TER sur la nature chaotique persistente des épidémies de rougeole aux Etats-Unis et le modèle TSIR

Koen Aristote

7/20/2020

Table des matières

Introduction

Dans ce travail de recherche nous nous intéressons au travail de Benjamin D.Dalziel sur la nature chaotique et persistente des cycles de la rougeole aux Etats-Unis. En effet, alors que les cycles de la rougeole se sont montrés stables au Royaume-Uni (des épidémies tous les 2 ans), il semble que des différences sur la période de basse transmission de la rougeole aux Etats-unis puisse entrainer une divergence du système vers une plus haute periodicité moyenne (D.Dalziel 2016). De plus, il a été observé que les épidémies de rougeole étaient très irrégulières au Niger où de fortes perturbations démographiques cassaient les chaînes de transmission, entraînant une extinction des cas infectés durant certaines périodes (Ferrari MJ 2008). Ainsi, on pensait que des divergences des cycles de la rougeole de ceux observés au Royaume-Uni entraînaient automatiquement des ruptures des cycles et des épidémies plus épisodiques. Mais l'article que nous étudions montre qu'une faible perturbation sur la période annuelle de basse transmission de la maladie entraîne une divergence des cycles observés au Royaume-Uni vers des cycles irréguliers mais sans rupture de la chaîne de transmission; c'est à dire sans extinction de la maladie (D.Dalziel 2016).

Afin de vérifier cela nous nous intéresserons à la reproduction des résultats publiés par l'auteur et à la modélisation des épidémies par le modèle TSIR: Time-Susceptible Infected Recovered.

Nous commencerons d'abord par introduire les données et effectuer une analyse descriptive des données, nous introduirons introduirons ensuite le modèle TSIR pour enfin présenter les résultats que nous avons reproduit ou approfondi de l'article de référence.

Vous trouverez une annexe contenant le code que nous avons écrit pour reproduire ces résultats. Pour executer ou accéder à l'intégralité du scripte, nous vous invitons à consulter le fichier .Rmd

1-Les données

Afin d'étudier ces cycles et de modéliser la population infectée, nous comparerons 40 villes aux Etats-Unis entre 1920 et 1940 et 40 villes au Royaume-Uni entre 1945 et 1965, c'est à dire sur 20 ans. Nous comparons sur des périodes différentes car les données ne sont pas disponibles pour les deux pays sur la même periode. Cependant la modélisation tient compte des variations de la population et de la natalité, et il a été montré dans d'autres travaux que les épidémies de la rougeole au Royaume-Uni sont restées stables (periode de 1 ou 2 ans) entre 1920 et 1940 , ce qui n'affecte donc pas notre analyse (D.Dalziel 2016). Les séries auxquelles nous aurons accès pour chacune des villes sont celles des naissances, de la taille de la population et des cas infectés toutes les deux semaines.

Les données sont celles utilisées dans l'article de référence et sont disponibles sur ce lien. Celles des Etats-unis, ont été obtenues à partir du Project Tycho qui recense les cas infectés de plusieurs maladies et à partir des données de recensement extrapolées de sorte à avoir des mesures bimensuelles. Pour le Royaume-Uni, les données utilisées ont été obtenues sur d'anciens travaux effectués sur la rougeole. (D.Dalziel 2016)

Voici le tracé des séries des cas infectés ainsi que des taux de natalité pour les villes de Londres, New York et Los Angeles et Spokane. On observe que contrairement au cas de Londres, les cycles d'épidémies de la rougeole dans les villes américaines sont moins réguliers.

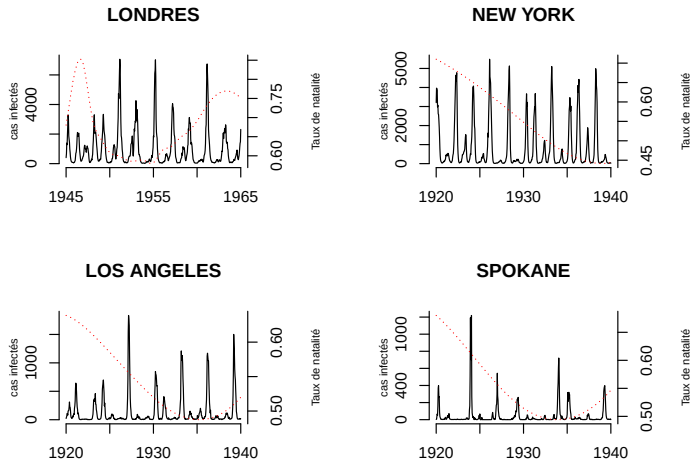

FIGURE 1 – Population infectée et taux de natalité (en rouge) dans 4 villes

En effet, en tracant le spectogramme de ces séries (Figure 2), nous pouvons observer que les 3 villes aux Etats-Unis présentent une périodicité plus élevée qu'à Londres où l'on observe des cycles réguliers de periode 1 an et 2 ans. De plus, ces cycles restent aussi réguliers et de même période pour la plupart des autres villes étudiées au Royaume-Uni (Grenfell BT 2002). Cependant aux Etats unis on observe des poids importants sur les fréquences de 5 ans à Spokane et 3 ans à Los Angeles, et ce phénomène est observé sur de nombreuses villes aux Etats-Unis.

Notre but sera donc de montrer que cette irrégularité est due à la nature chaotique des épidémies de la rougeole causée par une petite perturbation sur la periode de basse transmission de la maladie aux Etats-Unis à partir du modèle TSIR.

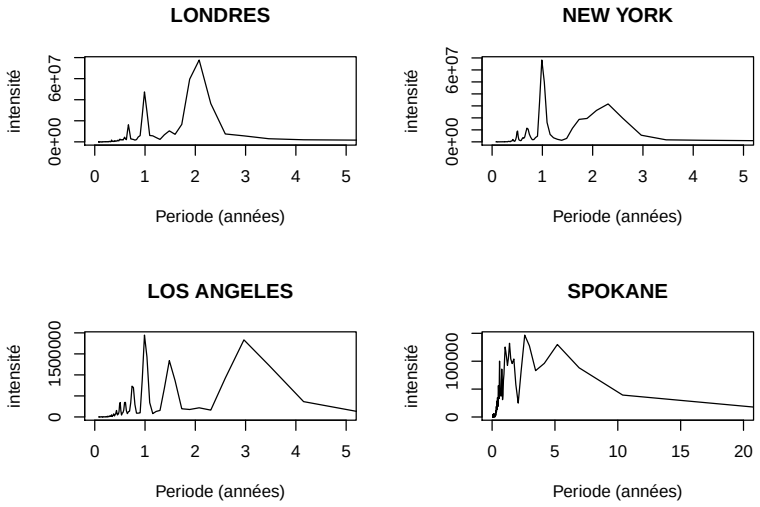

FIGURE 2 – Spectrogramme des 4 séries de la figure 1

2-Le modèle TSIR

Le modèle utilisé dans ce travail pour simuler les données est le modèle TSIR (Time-Series Susceptible Infected Removed) qui consiste à déterminer la population infectée à l'instant $t+1$ à travers l'espérance de la population infectée à l'instant $t+1$ puis de considérer les variations par rapport à cette espérance à travers la loi négative binomiale.

Définition du modèle:

Le modèle est établi comme suit. Soit:

- I_t la population infectée à l'instant t, où t représente une mesure toutes les 2 semaines
- B_t le nombre de naissances
- S_t la population susceptible à l'instant t
- N_t la taille de la population à l'instant t
- α un paramètre strictement positif et proche de 1
- $\beta_{t \mod 26}$ le taux de transmission à l'instant t , supposé saisonnier de période 1 an. Donc 26 valeurs

On modélise l'espérance de la population infectée et la population susceptible à l'instant suivant par:

$$\mathbb{E}[I_{t+1}] = \beta_t I_t^{\alpha} S_t N_t^{-1} \varepsilon_t \tag{1}$$

$$S_{t+1} = S_t + B_t - I_{t+1} + u_t$$

Où ε_t est un bruit tel que $\mathbb{E}\left[\log\left(\varepsilon_t\right)\right] = 0$ et $\mathbb{V}[\log(\varepsilon_t)] = \sigma^2$ que nous supposerons gaussien ($\log(\varepsilon_t)$) et u_t est tel que $\mathbb{E}\left[u_t\right] = 0$ et $\mathbb{V}[u_t] = \sigma_u^2$ (D.Dalziel 2016)(Barbel F. Finkenstadt 2000)(Grenfell BT 2002)(Bjørnstad ON 2002)

En passant au log dans la première équation on a alors:

$$\log(I_{t+1}) = \log(\beta_t) + \alpha \log(I_t) + \log(S_t) - \log(N_t) + \log(\varepsilon_t)$$

Ainsi on a:

$$\mathbb{E}[\log(I_{t+1})] = \log(\beta_t) + \alpha \log(I_t) + \log(S_t) - \log(N_t)$$

Le squelette déterministe correspond à la situation où $I_{t+1} = E[I_{t+1}]$ et le modèle stochastique suppose que la population infectée à l'instant suivant suit une loi Négative Binomiale:

$$I_{t+1} \sim \mathcal{NB}(\mathbb{E}[I_{t+1}], I_t)$$

On observe que si l'on connait la population susceptible S_t alors en passant au log nous avons un modèle linéaire gaussien. Cependant cette série n'est généralement pas connue et il faut donc la reconstruire avant de pouvoir estimer les paramètres β_t et α du modèle. De plus le nombre de cas reportés est souvent sous estimé et n'est généralement pas représentatif de la réalité. Il faut prendre cela en compte dans le modèle afin de stationnariser des séries que nous manipulerons durant l'estimation des paramètres et dans le but de s'approcher le plus possible de la réalité(Barbel F. Finkenstadt 2000).

