Faculté des Sciences et Technologies Département de Mathématiques

---

Master :Ingénierie, Statistique et Numérique - Data Sciences

Statistiques Extrêmes

Étude des minimax de températures pour la ville de Berlin

Réalisé par :

VERZURA Alexandre et ALASSANE ISSAKA Omar

Année universitaire : 2023-2024

Table des matières

Introduction

Les statistiques extrêmes constituent une branche des mathématiques visant à modéliser les comportements extrêmes ou rares d'un ensemble de données, dans le but d'estimer les probabilités d'événements rares situés dans la queue de la distribution.

Dans ce contexte, notre étude se focalise sur la distribution des températures maximales dans la ville de Berlin. Pour ce faire, nous allons représenter le comportement de ces maximums en utilisant la modélisation des distributions des valeurs extrêmes généralisées (GEV et GPD).

Ce rapport a pour objectif d'estimer la probabilité qu'une certaine température soit atteinte au moins une fois sur un large intervalle de temps. Il s'agit ainsi d'estimer, à un niveau de confiance de $\alpha\%$, la température maximale minimale atteinte sur un intervalle de plusieurs années.

Cette étude repose sur une base de données très fournie. En effet, le modèle prend en compte les températures quotidiennes enregistrées pour la ville de Berlin du 1er janvier 1876 au 31 décembre 2023, exprimées en degrés Celsius. Toutefois, pour des raisons historiques, les mesures météorologiques de l'année 1945 sont partiellement manquantes.

Le rapport se divise en trois parties. La première aborde les techniques de nettoyage et de sélection de données utilisées pour préparer les deux modèles. Ensuite, la deuxième partie présente la modélisation GEV de notre problème, aussi bien dans sa forme que dans sa distribution. Enfin, la troisième partie se concentre sur le modèle GPD, visant à encadrer les températures maximales atteintes pour la ville de Berlin en fonction des intervalles de temps considérés.

I Préparation des données

L'étude des températures maximales (ou minimales) de la ville de Berlin s'appuie donc sur la base de données explicitée dans l'introduction. On rappelle qu'elle représente les températures moyennes journalières dans la ville de Berlin entre les années 1876 et 2023. Elle comporte 4 colonnes : les trois premières sont les jours, mois et années de la mesure de température, tandis que la quatrième est la température maximale atteinte sur la journée.

Afin de modéliser les comportements "extrêmes" ou "rares" de la variable température au cours du temps, il est maladroit de supposer l'indépendance jour par jour de la température. On rappelle également que cette stationnarité se dédouane du pas de temps considéré, ainsi la stationnarité est conservée que l'on prenne la température maximale atteinte par jour ou la température maximale atteinte par mois (ou autres intervalles de temps) comme variable aléatoire.

La figure suivante représente l'ensemble de nos données de températures. Ce graphique semble représenter des variables stationnaires.

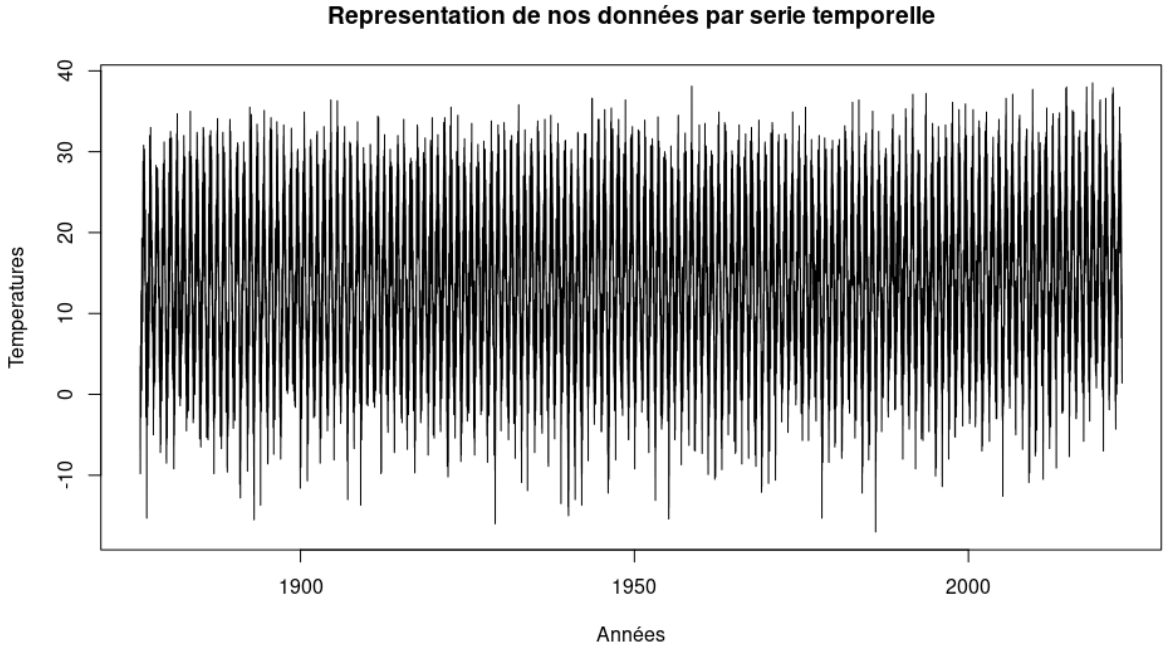

Figure 1 – Tracer de la série des températures

On décide donc, dans ce document, de construire un modèle GEV sur les températures maximales atteintes par mois dans la ville de Berlin. Malgré que nos données soient

incomplètes pour l'année 1945, ce qui pose un problème en termes d'hypothèse de stationnarité, nous souhaitons les conserver tout de même.

