = (54{,}838.X - 10786{,}1).10^9 + \varepsilon_{XY}$ avec $\rho = 96{,}1\%$
- $Z = 1{,}042Y - 0{,}74.10^9 + \varepsilon_{YZ}$ et $Y = 0{,}929.Z + 1{,}043.10^9\varepsilon_{ZY}$ avec $\rho = 98{,}4\ \%$.

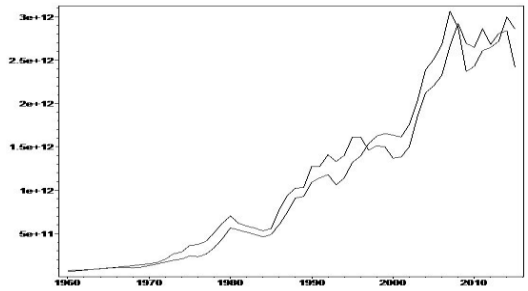
Ce dernier modèle montre que, pour chaque dollar créé par la production nationale anglaise pendant toute la période, il y a 1,042 dollar créé concurremment en France, quoiqu'un décalage de PIB ait continué d'exister entre les deux pays. Ainsi la productivité de la France s'avère-t-elle nettement plus grande que celle du Royaume-Uni. D'ailleurs, chaque année nouvelle voit un surcroît de plus de 57 milliards en France, contre un surcroît annuel de PIB de presque 55 milliards dans le Royaume-Uni.



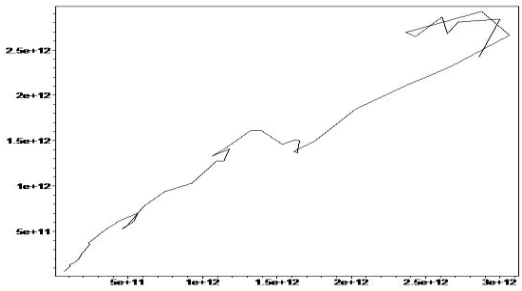
Le PIB de la France pendant 60ans.



Le PIB du Royaume-Uni pendant 60ans.



Les PIB des deux pays variant ensemble.



Nuage de points formé avec les PIB des deux pays.

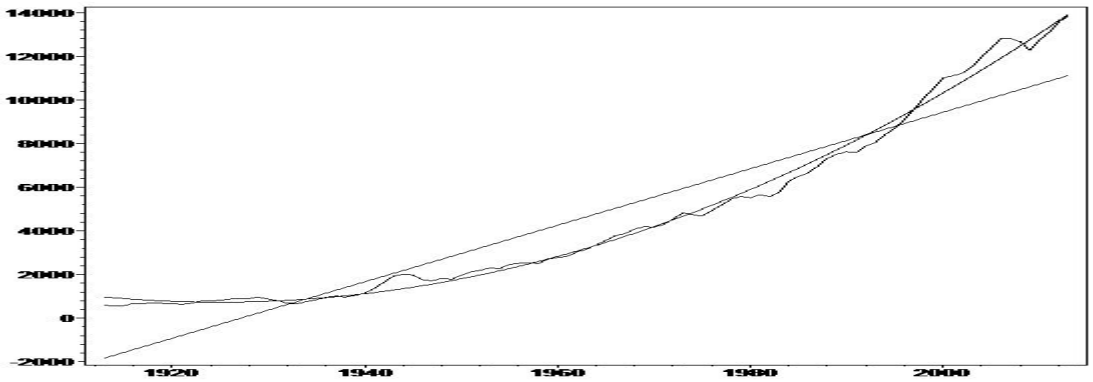
L'étude des auto-corrélations, là encore, est instructive. Il n'y a que huit décalages d'années, pendant cette longue période d'un demi-siècle, où les auto-corrélations des deux séries de PIB, aussi bien en France qu'au Royaume-Uni, aient pu être très significativement liées. Ce sont les décalages 1, 2, 3, 4, 12, 13, 15 ou enfin 16 ans. Le graphique ci-dessous donne les courbes d'auto-corrélation des deux pays, la courbe du Royaume-Uni étant en général au-dessus de celle de la France.

10. Le revenu national américain pendant un siècle.

On étudie le revenu national américain pendant un siècle. On note $Y(t)$ le revenu national des USA produit pendant l'année t.

Année	$Y(t)$	Année	$Y(t)$	Année	$Y(t)$	Année	$Y(t)$	Année	$Y(t)$
1913	602	**1933**	671	**1953**	2295	**1973**	4819	**1993**	8054
1914	557	**1934**	761	**1954**	2272	**1974**	4753	**1994**	8414
1915	578	**1935**	849	**1955**	2455	**1975**	4672	**1995**	8698
1916	670	**1936**	950	**1956**	2520	**1976**	4929	**1996**	9096
1917	669	**1937**	1020	**1957**	2548	**1977**	5170	**1997**	9565
1918	703	**1938**	958	**1958**	2503	**1978**	5452	**1998**	10074
1929	963	**1949**	1783	**1969**	4197	**1989**	7510	**2009**	12273
1930	884	**1950**	1966	**1970**	4167	**1990**	7625	**2010**	12740
1931	799	**1951**	2124	**1971**	4290	**1991**	7600	**2011**	13095
1932	683	**1952**	2204	**1972**	4540	**1992**	7880	**2012**	13561

Le graphique montre bien que l'évolution du revenu national américain est mal prédite par la droite de régression d'équation Y=129,48X-249521,82 et, cependant, le coefficient de corrélation qui vaut 94,3 % est loin d'être mauvais ! On tire un meilleur modèle par l'ajustement parabolique, avec la courbe régressive d'équation $Y=1{,}699432959X^2-6542{,}49X+6297575.944$, avec un coefficient de corrélation de 99,38 %.



