unisanté
Centre universitaire
de médecine générale
et santé publique · Lausanne

Extreme value theory in public health

Danyu Li, PhD student

DESS / Division de biostatistique
Unisanté

danyu.li@unisante.ch

Lausanne, mai 2024

Table of contents

01

Introduction

Introduce extreme events and the importance

1. Introduction

- Resources planning is a central concern in public health and it involves anticipating the possibility of **rare** or **extreme events** occurring in the foreseeable future.

- community epidemics
- significant heatwaves
- extreme air pollution periods
- unusually large flooding events
- accidental toxic exposures
- tornados outbreaks
- financial crises
- …

1. Introduction

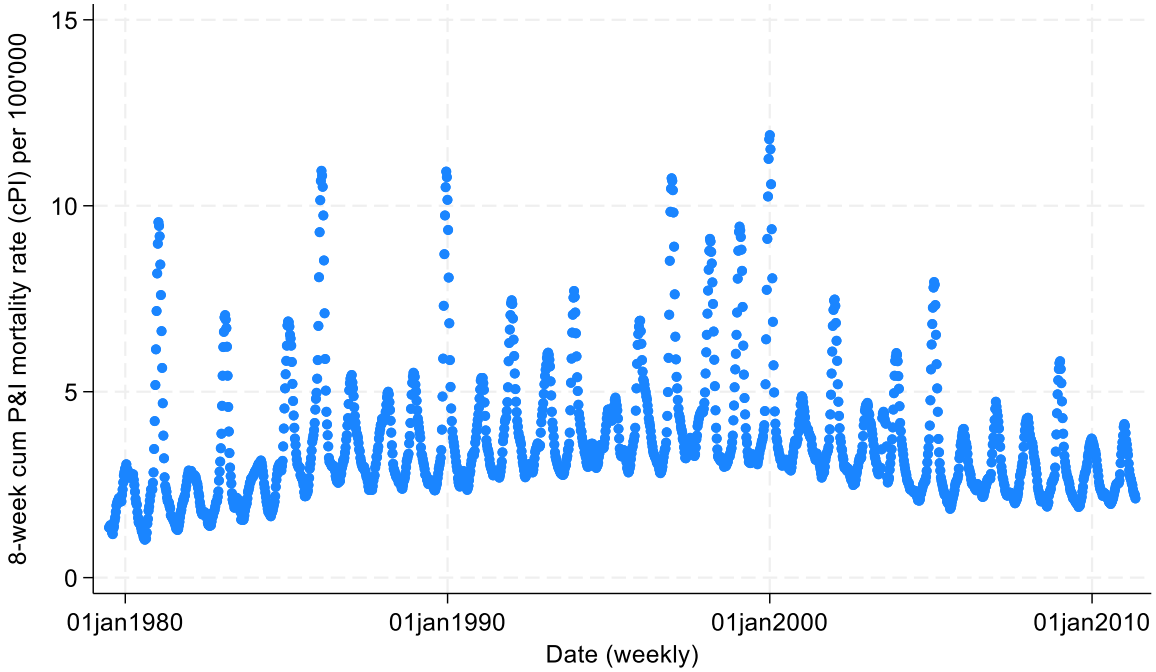
- **Extreme Value Theory (EVT)** is a branch of statistics concerned with the extreme values of probability distributions. It was developed by Emil Julius Gumbel in 1920.

- Hydrology: predict floods
- Oceanography: study rogue waves
- Epidemiology: identify emerging diseases
- Demography: predict the probability distribution of the maximum age that humans will be able to achieve
- Insurance: predict major disasters
- Finance: predict financial crisis
- Climatology: exceedances of heatwave records

1. Introduction

- **Objective of EVT**: is to evaluate the likelihood of events exceeding previously recorded extremes based on a series of observations.
- **Advantage** of using EVT is that: it is possible to predict the occurrence of a still unobserved event which is more extreme than those that have been observed up to now.
- **In this presentation**: we showcase the application of EVT in a public health context. For this, we will demonstrate its utilization in predicting the occurrence of a future extreme pneumonia and influenza mortality episode as large or even larger than what has been observed in the data.

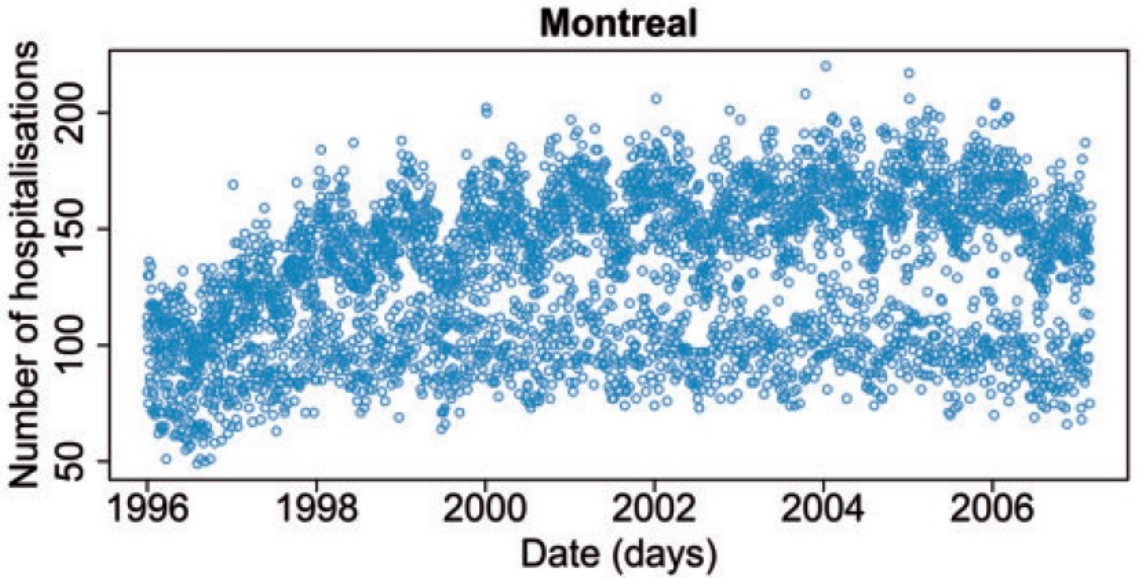
1. Introduction



Source: Thomas et al. Applications of Extreme Value Theory in Public Health. *PLoS ONE* 2016; 11(7):e0159312.

1. Introduction

Another public health data example is the daily number of hospitalizations (due to cardiovascular diseases) in Montreal



Source: Chiu et al. Mortality and morbidity peaks modeling: an extreme value theory approach. Statistical Methods in Medical Research 2018; 27: 1498-1512.

1. Introduction

Why do we need special statistical methods to study extremes?

1. Introduction

- When working with complete datasets and employing **traditional statistical techniques**, the primary **focus** is often on **average occurrences**.

1. Introduction

Why do we need special statistical methods to study extremes?

In summary, while average-based analyses are valuable for understanding typical patterns, they may not capture the full picture when it comes to health outcome peaks. Researchers should explore statistical methods that specifically address extreme events and provide a more comprehensive understanding of healthcare dynamics.

