Valeurs Extrêmes

Devoir maison obligatoire (Dauphine)

Paul Hardouin

January 31, 2020

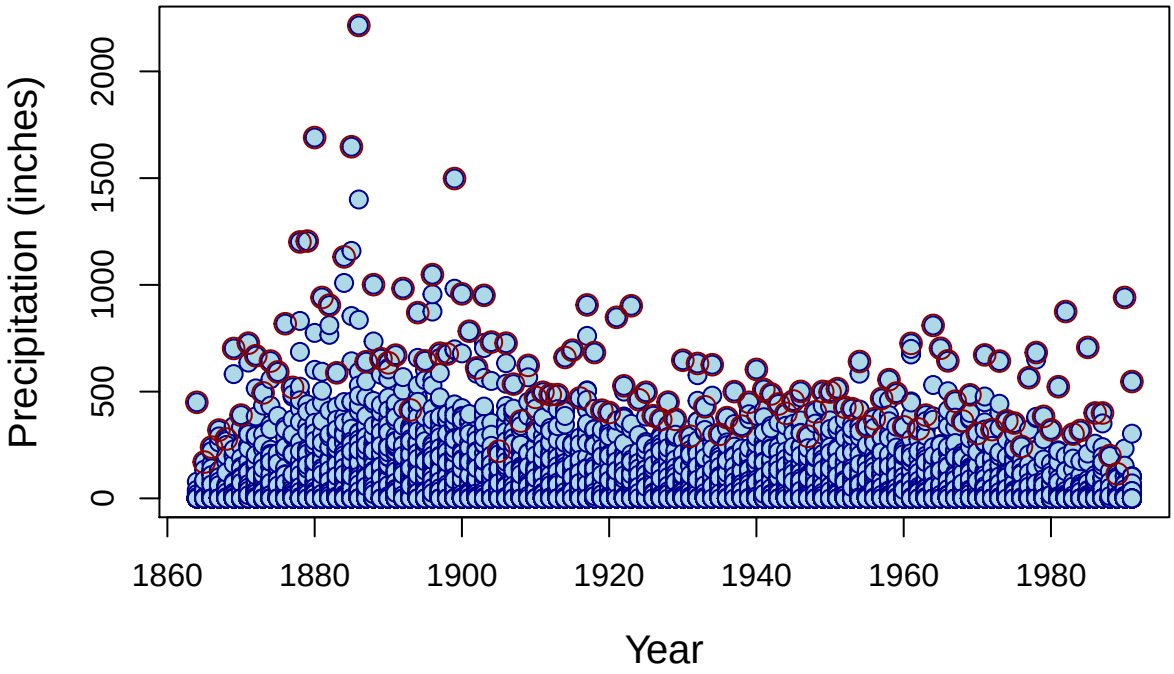
Contents

1. Etude du jeu de données “marseille”

Nous travaillons ici sur les données du fichier **Marseilles.txt**, lequel contient des relevés quotidiens du niveau de précipitation à Marseille en $10^{-1}mm$. Ces relevés ont ete effectués pendant 127 années [1864-1991], à partir du 1^{er} Août 1864. Tous les 29 février ont ete enlevés. On va analyser la distribution extrême de ce jeu de données, en utilisant 2 approches (GEV et GPD).

1.1. Exploration des données

On charge les librairies **extRemes et ismev** qui vont nous permettrent de réaliser notre étude. Ensuite, on lit les données. On les affiche (bleu), ainsi que les valeurs maximales par année (rouge).



Ce graphique semble indiquer que les précipitations marseillaises étaient globalement plus importantes à la fin du $19^{ème}$ siècle que pendant le $20^{ème}$ siècle. Cela peut potentiellement remettre en cause l'hypothèse d'une distribution identique des maxima annuels, nécessaire pour l'application du théorème fondamental des valeurs extrêmes. Afin de rester dans le cadre EVT, je fais l'hypothèse que cette tendance haussière des précipitations à la fin du $19^{ème}$ correspond à des valeurs extrêmes, et que la distribution des maxima annuels est bien IID.

D'autre part, ce graphique montre que beaucoup de valeurs élevées sont ne seront pas prises en compte avec l'approche GEV. En effet, ces valeurs sont relevées la même année que d'autres valeurs qui leur sont supérieures. Une approche GPD exploitera donc sans doute mieux l'ensemble de l'information disponible.

1.2. Approche GEV - MDA = "Fréchet"

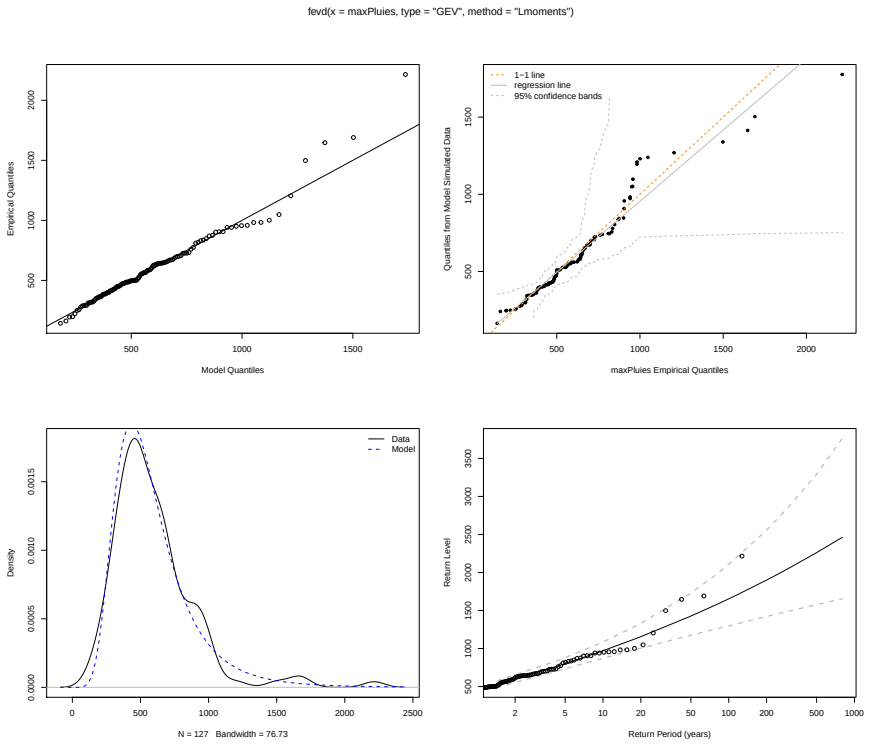
En premier lieu, nous faisons une approche GEV, par la méthode MLE et par la méthode des moments. En effet, comme nous n'avons que 127 observations, il se pourrait que l'optimisation MLE fonctionne mal, et nous voulons donc comparer. Finalement, nous tombons sur des valeurs proches, comme le montre le tableau ci-dessous.