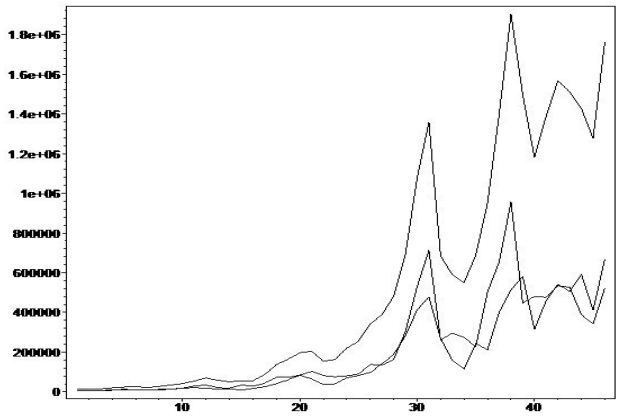
Ce modèle a été corroboré par les trois années qui le suivirent mais a été mis en brêche par les trois années du Covid. Il est notable qu'une telle fiabilité des données n'ait pas perduré, alors que l'époque couverte englobait non seulement les deux guerres mondiales, mais aussi la guerre du Vietnam, et la décennie de la conquête spatiale ! Le graphique ci-dessus donne les trois courbes, la série brute et les courbes auxquelles on a ajusté ces revenus des USA.

11. L'investissement direct entre les cinq continents entre 1970 et 2015.

On étudie l'investissement direct entre les cinq grandes parties du globe, l'Amérique, l'Europe, l'Asie, l'Afrique, le Monde entre 1970 et 2013. Le tableau ci-dessous ne donne pas les chiffres de l'Asie et de l'Afrique qui sont disponibles, parce que les fluctuations rencontrées dans les chiffres d'investissements de ces régions n'ont pas permis de dresser des conclusions intéressantes. Par souci de concision, la moitié environ des données a été reportée.

Année	Monde	Amérique	Europe	Année	Monde	Amérique	Europe
1970	13257,02	4681,822	5226,073	1993	220111,9	70529,89	80667,05
1971	14241,18	5083,817	5975,611	1994	254916,1	82291,32	89568,09
1972	14759,58	4450,931	6580,941	1995	341523,1	97534,84	138129
1984	56160,51	34319,75	8243,346	2007	1902244	513455,8	955969,5
1985	55830,83	28085,32	16851,51	2008	1497788	581722,9	446892,3
1988	164227,8	73818,87	60551,96	2011	1566839	542053,9	534318,9
1989	196936,4	83786,8	84070,72	2012	1510918	506420,8	527013,8
1990	204913,8	64929,19	102733,9	2013	1427181	591685,9	388493,4
1991	153981	37281,22	81922,69	2014	1276999	412200,9	342153,3
1992	162925,2	40085,06	75825,39	2015	1762155	667944,9	522999,5

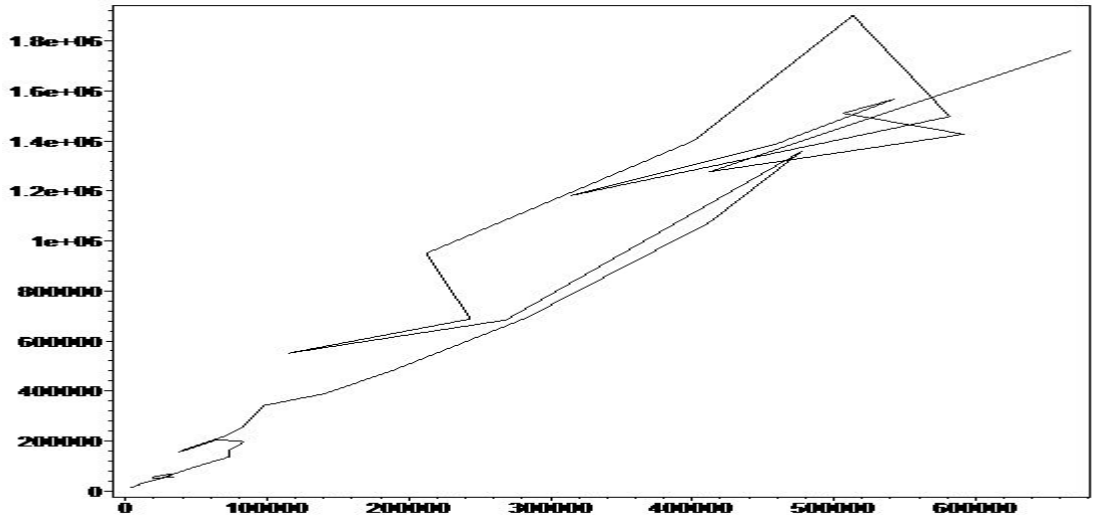
Tous les ajustements linéaires entre ces quatre séries donnent de piètres résultats, sauf un seul, conformément au graphique ci-dessous donnant l'évolution de ces investissements. Sur le graphique, les trois régions sont évidentes à deviner, l'Europe en dessous, l'Amérique, puis le Monde en haut.