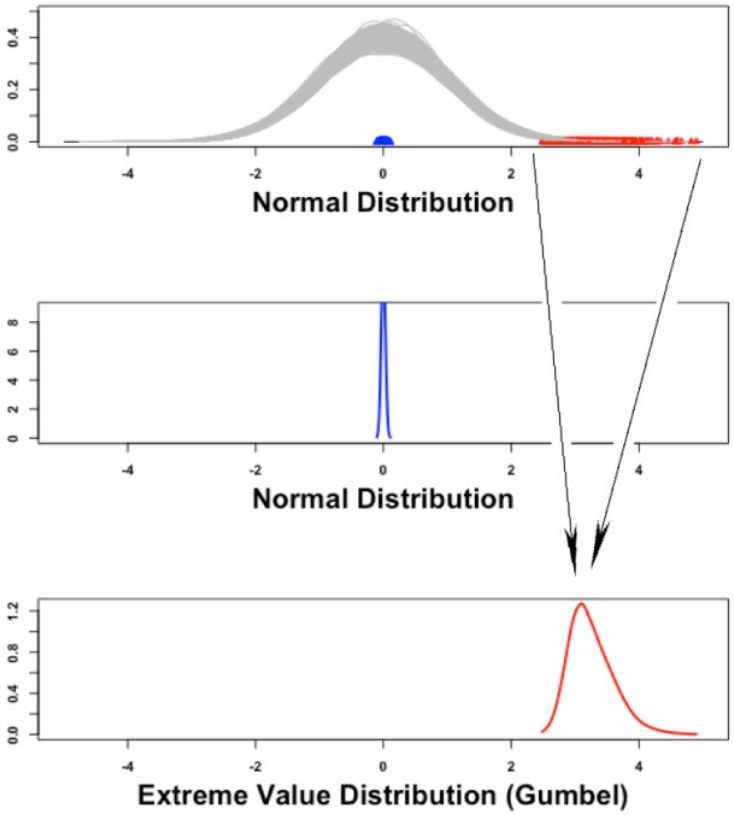
+02

Extreme Value Theory

Introduce two methods in extreme value theory

2. Extreme Value Theory

- **Extreme events** are **small probability** events, and when they occur, these small probability events can cause significant impact. Extreme value theory studies the behavior of the distribution function **in the tail**.



2. Extreme Value Theory

- When using extreme value theory in statistics there are, basically, two different approaches:

- The first is called the **block-maxima (BM) method**
- The second is the **peaks-over-threshold (POT) method**

2. Extreme Value Theory

BM method

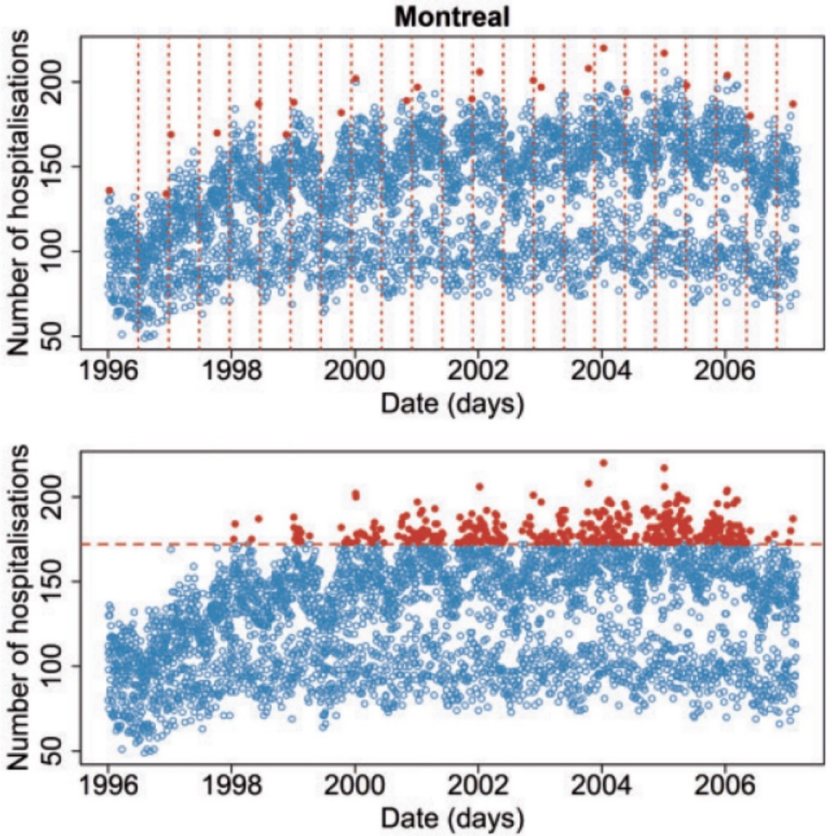
- Raw data are divided into blocks and the maximum observation is selected in each block, thereby forming a series of peaks to be analyzed.
- It can be shown that their limiting distribution is the **Generalized Extreme Value (GEV) distributioin**.
- Each block has the same size and the number of blocks determines the number of peaks.

POT method

- The POT method does not consider block maxima. Instead, all observations larger than a given threshold (typically a specific quantile) are selected for the analysis.
- It can be shown that their limiting distribution is the **Generalized Pareto distribution (GPD)**.

2. Extreme Value Theory

- Example: daily number of hospitalizations in Montreal
- This data set supplied from January 1996 to March 2007. This leads to a total of 4077 days of hospitalizations.
- The BM method uses a block size of 180 days while the POT method uses the 90% quantile (the peaks are in red):



2.1 The block-maxima (BM) method

1. The Gumbel Distribution: (representing distributions with lighter tails)

$$F(x;\mu,\sigma,0) = e^{-e^{-(x-\mu)/\sigma}} \quad \text{if} \quad \xi = 0$$

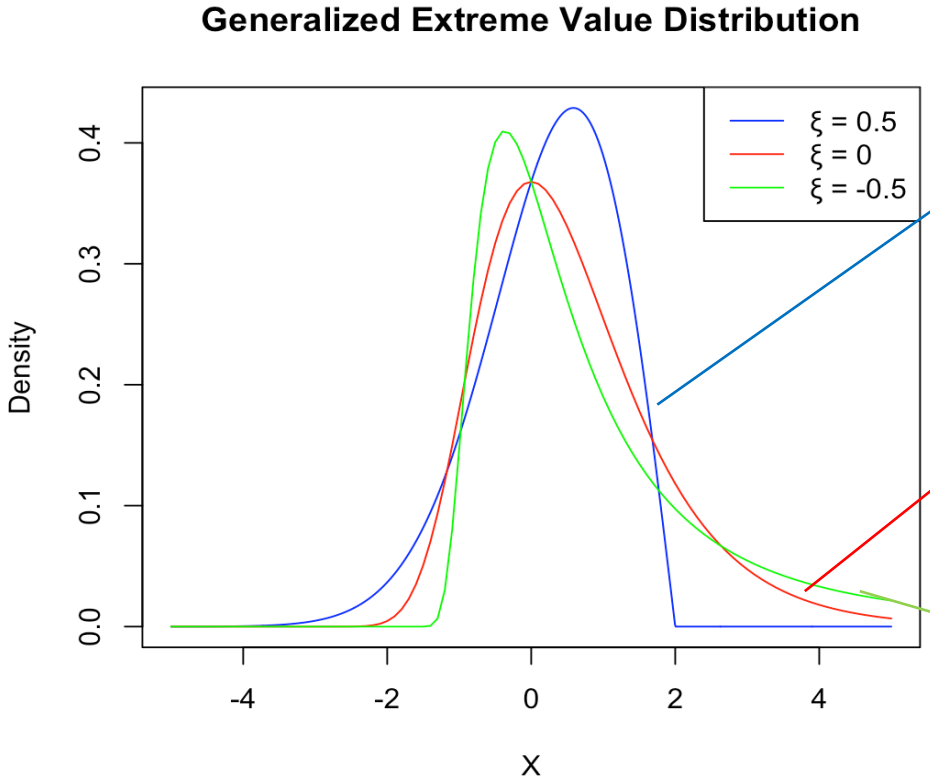
2. The Fréchet Distribution: (representing distributions with heavy tails)

$$F(x;\mu,\sigma,\xi) = \begin{cases} e^{-(\frac{x-\mu}{\sigma})^{-|\xi|}} & x > \mu \\ 0 & x \le \mu \end{cases} \qquad \text{if } \xi < 0$$