Type	Method	Location	Scale	Shape
GEV	MLE	459.3	196.4	0.100
GEV	Lmoments	456.7	192.5	0.124

L'observation des intervalles de confiance ne montrent pas d'incertitude sur le signe du paramètre de forme, qui est positif. Nous sommes donc dans le domaine d'attraction de **Fréchet**.

Type	Method	Shape 2.5%	Shape estimate	Shape 97.5%
GEV	MLE	-0.018657	0.1002232	0.2191034
GEV	Lmoments	-0.0140209	0.1241941	0.2572127

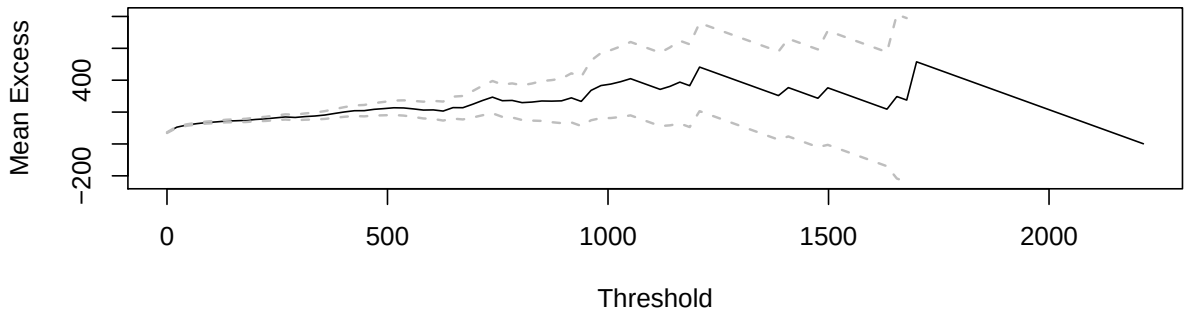
Ci-dessous, nous affichons les QQ-plots et les niveaux de retour liés à la méthode des moments (graphes quasi-identiques pour la méthode MLE). On observe une bonne adéquation du modèle aux maxima ayant un niveau de retour inférieur à 10 ans. Au-delà, en revanche, le modèle commence à s'écarter des observations, même si celles-ci restent dans l'intervalle de confiance.



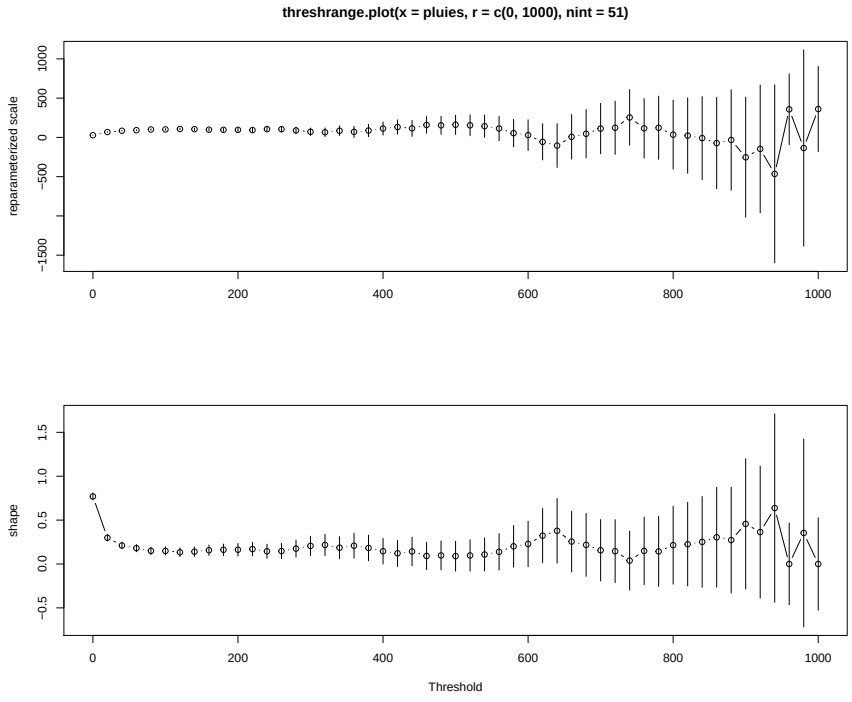
1.3. Approche GPD

1.3.1. Choix du seuil

En premier lieu, nous observons le **Mean Residual Life Plot**. On devine une zone linéaire à partir de $200.10^{-1}mm$ environ. La pente est clairement positive, ce qui nous permet d'intuiter que le paramètre de forme est positif, et que nous sommes dans le domaine d'attration de **Fréchet**.



Une estimation des paramètres en fonction du seuil nous permet d'affiner sa valeur (cf. figure ci-dessous). Les paramètres ont une dérivée quasi-nulle autour de la valeur $150.10^{-1}mm$, que nous retenons pour la suite. D'autre part, la valeur du paramètre de forme est nettement positive dans toute cette zone de stabilité, ce qui confirme une nouvelle fois le domaine d'attration de **Fréchet**.