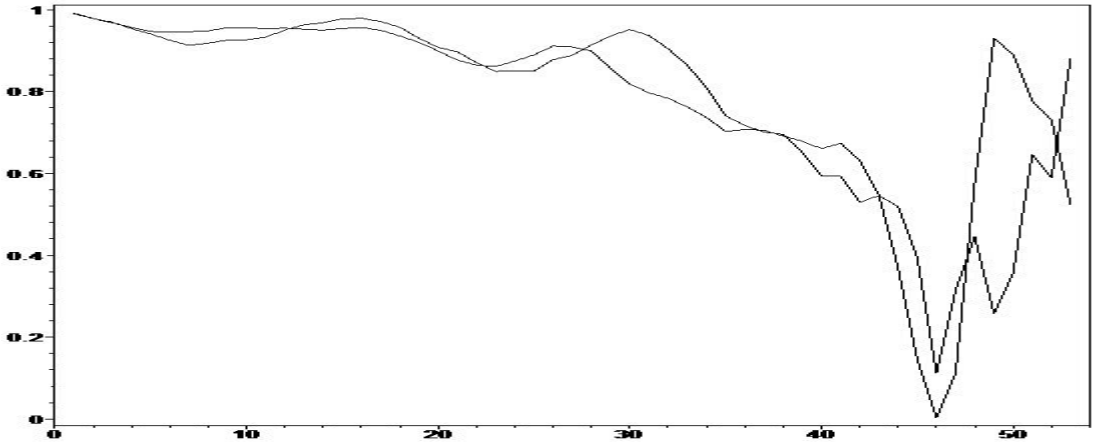
Seul, l'ajustement entre les investissements de l'Amérique et ceux du Monde entier est digne d'être mentionné. On trouve ainsi un seul modèle de qualité, qui s'écrit :

(Inv. Monde) = 2,874440982. (Inv Amérique) + 15077,07 avec $\rho = 97{,}8$ %.

Le nuage de points formé par les couples de valeurs de l'investissement de l'Amérique et du Monde, est reporté ci-dessous, et sa forme bizarre tient au fait que les points du nuage sont connectés quand l'un d'eux est à une année n, et l'autre à l'année n+1.



Dans ce nuage des points donnant les investissements de l'Amérique et ceux du Monde, entrants et sortants, les points sont connectés par la chronologie sous-jacente, et ceci crée un parcours très sinueux.



D'autres études de corrélation du PIB entre deux pays quelconques n'ont pas donné de corrélations si franches.

12. Passagers de l'aviation civile pendant 12 années consécutives.

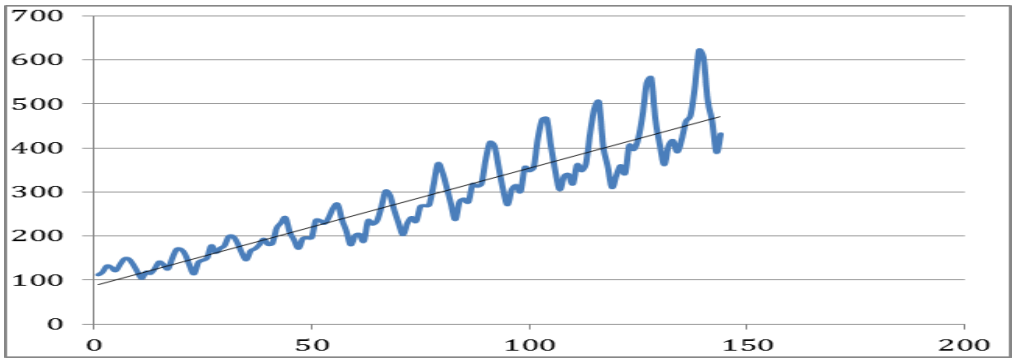
Dans ces trois dernières sections, on examine trois exemples pour lesquels l'ajustement « brutal » ne convient pas, mais néanmoins, en scrutant les données et en variant les méthodes, on peut fournir des modèles très précis pour les phénomènes étudiés, dans le cas de l'exemple de ce paragraphe, mais non dans les deux derniers exemples.

Une statistique sur les passagers de l'aviation civile aux USA a été faite pendant 144 mois consécutifs, allant de janvier 1949 à décembre 1960. Voici la suite de ces nombres de passagers en milliers pendant les 144 mois consécutifs, donnée comme une suite et non un tableau pour des raisons de concision :

[112, 118, 132, 129, 121, 135, 148, 148, 136, 119, 104, 118, 115, 126, 141, 135, 125, 149, 170, 170, 158, 133, 114, 140, 145, 150, 178, 163, 172, 178, 199, 199, 184, 162, 146, 166, 171, 180, 193, 181, 183, 218, 230, 242, 209, 191, 172, 194, 196, 196, 236, 235, 229, 243, 264, 272, 237, 211, 180, 201, 204, 188, 235, 227, 234, 264, 302, 293, 259, 229, 203, 229, 242, 233, 267, 269, 270, 315, 364, 347,

312, 274, 237, 278, 284, 277, 317, 313, 318, 374, 413, 405, 355, 306, 271, 306, 315, 301, 356, 348, 355, 422, 465, 467, 404, 347, 305, 336, 340, 318, 362, 348, 363, 435, 491, 505, 404, 359, 310, 337, 360, 342, 406, 396, 420, 472, 548, 559, 463, 407, 362, 405, 417, 391, 419, 461, 472, 535, 622, 606, 508, 461, 390, 432].