3. The Weibull Distribution: (representing distributions with finite tails)

$$F(x;\mu,\sigma,\xi) = \begin{cases} e^{-(-\frac{x-\mu}{\sigma})^{\xi}} & x < \mu \\ 1 & x \ge \mu \end{cases} \qquad \text{if } \xi > 0$$

2.1 The block-maxima (BM) method



Weibull distribution:
has a finite tail
Uniform distribution;
Beta distribution

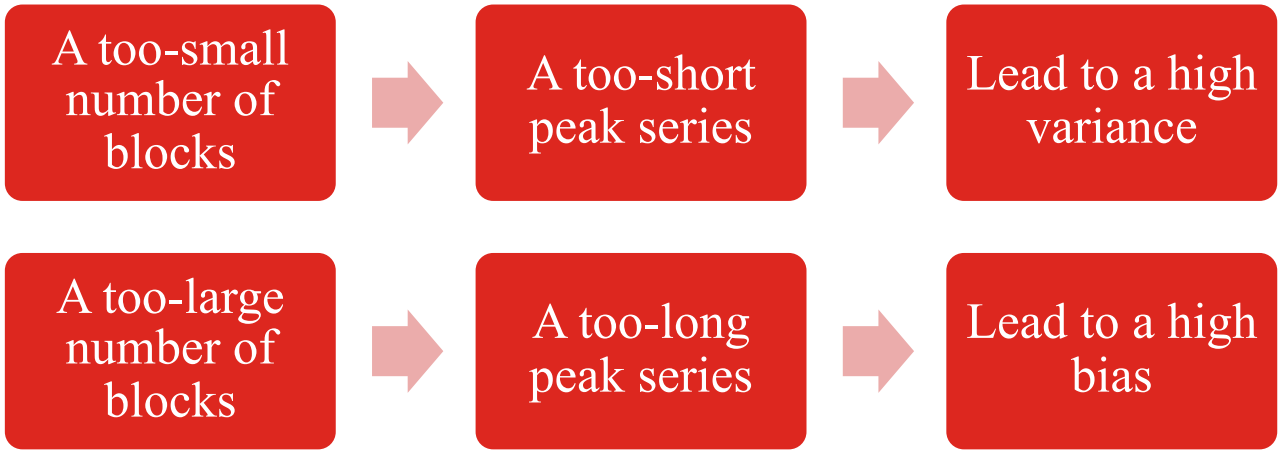
Gumbel distribution:
has a light tail
Normal distribution;
Log-normal distribution;
Exponential distribution

Frechet distribution:
has a heavy tail
t-distribution; Cauchy distribution

2.1 The block-maxima (BM) method

How to set blocks?

- The choice of block size is critical in the BM method. It amounts to a trade-off between bias and variance.



2.1 The block-maxima (BM) method

Return Level

From the fitted distribution, one can estimate the probability of the occurrence of an **extreme quantile** (called "**return level**") over a certain **return period *T***. Return level is one of the main outputs of Extreme Value Theory. More precisely, the return level is defined as the value that is expected to be equaled or exceeded on average once every interval of time T (with a probability of $1/T$).

2.1 The block-maxima (BM) method

Return Level

Once the parameters of GEV distribution have been estimated, the return level z_p associated with return period T can be computed (by inverting the GEV distribution):

$$z_p = \mu - \frac{\sigma}{\xi}\{1 - [-\ln(1-p)]^{-\xi}\} \quad \text{for } \xi \neq 0$$

$$= \mu - \sigma \ln[-\ln(1-p)] \quad \text{for } \xi = 0$$

2.2 The peak-over-threshold (POT) method

Could we do this?

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, having marginal distribution function F.

$$P(X > u + y \mid X > u) = \frac{1 - F(u + y)}{1 - F(u)}, \quad y > 0$$

If the parent distribution F were known, the distribution of threshold exceedances would also be known.

2.2 The peak-over-threshold (POT) method

- **Generalized Pareto distribution (GPD)**

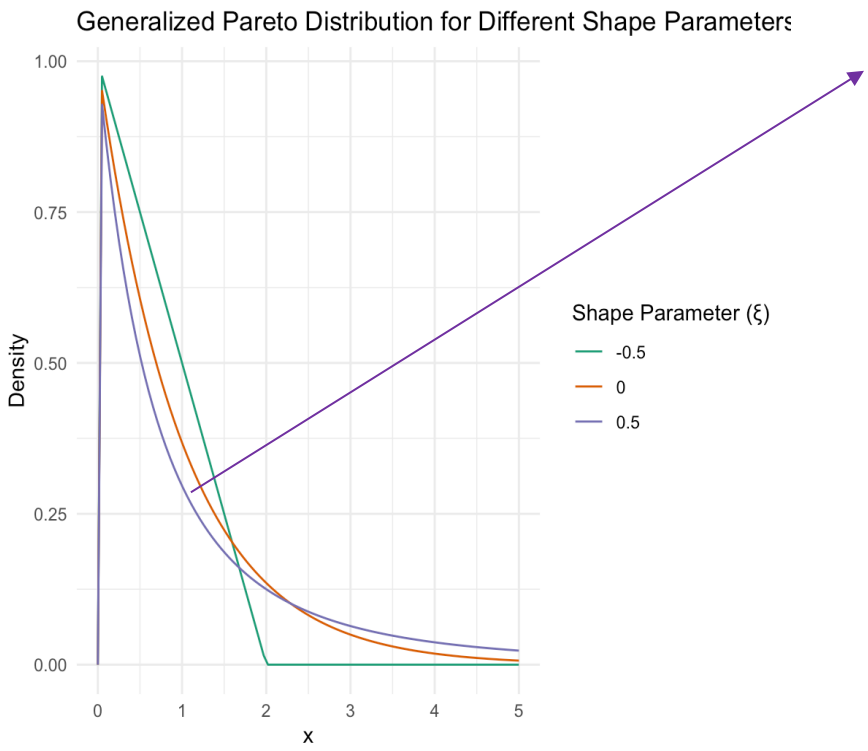
Let X be the raw data variable, the exceedances of X over the threshold u are then expressed as $Y = X - u$. The GPD distribution function is defined by:

$$H(y;\sigma,\xi) = 1-(1+\frac{\xi y}{\sigma})^{-\frac{1}{\xi}}, \quad \text{for } \xi \neq 0$$