Réecriture du modèle et estimation

Si I_t est le nombre de cas réel alors on modélise $I_t = \rho_t C_t$ où C_t est le nombre de cas reportés (les données) et $\rho_t > 0$ le taux de rapport. Si la population est sous estimée alors ρ_t sera supérieur à 1 si il est sur-estimé alors on aura $0 < \rho_t < 1$. Ainsi en remplacant dans l'équation régissant la population susceptible on obtient

$$S_{t+1} = S_t + B_t - \rho_{t+1} C_{t+1} + u_t$$

Puis en écrivant S_t comme $S_t = \bar{S} N_t + D_t$ c'est à dire comme les déviations D_t de la population susceptible par rapport à sa population susceptible moyenne $\bar{S} N_t$ (où $\bar{S}$ est la proportion moyenne de la population susceptible) à chaque instant , si on remplaçe dans l'équation ci-dessus on a alors:

$$D_{t+1} = D_t + B_t - \rho_{t+1} C_{t+1} + \bar{S}(N_t - N_{t+1}) + u_t$$

$$\iff D_{t+1} = D_0 + \sum_{i=0}^{t} B_i - \sum_{i=0}^{t} \rho_{i+1} C_{i+1} + \sum_{i=0}^{t} u_i + \bar{S}(N_0 - N_{t+1})$$

$$\iff D_{t+1} = D_0 - \rho \sum_{i=0}^{t} C_{i+1} + \sum_{i=0}^{t} B_i - \sum_{i=0}^{t} (\rho_{i+1} - \rho) C_{i+1} + \sum_{i=0}^{t} u_i + \bar{S}(N_0 - N_{t+1})$$

avec $\rho = \mathbb{E}[\rho_{t+1}]$

$$\iff \sum_{i=0}^{t} B_i = -D_0 + \rho \sum_{i=0}^{t} C_{i+1} + D_{t+1} + \sum_{i=0}^{t} (\rho_{i+1} - \rho) C_{i+1} - \sum_{i=0}^{t} u_i + \bar{S}(N_{t+1} - N_0)$$

Il est raisonnable d'ignorer le dernier terme de l'équation ci dessus car il est d'espérance nulle (D.Dalziel 2016), de plus en supposant un taux de rapport déviant peu de sa moyenne ρ et pour un bruit u_t quasiment nul on a une relation linéaire entre le cumul des naissances et des cas reportés et on peut extraire D_t par les résidus de la régression et la moyenne ρ comme la pente de régression.

Sinon nous pouvons extraire ρ_t comme les pentes de régressions locales(Barbel F. Finkenstadt 2000). En effet, en notant:

— $R_{t+1} = \sum_{i=0}^{t} (\rho_{i+1} - \rho) C_{i+1}$
— $U_{t+1} = \sum_{i=0}^{t} u_i$

L'équation ci-dessus devient

$$\sum_{i=0}^{t} B_i = -D_0 + \rho \sum_{i=0}^{t} C_{i+1} + D_{t+1} + R_{t+1} - U_{t+1} + \bar{S}(N_{t+1} - N_0)$$

On observe que

— $R_{t+1} = R_t + (\rho_{t+1} - \rho) C_{t+1}$
— $U_{t+1} = U_t + u_t$

Ainsi en remplacant dans l'équation:

$$\sum_{i=0}^{t} B_i = -D_0 + D_{t+1} + R_t + \rho \sum_{i=0}^{t} C_{i+1} + (\rho_{t+1} - \rho) C_{t+1} - U_t - u_t + \bar{S}(N_{t+1} - N_0)$$

Que l'on peut réecrire comme:

$$\sum_{i=0}^{t} B_i = -D_0 + D_{t+1} + R_t + \rho_{t+1} \sum_{i=0}^{t} C_{i+1} - (\rho_{t+1} - \rho) \sum_{i=0}^{t-1} C_{i+1} - U_t - u_t + \bar{S}(N_{t+1} - N_0)$$

En passant à l'espérance conditionnellement à U_t, R_t on a alors:

$$\mathbb{E}\left[\sum_{i=0}^{t} B_i | R_t, U_t\right] = -D_0 + R_t - U_t + \rho_{t+1} \sum_{i=0}^{t} C_{i+1}$$

Car:

$$\mathbb{E}\left[R_{t+1} | R_t\right] = R_t$$

et

$$\mathbb{E}\left[U_{t+1} | U_t\right] = U_t$$

En regroupant des termes comme un intercept variant dans le temps $\eta_t = -D_0 + R_t - U_t$ (Barbel F. Finkenstadt 2000):

$$\sum_{i=0}^{t} B_i = \eta_t + \rho_{t+1} \sum_{i=0}^{t} C_{i+1} + D_{t+1} - u_t + \bar{S}(N_{t+1} - N_0)$$

En ignorant les deux derniers termes de l'équation ci-dessus on observe que l'on peut obtenir les valeurs de ρ_t comme les valeurs des pentes de régression locales entre le cumul des naissances et le cumul des cas infectés. Et les valeurs de D_t proviennent des résidus. De plus, il a été montré que le lissage de splines, fonctionne aussi et qu'il est plus robuste que la régression linéaire et nous utiliserons cette méthode pour estimer les valeurs de ρ_t en ajustant une fonction polynomiale par morceaux sur le nuage de points correspondant au cumul des cas infectés en fonction des naissances, où chaque polynôme est de degré q de sorte à obtenir une fonction de classe C^{q-1}. Dans notre cas, q=3, c'est à dire que nous ajusterons une spline cubique.

Ainsi en connaissant D_t il ne reste plus qu'à estimer $\bar{S}$ afin de reconstruire la population susceptible. La valeur moyenne de la population susceptible à l'instant t σ est égale à $\sigma = \bar{S}N_t$ où $\bar{S}$ est la proportion moyenne de susceptibles (D.Dalziel 2016). Donc on a

$$\mathbb{E}[\log(I_{t+1})] = \log(\beta_t) + \alpha \log(I_t) + \log(\bar{S}N_t + D_t) - \log(N_t)$$

Ainsi en estimant $\bar{S}$ par maximum de vraisemblance indépendamment des paramètres β_t et α, c'est à dire en calculant la vraisemblance pour plusieurs valeurs candidates de $\bar{S}$ et en conservant la valeur de $\bar{S}$ maximisant la vraisemblance, nous aboutissons à un modèle linéaire en les paramètres. Nous pourrons donc estimer les 27 paramètres restants α et β_t conditionnellement à $\bar{S}$.

Reconstruction de la population susceptible et estimation des paramètres

Afin d'illustrer le processus voila comment nous avons procédé sur les données de la ville de Londres.

Afin d'obtenir un taux de rapport variable nous avons estimé la pente d'un lissage par spline du cumul des naissances en fonction du cumul des cas infectés. (N.Bjørnstad 2018)

Tout d'abord nous effectuons l'estimation des paramètres à l'aide du package tsiR crée par un des co-auteurs de cet article qui permet d'estimer les paramètres automatiquement(Alexander D. Becker 2017). Ceci nous permettra de comparer nos résultats.

Voici la liste des paramètres obtenus par le package tsiR. Nous comparerons l'ajustement de notre modèle avec celui du package pour s'assurer de la cohérence des résultats obtenus.

```
##            alpha        mean beta         mean rho         mean sus
##         9.80e-01         3.51e-06         4.75e-01         3.25e+05
## prop. init. sus. prop. init. inf.
##         9.37e-02         2.55e-04
```

1ère méthode: Lissage par splines

Voici les réultats obtenus par par l'ajustement d'une spline cubique au cumul des cas infectés en fonction du cumul des naissances. En effet il est plus précis d'effectuer la régression sur le cumul des cas infectés en fonction des naissances car le nombre de cas fluctue beaucoup plus que les naissances, surtout aux Etats-Unis, et effectuer la régression dans le cas inverse provoque un lissage exagéré (D.Dalziel 2016). Ainsi nous pouvons extraire les résidus et les pentes locales de la spline ajustée en chaque point, c'est à dire ρ_t. (c.f annexe pour voir le code)

On obtient bien des résultats proches de celui du package réglé de sorte à utiliser la méthode de lissage par splines pour l'obtention de rho.