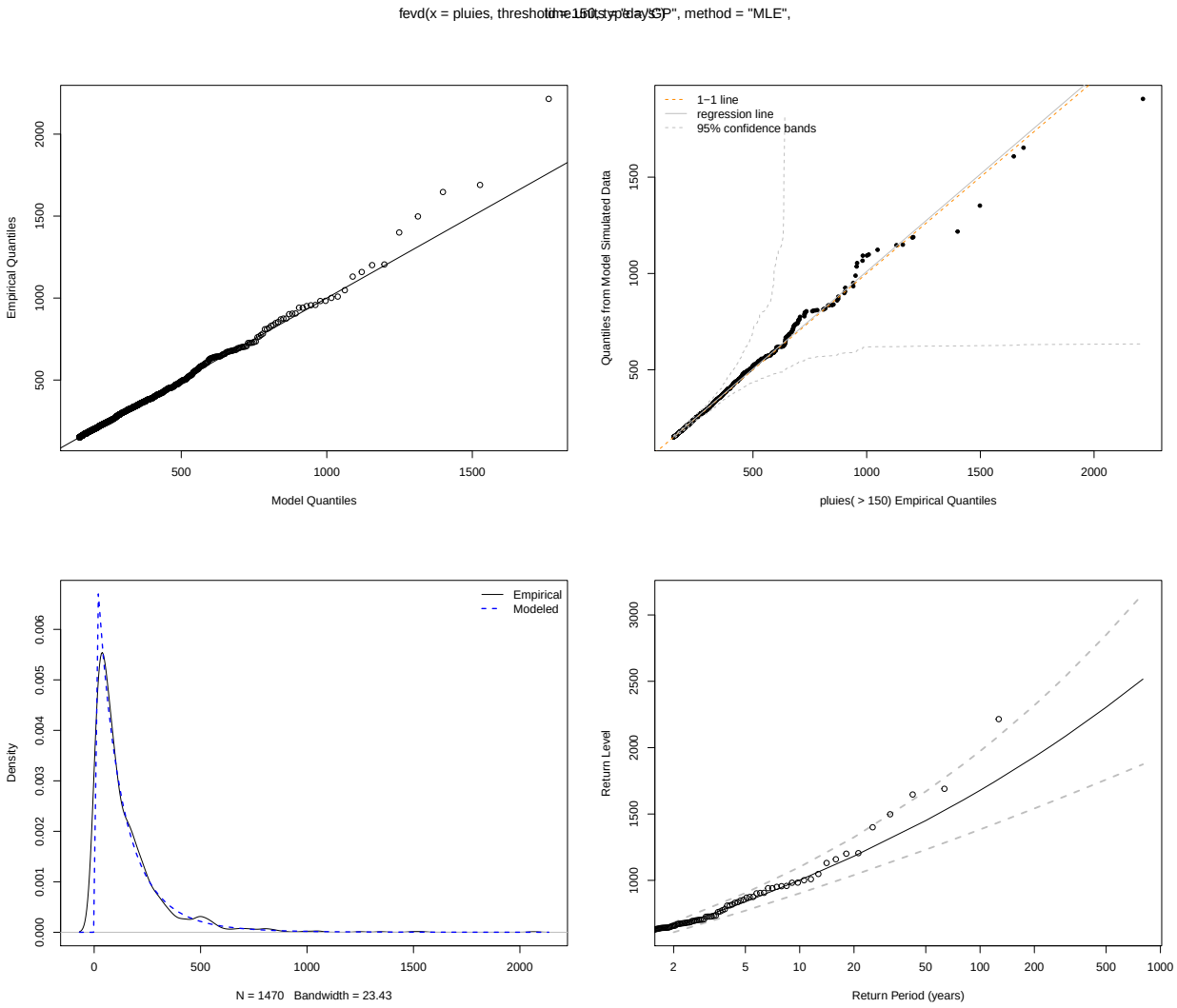


1.3.2. Estimation des paramètres

Pour l'estimation des paramètres, on cherche un modèle GPD avec la méthode des moments et la méthode MLE. Dans les 2 cas, le fait d'avoir utilisé une approche GPD a fortement resserré les intervalles de confiance par rapport à une approche GEV. Le paramètre de forme se trouve désormais dans un intervalle de confiance purement positif, ce qui confirme encore une fois le domaine d'attraction de Fréchet. De plus, on trouve ici 2 estimations très proches des paramètres.

Type	Method	scale	shape	shape 95% Confidence Interval
GPD	MLE	124.6	0.144	(0.0878 , 0.2011)
GPD	Lmoments	125.8	0.136	(0.0714 , 0.1989)

On affiche ci-dessous les QQ-plots et les niveaux de retour liés à la méthode MLE (les résultats sont équivalents pour la méthode des moments). Par rapport à l'approche GEV, on observe un modèle visuellement plus proche des observations.



1.4. Comparaison

Niveau de retour à 100 ans

C'est le modèle **GPD - MLE** qui donne l'intervalle de confiance le plus fin. Son niveau de retour à 100 ans vaut $0.1651m$.

Type	Method	100-year return level	95% Confidence Interval
GEV	MLE	1607.505	(1247.4577, 1967.5528)
GEV	Lmoments	1651.503	(1274.9765, 2276.4579)
GPD	MLE	1678.429	(1383.3087, 1973.5490)
GPD	Lmoments	1642.874	(1372.4279, 2017.0893)

Niveau de retour à 1000 ans

C'est le modèle **GPD - Lmoments** qui donne l'intervalle de confiance le plus fin. Son niveau de retour à 1000 ans vaut $0.2534m$.

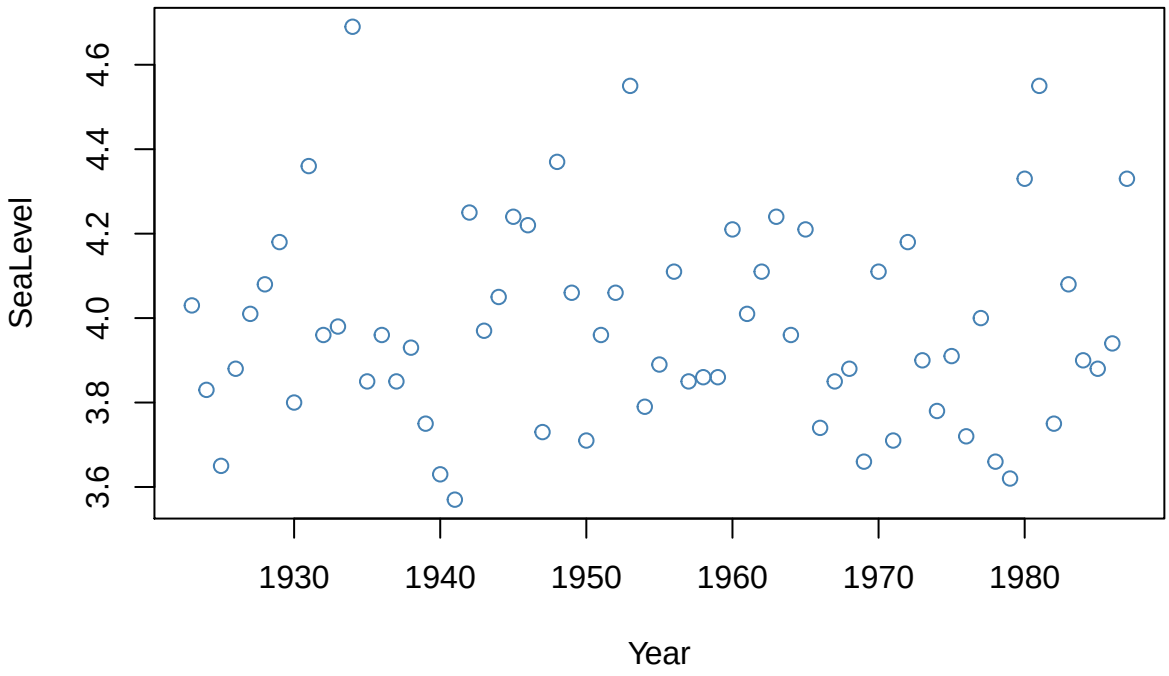
Type	Method	1000-year return level	95% Confidence Interval
GEV	MLE	2416.156	(1510.9467, 3321.3655)
GEV	Lmoments	2562.38	(1719.2893, 4220.2312)
GPD	MLE	2622.336	(1929.9369, 3314.7355)
GPD	Lmoments	2534.285	(1970.5454, 3310.1771)