$$= 1-\exp(-\frac{y}{\sigma}), \quad \text{for } \xi = 0$$

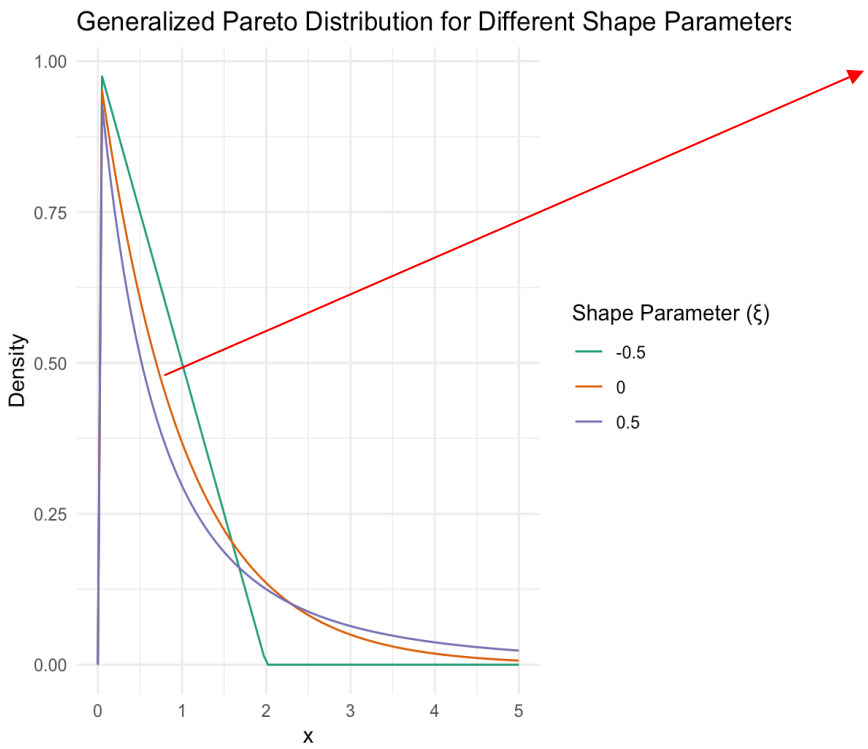
on the interval $\{y : y > 0 \text{ and } (1+\frac{\xi y}{\sigma}) > 0\}$ where $\sigma > 0$ is the scale parameter and $-\infty < \xi < \infty$ is the shape parameter.

2.2 The peak-over-threshold (POT) method



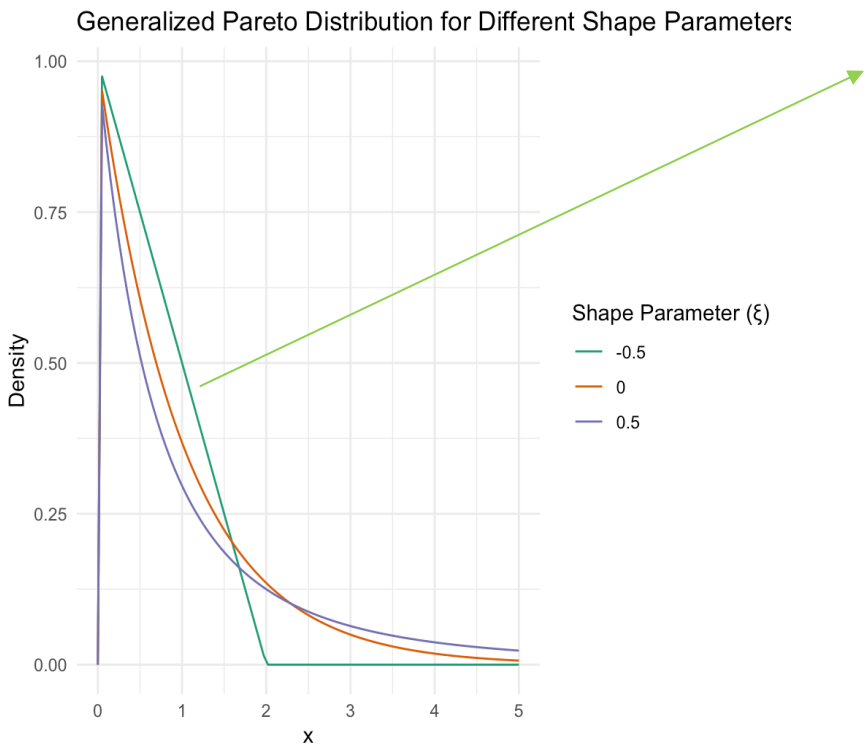
- **Heavy tail**: the heavier tail means the probability of observing extremely large values is higher compared to other distributions.
- The tail of the distribution does not have an upper bound, which makes it suitable for modeling process where extremely large values can be observed without a theoretical upper limit.
- Used to model financial losses from rare, catastrophic events where losses can be exceptionally high, such as natural disasters.

2.2 The peak-over-threshold (POT) method



- **Exponential Tail:** It represents a "lighter-tailed" distribution.
- It is useful for modeling decay processes.
- It can be applied in modeling the inter-arrival times in queues or the reliability of mechanical systems where failures follow an exponential pattern.

2.2 The peak-over-threshold (POT) method

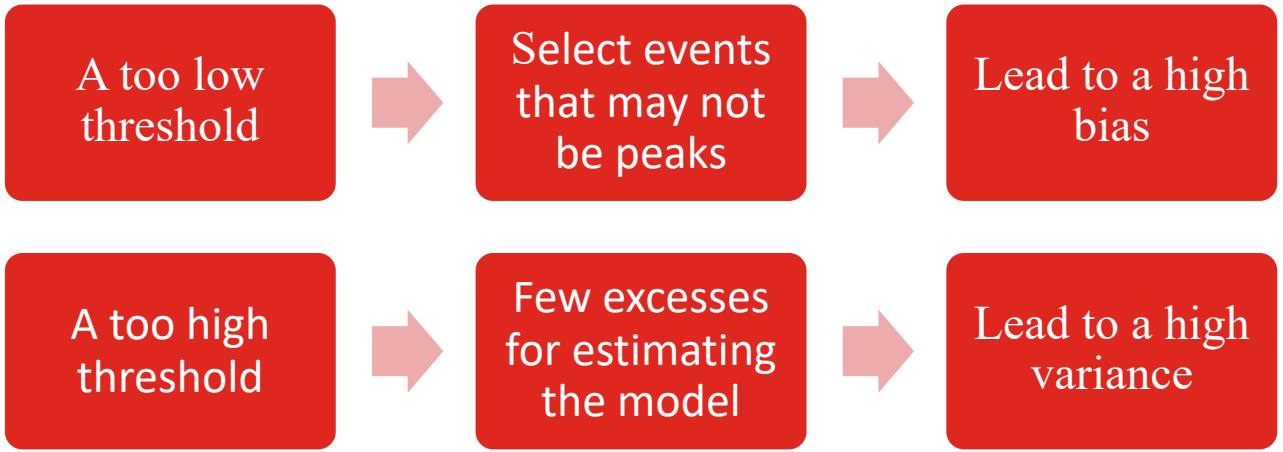


- **Short tail**: in this scenario, the distribution has an upper limit.
- This means the distribution is bounded above, which is useful for modeling quantities that have a natural upper limit.
- It is suitable for modeling phenomena like maximum capacity scenarios where values cannot exceed a certain limit.

2.2 The peak-over-threshold (POT) method

How to set threshold?

- The choice of threshold is critical in the POT method. It amounts to a trade-off between bias and variance.



2.2 The peak-over-threshold (POT) method

Return Level

Suppose that a Generalized Pareto distribution is a suitable distribution for modeling exceedances over the threshold u by variable X:

$$\Pr(X > x \mid X > u) = [1 - \xi(\frac{x-u}{\sigma})]^{-\frac{1}{\xi}}$$

It follows that

$$\Pr(X > x) = \zeta_u [1 - \xi(\frac{x-u}{\sigma})]^{-\frac{1}{\xi}}$$

where $\zeta_u = \Pr(X > u)$.