En effet, la distribution des températures peut admettre une tendance. Cependant, on suppose que cette tendance est très faible, voire inexistante. Un test de Student pour comparer les moyennes de températures sur différentes plages de temps entre les années 1944 et 1946 met en lumière que la tendance n'a pas d'impact significatif, du moins sur une période courte (1 an) (la p-valeur du test de Student est de 0,6804 $>$ 0,05). Nous sommes donc très confiants pour conclure que l'hypothèse nulle est valable, c'est-à-dire que les moyennes ne sont pas significativement différentes.

Cela nous conduit à réaliser deux tests de stationnarité sur nos données (Augmented Dickey-Fuller et Test de Racine Unitaire). Les p-valeurs de ces deux tests valident nos hypothèses respectivement $\leq 10^{16}$ et $= 0.01$).

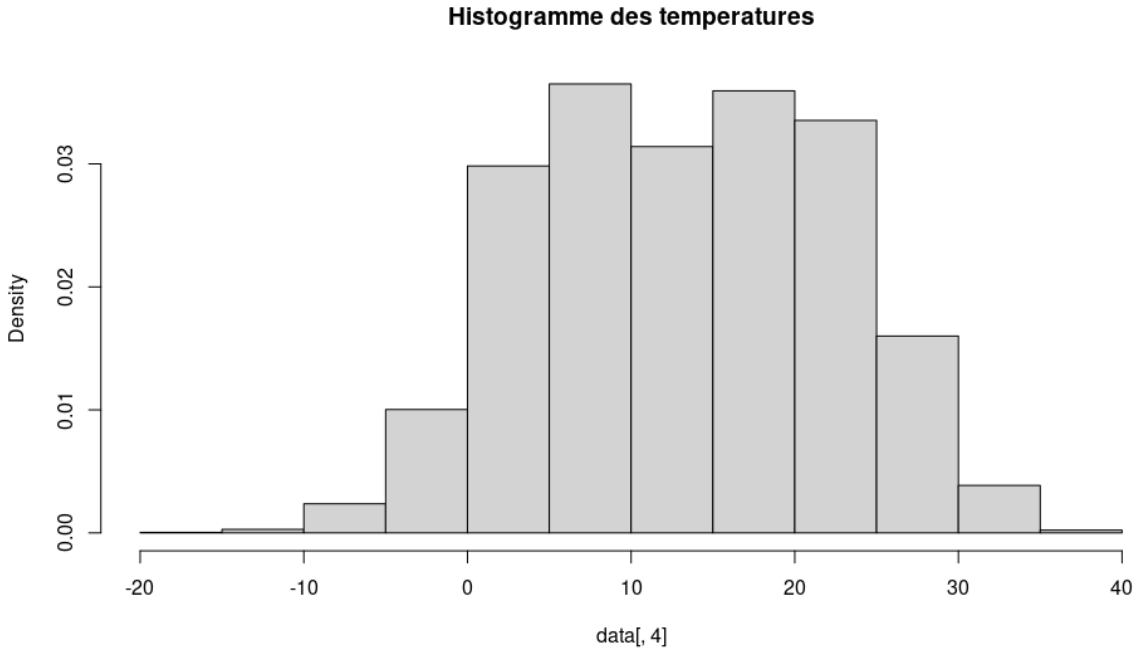

Figure 2 – Tracer du histogramme des temperatures

Après avoir vérifié les différentes hypothèses sur nos données, on rappelle que l'étude porte sur la prédiction des valeurs présentes dans les queues de distribution (température supérieure à 35°C et température inférieure à -10°C, représentées à la figure 2). Les parties précédentes explicitent les deux modèles utilisés, le modèle GEV et le modèle GPD.

II Présentation du modèle GEV

Nous avons vu précédemment que les données de températures sont hypothétiquement stationnaires. Ainsi, sous cette hypothèse, et avec une quantité de données suffisantes, on réduit les données à leur maximum par année. Ces maximums sont donc indépendants, on peut appliquer le modèle GEV. On représente dans la figure ci-dessous les maximums de température par an. Voici le graphe des maximums retenus par année :

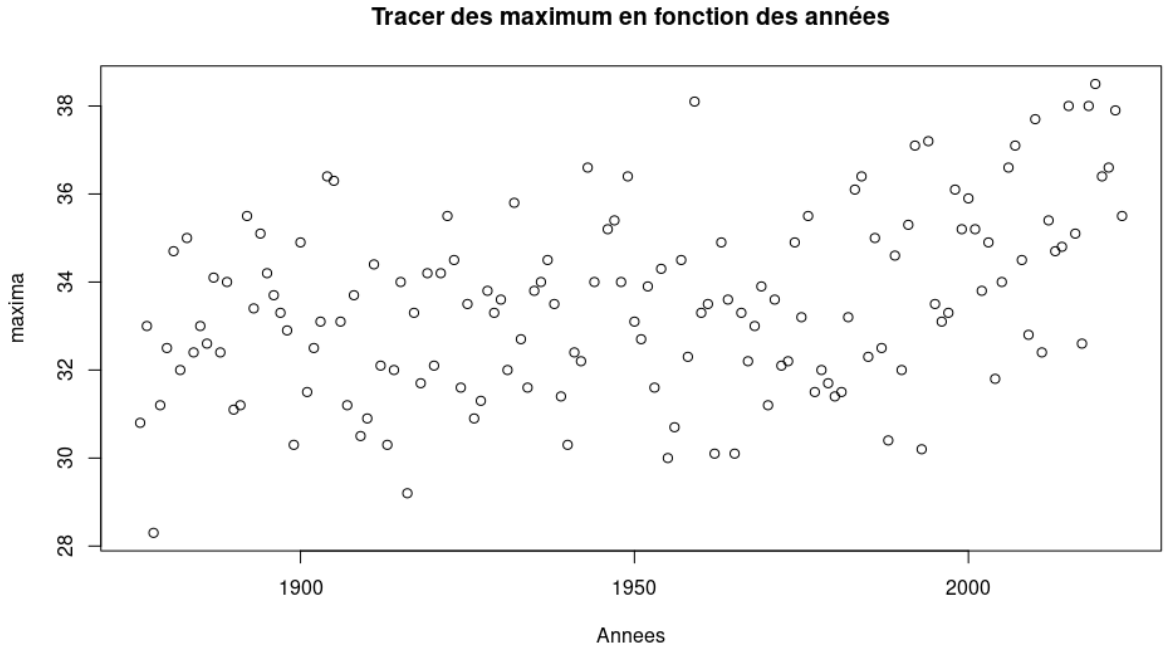

Figure 3 – Tracé des maximums sur chaque année en fonction des années

II.1 Estimation des paramètres de notre modèle

Les sorties de codes pour l'estimation des paramètres sont incluses dans l'annexe du document. On précise seulement dans cette partie que la forme ("shape" dans l'annexe) est négative. Notre modèle suit donc une loi de Weibull.

