Extremes and data assimilation in epidemiology:

two influenza case studies

Hans Wackernagel[1],Christian Lajaunie[1]
Cyrille Jégat[1,2], Huey Chyi Lee[1], Dag Johan Steinskog[1]
Fabrice Carrat[2], Magali Lemaître[2], Mark Wilson[3]

[1]Equipe de Géostatistique — MINES ParisTech
[2]INSERM UMR-S 707 [3]University of Michigan

Influenza epidemics

1) Short term (morbidity data):

- data assimilation by particle filtering

2) Long term (mortality data):

- extreme value analysis of epidemics

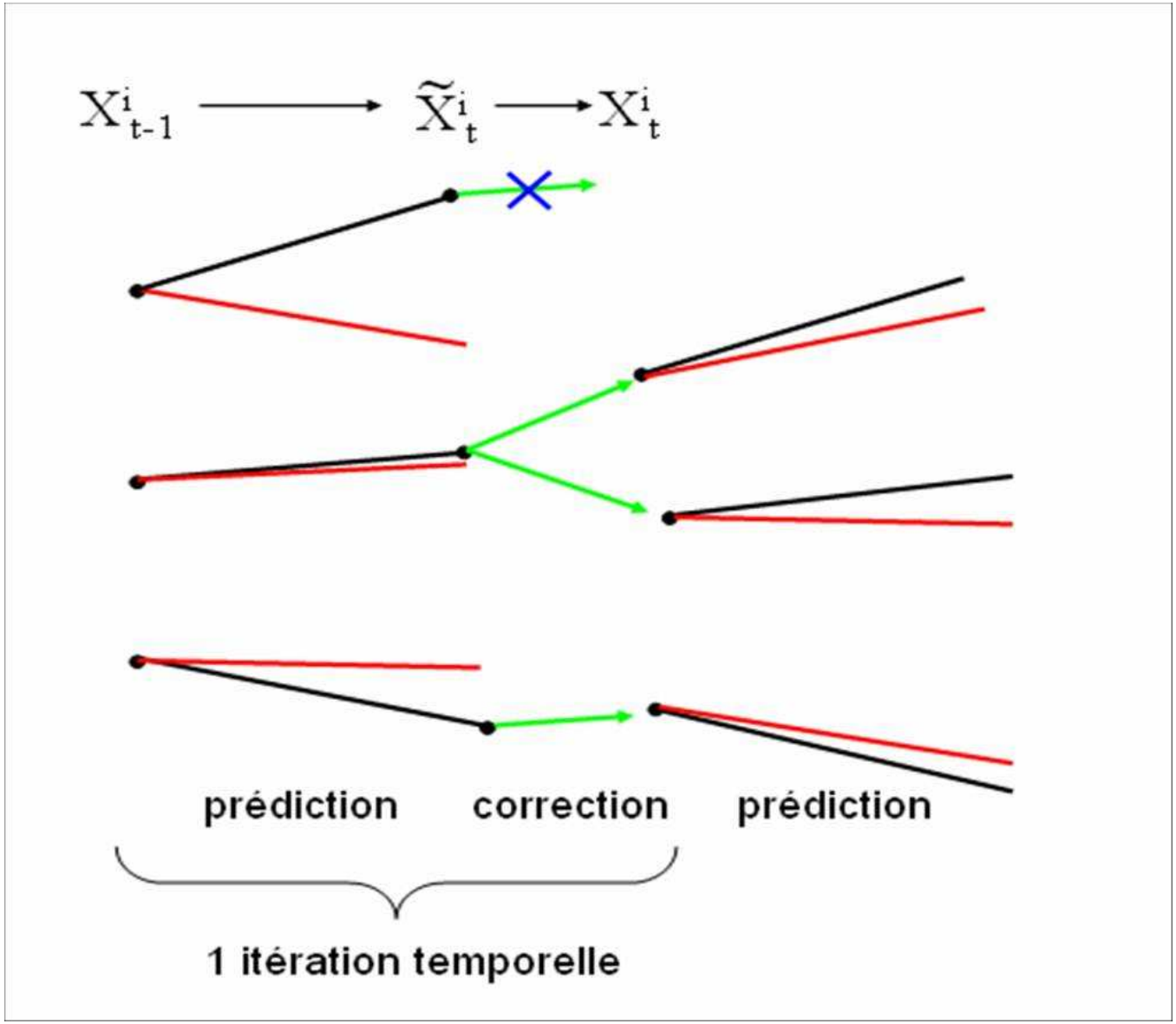
Outline: Particle filtering

Particle Filters

The different Kalman filters involve a Gaussian distribution assumption — at least for the update.
This not the case for particle filters.