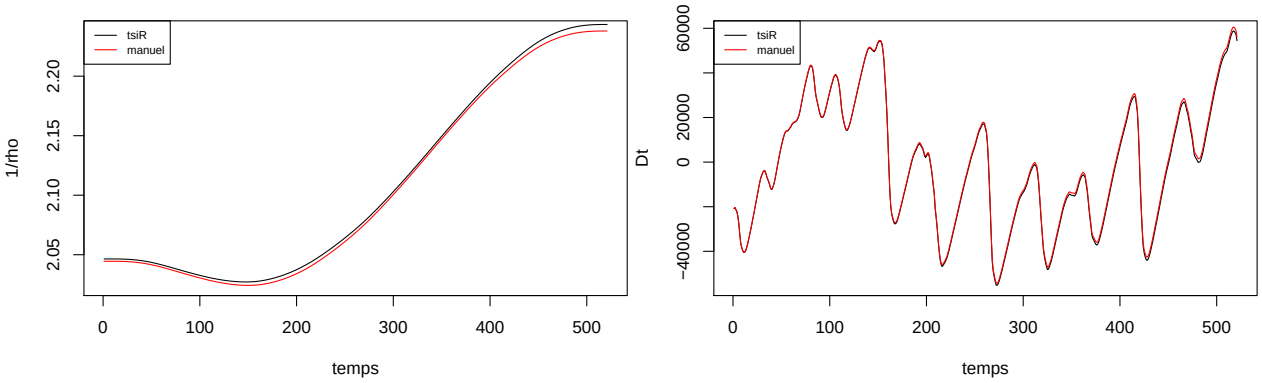

FIGURE 3 – Comparaison de rho et des résidus

2ème méthode: régression linéaire:

Nous pouvons aussi obtenir D et ρ par régression linéaire comme nous l'avons décrit ci-dessus et nous obtiendrons donc un ρ constant. (c.f annexe pour voir le code)

On règle la fonction du package tsiR afin de calculer rho à partir d'une régression linéaire. (c.f annexe pour voir le code)

```
##            alpha        mean beta         mean rho         mean sus
##         9.60e-01         3.23e-06         4.78e-01         4.00e+05
## prop. init. sus. prop. init. inf.
##         1.19e-01         2.60e-04
```

Et on observe que les résidus dans les deux méthodes sont très proches et que les résultats sont les mêmes avec le package. De plus les valeurs de rho s'alignent aussi toutes bien.

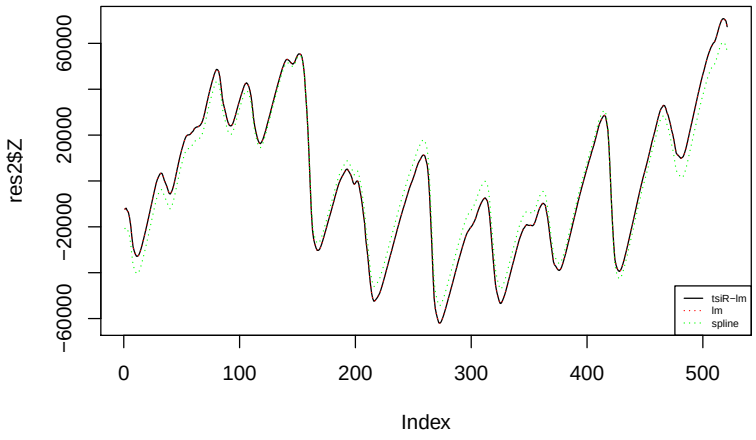

FIGURE 4 – Comparaison des deux méthodes sur les résidus

```
## [1] "rho-tsiR-lm= 0.478"
```

```
## [1] "rho_lm= 0.477"
```

```
## [1] "mean-rho-spline 0.476"
```

On ajuste ensuite les cas reportés par le taux de rapport puis on crée une série avec un retard d'une mesure de temps et une en avance pour pouvoir représenter I_t et I_{t+1} respectivement.

Il reste alors à estimer les paramètres α, β, $\bar{S}$. On suppose dans le modèle que β_t,le taux de transmission, varie de manière saisonnière (Barbel F. Finkenstadt 2000). Ainsi il suffit d'estimer 26 coefficients qui se répèteront chaque année.

On crée un vecteur de la taille de notre série avec 26 niveaux pour la saisonnalité annuelle, un vecteur de candidats pour $\bar{S}$ parmi lequel nous choisirons sa valeur par maximum de vraisemblance puis un vecteur où nous stockerons les valeurs de la vraisemblance pour chaque valeur de sigma. De plus nous passons au log pour les séries I_t, I_{t+1}, N_t. (c.f annexe pour accéder au code)

Nous ajustons alors un modèle linéaire gaussien pour chaque valeur de $\bar{S}$ et calculons la vraisemblance. Notre estimation de σ sera donc la valeur de $\bar{S}$ maximisant la vraisemblance. Nous utilisons la fonction glm avec lien 'identity' et family='gaussian', c'est à dire un modèle linéaire gaussien car elle nous retourne la valeur de la déviance $= -2log(l(\bar{S}))$ contrairement à la fonction lm, ce qui nous permet de facilement accéder à la vraisemblance. (c.f annexe pour accéder au code)

Voici un graphe de l'opposé de la log vraisemblance obtenu en fonction de sigma et celui obtenu par le package en fonction de $\bar{S} * \bar{N}$ ($\bar{N}$ la population moyenne) à droite.

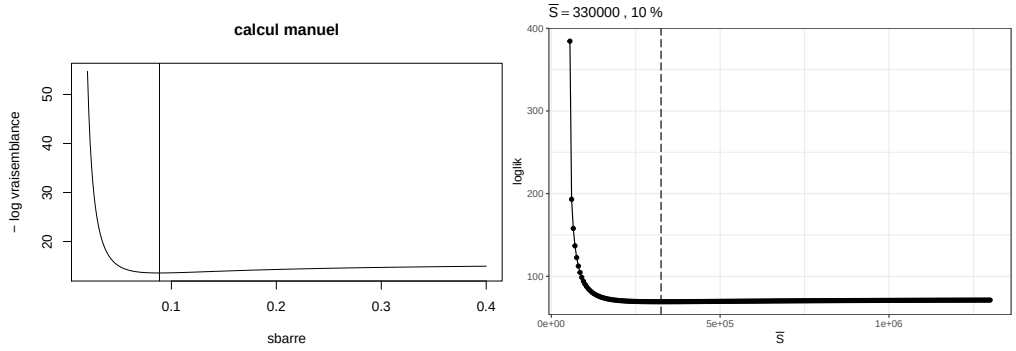

FIGURE 5 – Vraisemblance calculée vs package tsiR

Ainsi on obtient une proportion moyenne de la population susceptible de 0.088 vs 0.1 dans le package tsiR. Cette différence avec le package TSIR provient probablement du fait qu'il ne semble pas tenir compte de la taille de la population durant l'estimation des paramètres.

```
## [1] "sbarre-tsiR= 0.1"
```

```
## [1] "sbarre= 0.089"
```

Nous pouvons maintenant reconstruire S et estimer α et β (c.f annexe pour accéder au code):

On obtient les valeurs estimées de β_t et α et nous pouvons donc obtenir les valeurs ajustées de la série des populations infectées. On observe que notre modèle s'ajuste bien aux données sur le plot des valeurs ajustées comparées aux données ci dessous et de même pour le résultat obtenu par package tsiR. De plus, on observe que les résidus de la régression semblent bien normalement distribués au regard du qqplot et sur l'histogramme.

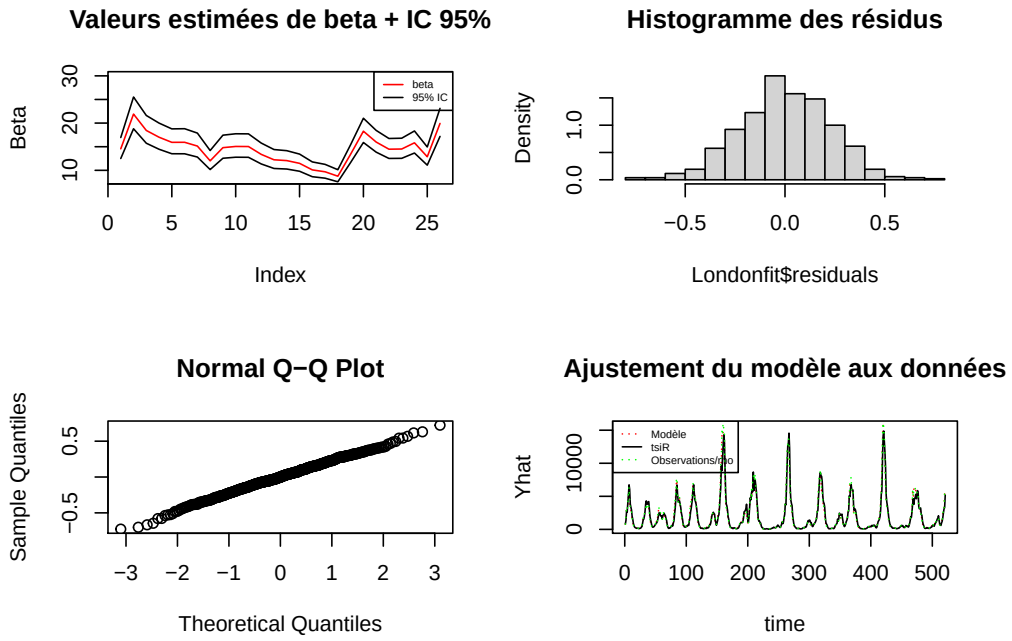

FIGURE 6 – Ajustement du modèle aux données et estimation

```
## [1] "alpha= 0.967"
```

Comment simuler à partir des paramètres estimés:

Ayant estimé les paramètres nous pouvons donc maintenant simuler les cycles de la rougeole à partir de conditions initiales sur la population susceptible et le nombre d'infectés. Nous créons une fonction qui calculera étant donnés β_t, α, B_t et N_t, la population infectée en partant de conditions initiales S_0 et I_0 (c.f annexe pour accéder au code). Nous pourrons simuler selon le modèle déteministe qui considère que $I_{t+1} = \mathbb{E}[I_{t+1}]$ ou selon le modèle stochastique négatif binomial décrit ci-dessus.