Traçons les profils de vraisemblance pour nos modèles.

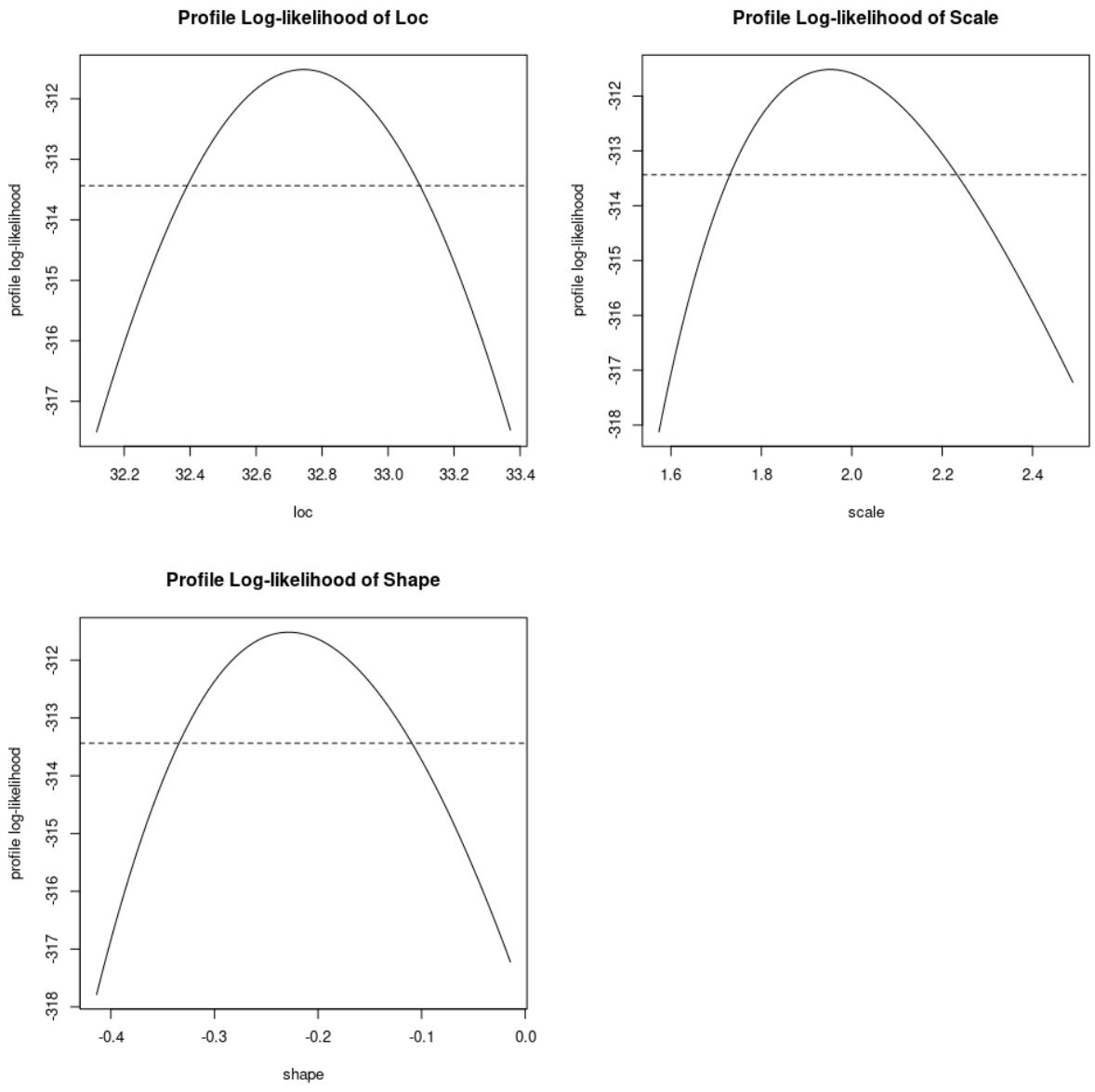

Figure 4 – Tracé des profils de vraisemblance

L'intervalle formé par l'intersection de la ligne en pointillé et la courbe forme l'intervalle de confiance de notre paramètre.

On voit que la distribution empirique et de la distribution estimée collent parfaitement.