- In particle filters the update step is based on sequential importance resampling.

Sequential importance resampling



Influenza epidemics

- Influenza infects from 5% to 20% of the population during an epidemic.
- Inserm *Sentinelles* network:
  1200 general practitioners reporting once a week.
- The conditions leading to an outbreak are not well understood.
- Early detection of an epidemic could limit its impact.

Data: reported cases by general practitioners

- No virological analysis:
  diagnostic of influenza-like illness (ILI).
- Many infected people do not consult (~ 50%).
- Irregular logging in by doctors:
  absence of cases is not reported,
  accumulation of cases, etc.
  $\Longrightarrow$ need for data pre-processing.

Stochastic epidemiological model

Susceptible: the people free from the disease and without specific immunity,

Infected: the people infected by the virus and who are still infectious,

Removed: the people who have recovered from their illness, are no longer infectious and are immunized.

- SIR model: a partition (varying with time) of the total population (which is assumed constant),

$$N(\mathbf{x}) = S_t(\mathbf{x}) + I_t(\mathbf{x}) + R_t(\mathbf{x})$$

where $\mathbf{x}$ is location in geographical space.

Epidemic state: a hidden Markov chain

- Two epidemic states: $Y_t \in \{1, 2\}$
- Transition probabilities: $P(Y_t = j \mid Y_{t-1} = i) = P_{ij}$
- Contamination probabilities for S-I contact: $\tau_t \in \tau_1, \tau_2$
  (change of dynamics during epidemics)

Regionalized SIR model

- The new cases in a region **x** result from contacts between **susceptibles** at **x** and **infectives** from all regions **x**′.
- Each **susceptible** avoids contamination during a time interval with probability:

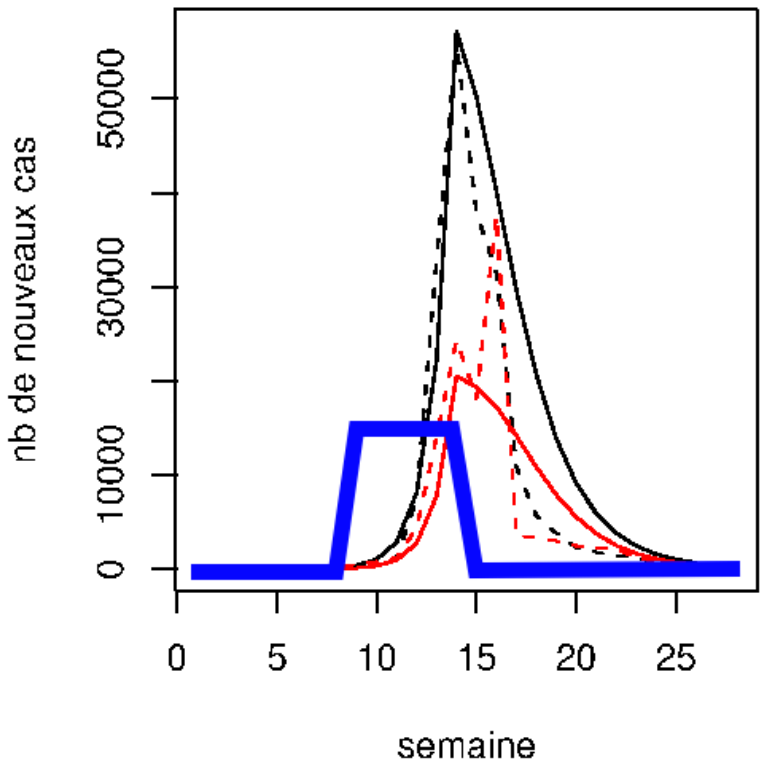
$$q(\mathbf{x}, t) = \prod_{\mathbf{x}'} \left(1 - \tau_t \frac{I_t(\mathbf{x}')}{N(\mathbf{x}')}\right)^{a_t(\mathbf{x},\mathbf{x}')}$$

  where $a_t(\mathbf{x}, \mathbf{x}')$ is the contact rate, known from the population fluxes.
- Number of new **infectives** at **x**:
  binomial with parameters $S(\mathbf{x})$ and $1 - q(\mathbf{x}, t)$.