2. Etude du jeu de données “portpirie”

Ce jeu de données présente l'évolution des maxima annuels du niveau de la mer à Port Pirie, un lieu juste au Nord d'Adelaide, dans le sud de l'Australie, pendant la période 1923-1987. A partie de telles données, nous voudrions obtenir une estimation du niveau maximum que la mer pourrait prendre dans la région sur une période de 100 ou de 1000 ans. Il semble raisonnable de supposer que, comme le motif de variation est resté constant tout au long de la période d'observation, nous pouvons modéliser les données comme des observations indépendantes de la distribution GEV.

2.1. Exploration des données et choix du modèle : GEV

En premier lieu, nous chargeons les données et nous les affichons. La distribution du maximum annuel semble stable, ce qui renforce l'hypothèse d'indépendance nécessaire à la distribution par un modèle GEV.



2.2. Estimation des paramètres - MDA = “Gumbel”

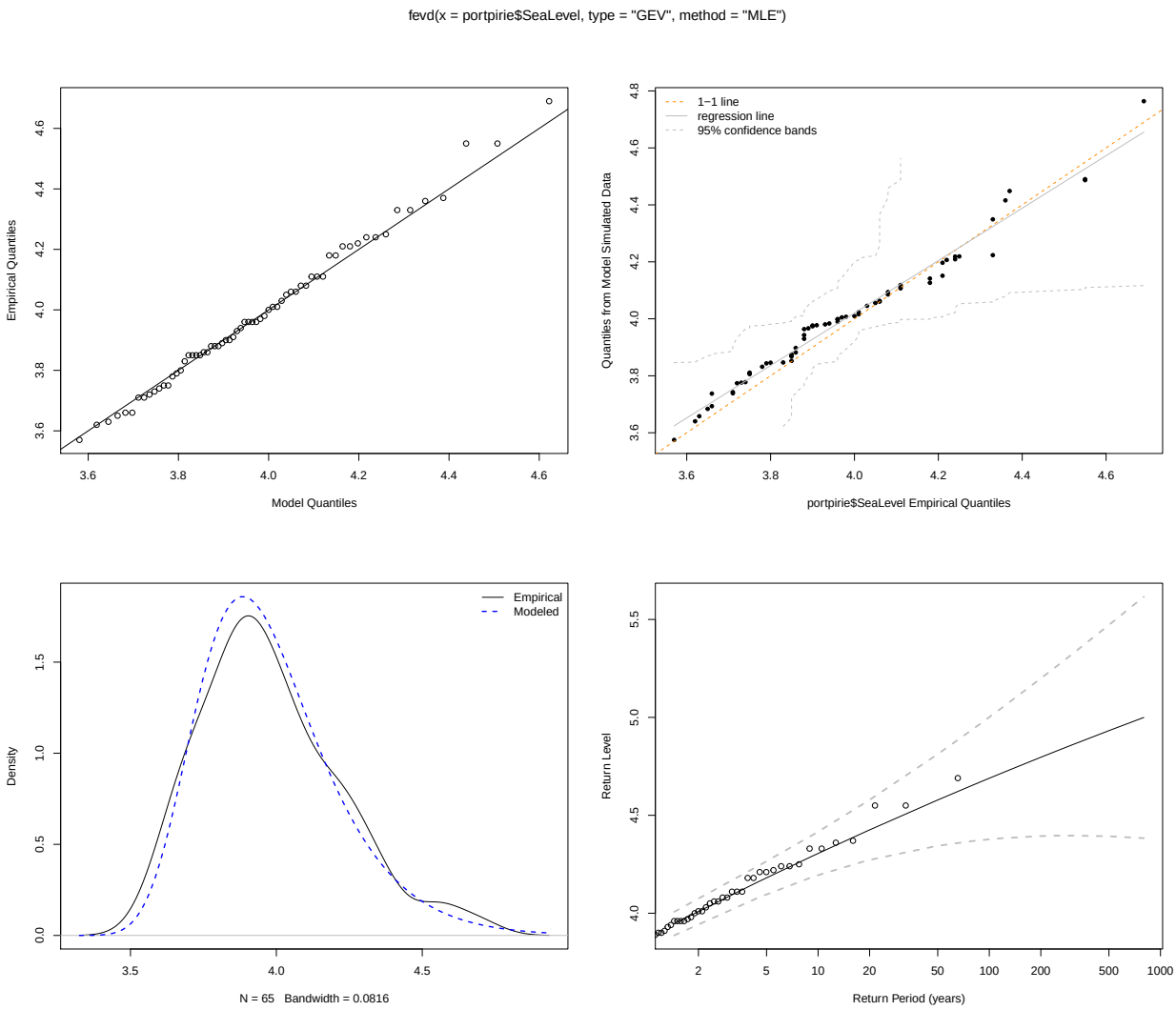
En premier lieu, nous faisons une approche GEV, par la méthode MLE et par la méthode des moments. En effet, comme nous n'avons que 65 observations, il se pourrait que l'optimisation MLE fonctionne mal, et nous voulons donc comparer. Finalement, nous tombons sur des valeurs très proches, comme le montre le tableau ci-dessous.

Type	Method	Location	Scale	Shape
GEV	MLE	3.874	0.198	-0.050
GEV	Lmoments	3.873	0.203	-0.051

En revanche, les intervalles de confiance montrent une forte incertitude sur le signe du paramètre de forme.