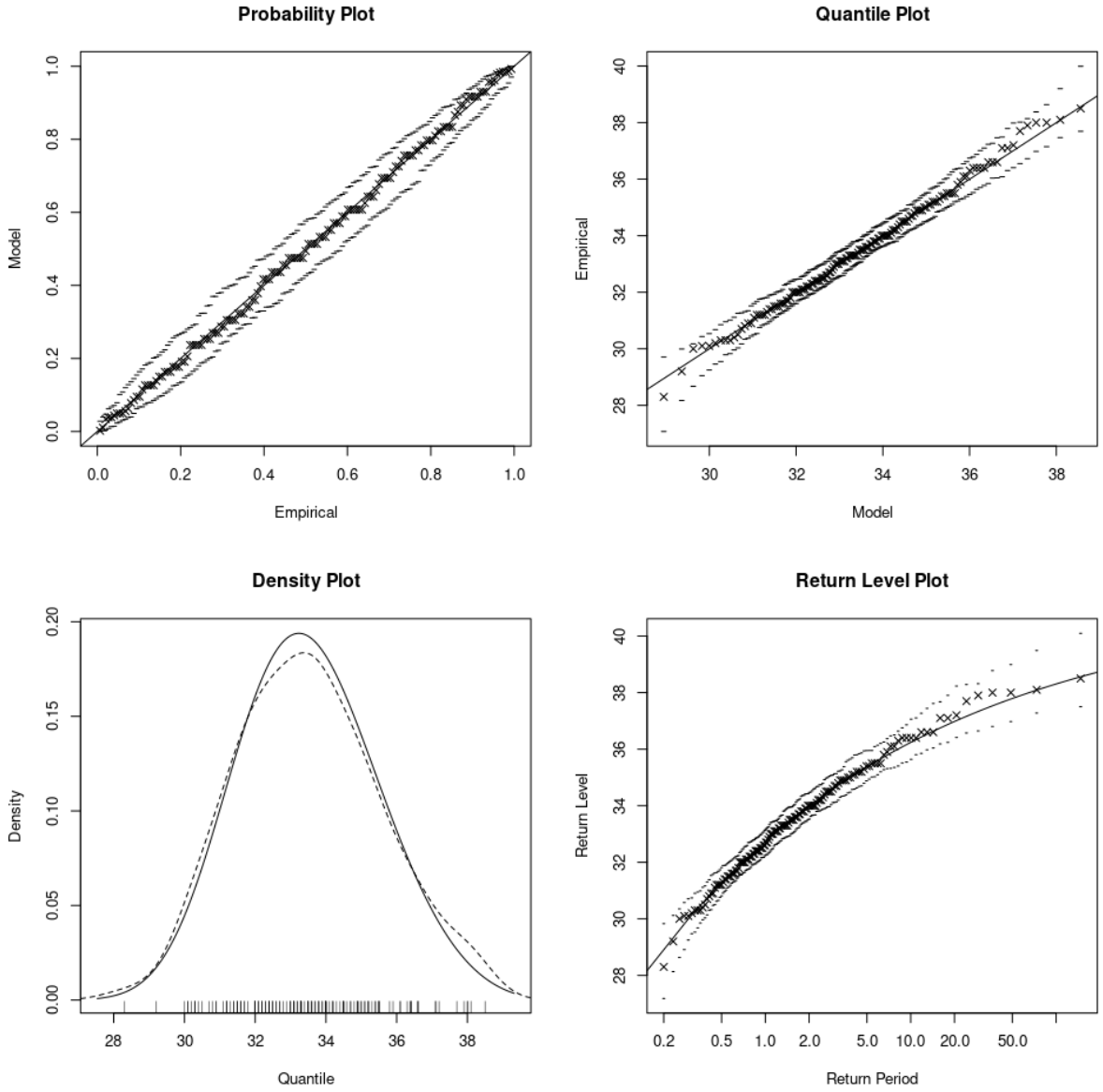

Figure 5 – Validation de notre modèle GEV

II.2 Niveau de retour

Voici les niveau de retour que nous avons calculé :

Sur 10 ans :
niveau de retour = 36.17°C avec un intervalle de confiance de [35.23 ; 37.12] (en °C).

Sur 100 ans :
niveau de retour = 38.30°C avec un intervalle de confiance de [36.95 ; 39.65] (en °C).

III Présentation du modèle GPD

Comme indiqué dans la partie 1, nos données sont stationnaires mais présentent une dépendance temporelle. Afin d'appliquer la méthode GPD, nous devons procéder à une déclusterisation. Ensuite, nous supposerons que les maxima par cluster sont indépendants avant d'appliquer notre modèle GPD.

Nous allons examiner les températures maximales. Tout d'abord, nous tenterons de représenter nos données en fixant un seuil u pour avoir une idée approximative de son emplacement.

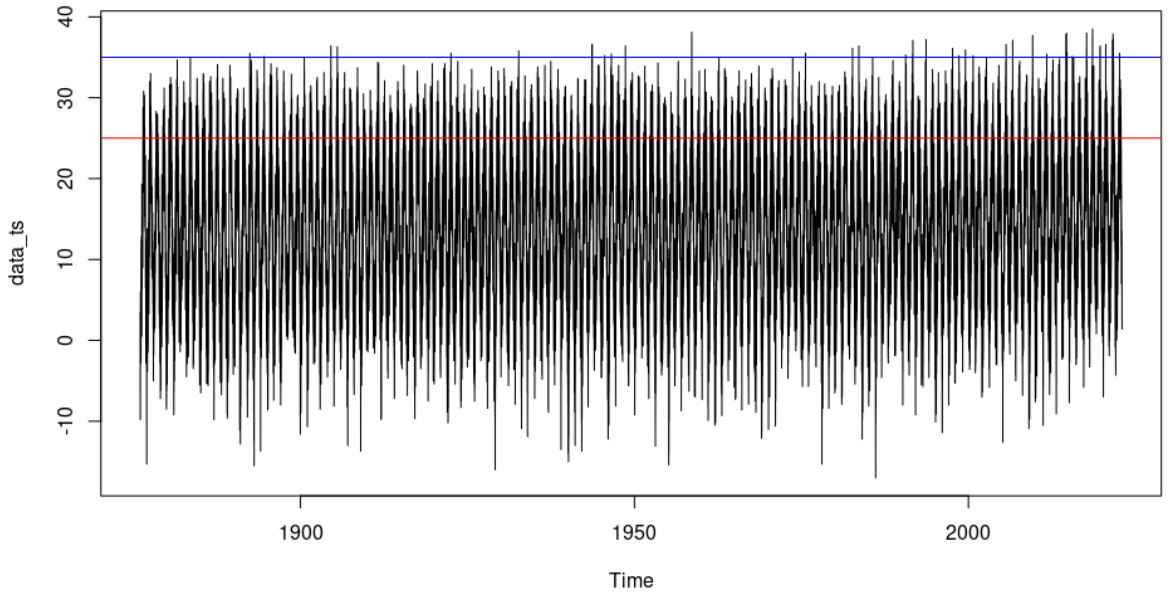

Figure 6 – Tracer des seuils de u =25°C en rouge et u=35°C en bleu

En effet, la température maximale sur nos données est de 38 °C.