Hence, the level x_m that is exceeded on average once every m observations is the solution of

$$\zeta_u [1 + \xi(\frac{x_m - u}{\sigma})]^{-\frac{1}{\xi}} = \frac{1}{m}$$

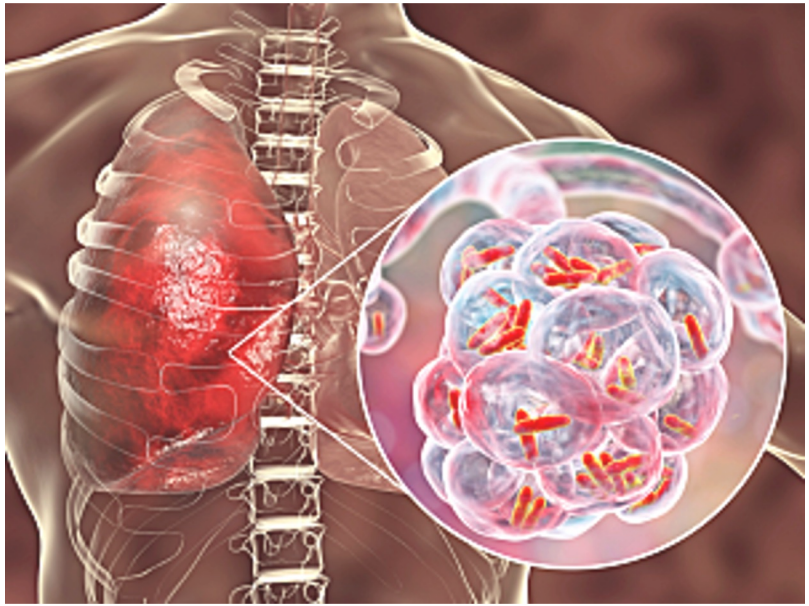
+03

Application on Public health

Apply two methods on Pneumonia and Influenza (P&I) mortality data

3.1 Description of the problem

- Pneumonia and Influenza (P&I) mortality refers to deaths caused by pneumonia and influenza viruses.
- Seasonal influenza occurs annually and varies in severity
- Understanding and predicting P&I mortality rate peaks is crucial for planning health services and interventions.



3.2 Description of the data

- We used the weekly cumulative number of P&I deaths in France from July 1979 to June 2011 (Thomas et al. Applications of Extreme Value Theory in Public Health. *PLoS ONE* 2016; 11(7):e0159312).

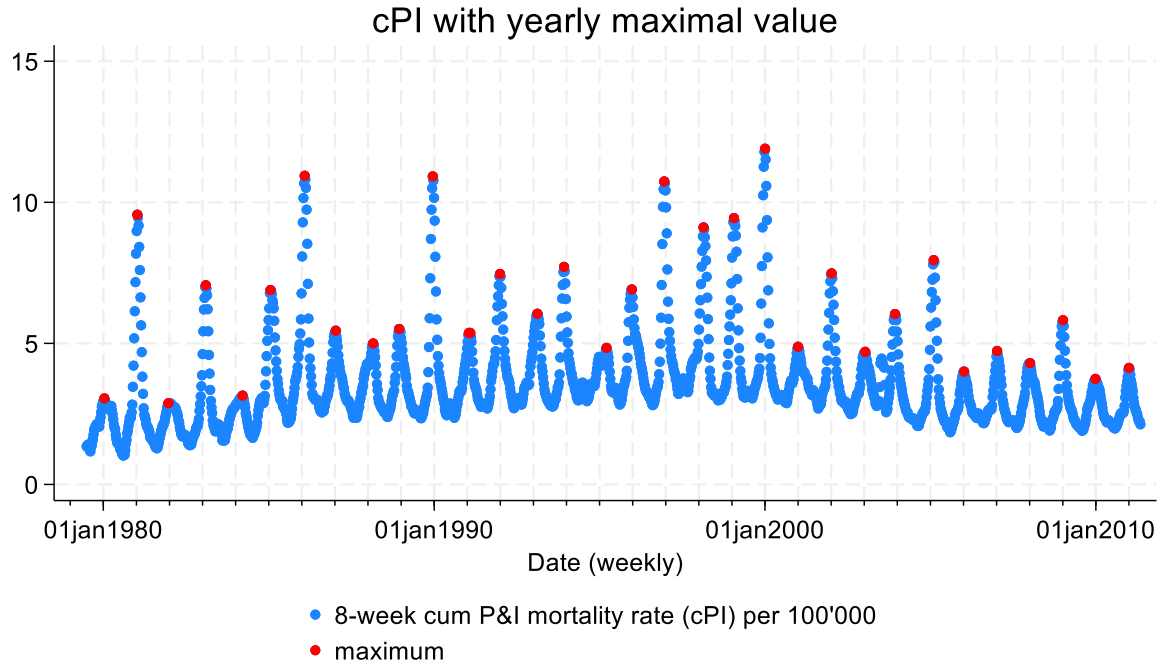
3.2 Description of the data

- Here is look at the data set:

observation	flu_season	week	cPI
1	1980	1979-07-01	1.35
2	1980	1979-07-08	1.34
3	1980	1979-07-15	1.41
4	1980	1979-07-22	1.34
5	1980	1979-07-29	1.27
...	...	...	...
1659	2011	2011-04-10	2.37
1660	2011	2011-04-17	2.30
1661	2011	2011-04-24	2.23
1662	2011	2011-05-01	2.23
1663	2011	2011-05-08	2.13

3.3 Application of the Block Maxima method

- The yearly maxima within each respiratory year (that is, from July to June to encompass annual influenza epidemics) have been extracted, generating a series of 32 annual maxima:



3.3.1 Some tests before modeling

In the classical approach of EVT, several assumptions need to be validated before proceeding to the fitting of the EVD to the peak series:

- **Independence**: The data should be independent of each other.
- **Stationarity**: There should be no trend.
- **Homogeneity**: Observations must come from the same distribution.

3.3.1 Some tests before modeling

In the classical approach of EVT, several assumptions need to be validated before proceeding to the fitting of the EVD to the peak series:

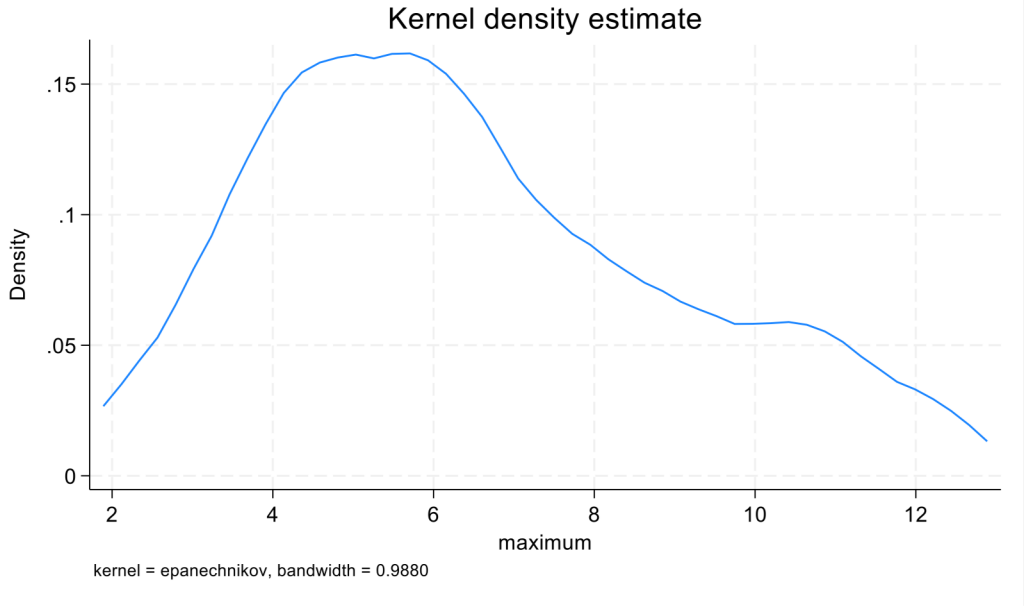
- **Independence**: Ljung-Box autocorrelation test (H0: The data are independently distributed)
- **Stationarity**: Mann-Kendall test (H0: Data have no trend)
- **Homogeneity**: Kolmogorov-Smirnov Test (H0: The data have the same distribution)