L'évolution du nombre de passagers dans l'aviation civile est manifestée dans le graphique ci-dessous où l'on voit la droite régressive, qui est évidemment un modèle pauvre en ce cas. La méthode des moyennes mobiles ne donne pas de résultats convaincants.

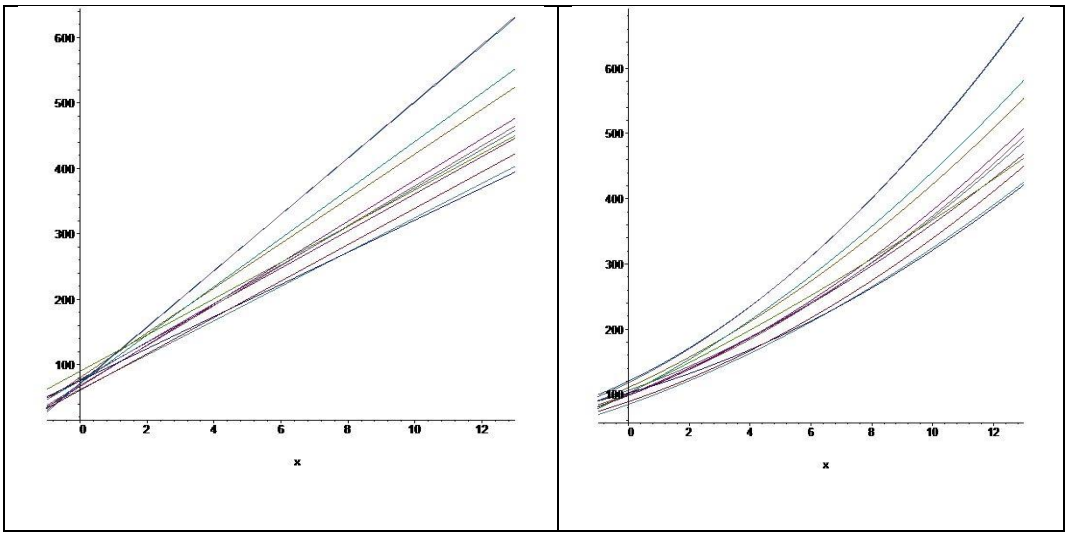


Si on applique la régression parabolique à la série de 144 mois, on obtient une matrice de moments très mal conditionnée avec un conditionnement de 4,9.10^11 et un déterminant de 3,7.10^12. Les coefficients a, b, c ont des tailles excessives. De même, la méthode des moyennes mobiles n'est pas couronnée de succès. On procède alors en examinant les chiffres correspondants aux différents mois de l'année. On trouve ainsi douze droites régressives d'équations similaires :

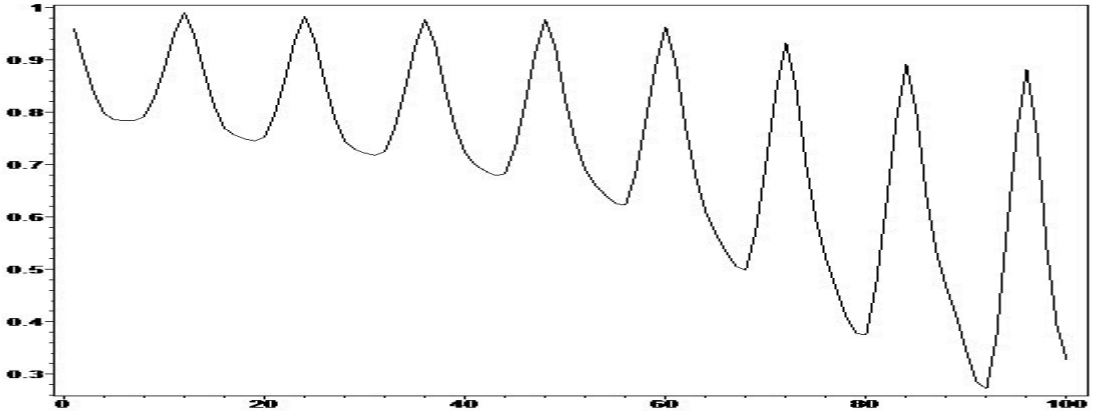
$$[\,27.79\,x + 61.14,\ 24.53\,x + 75.55,\ 27.69\,x + 90.17,\ 29.40\,x + 76.02,\ 31.52\,x + 66.92,$$
$$36.96\,x + 71.44,\ 43.16\,x + 70.79,\ 42.86\,x + 72.47,\ 34.12\,x + 80.62,$$
$$30.48\,x + 68.47,\ 26.22\,x + 62.42,\ 28.36\,x + 77.52\,]$$

Il est extraordinaire que tous les coefficients de corrélation mensuelle soient supérieurs à 99 % sauf celui des mois de Février qui vaut 98,7 %. Si on teste une régression quadratique, tous les coefficients de corrélation avoisinent la valeur de 99,5 %. Les deux graphiques ci-dessous fournissent les courbes d'ajustement soit affine soit parabolique, pour les différents mois de l'année.

Prévision de l'évolution du nombre de passagers dans l'aviation civile aux USA suivant les mois de l'année, par ajustement linéaire à gauche ou ajustement parabolique à droite.



La qualité de ces ajustements est mesurée en premier chef par le coefficient de variation des séries de résidus de la régression. Pour le cas linéaire, les coefficients de variation s'étalent entre 11 % et 13 %, les deux meilleurs mois de prévision étant Juillet et Août, avec deux exceptions en Février et Mai. Pour le cas quadratique, les coefficients de variation s'améliorent nettement et s'étalent entre 5 % et 8 %, les deux meilleurs mois de prévision étant Juillet et Août, le pire étant encore Février. Néanmoins, la nécessité de mensualiser l'étude prévisionnelle apparaît également dans le calcul des auto-corrélations, qui sont représentées graphiquement dans la figure ci-dessous.