Afin de déterminer un seuil approprié pour notre modèle GPD, nous avons choisi une plage de seuils entre 25 et 35. Nous examinerons comment notre variance et notre

paramètre de forme évoluent en fonction du seuil u, afin d'évaluer la stabilité de nos estimateurs en fonction de u.

Voici les résultats que nous avons obtenus :

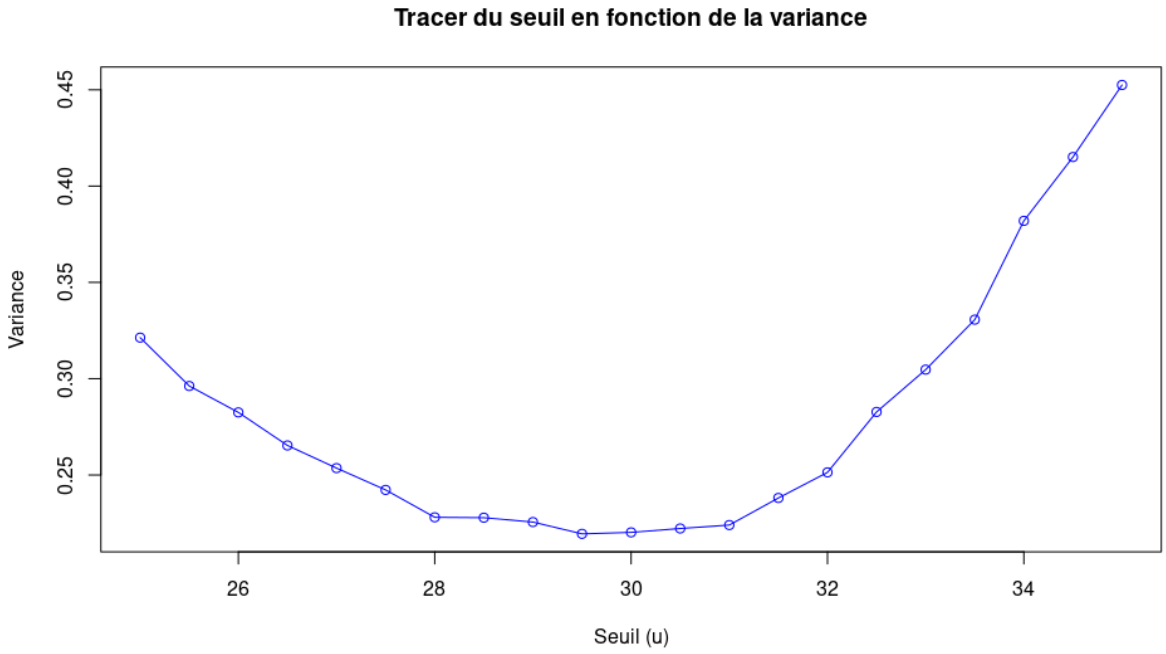

Figure 7 – représentation de la variance en fonction du seuil (u)

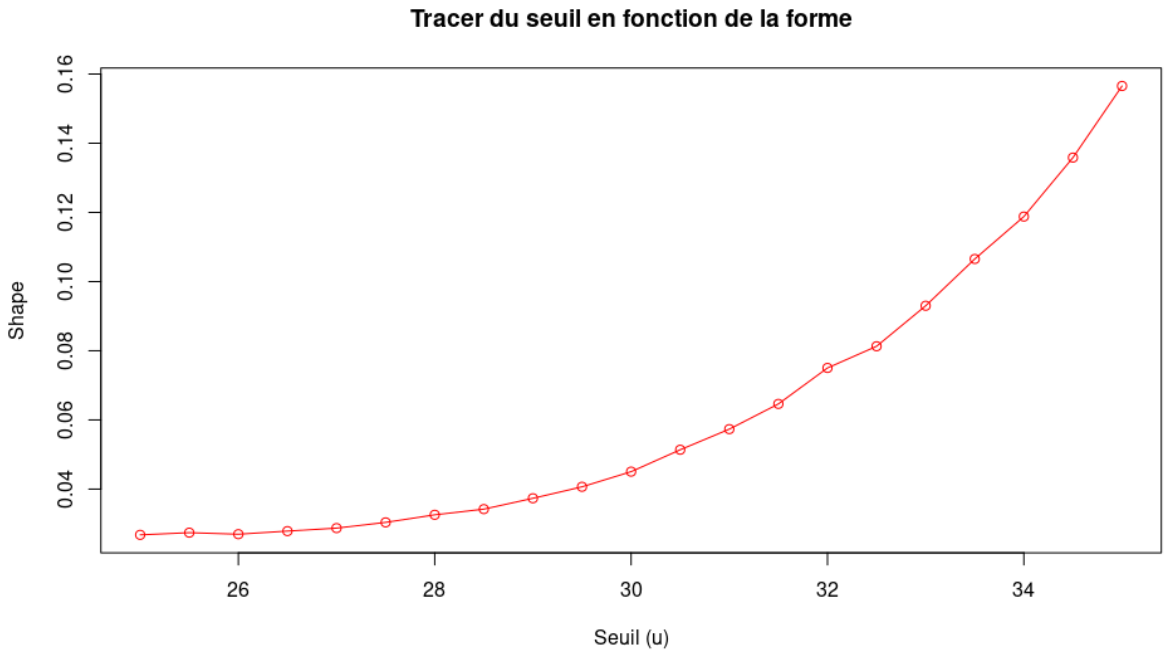

FIGURE 8 – Représentation de shape en fonction du seuil (u)

En réalité, uu doit être suffisamment grand pour que l'approximation définie par la distribution généralisée de Pareto (GPD) soit valide, mais pas excessivement élevé afin de conserver un nombre suffisant de dépassements pour estimer les paramètres du modèle. Le choix du seuil nécessite donc de trouver un compromis traditionnel en statistiques entre le biais et la variance.