3.3.2 Results of the tests

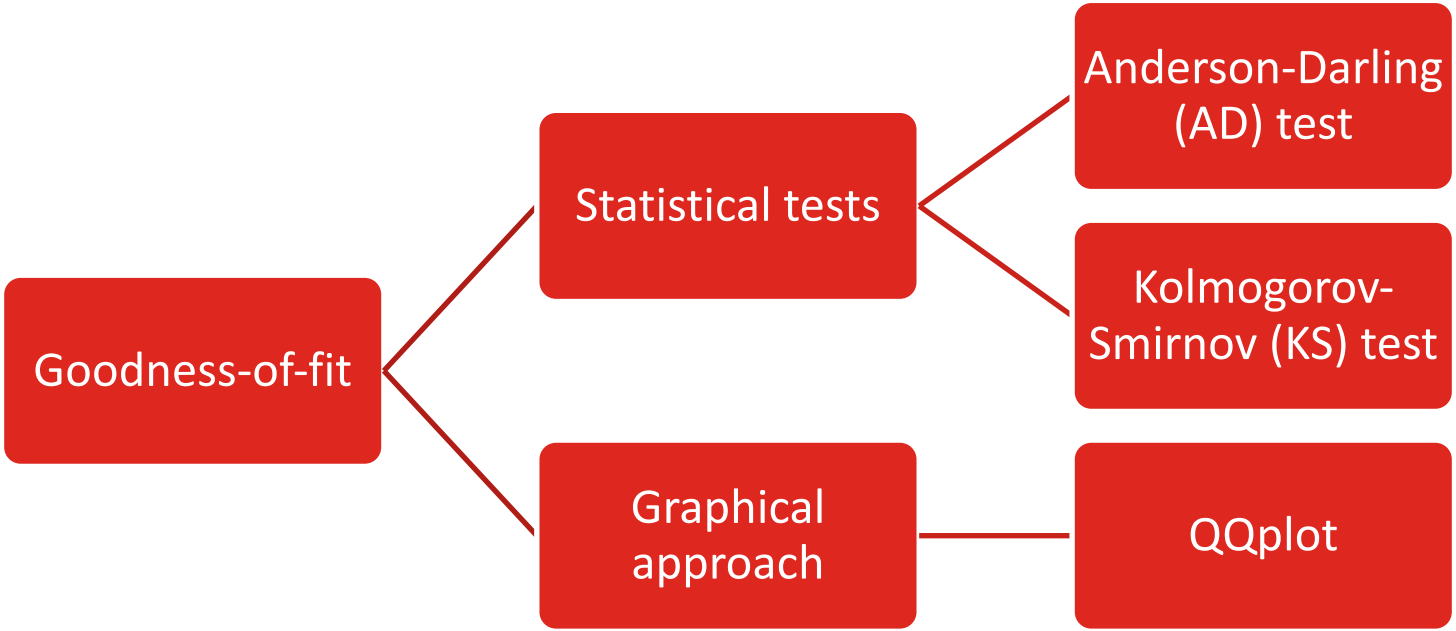
- **Independence**: The p-value = 0.720 of Ljung- Box test shows that the assumption of independence is not rejected (i.e. there is no autocorrelation)
- **Stationarity**: The Mann-Kendall test (p-value = 0.446) shows that the assumption of no trend is not rejected
- **Homogeneity**: The Kolmogorov-Smirnov test (p-value = 0.952) shows that the homogeneity assumption is not rejected.

3.3.3 Model estimation and goodness-of-fit

Maximum-likelihood estimation of the model parameters resulted in a location parameter μ = 5.33, 95%CI (4.51;6.14), scale parameter σ = 1.97, 95%CI (1.35;2.59), and shape parameter ξ = 0.004, 95%CI (-0.36;0.37).

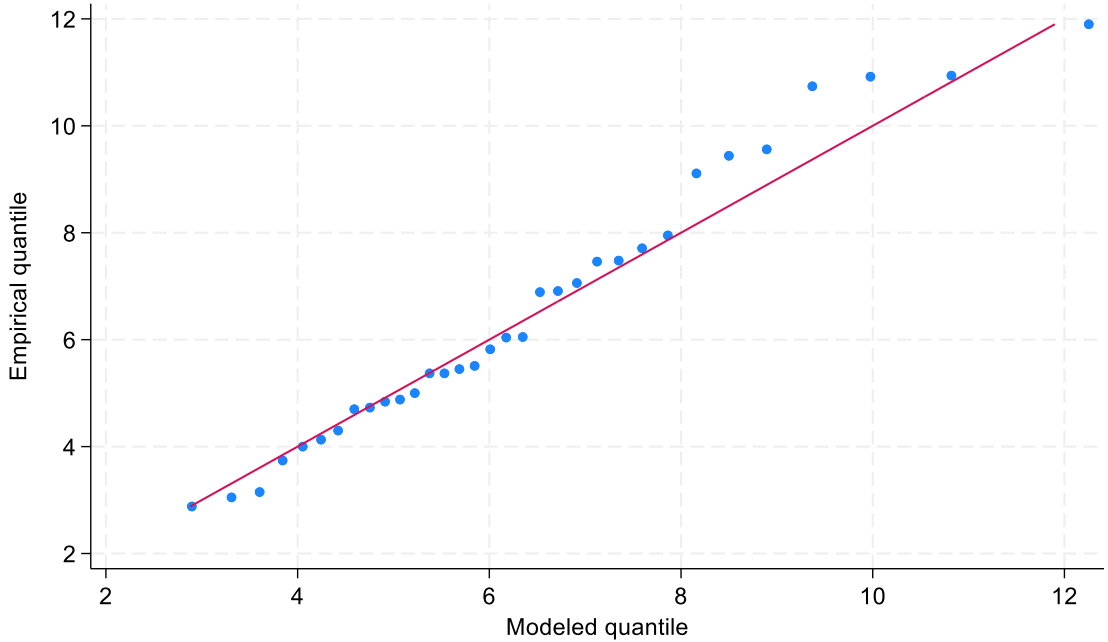


3.3.3 Model estimation and goodness-of-fit



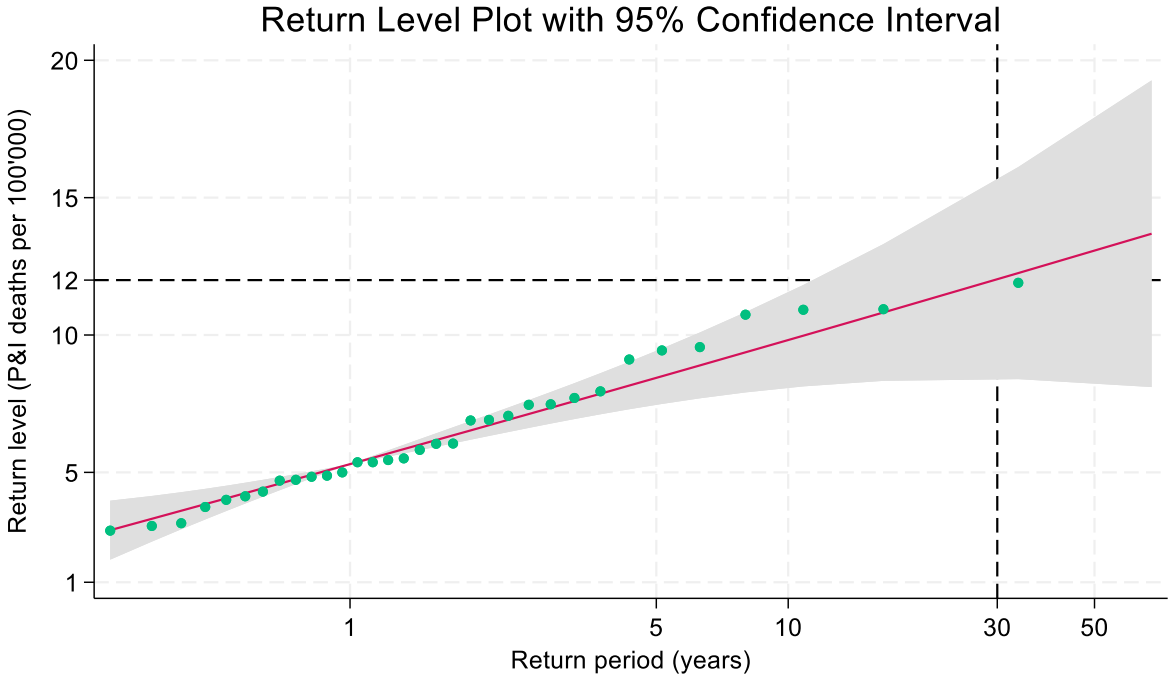
3.3.3 Model estimation and goodness-of-fit