On observe par le calcul que les auto-corrélations sont très appuyées aux différents décalages mensuels suivants : [p=1, $\rho = 0{,}9603$], [p=12, $\rho = 0{,}9903$], [p=24, $\rho = 0{,}9833$], [p=36, $\rho = 0{,}9775$], [p=48, $\rho = 0{,}9772$] et enfin [p=60, $\rho = 0{,}9636$]. Toutes les autres valeurs sont < 95 %. Il existe 35 décalages produisant des séries très décorrélées, ces décalages formant des séquences respectant le retour d'un an.

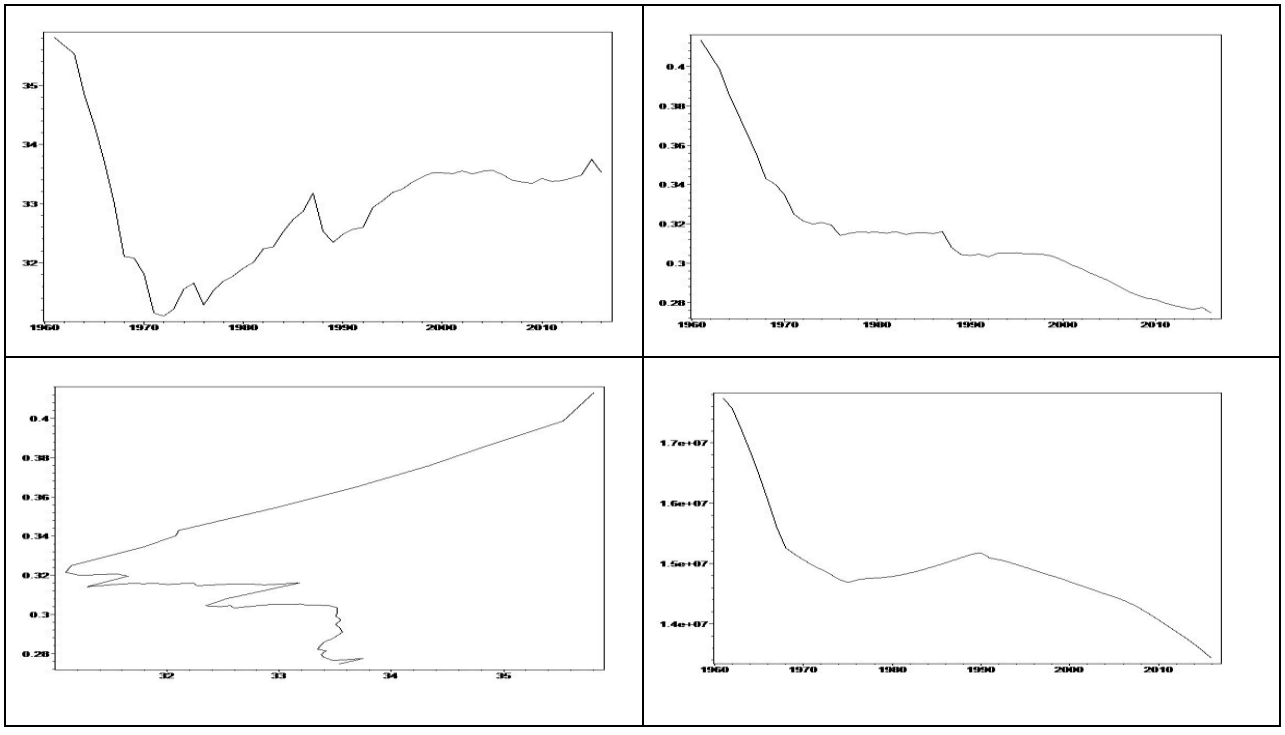
13. Prévision de séries temporelles liées à l'agro-alimentaire.

Dans cette section, on donne un deuxième exemple de prévision insatisfaisante, en dépit de la conviction a priori que les variables étudiées sont fort corrélées, parce qu'elles participent du même

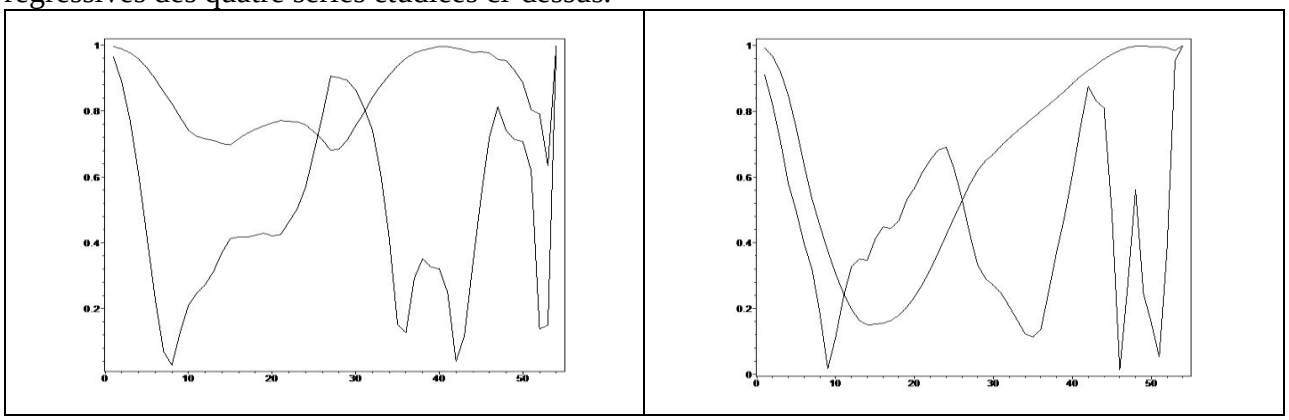
phénomène. La Banque mondiale fournit des statistiques de toutes espèces sur l'activité économique des pays. Intéressons-nous ici à quelques indicateurs de l'industrie agro-alimentaire et de la géographie en France pendant les 56 années entre 1961 et 2016. Les années forment une série X. On note A le pourcentage du territoire métropolitain occupé par les terres arables, T le nombre d'hectares de terres arables par personne, R la population rurale. On note également P la production totale de la pêche en millions d'Euros. Le tableau ci-dessous ne donne que les quatre premières et les quatre dernières années de l'époque considérée.