Regionalized SIR model

- Sharp increase of the probability of contamination during epidemic episodes, modeled **for each region**: $\tau_t(\mathbf{x})$
- Infected $I_t(\mathbf{x})$ split according to time since infection: $I_t^0(\mathbf{x}), ..., I_t^6(\mathbf{x})$
- **Recovery** probability for each category at a time step is: $p_{rec}^k; k = 1, \cdots, 7$ with $p_{rec}^7 = 1$ (since infection cannot last more than 7 days).
- Relation between **declarations** and the **epidemic state**: the number of declarations in $\mathbf{x}$ is conditionally binomial, with parameters $S_{t-1}(\mathbf{x}) - S_t(\mathbf{x})$ and $p_{decl}(\mathbf{x})$
- **Population fluxes**: described using an **origin-destination matrix** based on a gravitational model (in analogy with Newton's laws).

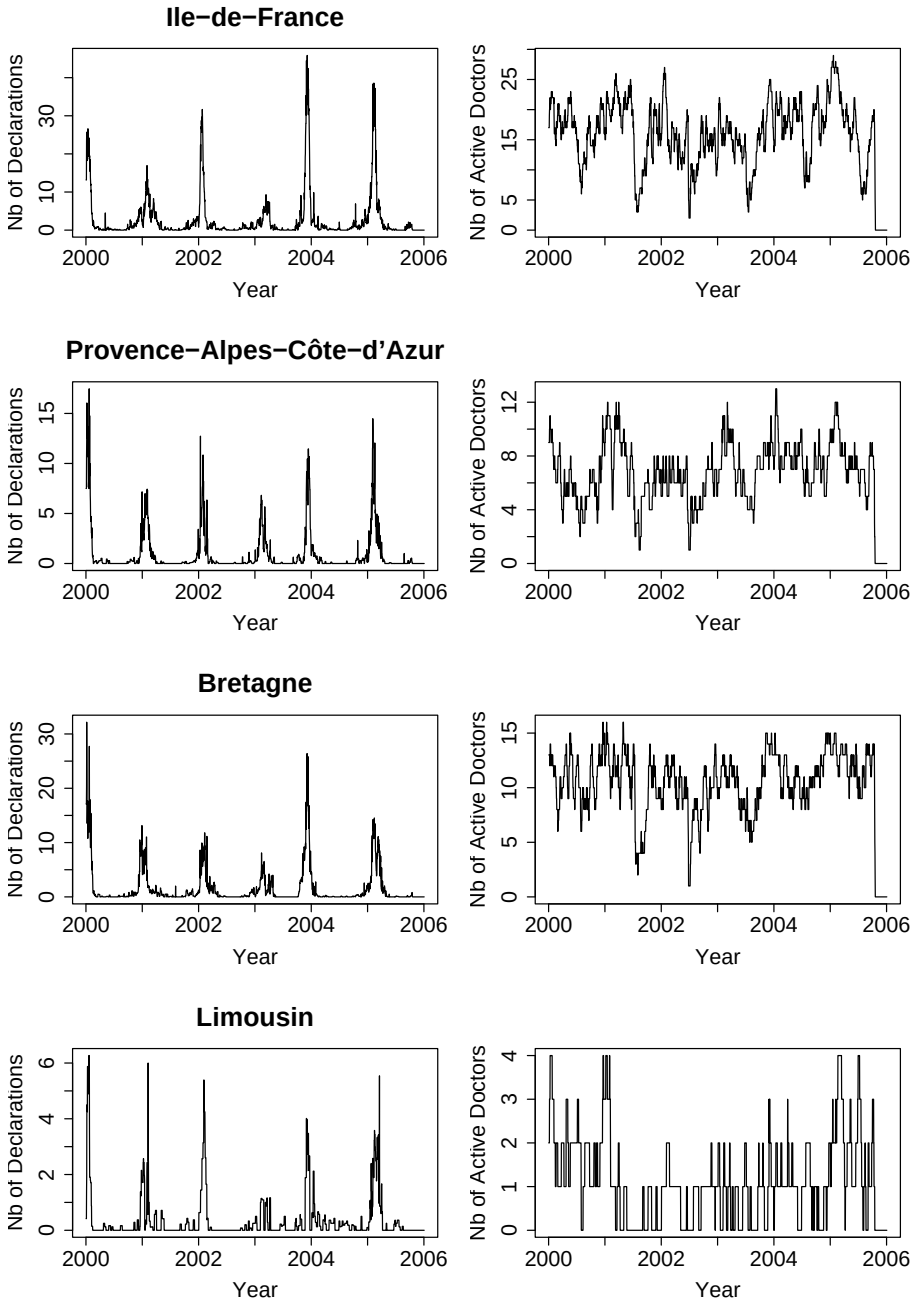
Early detection problem



Reported cases: ■ Paris (75) ■ Hauts de Seine (92)

Epidemic state: ■ common to both departements

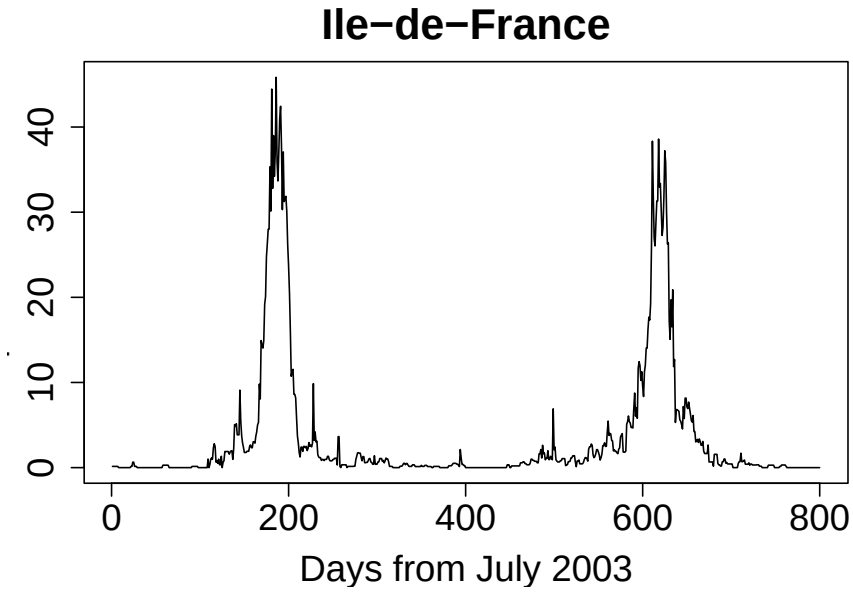
Sentinelles data (example)