3-Résultats

Pour résoudre notre problématique sur la nature chaotique des cycles de la rougeole nous souhaitons montrer empiriquement que des perturbation de la periode de basse transmission entraînent de grandes différences sur les cycles de la rougeole en reproduisant les résultats de l'article de Dalziel.

Précision et comportement du modèle

Dans cette partie nous reproduisons les résultats traitant de la précision du modèle TSIR et effectuons des prédictions.

Pour cela, nous comparons les séries de New York, Boston et Londres avec les simulations puis avec les prédictions.

Ensuite nous comparons le spectrogramme des données observées pour ces trois villes avec les spectogrammes de 100 simulations du modèle stochastique par ville à partir des paramètres éstimés par la procédure décrite précédemment et des conditions initiales.

Afin de pouvoir traiter différentes villes de manière plus systématique nous avons créé une fonction param_estim qui estime les paramètres du modèle pour n'importe quel jeu de données de la même manière que décrite dans la partie précédente.

Données manquantes

Avant de pouvoir estimer et simuler massivement, nous avons remarqué que la majorité des séries sur la population infectée aux Etats-Unis contenait des données manquantes. Cependant le nombre de données manquantes est relativement faible (maximum de 42 données manquantes pour la série de Milwaukee) comparé à la longeur des séries étudiées de 547 mesures. Ainsi, nous les avons complété par interpolation linéaire en vérifiant que la reconstruction est cohérente avec la réalité. Ci dessous un exemple de complétion des données manquantes pour Buffalo.

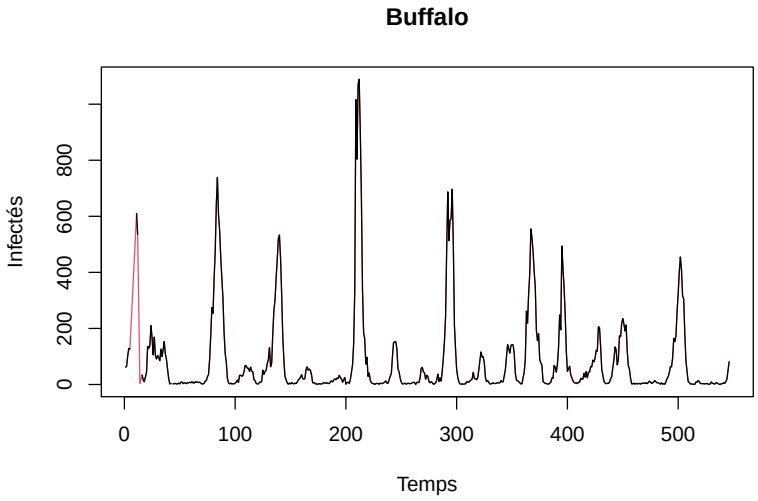

FIGURE 7 – Interpolation des données pour Buffalo

Une fois les données manquantes complétées nous avons créé deux listes de jeux de données (une pour les Etats-Unis et une pour le Royaume-Uni, puis nous avons remplacé les valeurs correspondant à 0 cas infectés par 0.1 afin de pouvoir passer au log durant l'estimation des paramètres de chaque série.

Ensuite nous créeons une fonction qui pourra effectuer des simulations pour plusieurs jeux de paramètres et de conditions initiales à la fois. (c.f. annexe)

Simulations

Ainsi nous pouvons désormais estimer les paramètres pour chaque jeu de données et effectuer des simulations pour chaque jeu de paramètres estimés.

Dans la suite, nous fixerons $\bar{S} = 0.035$ et $\alpha = 0.975$ et nous n'estimerons donc que β_t comme dans l'article par souci de comparaison. Il a été montré que la modification de $\bar{S}$ n'affecte pas les résultats et que le choix de $\alpha = 0.975$ donne de bonnes performances (D.Dalziel 2016). Cependant, il est important de noter que d'après notre experience, les résultats se sont néanmoins montrés très sensibles à la valeur de α.

Afin de tester la fonction simulant la population infectée, voici le résultat de la simulation déterministe à partir de la première valeur I_0 observée sur la série de Londres et de la première valeur de la population susceptible estimée comme condition initiale. Nous évaluons la précision des simulations et des prédictions en calculant le MSE entre entre la série simulée et les données observées, c'est à dire:

$$MSE = \frac{1}{t}\sum_{i=0}^{t}(I_i - \hat{I}_i)^2$$

Avec $\hat{I}_t$ la série simulée à l'instant t.

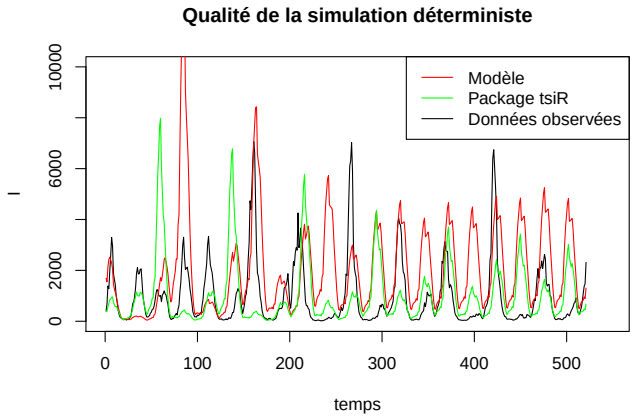

FIGURE 8 – Comparaison des simulations aux données

```
## [1] "MSE_déterministe_tsiR = 2881878"
```

```
## [1] "MSE_déterministe= 1571265"
```

On observe que la simulation déterministe capte bien la périodicité annuelle des épidémies de la rougeole à Londres, cependant certaines périodes où il n'y a pas eu d'épidémies ne sont pas pris en compte par le modèle déterministe. On peut aussi observer que la simulation du package tsiR capte moins bien certains pics que la simulation que nous avons effectuée, d'où la meilleure performance de notre simulation. Afin d'évaluer précisément la qualité du modèle stochastique, nous procédons par méthode de Monte Carlo en effectuant 100 simulations stochastiques avec notre propre fonction et celle du package tsiR. Puis nous moyennons les 100 valeurs du MSE (Mean Squared Error).

```
## [1] "MSE de la série moyénée tsiR =  281165891"
```

```
## [1] "MSE de la série moyénée  =  166344226"
```