Année	A	T	R	P
1961	35.806	.413	17741016	791620
1962	35.667	.406	17568965	761211
1963	35.530	.399	17226723	674199
⋮	⋮	⋮	⋮	⋮
2012	33.432	.277	13770620	740550
2014	33.482	.276	13665094	736397
2015	33.748	.277	13548420	669030
2016	33.523	.274	13427447	727814

Les scénarii de modélisation en fonction affine ou parabolique de X des variables économiques A, T, P sont décevants. De même, est très mauvaise la qualité de la liaison entre A et T, étrangement, ou bien entre A et R ou encore entre R et P. Le seul modèle qui reçoit grâce à nos yeux lie les deux variables T et R et s'écrit $R = 25774908.44T + 6818099.313 + \varepsilon$ avec un coefficient $\rho = 95{,}6$ %. Le graphique ci-dessous représente quatre nuages de points. En haut, à gauche on donne le pourcentage de terres arables en fonction des années, en haut à droite, le nombre d'hectares de terres arables par personne en fonction des années, autrement dit pour ces deux courbes les nuages (X, A) et (X, T). En dessous**,** ce sont les nuages (X, R) de la population rurale et (X, P) de la production liée à la pêche.



Dans les deux graphiques ci-après nous donnons sur chaque figure les deux courbes auto-régressives des quatre séries étudiées ci-dessus.



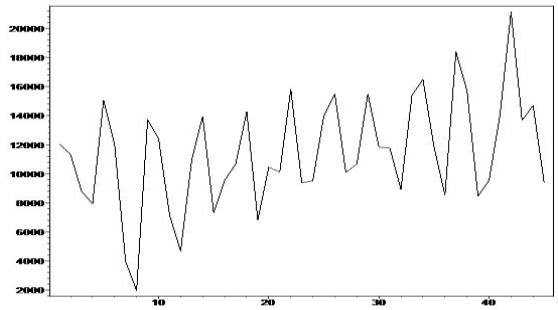
La figure de gauche concerne les variables A, T et celle de droite les variables R, P. L'auto-corrélation de la série de la pêche est la plus défaillante, car aucun décalage ne permet de trouver un seuil d'au moins 92 %. Pour ce secteur, il est clair que chaque année est nouvelle et ne s'appuie nullement sur les résultats antérieurs ! Les niveaux des deux séries T et R, qui sont en quelque sorte de nature démographique plus qu'économique, restent très corrélés aux niveaux à un an ou bien deux ans d'écart, mais néanmoins, aucune vraie continuité ne se dégage de ces quatre séries temporelles.

14. Vente de rhum en France.

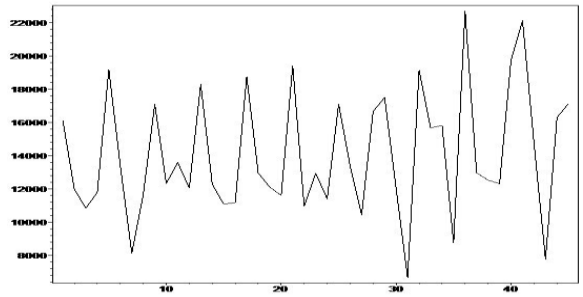
Dans cette section, on examine un exemple récalcitrant aux méthodes de prévision, et sur lequel échouent les techniques habituelles. Les données ci-dessous ont été compilées par le Sénat et par l'Assemblée Nationale en 2000, pour des questions de fiscalité sur les rhums importés en France, et qui ne sont pas produits par les quatre régions productrices, la Martinique, la Guadeloupe, la Réunion, et à moindre titre, la Corse. Par souci de concision nous ne reportons qu'une petite moitié des données sur les 45 trimestres de l'étude.