Type	Method	Shape 2.5%	Shape estimate	Shape 97.5%
GEV	MLE	-0.2426841	-0.0501095	0.1424651
GEV	Lmoments	-0.2494237	-0.0514771	0.0953777

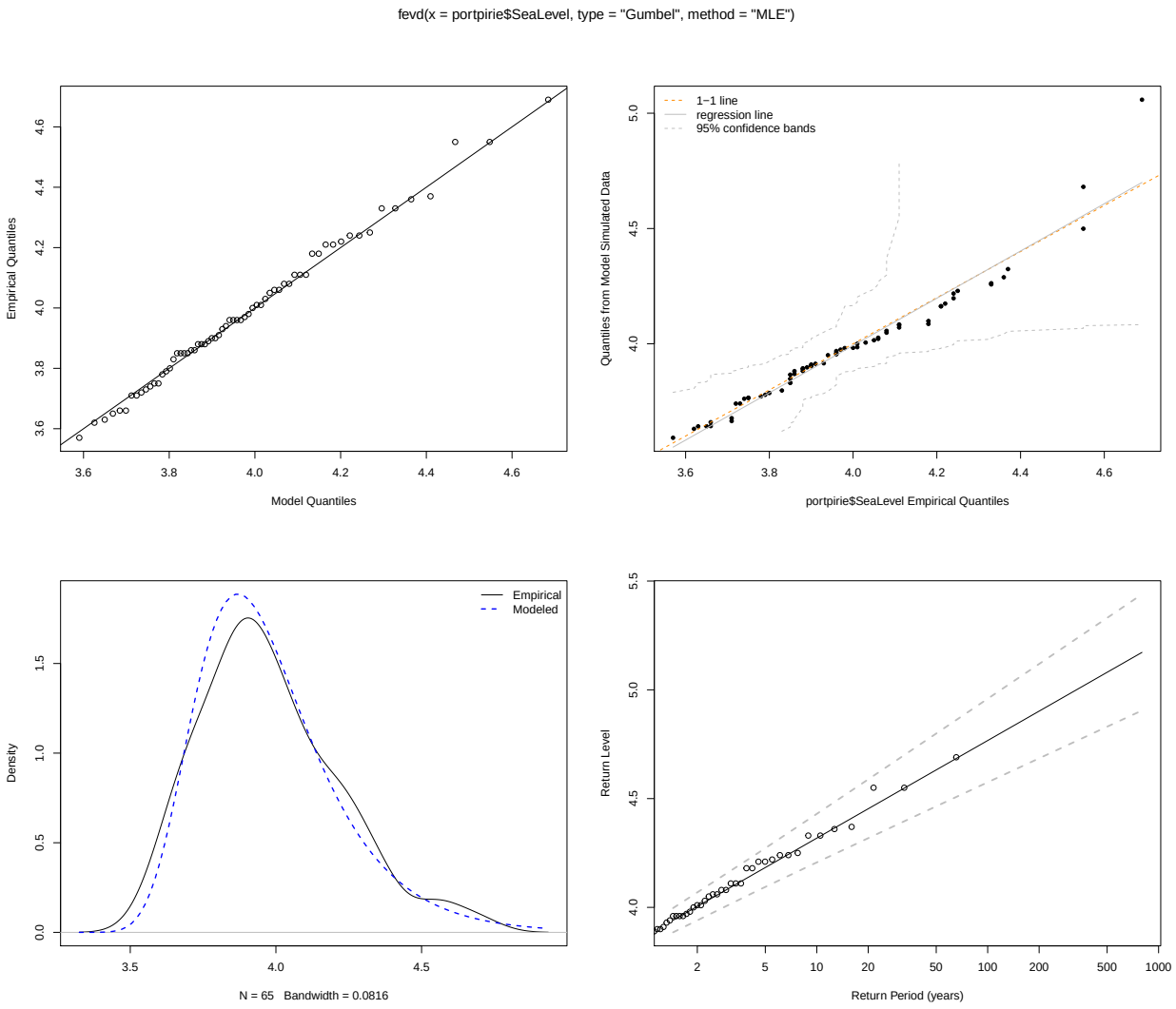
Ci-dessous, nous affichons les QQ-plots et les niveaux de retour liés à la méthode MLE (graphes quasi-identiques pour la méthode des moments). On observe une bonne adéquation du modèle aux maxima ayant un niveau de retour inférieur à 20 ans. Au-delà, en revanche, le modèle commence à s'écarter des observations.



Entre l'incertitude sur le signe du paramètre de forme, et l'écart du modèle GEV aux observations extrêmes, on se dirige alors vers une modélisation de type **Gumbel** (shape = 0). On trouve des estimations des paramètres **location** et **scale** très proches de celles trouvées lors des modélisations précédentes.

Type	Method	Location	Scale	Shape
GEV	MLE	3.874	0.198	-0.050
GEV	Lmoments	3.873	0.203	-0.051
Gumbel	MLE	3.869	0.194	0

Concernant, l'adéquation du modèle aux valeurs extrêmes, cette modélisation est en revanche beaucoup pertinente, comme le montrent les QQ-plots et les niveaux de retour ci-dessous. Les intervalles de confiance sont d'ailleurs nettement plus étroits qu'avant.



2.3. Estimations des niveaux de retour extrêmes

Niveau de retour à 100 ans

La modélisation de Gumble nous donne **4.766 m**. C'est environ 0.070 m de plus que les 2 autres modélisations. Pour autant, l'intervalle de confiance est plus fin, et la valeur maximale à 95% est plus faible.

Type	Method	100-year return level	95% Confidence Interval
GEV	MLE	4.688	(4.3771, 4.9997)
GEV	Lmoments	4.706	(4.4468, 5.0317)
Gumbel	MLE	4.766	(4.5742, 4.9578)

Niveau de retour à 1000 ans

La modélisation de Gumble nous donne **5.216 m**. C'est environ 0.200 m de plus que les 2 autres modélisations. Pour autant, l'intervalle de confiance est plus fin, et la valeur maximale à 95% est plus faible.

Type	Method	1000-year return level	95% Confidence Interval
GEV	MLE	5.031	(4.3765, 5.6857)
GEV	Lmoments	5.055	(4.5784, 5.8023)
Gumbel	MLE	5.216	(4.9404, 5.4908)

Vu le prix d'une digue anti-inondation, il est important d'avoir l'estimation la plus fiable possible de ces niveaux de retour, et la modélisation de Gumbel apparait clairement comme étant la plus pertinente.