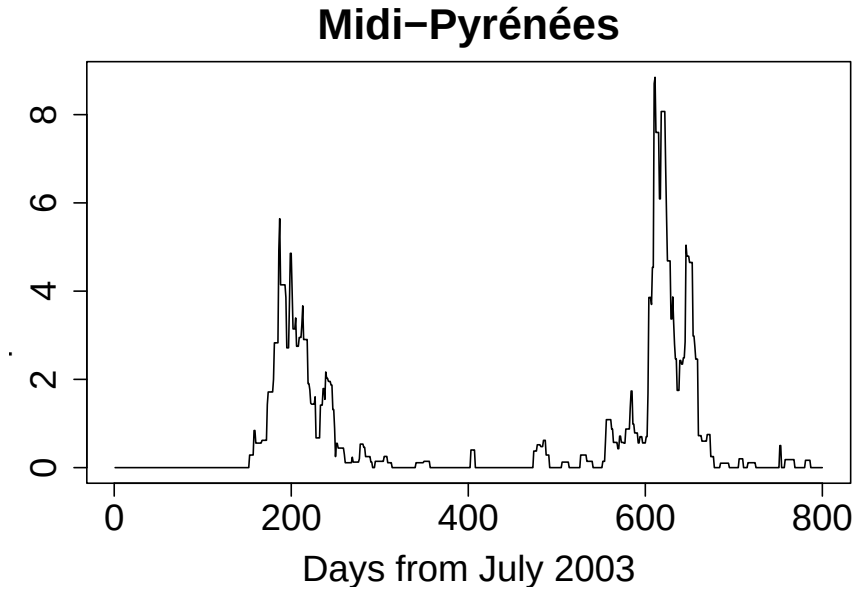


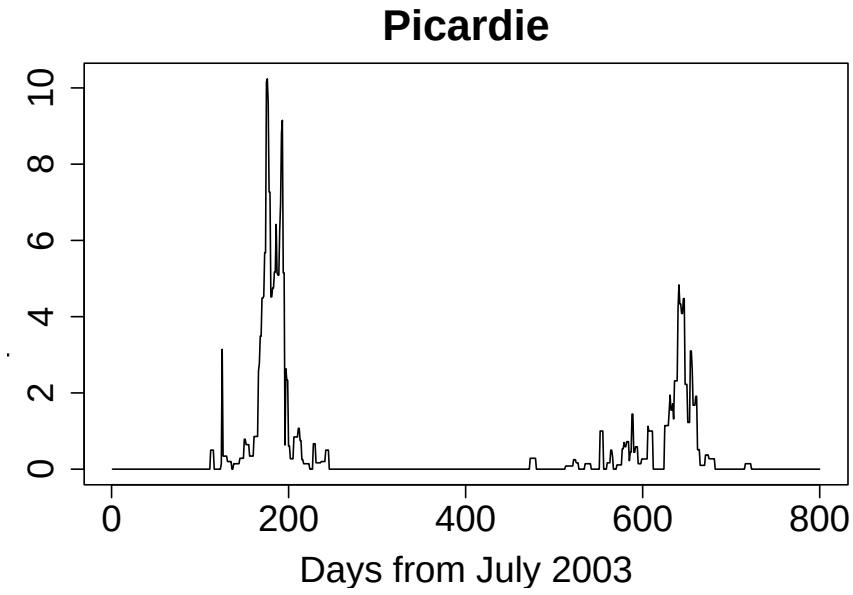
Reported cases / week (left column)

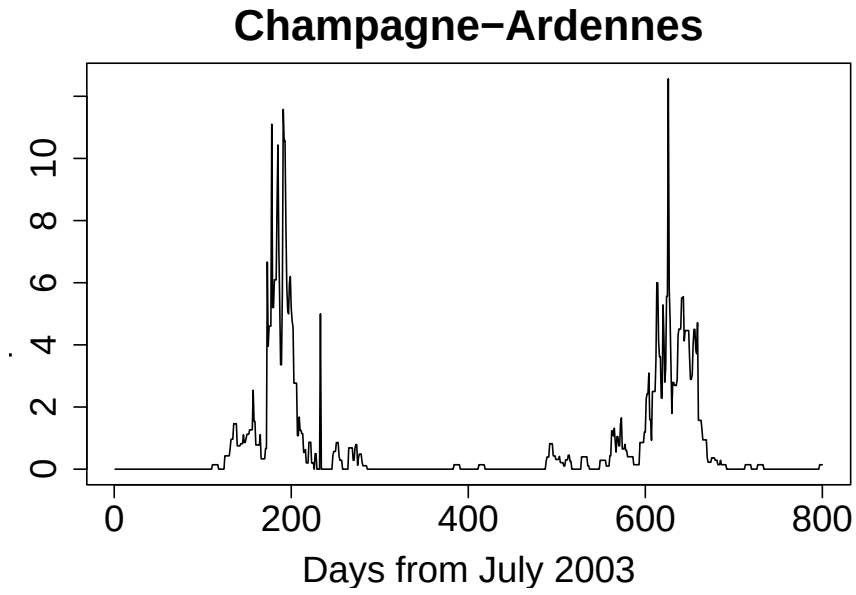
Active doctors / week (right column)