Dans notre analyse, nous constatons que 30 semble être un bon candidat pour uu. Au cours de nos itérations pour tracer ces graphes, nous avons choisi une taille de cluster de 10 (r = 10).

Étant donné que le choix de la taille du cluster influence également le choix du seuil, nous allons comparer le modèle avec un seuil plus élevé mais avec des tailles de cluster plus petites ($u_2 = 34$ et $r_2 = 2$), sachant que plus u est élevé, moins nous avons d'éléments qui dépassent le seuil.

III.1 Représentation des clusters

Etant donnée que les données sont stationnaires, si la température le jour j dépasse le seuil que l'on a fixé, les températures des jours j+1 et j-1 ont naturellement plus de chance

d'être aussi supérieur au seuil. Ceci est en désaccord avec notre hypothése d'indépendance entre les maximums de températures.

Afin de récupérer notre hypothèse, on décide de créer des clusters de températures, leur taille seront explicité dans la partie III-2.

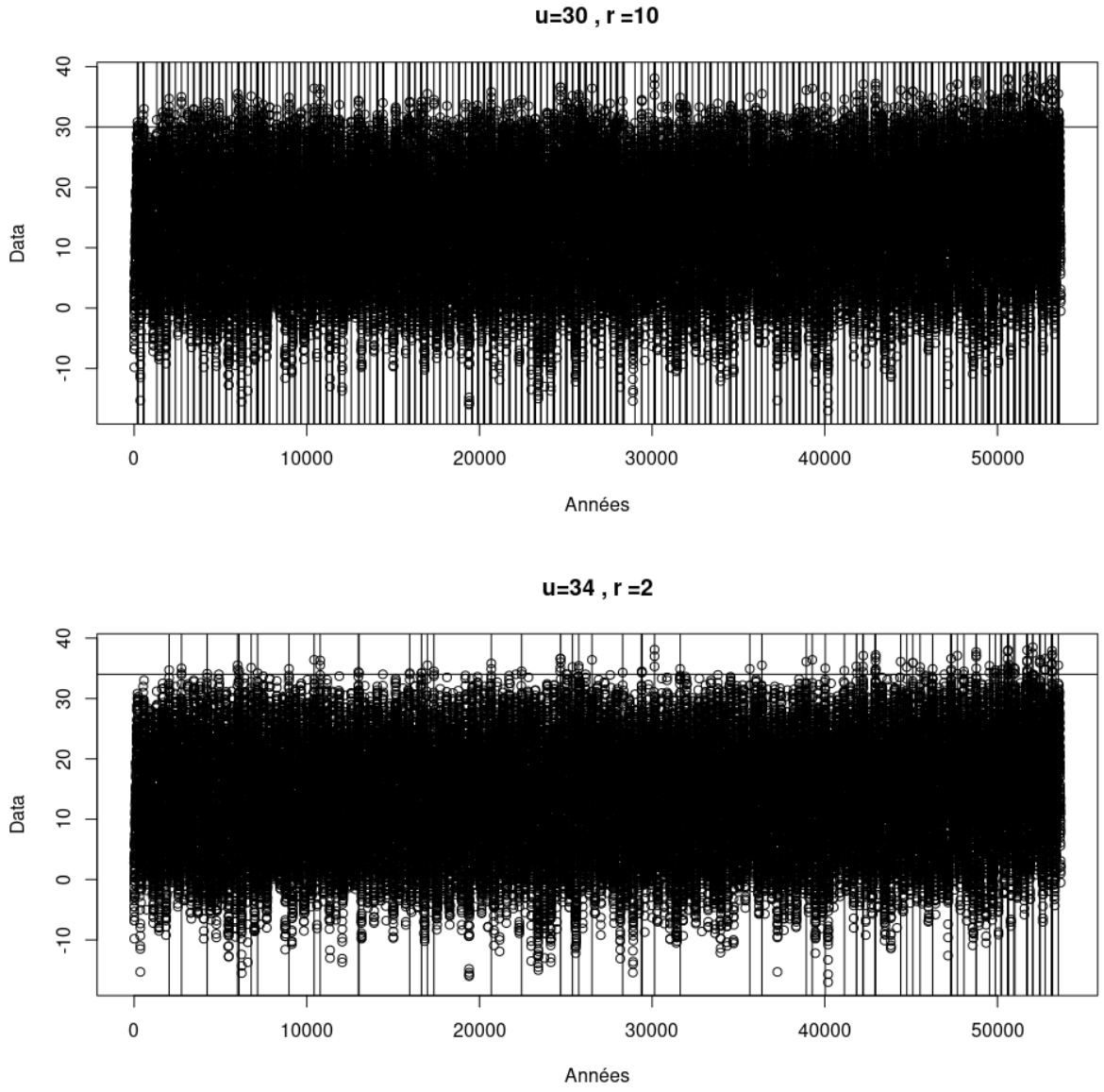

Figure 9 – Représentation des clusters

III.2 Estimation de nos paramètres

Les sorties de codes pour l'estimation des paramètres sont inclus dans l'annexe du document. On précise seulement dans cette partie que la forme ("shape" dans l'annexe) est positive pour le modèle 1, il suit donc une loi de Fréchet. Tandis que la forme est négative dans le modèle 2 (ce qui suppose que cela suit une loi de Weibull.)