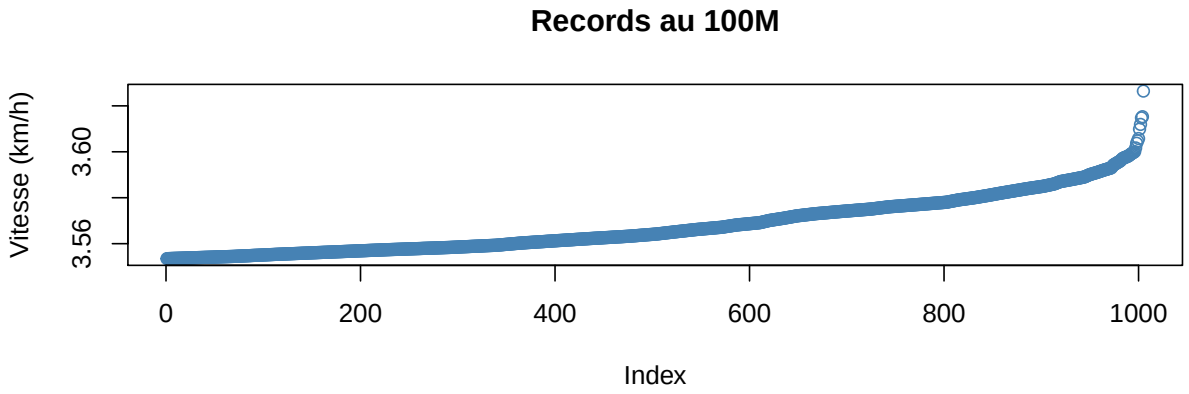
3. Etude du jeu de données “temps100m”

Ce jeu de données présente les records personnels des meilleurs athètes mondiaux sur le 100m. Il ont été mesurés dans les compétitions officielles entre janvier 1991 et avril 2017. A partir de ces données, nous souhaitons estimer s'il existe un record absolu, et quelle est sa valeur.

3.1. Exploration des données et choix du modèle : GPD

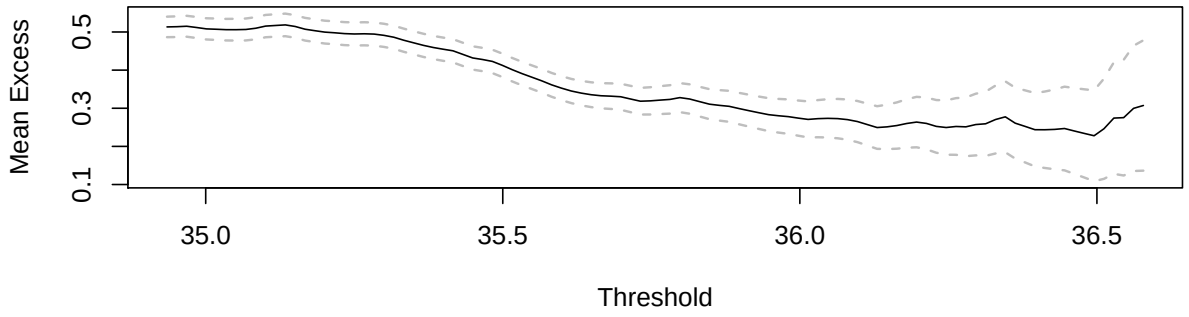
En premier lieu, nous chargeons les données. Celles-ci se présentent sous la forme d'une liste de valeurs non labellées en temps. Comme nous étudions ces données dans le cadre des valeurs extrêmes, nous devons convertir les meilleures performances en grandes valeurs. C'est pourquoi nous convertissons les temps au 100m en vitesses moyenne sur la course (en km/h). L'affichage de ces valeurs montre en particulier quelques outliers qui se démarquent de la tendance globale des records personnels.

Pour cette étude, sachant que l'ont a accès à l'ensemble des mesures, et qu'en plus elles ne sont pas labellées en date de mesure, nous optons pour une approche GPD.