Reported cases of ILI: 4 regions



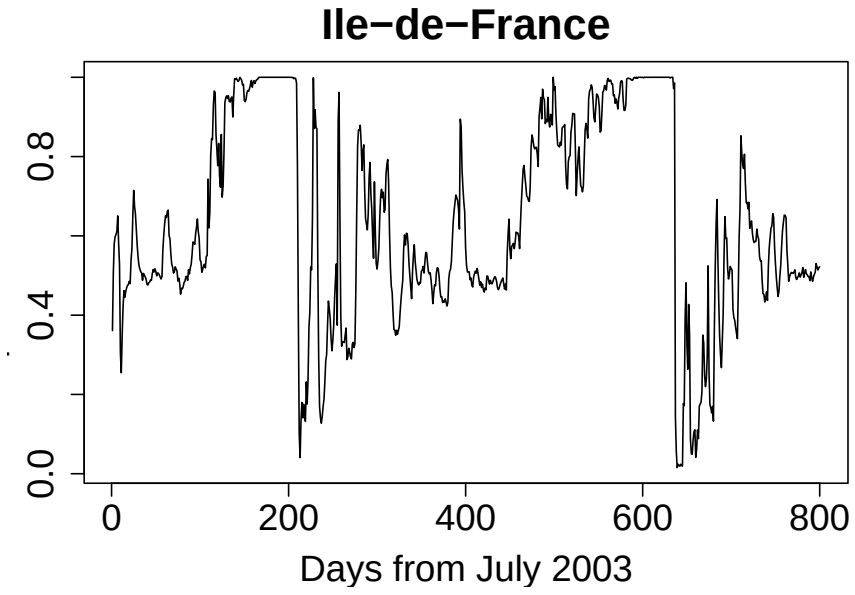


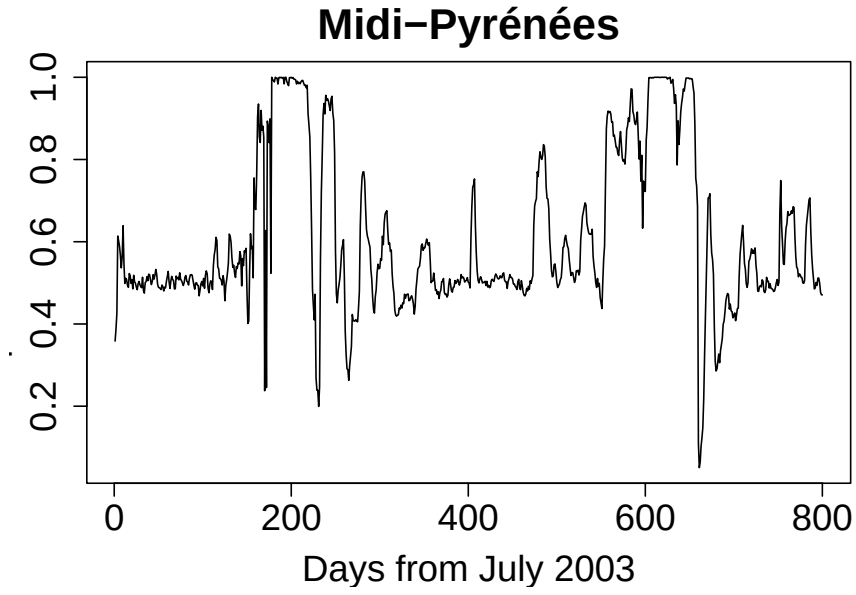


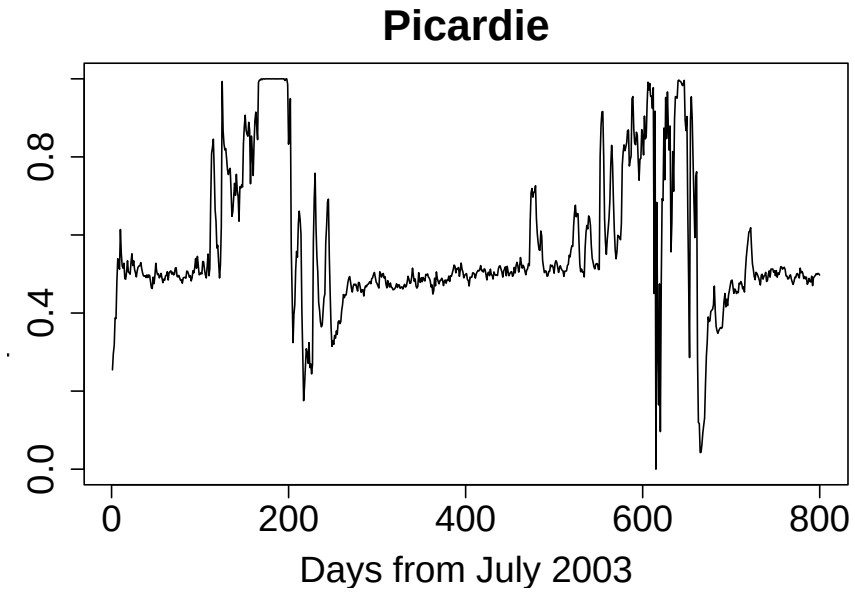