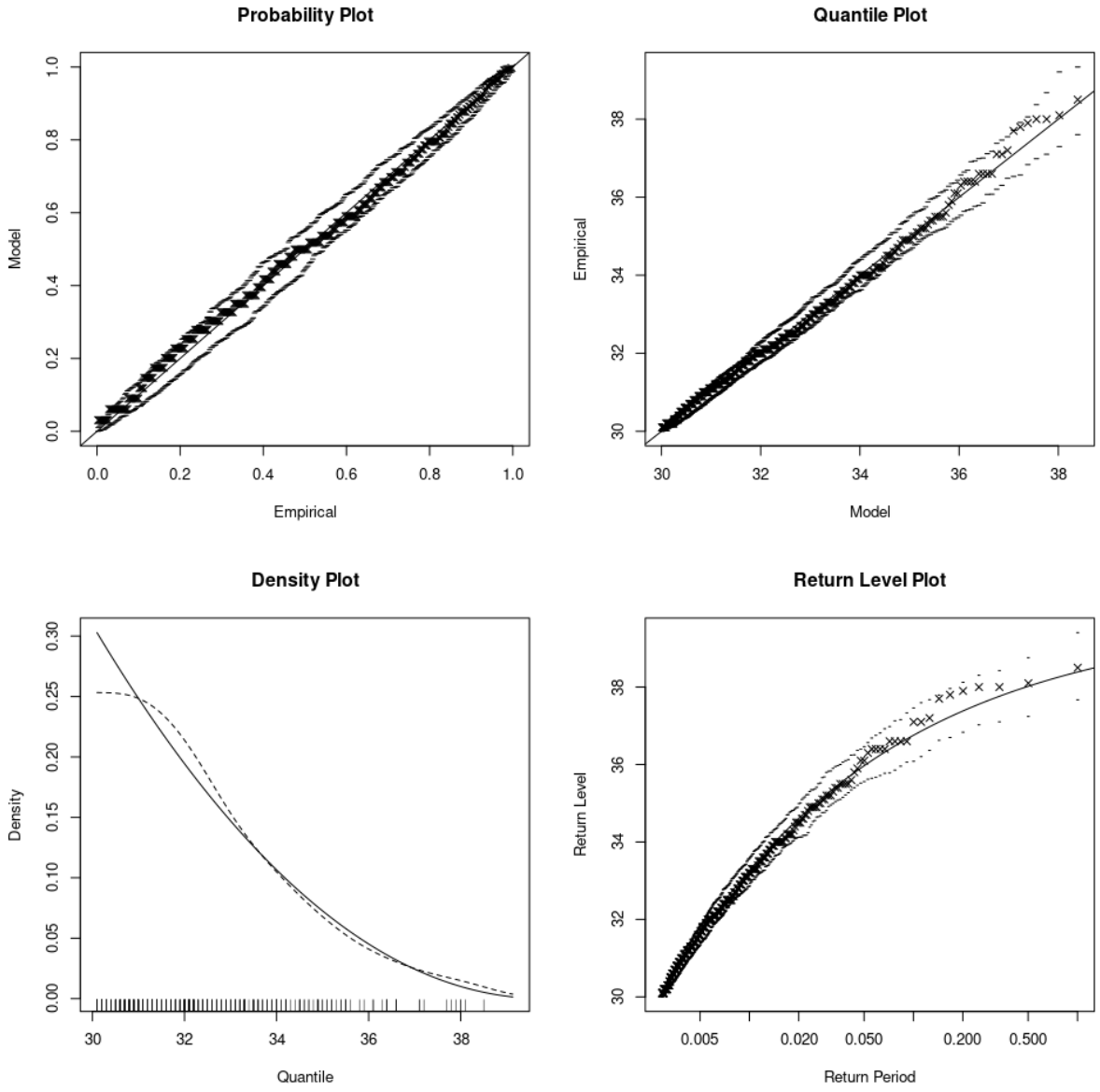

Figure 10 – Validation du modele u_1= 30 et r_1 = 10

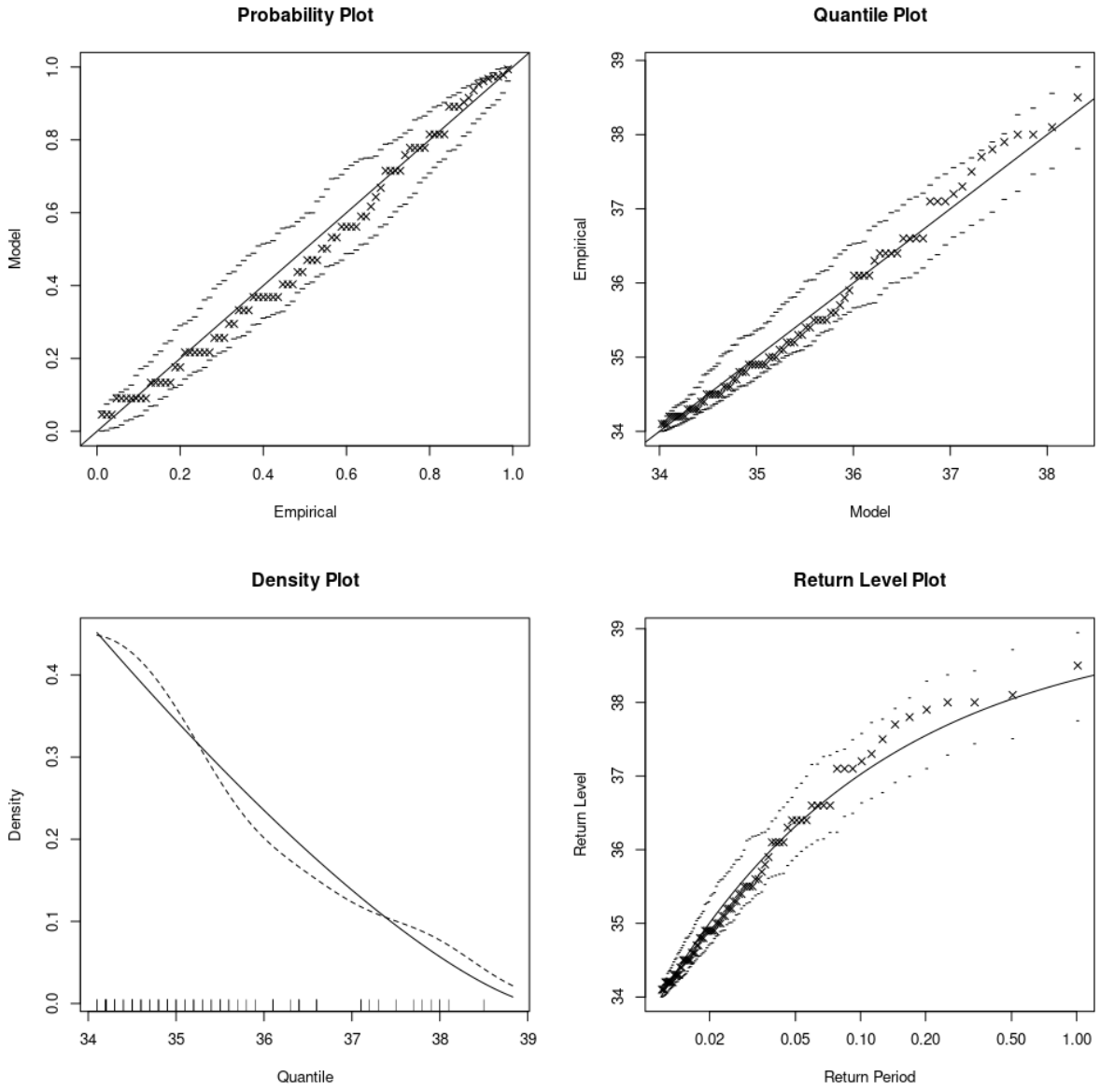

FIGURE 11 – Validation du modele $u_2 = 34$ et $r_2 = 2$

En comparant la figure 10 et la figure 11, il est immédiat que le modèle avec le seuil $u_1 = 30$ et $r_1 = 10$ est plus adapté à nos données. En effet, les courbes de densité et de probabilité se confondent quasiment dans la figure 10 contrairement à la figure 11. On conserve donc le modèle $u_1 = 30$ et $r_1 = 10$.

III.3 Niveau de retour

Nous allons calculer les niveaux de retour pour 10 et 100 ans, comme mentionné précédemment, ainsi que les niveaux de retour sur 100 ans.

Sur 10 ans, le niveau de retour est de 39.15 avec un intervalle de confiance de [37.69, 40.61].

Sur 100 ans, le niveau de retour est de 39.51 avec un intervalle de confiance de [37.91, 41.11].

En effet, notre modèle ne prend pas en compte le réchauffement climatique. L'hypothèse de stationnarité, nécessaire pour notre modèle, suppose l'absence de tendance. On en déduit qu'il est plus raisonnable de calculer un niveau de retour sur une courte période, donc sur 10 ans. C'est d'ailleurs ce que l'on peut remarquer dans nos résultats. La différence entre le niveau de retour sur 10 ans et le niveau de retour sur 100 ans est peu "significative".

IV Modèle avec les températures minimales

Nous avons également modélisé les données avec les températures minimales. La modélisation et les niveaux de retour se trouvent dans le fichier R. Nous avons décidé de ne pas inclure cette partie dans le rapport, car la modélisation est pratiquement la même, afin de ne pas allourdir notre rapport.
Les résultats seront cependant présenté dans le tableau récapitulatif dans la conslusion.

V Conclusion