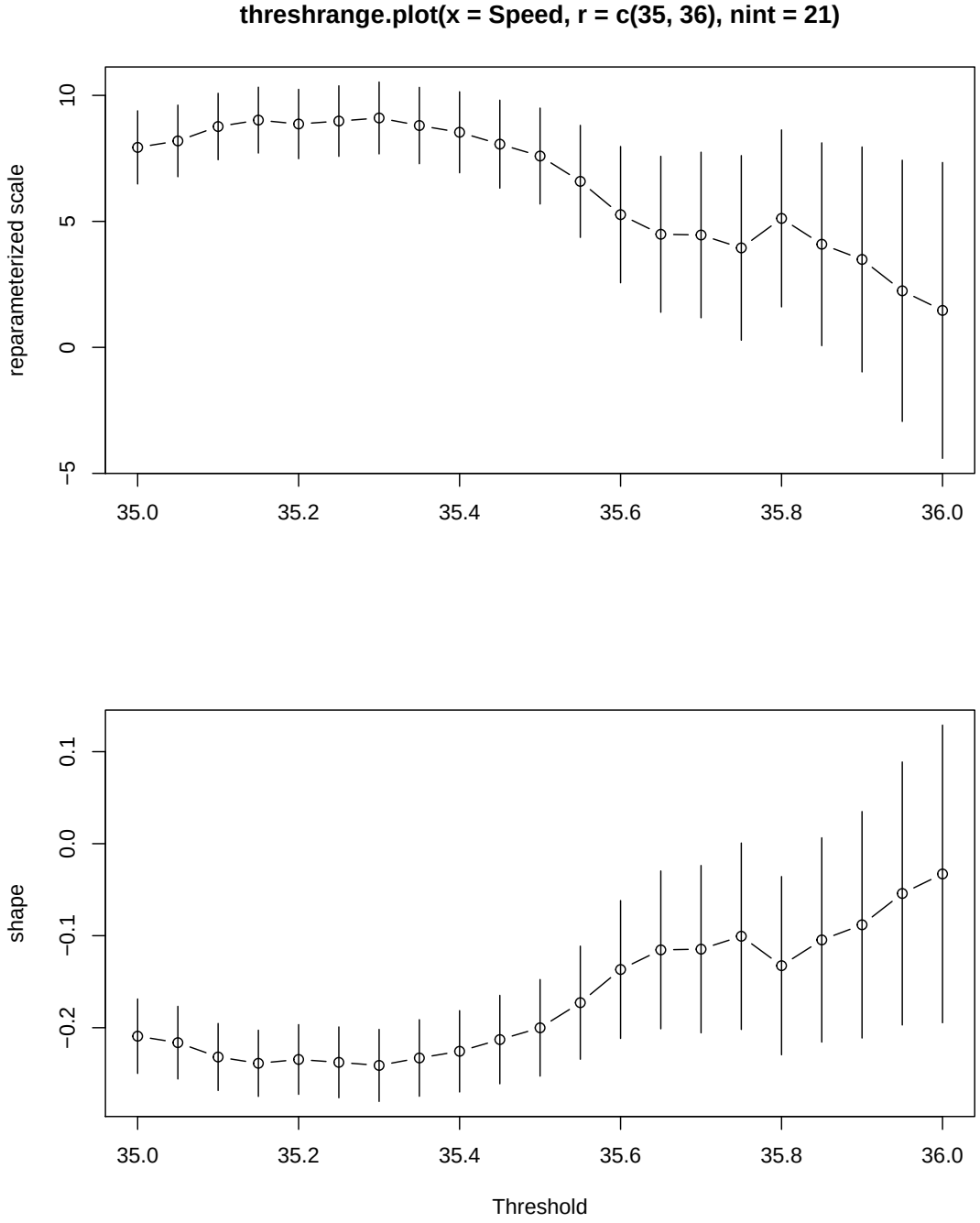


3.2. Choix du seuil - 35.2 km/h - MDA = “Weibull”

En premier lieu, nous observons le **Mean Residual Life Plot**. On devine une zone linéaire à partir de 35.3 km/h environ. La pente est clairement négative, ce qui nous permet d'intuiter que le paramètre de forme est négatif, et que nous sommes dans le domaine d'attration de **Weibull**. Cela va dans le sens de notre étude, puisque le domaine d'attraction de Weibull est celui qui propose une valeur extrême absolue dans sa distribution.



Une estimation des paramètres en fonction du seuil nous permet d'affiner sa valeur (cf. figure ci-dessous). Les paramètres ont une dérivée quasi-nulle autour de la valeur **35.2 km/h**, que nous retenons pour la suite. D'autre part, la valeur du paramètre de forme est nettement négative dans toute cette zone de stabilité, ce qui confirme une nouvelle fois le domaine d'attration de **Weibull**.



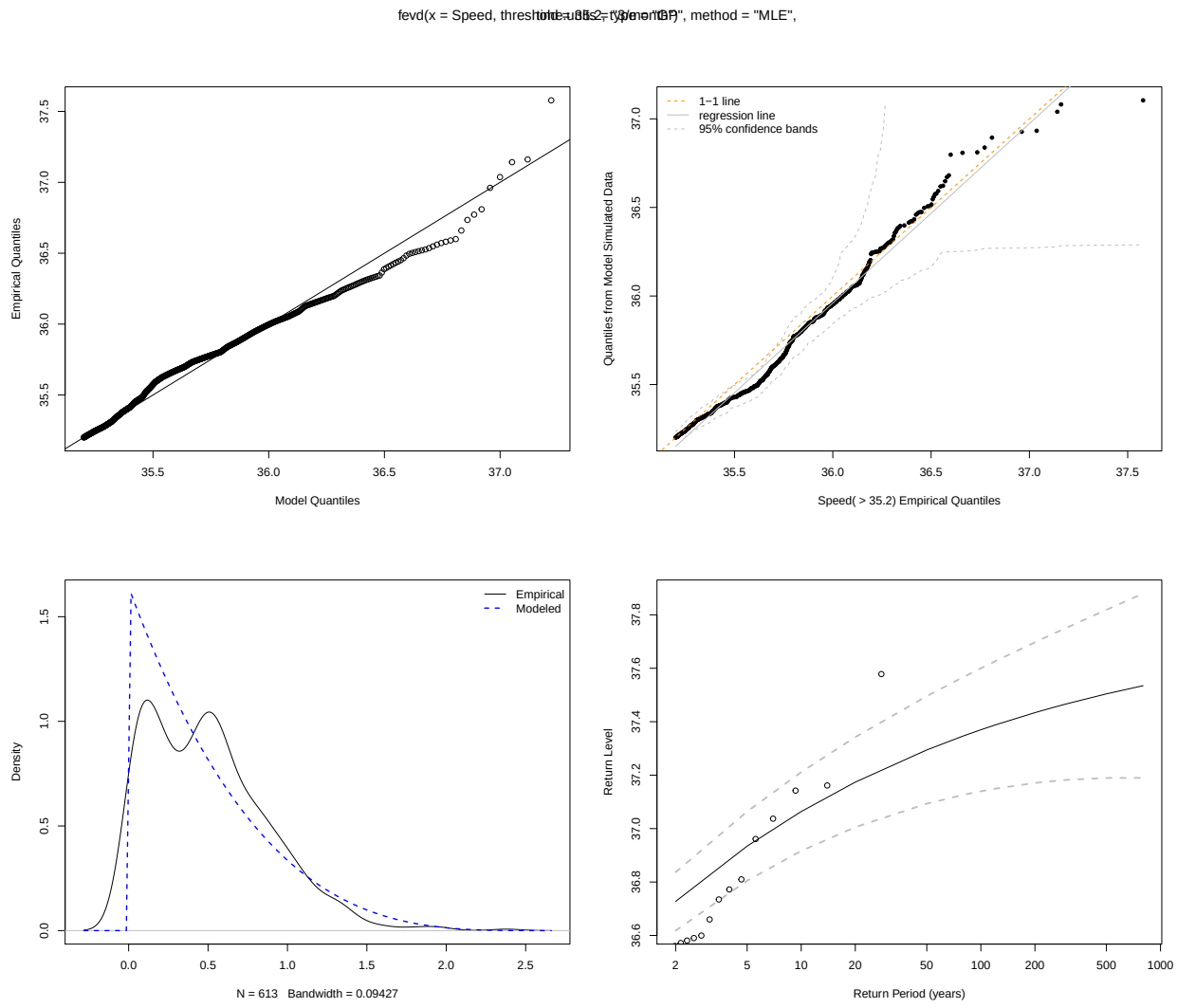
3.3. Estimation des paramètres

Pour l'estimation des paramètres, on cherche un modèle GPD avec la méthode des moments et la méthode MLE. Dans les 2 cas, le paramètre de forme se trouve dans un intervalle de confiance purement négatif, ce qui confirme encore une fois le domaine d'attraction de Weibull. En revanche, on trouve des estimations très différentes des paramètres.