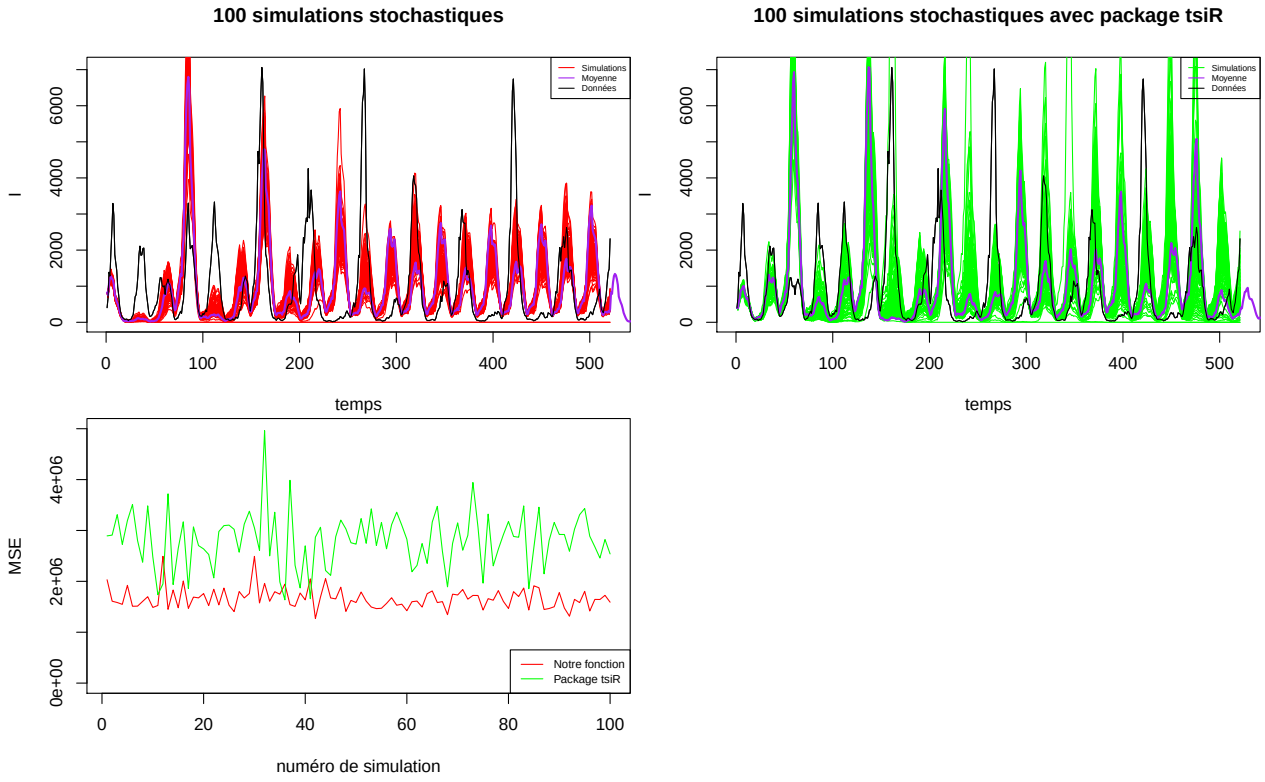

FIGURE 9 – Comparaison des simulations stochastiques

On observe que notre implémentation du modèle stochastique s'ajuste mieux aux données que celle du package tsiR aussi bien au niveau stochastique que au niveau déterministe. En effet nous obtenons un meilleur MSE sur la série moyénnant les 100 simulations stochastiques, c'est à dire:

$$\frac{1}{t}\sum_{i=0}^{t}(\hat{Y}_t - I_t)^2$$

avec

$$\hat{Y}_t = \frac{1}{n}\sum_{i=0}^{n}\hat{I}_t^i$$

où $\hat{I}_t^i$ correspond à la i-ème simulation stochastique à l'instant t et un meilleur MSE sur la simulation déterministe. On observe aussi sur le troisième graphe ci-dessus que la quasi-totalité des 100 simulations stochastiques que nous avons implementé a un meilleur MSE que ceux des simulations stochastiques du package tsiR. Enfin on observe que les deux méthodes captent bien les pics d'épidémie de la rougeole et les cycles bienniaux.

Cependant, cela n'est pas une prédiction, car nous estimons les paramètres sur la totalité des données observées, cela montre juste que l'on explique bien les données. Ainsi, les travaux publiés dans l'article de Dalziel sur la précision du modèle par rapport aux données observées n'est pas une vraie prédiction. Les auteurs en sont conscient, en effet Otto Bjørnstad qualifie la comparaison des simulations aux données observées de "postdiction" et estime que le fait que le modèle capture bien les pics d'épidémies et la periodicité sur les données est déjà remarquable.(Grenfell BT 2002)

Prédictions

Nous avons néanmoins tenté d'effectuer une réelle prédiction en estimant les paramètres sur les 300 premières observations du jeu de données de Londres puis de prédire les valeurs suivantes. Ceci n'a pas été effectué sur

l'article de référence. Les valeurs α et $\bar{S}$ étant fixées les seuls paramètres estimés sont les 26 valeurs de β_t. Voici les résultats de la prédiction (rouge) obtenus pour le modèle stochastique et le modèle déterministe comparés aux données observées:

```
## [1] "MSE_deterministe= 3963664"
```

```
## [1] "MSE_stochastique= 3898455"
```

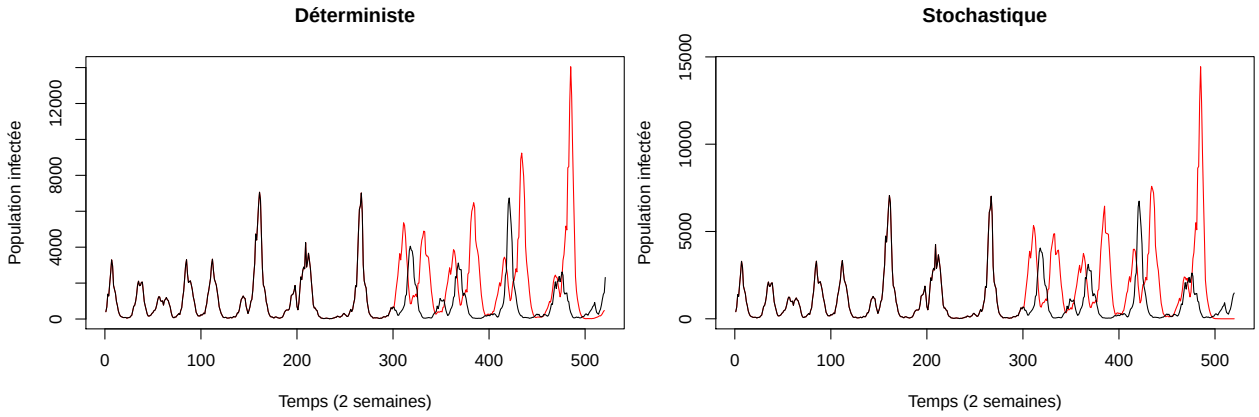

FIGURE 10 – Comparaison des prédictions du modèle

On observe que les cycles semblent respectés, cependant les pics sont fortement décalés.

Périodogrammes

Ayant désormais estimé les paramètres pour les 80 jeux de données, nous effectuons 100 simulations selon le modèle stochastique afin de comparer l'ajustement des périodogrammes avec celui des données observées. En effet, nous souhaitons observer si les simulations stochastiques capturent bien la periodicité des épidémies par rapport aux données observées.

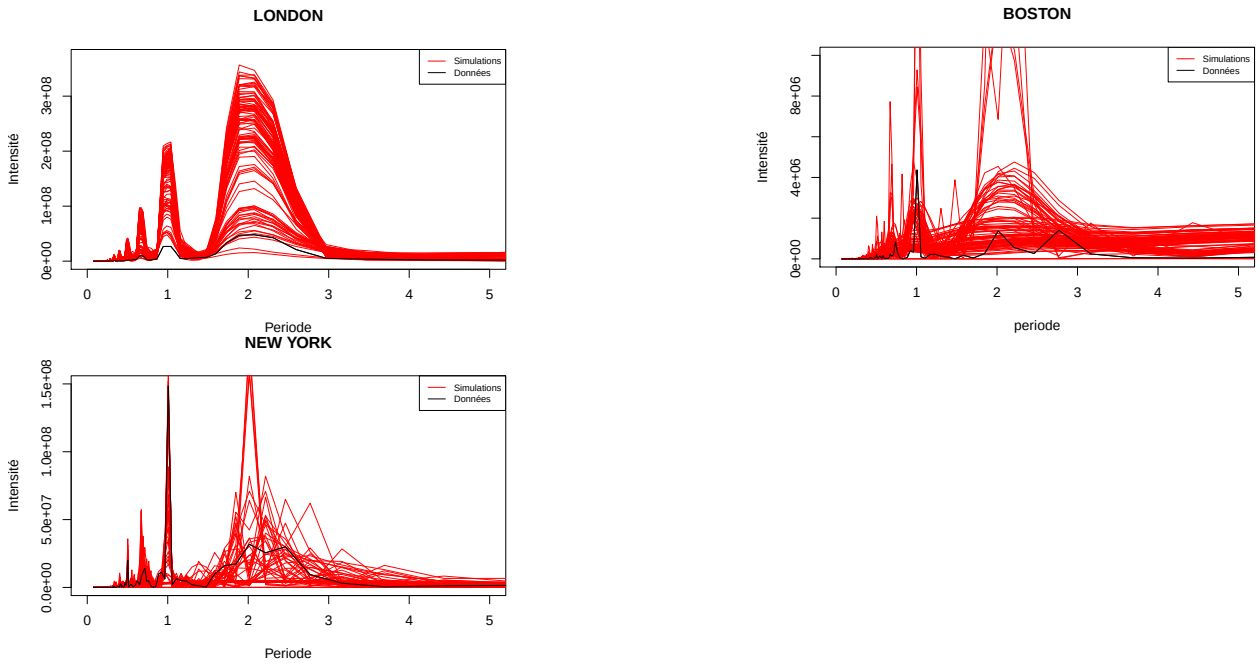

FIGURE 11 – Ajustement des Spectrogrammes de 100 simulations stochastiques

On observe que malgré des différences au niveau de l'intensité, les periodogrammes s'alignent bien au niveau des fréquences prépondérantes. Ceci signifie que le modèle capture bien la periodicité des épidémies dans chaque ville. On observe cependant un bien meilleur ajustement des spectogrammes à Londres. En effet, les periodogrammes des simulations stochastiques pour les villes américaines dévient beaucoup plus du spéctogramme observé pour les fréquences élevées que par rapport au cas de Londres, ce qui peut laisser supposer que les cycles sont plus instables aux Etats-Unis car les cycles y sont moins réguliers et de periodicité moyenne plus importante.