- We obtained p-value = 0.965 for the Anderson-Darling test and p-value = 0.974 for the Kolmogorov-Smirnov test, and inspection of the QQplot:

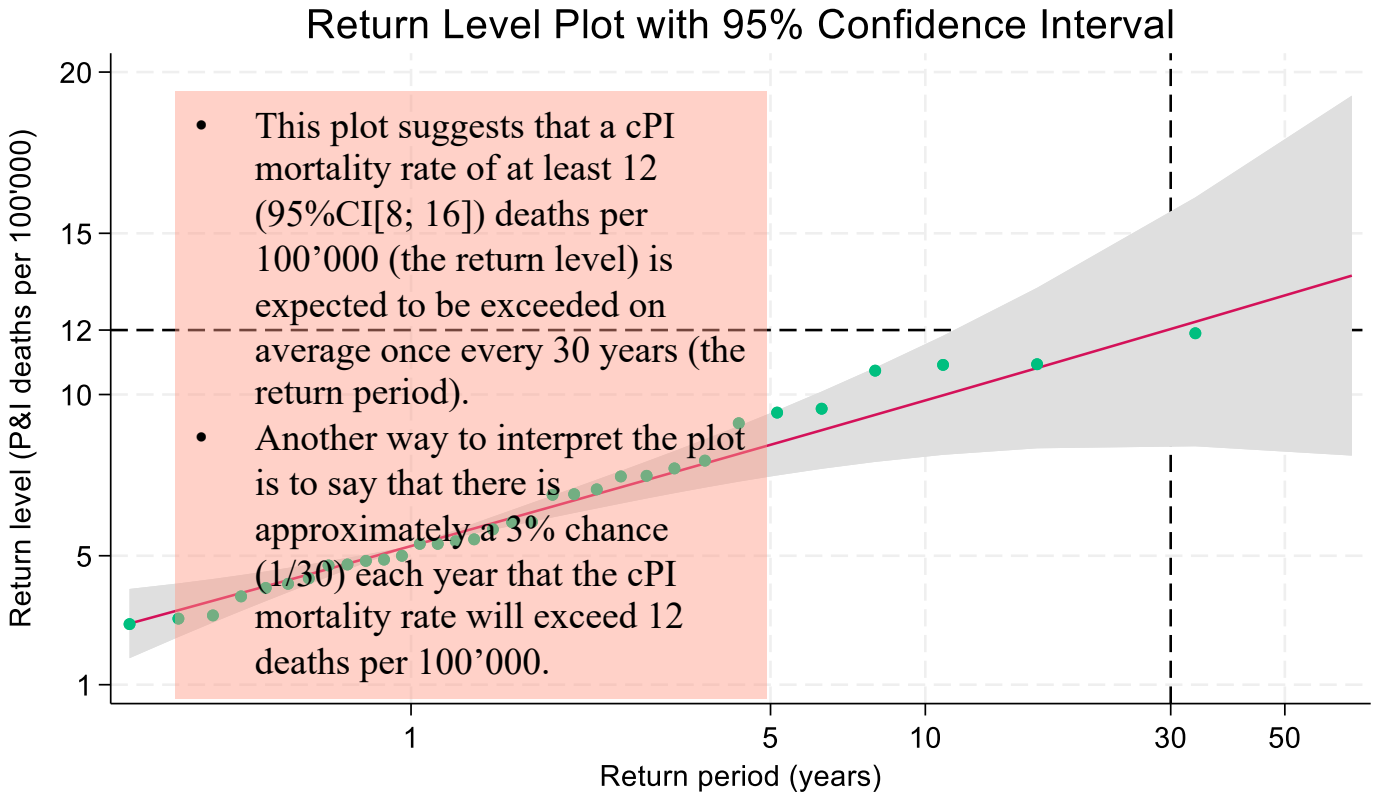


3.3.4 The Return Level plot

- Based on the estimated model parameters, the **Return Level plot** can be computed:



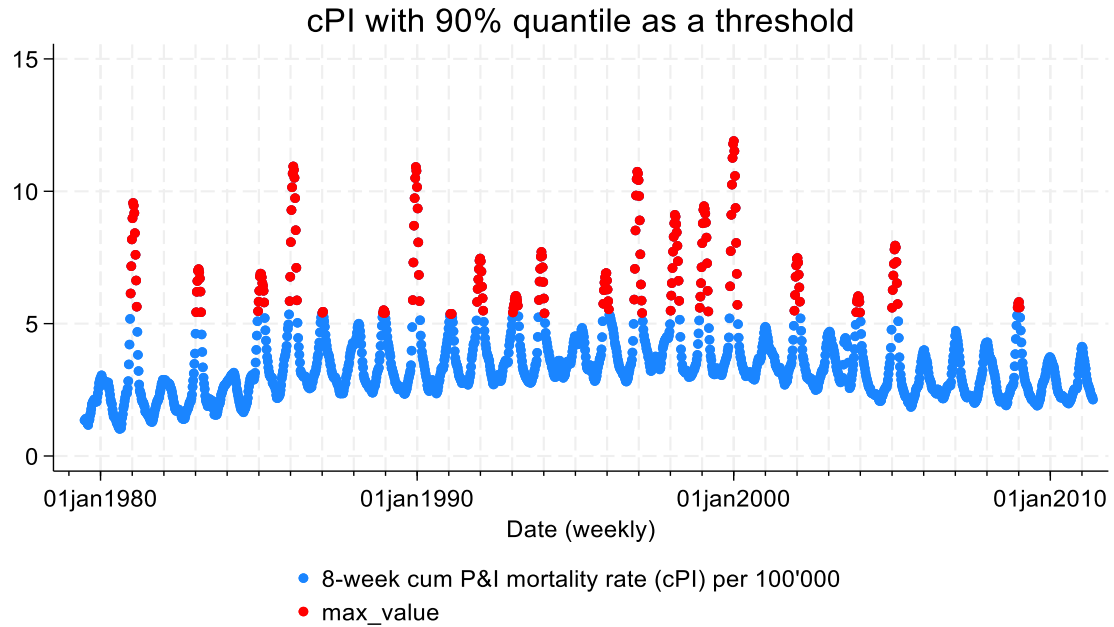
3.3.4 The Return Level plot



3.4 Application of the POT method

Several thresholds 0.75, 0.8, 0.85, 0.9, 0.95, 0.975, and 0.99 have been investigated.