Trimestre	no.	Martinique	Guadeloupe	Trimestre	no.	Martinique	Guadeloupe
2000T2	1	16123	12051	**2005T4**	23	12949	9395
2000T3	2	11984	11336	**2006T1**	24	11414	9514
2000T4	3	10846	8818	**2006T2**	25	17112	13973
2001T1	4	11790	7936	**2006T3**	26	13382	15500
2004T4	19	12115	6810	**2010T2**	41	22143	14066
2005T1	20	11608	10464	**2010T3**	42	14741	21157
2005T2	21	19427	10119	**2010T4**	43	7760	13677

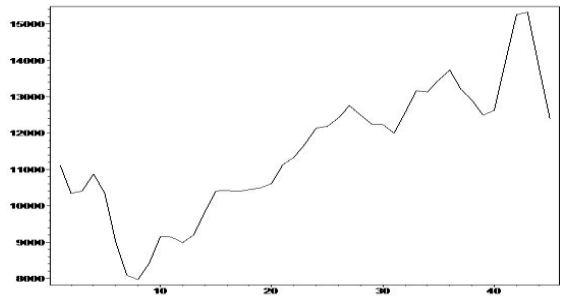
On donne ci-dessous les séries brutes de la Guadeloupe et de la Martinique puis les deux séries des moyennes mobiles de la Guadeloupe et de la Martinique.



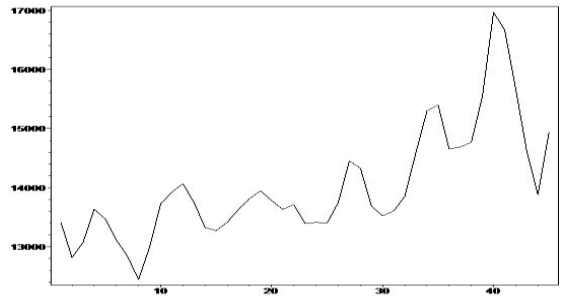
Série brute de la Guadeloupe.



Série brute de la Martinique.

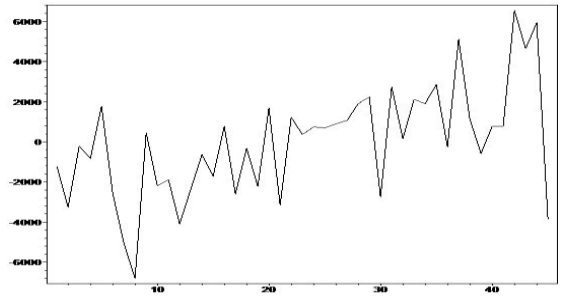


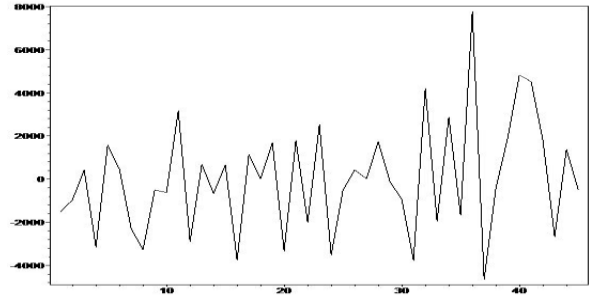
Moyennes mobiles de la Guadeloupe.



Moyennes mobiles de la Martinique.

Les résidus des deux modèles saisonniers sont ci-dessous, Guadeloupe à gauche, Martinique à droite.

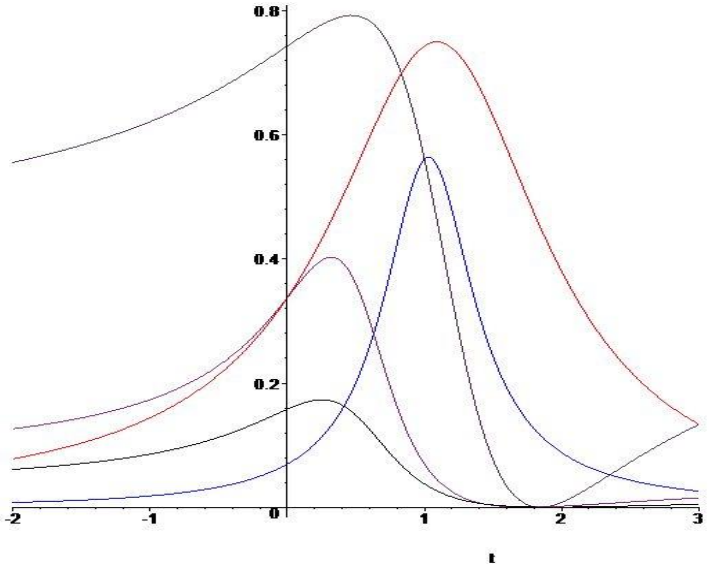




Voici différents scenarii où on donne le coefficient de corrélation linéaire ρ en ordonnée pour plusieurs modèles ayant chacun un paramètre libre t reporté en abscisse. Le premier modèle est le plus naïf, on fait une combinaison convexe de la série M de la Martinique et de la série G de la Guadeloupe, autrement dit t.M(t)+(1-t).G(t). Le coefficient de corrélation est visible en ordonnée sur la courbe en noir. Dans les quatre modèles suivants, on procède de même, mais en employant de surcroît, d'une part, les moyennes mobiles MA (moving averages) et, d'autre part, le modèle saisonnier à quatre saisons SM (season model), puisque les données de l'Assemblée Nationale répartissent les volumes importés sur la métropole en suivant les quatre trimestres (qu'on nomme injustement ci-après les saisons), pour les deux DOM. Sans entrer dans le détail technique des coefficients saisonniers, on forme donc ainsi au préalable deux nouvelles séries de moyennes mobiles MA(M(t)) et MA(G(t)), et deux nouvelles séries désaisonnalisées SM(M(t)) et SM(G(t)). Cela étant, le deuxième modèle est celui de la combinaison convexe t.MA(M(t))+(1-t).SM(M(t)), donnant la

courbe en bleu pour la Martinique, le troisième est celui de t.MA(G(t))+(1-t).SM(G(t))) donnant la courbe en rouge pour la Guadeloupe, le quatrième est celui de la combinaison de moyennes mobiles de t.MA(M(t))+(1-t).MA(G(t))) donnant la courbe en violet, le cinquième enfin est celui de la combinaison de séries désaisonnalisées t.SM(M(t))+(1-t).SM(G(t))) donnant la courbe en marron.