Nature chaotique des épidémies causée par des perturbations sur la période de basse transmission:

En tracant les boxplots des 26 valeurs estimées β_t de chaque villes aux Etats-Unis et au Royaume Uni et en les comparant, on observe une periode de basse transmission s'étendant environ de fin juillet à fin septembre (soit 2 mois) au Royaume-uni. Cependant on observe qu'aux Etats-Unis, la période de basse transmission est entre mi-mai et fin septembre (soit envrion 5 mois) et qu'elle est de plus forte amplitude. On observe donc une plus longue periode de basse transmission aux Etats-unis ainsi qu'une plus forte amplitude qu'au Royaume-Uni.

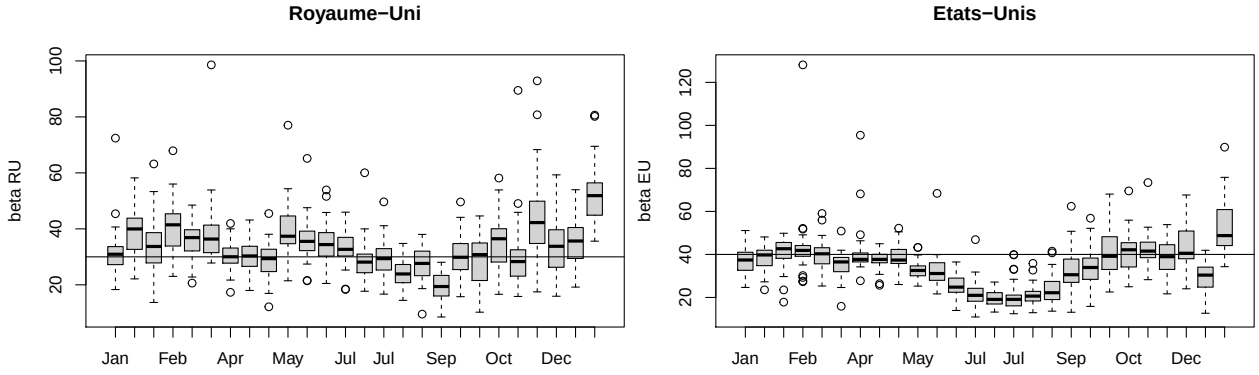

FIGURE 12 – Boxplot de beta au Royaume Uni (Gauche) et aux EU (Droite)

Comme expliqué précédement, nous souhaitons montrer que cette différence sur la période de basse transmission affecte les cycles d'épidémies de la rougeole aux Etats-Unis résultant en une périodicité plus élevée et des cycles plus irréguliers c'est à dire instables.

Exposants globaux de Lyapunov

Afin de le montrer, nous reproduisons les résultats de l'article de référence en calculant les exposants globaux de Lyapunov sur des simulations du modèle déterministe 100 ans en avant avec 100 ans de 'burn phase', c'est à dire simuler sur 200 ans en ne tenant compte que des 100 dernières années.

Les coefficients de Lyapunov mesurent la stabilité d'un système à travers le calcul du log du taux moyen d'accroissement, c'est une mesure de la propagation de l'erreur donc de la stabilité d'un système dynamique. Un coefficient négatif indique que le système est stable tandis que si il est positif cela signifie que le système est instable.

Dans le cas d'une fonction $f : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}}$

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \log(\|J_t U_0\|)$$

(D.Dalziel 2016) avec

$$U0 = (1, 0, ..., 0) \text{ et } J_t \text{ la jacobienne à l'instant t}$$

Dans notre cas on a:

$$(S_{t+1}, I_{t+1}) = (S_t + B_t - \beta_s I_t^\alpha S_t N_t^{-1}, \beta_s I_t^\alpha S_t N_t^{-1}) = f(S_t, I_t)$$

et donc $U_0 = (1, 0)$ la jacobienne vaut:

$$\begin{pmatrix} 1 - \beta_s I^\alpha / N_t & -\beta_s S_t (I_t^{\alpha-1} \alpha / N_t) \\ \beta_s I_t^\alpha / N_t & \beta_s S_t (I_t^{\alpha-1} \alpha) / N_t \end{pmatrix}$$

Nous créeons donc une fonction `lyapunov` calculant le coefficient associé à un sytème étant donné les paramètres estimés et les séries de la population susceptible et des cas infectés pour un ville donnée. (c.f annexe pour l'implémentation de la fonction lyapunov)

Dans le but de tester cette fonction nous effectuons une simulation de 100 ans à partir du modèle deterministe pour New York en gardant la taille de la population et la natalité constante.(c.f annexe pour accéder au code) On obtient un coefficient de Lyapunov positif ce qui est donc cohérent avec l'hypothèse que les cycles de la rougeole aux Etats-unis sont instables.

```
## [1] "Exposant de Lyapunov pour New York =  0.014"
```

Nous nous intéréssons maintenant à calculer les coefficients de Lyapunov pour des simulations du modèle déterministe 100 ans en avant pour chaque ville. Nous stockons les conditions initiales (I_0 comme la valeur initiale des infectés de la série observée et S_0 comme la première valeur de la série des susceptibles reconstruite).

Ensuite nous simulons selon le modèle déterministe pour chaque jeu de paramètres puis nous calculons les coefficients de Lyapunov pour chaque simulation. Enfin nous traçons les coefficients de Lyapunov et la proportion de periodes de deux semaines avec strictement moins de 1 cas en fonction de la taille moyenne de la population pour chaque ville. Nous choisissons de montrer la proportions de périodes avec moins de 1 cas car pour pouvoir passer au log nous avons ajouté 0.1 à toutes les valeurs de la série des cas infectés. Voici les résultats obtenus:

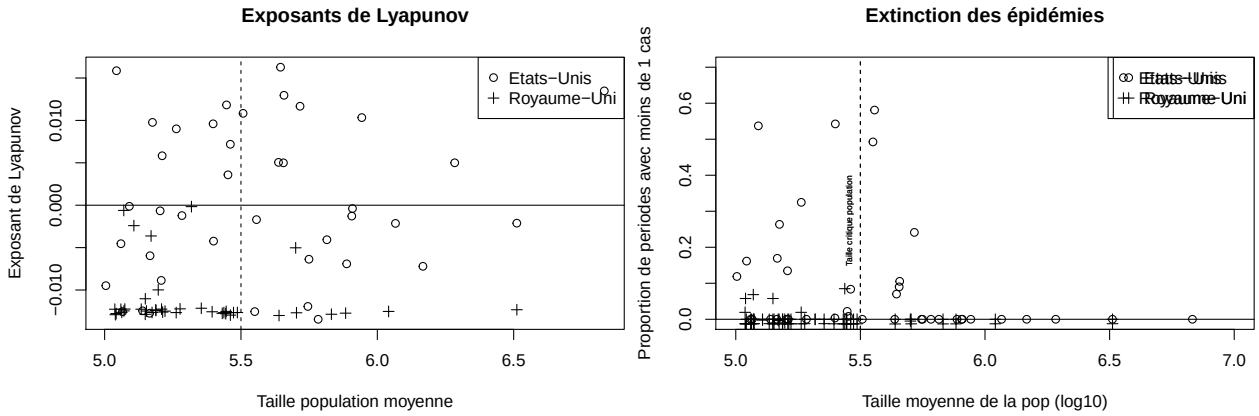

FIGURE 13 – Exposants de Lyapunov (gauche) et proportion des cas <1 (droite)

On observe que la totalité des exposants de Lyapunov correspondant aux villes du Royaume-Uni sont négatifs, ce qui correspond donc à des cycles stables de la rougeole alors que plus de 18 coefficients, chacun associé à une ville américaine, sont supérieurs ou égaux à 0. Il est donc clair que les cycle d'épidémies de rougeole aux Etats-Unis présentent une instabilité.

De plus sur la deuxième figure on observe cependant que la proportion de periodes de deux semaines avec moins de 1 cas reste proche de celle des villes du Royaume-Uni mis à part un faible nombre d'exceptions pour des faibles populations. Au dessus de la taille de population critique de $10^{5.5}$, on observe que malgré l'instabilité des cycles de la rougeole aux Etats-Unis, il n'y a que 3 villes américaines avec plus 30% de periodes avec moins de 1 cas. Ceci est intéressant car il était pensé qu'une déviation des cycles stables observés au Royaume-Uni résultait automatiquement à des épidémies épisodiques à forte périodicité moyenne et donc avec de nombreuses périodes d'extinction de la maladie. C'est ce qui avait été nottament observé à Niamey au Niger où de fortes perturbations démographiques entrainaient des extinctions de l'épidémie. Or ici, on observe que malgré des cycles instables, les épidémies persistent.

Bifurcations en fonction de l'amplitude ou de la durée de la periode de basse transmission

Enfin, pour confirmer l'hypothèse que ce sont les différences sur la période de basse transmission de la maladie qui affectent les cycles de la rougeole aux Etats-Unis. Nous effectuons plusieurs simulations déterministes pour la ville de Los Angeles en faisant varier la durée de cette période de basse transmission ou en faisant varier l'amplitude de la période de basse transmission à travers des fonctions synthétiques du taux de transmission β_t dans le but d'obtenir des tracés de bifurcation.