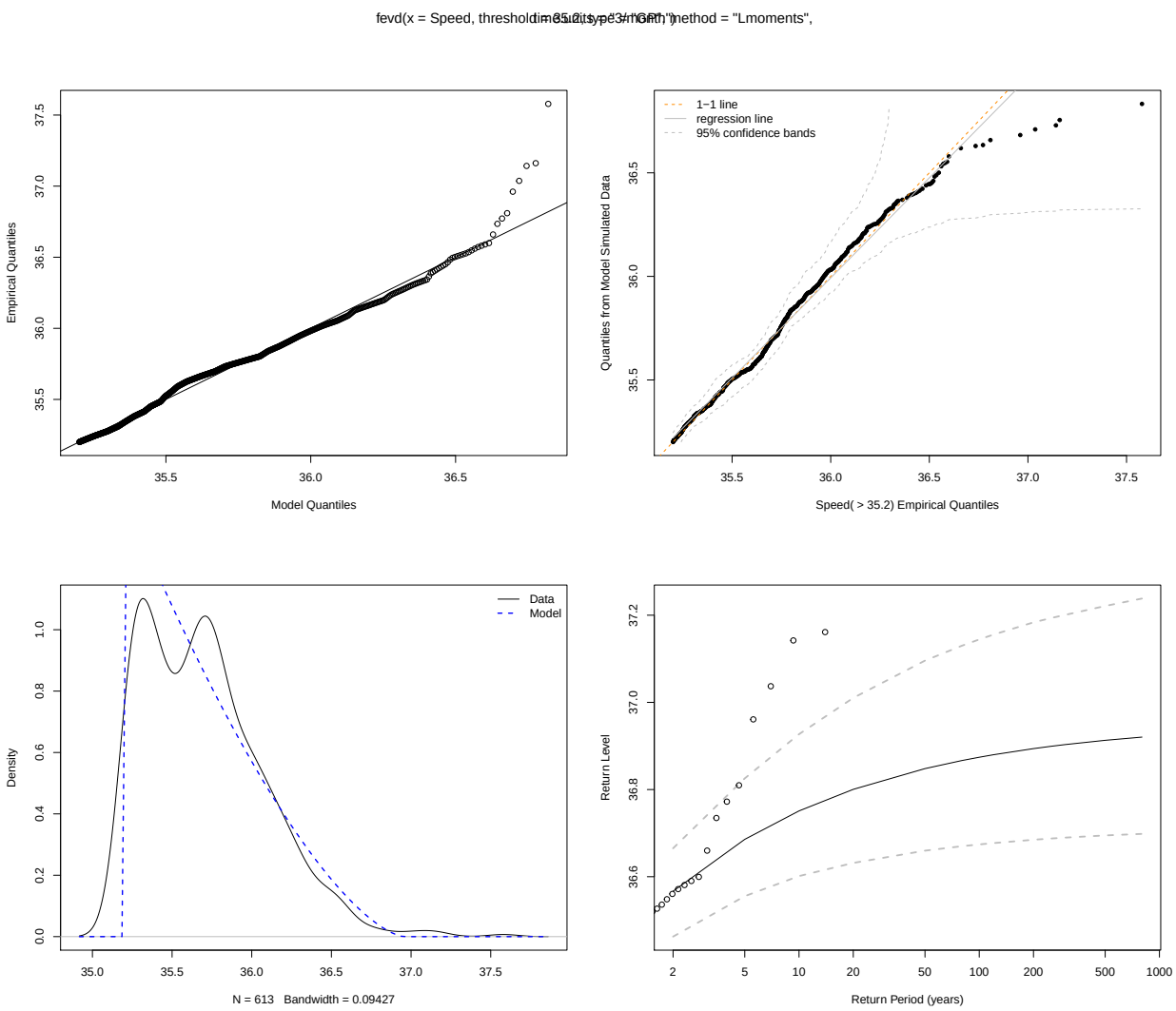
Type	Method	scale	shape	shape 95% Confidence Interval
GPD	MLE	0.609	-0.234	(-0.2724 , -0.1966)
GPD	Lmoments	0.699	-0.398	(-0.5166 , -0.2877)

Pour mieux comprendre, on affiche les QQ-plots et les niveaux de retour liés à chaque méthode.

```
plot(fevd(Speed, threshold = 35.2, type="GP", method="MLE" ,time.units = "3/month"))
```



```
plot(fevd(Speed, threshold = 35.2, type="GP", method="Lmoments" ,time.units = "3/month"))
```



La méthode MLE donne une bonne adéquation du modèle aux valeurs extrêmes (sauf le record du monde de Usain Bolt), mais n'est pas aligné avec la majorité des observations. La méthode des moments est aligné avec la majorité des observations, mais passe complètement à côté des observations extrêmes. **Le modèle trouvé avec la méthode MLE me semble malgré tout plus pertinent, au vu de notre objectif.**

3.4. Valeur extrême et niveau de retour

On peut estimer les records absolus avec la formule $\frac{360}{threshold - \frac{scale}{shape}}$. On voit alors que la méthode des moments propose des valeurs nettement supérieures au record du monde actuel, ce qui la remet fortement en cause. Concernant la méthode MLE, elle semble estimer un record absolu à 2 centièmes de moins que celui d'Usain Bolt seulement, ce qui semble peu.

Type	Method	100m duration
GPD	MLE	9.524 secondes
GPD	Lmoments	9.741 secondes

Niveau de retour à 100 ans

Type	Method	100-year return level	95% Confidence Interval
GEV	MLE	9.633 secondes	(9.6931, 9.5745)
GEV	Lmoments	9.763 secondes	(9.8196, 9.6958)

Niveau de retour à 1000 ans

Type	Method	1000-year return level	95% Confidence Interval
GEV	MLE	9.588 secondes	(9.6807, 9.4965)
GEV	Lmoments	9.750 secondes	(9.8103, 9.6728)

3.5. Conclusion

Ces niveaux de retour semblent peu crédibles au regard de leur valeurs élevées par rapport aux performances enregistrées les 10 dernières années. Ces résultats invitent d'autant plus au sceptiscisme, que les QQ-plot et les courbes de niveaux de retour n'étaient pas complètement en adéquation avec les données observées.

Cela m'incite à remettre en cause le respect par ces données des conditions d'application de la théorie des valeurs extrêmes. En particulier, la distribution IID des variables mesurées entre janvier 1991 et avril 2017.

En effet, les records sont publics et sont fait pour être battus. Cela infuence les efforts réalisés dans les entraînement des athlètes, si bien que la distribution des records personnels tend certainement à diminuer au cours du temps, et donc les observations ne sont plus IID. D'autre part, les différents athlètes viennent de pays différents avec des conditions différentes de relief et de climat, qui ont permis le développement d'aptitudes physiques naturelles différentes d'un pays a l'autre. Ce deuxième point remet également en question la nature IID de nos observations.