Pour conclure, nous souhaitons rappeler les différents résultats présentés tout au long du document. On rappelle que le modèle GEV est plus restrictif au niveau des résultats que le modèle GPD. Les résultats sont donc présentés dans le tableau suivant :

Données en °C	Niv retour 10 ans	I.C. 10 ans	Niv retour 100 ans	I.C. 100 ans
GEV (max)	36.17	[35.23 ; 37.12]	38.30	[36.95 ; 39.65]
GPD (max)	39.15	[37.69 ; 40.61]	39.51	[37.91 ; 41.11]
GEV (min)	-12.34	[-13.65 ; -11.04]	-16.47	[-18.39 ; -14.55]
GPD (min)	-19.34	[-16.58 ; -22.10]	-20.65	[-24.23 ; -20.31]

Figure 12 – Tableau récapitulatif des résultats

Annexe

Sortie code GEV

```
> modele = fgev(maxima)
> modele

Call: fgev(x = maxima)
Deviance: 623.0298

Estimates
    loc    scale    shape
32.7437   1.9523  -0.2283

Standard Errors
    loc    scale    shape
0.17920  0.12630  0.05707
```

Sortie de code modèle GPD (1)

Pour le modèle avec u1 = 30 et r1=10 on a :

```
> model_GPD1 = fpot(data[,4],u_1, r = r1, cmax=TRUE)
> model_GPD1

Call: fpot(x = data[, 4], threshold = u_1, cmax = TRUE, r = r1)
Deviance: 1261.785

Threshold: 30
Number Above: 1093
Proportion Above: 0.0204

Clustering Interval: 10
Number of Clusters: 342
Extremal Index: 0.3129

Estimates
  scale    shape
```

```
 3.2335  -0.3289

Standard Errors
  scale    shape
0.22020  0.04505
```

Sortie de code modèle GPD (2)

Pour le modèle $u_2 = 34$ et $r_2 = 2$ on a :

```
> model_GPD2 = fpot(data[,4],u_2, r = r2, cmax=TRUE)
> model_GPD2

Call: fpot(x = data[, 4], threshold = u_2, cmax = TRUE, r = r2)
Deviance: 225.9054

Threshold: 34
Number Above: 120
Proportion Above: 0.0022

Clustering Interval: 2
Number of Clusters: 84
Extremal Index: 0.7

Estimates
  scale    shape
 2.1561  -0.4236

Standard Errors
 scale   shape
0.3240  0.1132
```