Les coefficients de corrélation linéaire pour cinq modèles à paramètres.



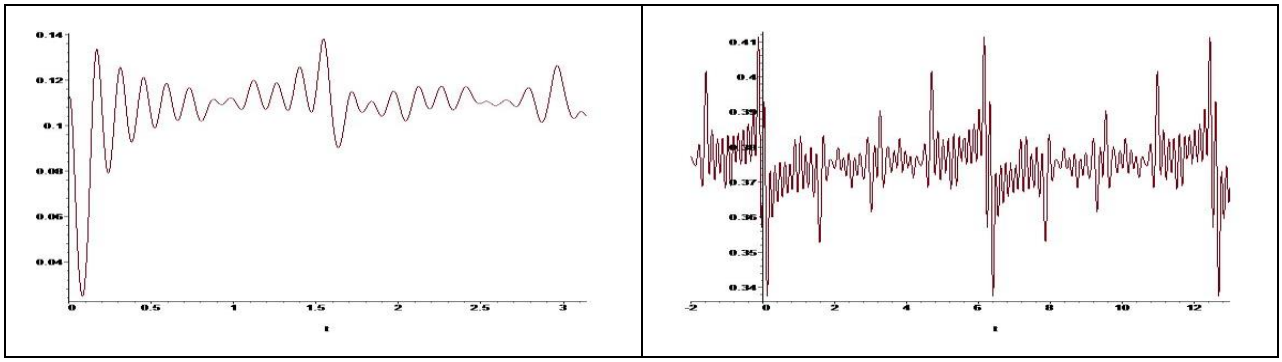
De tous ces modèles, le meilleur est le quatrième, pour 0,4673704162, à savoir :

0,4673704162 . MA(M(t)) +0,5326295838 . MM(G(t)).

Mais néanmoins, sa pertinence est insatisfaisante, comme le montre l'analyse auto-régressive. Aussi, nous avons également introduit des modèles à paramètres donnant plus de satisfaction. Pour cela, introduisons deux séries courtes de Fourier, avec une fréquence inconnue et à déterminer, donnant les résidus d'une approximation « harmonique » des séries brutes. Autrement dit on pose :

Ym(t)=M(t)-am*t-bm-cm*cos(km*t)-dm*sin(km*t)
et
Yg(t)=G(t)-ag*t-bg-cg*cos(kg*t)-dg*sin(kg*t),

où les huit coefficients am, bm, cm, dm pour la Martinique et ag, bg, cg, dg pour la Guadeloupe sont calculés par la méthode des moindres carrés. Ce passage se fait avec Maple et semble très complexe à obtenir avec n'importe quel tableur, y compris en le programmant. Lorsque les huit coefficients ci-dessus sont déterminés (de façon exacte avec le logiciel Maple), on peut reporter ces huit expressions dans les deux séries et reprendre le calcul du $R2 = \rho^2$ comme expression explicite (énorme) de la fréquence km ou kg. Sans donner les détails, on peut examiner enfin ρ comme fonction des fréquences km et kg apparaissant dans les séries trigonométriques ci-dessus et on obtient les deux courbes attendues pour les corrélations linéaires de la Martinique (à gauche) et de la Guadeloupe (à droite).



15. Conclusion et perspectives.

Au terme de cet ensemble de séries chronologiques véritables, on observe que des *pépites* dans les comportements des variables économiques ou démographiques peuvent surgir, pour des séries comportant parfois de grands nombres d'observations. Lorsque tel est le cas, on ne peut parler pour autant de loi macro-économique ou de loi démographique, car les corrélations dégagées par ces analyses ne reposent pas sur un modèle a priori, mais plutôt sur le fruit du hasard, ou plus exactement sur l'opiniâtreté avec laquelle on doit s'attaquer aux observations pour qu'elles deviennent signifiantes. Nous avons dit en introduction que les corrélations temporelles fortes entre diverses variables peuvent, par transitivité, amener des raisonnements infondés. La dépendance fonctionnelle entre variables statistiques est d'une autre nature. L'étude de la causalité des phénomènes est amorcée dans cet article, mais pour garantir la dépendance entre différentes variables il faut aussi employer d'autres indicateurs statistiques, notamment la suite des moments centrés ou non-centrés, afin de dégager des liens fonctionnels. Dans une suite de cet article, d'autres exemples plus résistants de données économiques, non nécessairement temporelles, seront ainsi démontrés. Nous examinerons également quelques exemples remarquables de modèles à plusieurs paramètres en économie et en démographie, à la fois *suffisamment simples et particulièrement précis.*

Bibliographie

1. Dodge Y., 1999. *Premiers pas en statistique*. Springer Verlag, Paris, 427 p.
2. Gourieroux, C., Montfort, A. (1995), *Séries temporelles et modèles dynamiques*, 2ème édition, Economica, Paris.
3. Gourieroux, C., Montfort, A. (1999), *Statistiques et Modèles économétriques, notions générales, estimations, prévisions, algorithmes*, 1ère édition, Economica, Paris.
4. Saporta, G., (2006*), Probabilités, analyse des données et statistique*, Technip, Paris.