For example, using the 90% quantile as a threshold, one obtains:



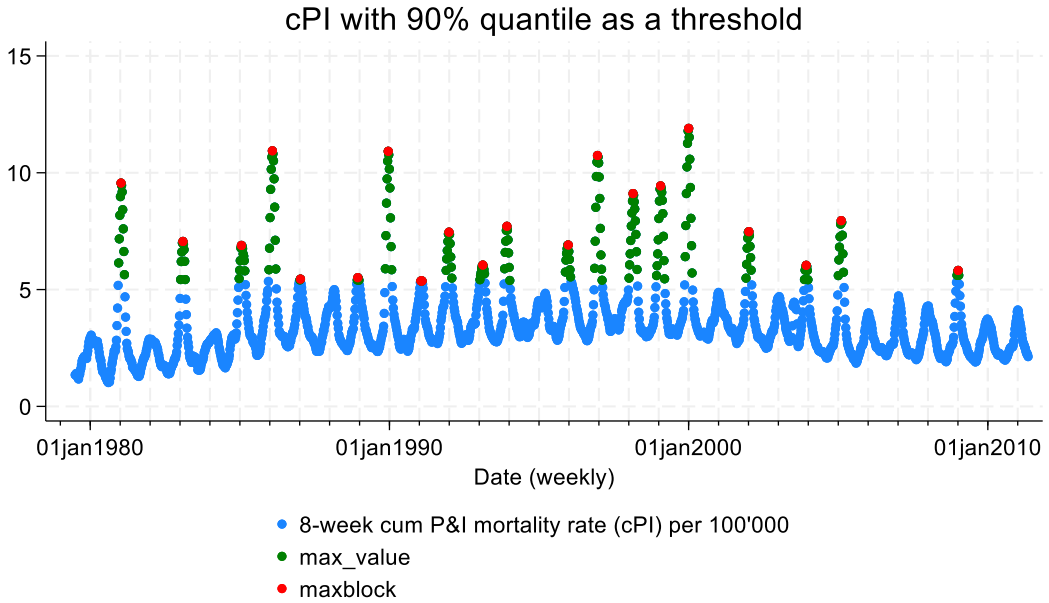
3.4 Application of the POT method

Results of the independence, stationarity, and homogeneity tests are provided in the table below for the various thresholds considered:

GPD		Assumption Test			Estimated Parameter		Goodness of Fit	
Threshold (%)	n	Box-Ljung test	M-K test	K-S test	$\hat{\xi}$	$\hat{\sigma}$	KS	AD
0.75	416	0	0.02	0.65	0.04	1.66	0	0
0.8	332	0	0.26	0.51	-0.06	1.92	0	0
0.85	250	0	0.49	0.20	-0.21	2.34	0	0
0.9	166	0	0.89	0.13	-0.36	2.72	0	0
0.95	84	0	0.26	0.11	-0.54	2.76	0	0
0.975	42	0	0.95	0.04	-0.60	2.00	0	0
0.99	17	0.07	0.68	0.28	-0.60	1.11	0	0

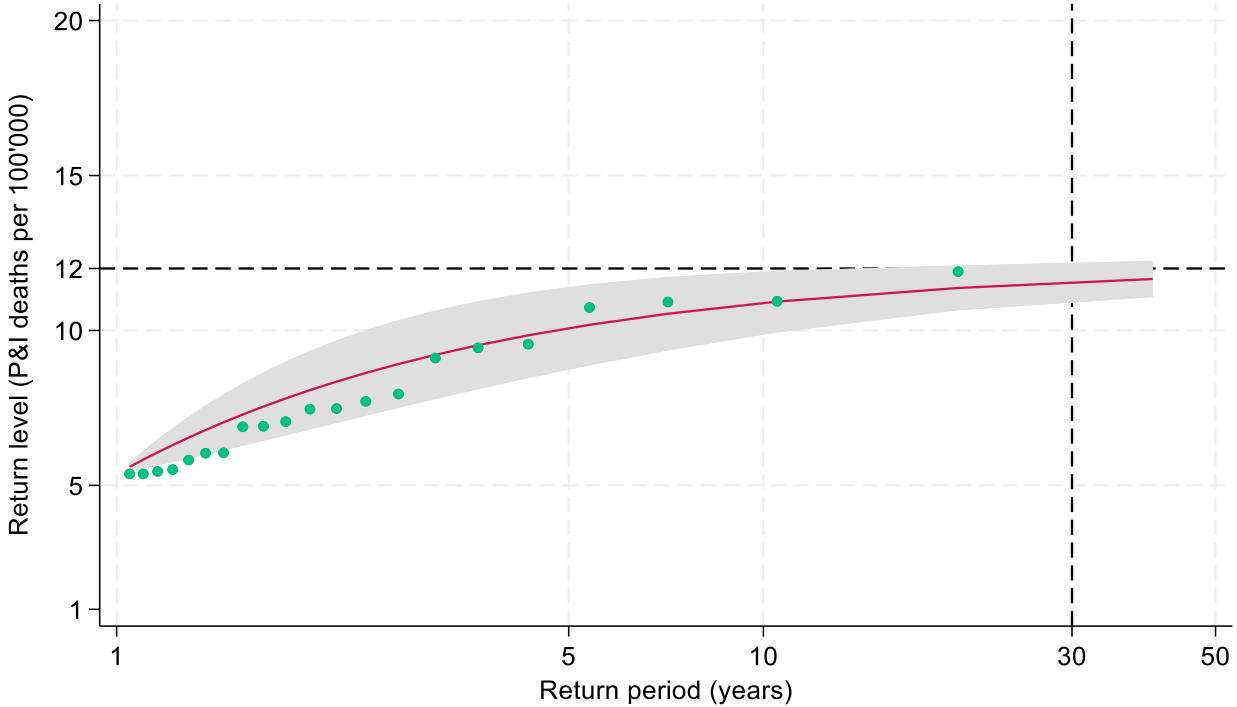
3.4 Application of the POT method

- Actually, **peaks tend to occur in clusters**, especially in the POT method. This compromises the independence assumption.
- After decluster:



3.4 Application of the POT method

- The return level plot looks almost the same as that from BM method:



04

Expansion

Links to my master thesis

4. Links to my Master thesis

- In this presentation, we considered the simple univariate setting. In my master thesis, I explored the application of the multivariate version of the Extreme Value Theory in the context of structural models (only the peak-over-threshold method).

05

Conclusion

Comments on two methods and extreme value theory

5.1 Compare two methods

Block Maxima Method

1. Easy to ignore some valuable data.
2. The trend is consequently kept in the peak series.

Peak Over Threshold Method

1. May miss some information caused by time.
2. Peaks tend to occur in clusters.

5.2 Some advantages of using extreme value theory

- Extreme value theory does not make any assumptions about the distribution of the overall data
- Extreme value theory focuses on the tails of the distribution and thus provides a more accurate measure of the loss of extreme events.

Extreme Value Theory

References

- Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. 2001 Springer.
- Guillou et al. An extreme value theory approach for early detection of time clusters. Asimulation-bsed assessment and an illustration to the surveillance of Salmonella. *Statistics in Medicine* 2014; 33: 5015-5027.
- Thomas et al. Application of extreme value theory in public health. *PLoS ONE* 2016; 11:e0159312.
- Chiu et al. Mortality and morbidity peaks modelling: an extreme value theory appeoach. *Statistical Methods in Medical Research* 2018; 27: 1498-1512.
- Caetano et al. Modeling large values of systolic blood pressure in the Portuguese population. *REVSTAT* 2019; 17: 163-186.

Thank you for your listening

Merci pour votre attention ☺