Afin d'effectuer cela, nous calculons plusieurs séries synthétiques des 26 valeurs du paramètre β_t de sorte à obtenir des periodes de basse transmission plus ou moins longues ou plus ou moins amples. Nous créeons ces valeurs synthétique à partir de la fonction synthétique:

$$\beta_t = \begin{cases} \beta_- & \text{si } a \leq t \leq b \\ \beta_+ & \text{sinon.} \end{cases}$$

où a est le début de la période de basse transmission et b la fin de la période de basse transmission et β_- et β_+, les valeurs minimales et maximales atteintes par la fonction de transmission (D.Dalziel 2016). Nous ajoutons ensuite une constante afin d'obtenir une moyenne des valeurs de beta égale à la moyenne des valeurs de β obtenues lors de l'estimation des paramètres pour les jeux de données.

Voila le tracé d'une fonction synthétique générée pour une période de de basse transmission entre la 9ème bisemaine et la 15ème bisemaine, pour une amplitude de 20 et pour une moyenne désirée de 40. (c.f annexe pour accéder au code)

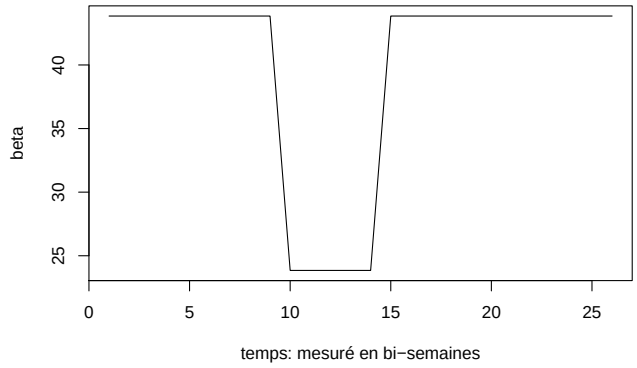

FIGURE 14 – Exemple d'une fonction synthétique de beta

Nous traçons deux bifurcations, une en fonction de la durée de la période de basse transmission pour une amplitude fixée puis une en fonction de l'amplitude de la période de basse transmission pour une durée fixée.

Afin d'y parvenir pour le premier tracé:

Nous créeons plusieurs séries synthétiques du taux de transmission en faisant varier la durée de la période de basse transmission sans modifier l'amplitude (que l'on paramètre comme égale à l'amplitude de l'estimation de β_t sur les données observées) Pour chaque série synthétique β_t nous effectuons une simulation 100 ans en avant en maintenant la taille de la population et la natalité constante égale à la moyenne observée. *Nous stockons les valeurs simulées de la population infectée tous les mois de mai c'est à dire à la 10 ème bi-semaine (correspondant en général au pic annuel des cas de rougeole).* Pour chaque durée de la période de basse transmission de chaque série β_t synthétisée, nous traçons les valeurs de la population infectée à chaque moi de mai, soit 100 valeurs.

De même pour le deuxième tracé:

Nous créons plusieurs séries synthétiques du taux de transmission β_t en faisant varier l'amplitude pour une durée constante égale à celle observée sur les données Nous simulons pour chaque série β_t synthétisée, la population infectée 100 ans en avant avec la population et le nombre de naissance constant. *Nous tracons ensuite les valeurs obtenues de la population infectée tous les mois de mai en fonction de l'amplitude de la série synthétisée.

Sachant que les mesures sont effectuées toutes les deux semaines, donc que les séries du taux de transmission ne contiennent que 26 valeurs et que nous devons centrer les intervalles de temps pour la période de basse transmission au niveau du mois de juin pour être cohérent avec les observations, nous ne pouvons donc pas créer plus de 11 séries synthétiques pour le premier tracé. Il aurait été intéressant de pouvoir en créer plus pour bien observer les bifurcations. Cependant nous obtenons quand même de bons résultats et nous pouvons créer de nombreuses séries synthétiques pour le deuxième tracé qui lui est en fonction de l'amplitude. (c.f annexe pour accéder au code)

Voici les résultats obtenus:

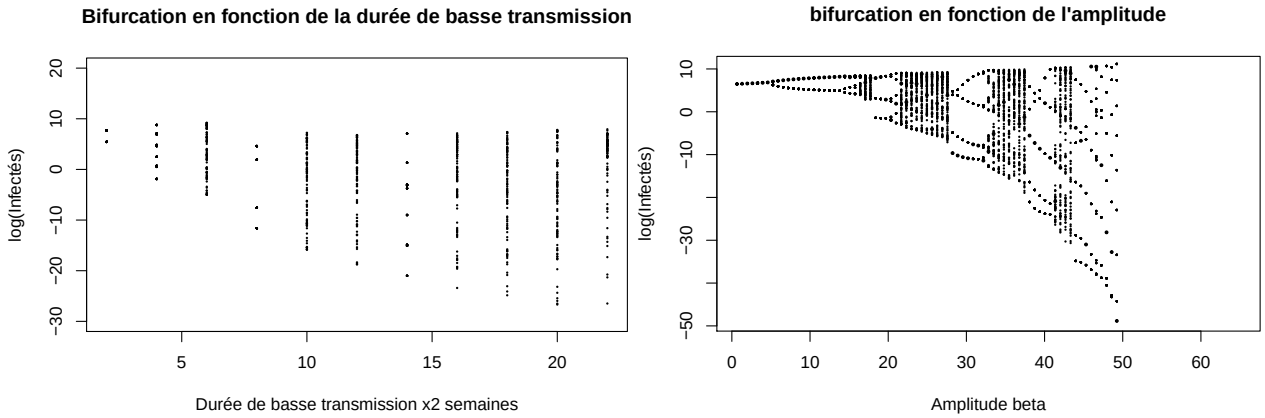

FIGURE 15 – Tracés de Bifurcation

Dans le premier tracé on observe que pour une très faible durée de la periode de basse transmission, le nombre de cas infectés sur les 100 mois de mai ne prend que deux valeurs et plus la durée de la période de basse transmission augmente plus le nombre de valeurs prises ainsi que l'intervalle de valeurs prises par le nombre de cas infectés à chaque mois de mai augmente. De plus pour des durées de plus de 8 "bisemaines" (16 semaines) comme celle observée aux Etats-Unis (20 semaines) on atteint parfois des valeurs proches de zéro et d'autres supérieures à 1000 ce qui montre que les cycles ne sont plus stables du tout.

Dans le deuxième tracé de bifurcation on observe bien un nombre croissant de trajectoires possible en fonction de l'amplitude de la periode de basse transmission.

Ainsi on peut empiriquement conclure que des perturbations sur la periode de basse transmission entrainent une instabilité des cycles de la rougeole malgré une persistance des épidémies.

Conclusion

En conclusion, le modèle TSIR semble bien capter les cycles de la rougeole. Malgré des prédictions de qualité médiocre, les postdictions restent précises. Il ne fait pas de doute que les cycles d'épidémies aux Etats-Unis sont instables, et les simulations montrent que la faible perturbation sur la période de basse transmission semble entraîner une déviation des cycles stables observés au Royaume-Uni vers des cycles instables et des épidémies persistantes. Nous regrettons de ne pas avoir pu effectuer des simulations en faisant varier les conditions initiales afin de montrer une plus forte sensibilité aux conditions initiales pour les cycles de la rougeole aux Etats-Unis afin de souligner encore plus la nature chaotique de ces épidémies, faute de temps. De plus, nous aurions souhaité effectuer des analyses sur plus de villes et de pouvoir plus approfondir les aspects théoriques associés au modèle TSIR, par exemple discuter de l'identifiabilité du modèle, de la modélisation comme un processus auto-régressif, ou de l'estimation des paramètres par méthodes Bayesiennes (sujet sur lequel les auteurs de l'article sont entrain de travailler). Cependant nous sommes convaincu des résultats obtenus par D.Dalziel qui semblent cohérents avec ceux que nous avons reproduit malgré des doutes sur le fait de fixer la valeur du paramètre α pour toutes les villes.

Annexe

Ci dessous vous trouverez l'implémentation des étapes cruciales de ce TER, si vous souhaitez executer tout le code nous vous invitons à consulter le fichier .Rmd qui contient le scripte executable intégralement.

```r
#Regression cumul/cumul
regression = smooth.spline(x=cumsum(Births),y=cumsum(modeldata$cases),df=2.5)

## Calcul de rho comme la pente en chaque point
rho = predict(regression,cumsum(Births),deriv=1)$y
# Extraction des résidus, ajustement par rho.
D = -resid(regression)/rho
```

```r
regression.lin = lm(cumsum(modeldata$cases)~cumsum(Births))
rho_2 = regression.lin$coefficients[2]
D2 = -regression.lin$residuals/rho_2
```

```r
res2 = runtsir(data = modeldata,xreg="cumbirths",method = "negbin",regtype='lm')
```

```r
I= modeldata$cases/rho
I_t=I[1:546]
I_tp1=I[2:547]
```

```r
## Estimation de Sbarre, beta et alpha
saison= rep(1:26,21) #dummy variable pour beta
log_Itp1= log(I_tp1) # It+1 en log
log_It= log(I_t) # It en log
D_t= D[1:546] # résidus même longueur
sbar = seq(0.02,0.4,length=300) #vecteurs des candidats de sbarre
offsetN = -log(modeldata$pop[1:546]) #log de la population
vraisemblance = rep(NA, length(sbar)) # vecteur contenant les valeurs de la vraisemblance
```

```r
for(i in 1:length(sbar)){
St = log(sbar[i]*modeldata$pop[1:546] + D_t) #reconstruction pour chaque valeur de sbarre
glmfit = glm(log_Itp1~ -1 +as.factor(saison) +log_It +offset(St)+offset(offsetN)
,family=gaussian(link="identity")) #on estime les paramètres du modèle pour Sbarre
vraisemblance[i] = glmfit$deviance/2 #stockage de la vraisemblance pour sbarre
}
```

```r
Susc_ldn_rec =  log(sbar_hat*modeldata$pop[1:546]+ D_t) #reconstruction de S
Londonfit = glm(log_Itp1~ -1+ as.factor(saison) + log_It + offset(offsetN)
                + offset(Susc_ldn_rec)) #estimation des paramètres
```

```r
#Fonction simulant une série à partir des paramètres
#du modèle et de conditions initiales, mode stochastique ou deterministe
SimTsir2=function(beta, alpha, B, N, inits = list(Snull = 0, Inull = 0), type = 1)
  {
IT = length(B)
s = length(beta)
lambda = rep(NA, IT)
I = rep(NA, IT)
S = rep(NA, IT)
I[1] = inits$Inull
lambda[1] = inits$Inull
S[1] = inits$Snull


for(i in 2:IT)
  {
lambda[i] = beta[((i-2) %% s)+1]*S[i - 1]*(I[i - 1]^alpha)/N[i-1]
```

```
if(type == 2)
    {I[i] = rnbinom(1,mu=lambda[i],size=I[i-1]+1e-10)}
if(type == 1)
{I[i] = lambda[i]}



S[i] =max(S[i - 1] + B[i-1] - I[i],1)

  }
return(list(I = I, S = S))
}
```

```
lyapunov=function(I,S,alpha,b,N){
    len=length(I)
    s = length(b)
    j11=rep(0,len)
    j12 = rep(0,len)
    j21 = rep(0,len)
    j22 = rep(0,len)
    J= matrix(c(1,0),ncol=1)

    for(i in 1:len){
      j11[i]=1 - b[(i-1)%%s +1]  * (I[i]^alpha)/N
      j12[i]= - b[(i-1)%%s+1]*S[i]*(I[i]^(alpha-1))*alpha/N
      j21[i]= b[(i-1)%%s+1]*(I[i]^alpha)/N
      j22[i] = b[(i-1)%%s+1]*S[i]*((I[i]^(alpha-1))*alpha)/N
      J= matrix(c(j11[i],j12[i],j21[i],j22[i]),ncol=2,byrow = T)%*%J
    }
    res= list(lypunov_exp=log(norm(J))/len,j11_=j11,j12_=j12,j21_=j21,
              j22_=j22,S=S,I=I,alpha=alpha)
    return(res)
}
regression$fit$range
```

```
param_NY = param_estim2(split_df_no0$`NEW YORK`)
res_tsir_ny= SimTsir2(param_NY$beta_,0.975,rep(mean(split_df_no0$`NEW YORK`$births),5201)
                      ,rep(mean(split_df_no0$`NEW YORK`$pop),5201),inits=list(Snull=param_NY$Suc[1],
  Inull=split_df_no0$`NEW YORK`$cases[1]/param_NY$rho_hat[1]),type=1)


lyapNY= lyapunov(res_tsir_ny$I[2601:5200],res_tsir_ny$S[2601:5200],
                 alpha = 0.975,param_NY$beta_,
                 N=mean(split_df_no0$`NEW YORK`$pop))
paste("Exposant de Lyapunov pour New York = ",round(lyapNY$lypunov_exp,digits= 3))
```

```
mean_bet = mean(params_US$`LOS ANGELES`$beta_)
generate_beta = function(a,b,beta_minus,beta_plus,mean_b)
{
  bet=c(1:26)
  sequence = seq(1,26,by=1)
  for(i in sequence)
  {
    if(i<b && i>a) bet[i]=beta_minus
```

```
    else bet[i] = beta_plus
  }
  diff= mean_b-mean(bet)
  bet = bet +diff
  return(bet)
}

plot(generate_beta(9,15,10,30,40),type='l',xlab='temps: mesuré en bi-semaines',ylab='beta')
```

```
grid_duration= data.frame(a=c(1:11),b=c(23:13))
beta_amp = range(params_US$`LOS ANGELES`$beta_)
bifurc_duration = vector(mode="list",length=length(grid_duration$a))
sequ= seq(10,5200,by=26)
for(i in c(1:length(grid_duration$a)))
{
  beta_var = generate_beta(grid_duration[i,1],grid_duration[i,2],beta_amp[1],beta_amp[2],mean_bet)
  res_tsir_LA= SimTsir2(beta_var,0.975,rep(mean(split_df_no0$`LOS ANGELES`$births),5201),rep(mean(split_

  res_tsir_LA = res_tsir_LA$I[sequ]
  res_tsir_LA = res_tsir_LA[100:200]
  res_tsir_LA = log(res_tsir_LA)
  bifurc_duration[[i]]=res_tsir_LA
}

bifurc_duration = as.data.frame(bifurc_duration)
names(bifurc_duration)= grid_duration$b-grid_duration$a
bifurc_duration=rev(bifurc_duration)
plot(rev(grid_duration$b-grid_duration$a),bifurc_duration[1,],ylim=c(-30,20),pch=16,cex=0.3,xlab="Durée
for(i in c(2:100))
{
  points(rev(grid_duration$b-grid_duration$a),bifurc_duration[i,],pch=16,cex=0.3)
}

beta_m = seq(5,mean(range(params_US$`LOS ANGELES`$beta_)),length=100)
beta_p = seq(mean(range(params_US$`LOS ANGELES`$beta_)),70,length=100 )

grid_amp=data.frame(b_= beta_m,b_p=rev(beta_p))
grid_amp=grid_amp[1:99,]


bifurc_amp = vector(mode="list",length=length(grid_amp$b_))
sequ= seq(10,5200,by=26)
for(i in c(1:length(grid_amp$b_)))
{
  beta_var = generate_beta(10,19,grid_amp[i,1],grid_amp[i,2],mean_bet)
  res_tsir_LA= SimTsir2(beta_var,0.975,rep(mean(split_df_no0$`LOS ANGELES`$births),5201),rep(mean(split_

  res_tsir_LA = res_tsir_LA$I[sequ]
  res_tsir_LA = res_tsir_LA[100:200]
  res_tsir_LA = log(res_tsir_LA)
  bifurc_amp[[i]]=res_tsir_LA
}
bifurc_amp = as.data.frame(bifurc_amp)
names(bifurc_amp)= grid_amp$b_p-grid_amp$b_
```

```
bifurc_amp=rev(bifurc_amp)
plot(rev(grid_amp$b_p-grid_amp$b_),bifurc_amp[1,],xlab="Amplitude beta",ylab="log(Infectés)",cex=0.5,pch
for(i in c(2:100))
{
  points(rev(grid_amp$b_p-grid_amp$b_),bifurc_amp[i,],pch=16,cex=0.3)
}
```

References

Alexander D. Becker, Bryan T. Grenfell. 2017. « tsiR: An R package for time-series Susceptible- Infected-Recovered models of epidemics ». *PLOS One* 12 (9): 891-921. doi: https://doi.org/10.1371/journal.pone.0185528.

Barbel F. Finkenstadt, Bryan T. Grenfell. 2000. « Time series modelling of childhood diseases: a dynamical systems approach ». *J R Stat Soc C. Blackwell Publishers Ltd* 49: 187-205.

Bjørnstad ON, Grenfell BT, Finkenstadt BF. 2002. « Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model ». *Ecological Monographs Eco Soc America* 72: 169-84.

D.Dalziel, Benjamin. 2016. « Persistent Chaos of Measles Epidemics in the Prevaccination United States Caused by a Small Change in Seasonal Transmission Patterns ». *PLOS Computational Biology.* doi:10.1371journal.pcbi.1004655.

Ferrari MJ, Bharti N, Grais RF. 2008. « The dynamics of measles in sub-Saharan Africa ». *Nature. Nature Publishing Group* 451 (10): 679-84. doi:10.1038/ nature06509 PMID: 18256664.

Grenfell BT, Bjørnstad ON. 2002. « Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model ». *Ecological Monographs.* 72: 185-202.

N.Bjørnstad, Ottar. 2018. *Epidemics, Models and Data using R.* Use R! Springer.