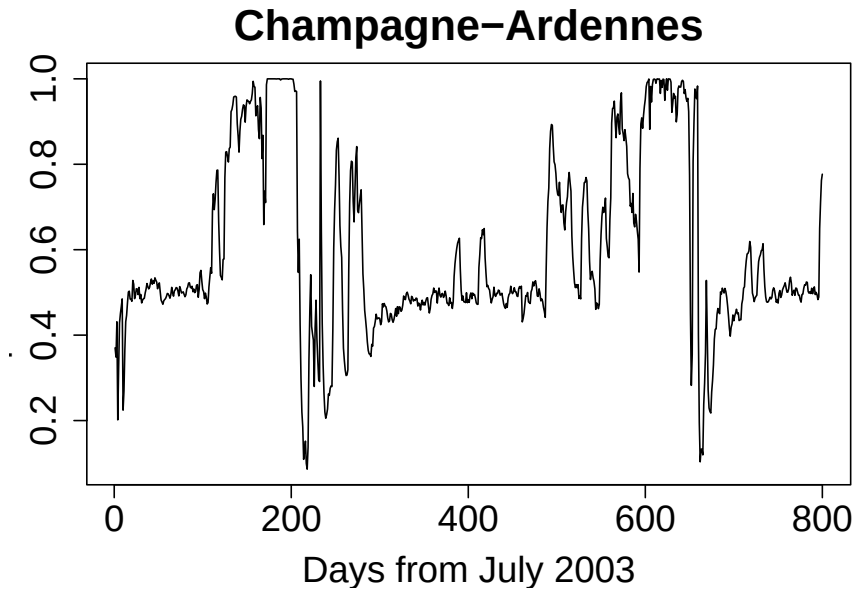
Assimilation of 2003-2005 data

Estimation of the two-state Markov chain



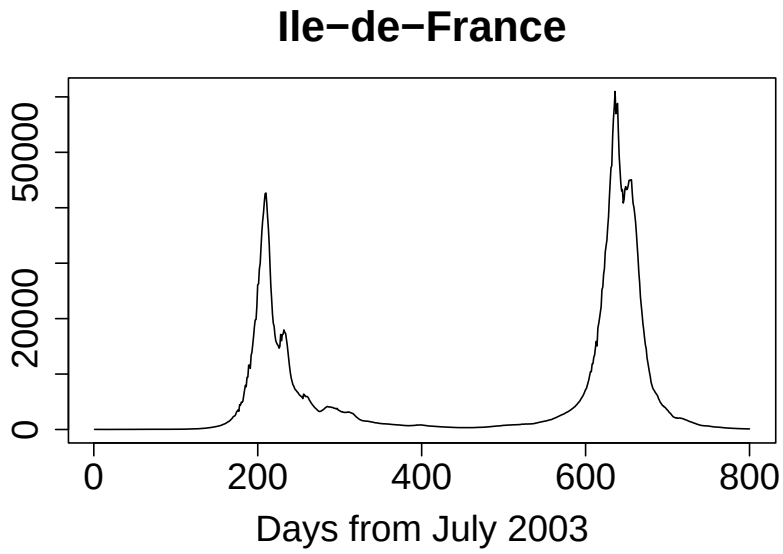


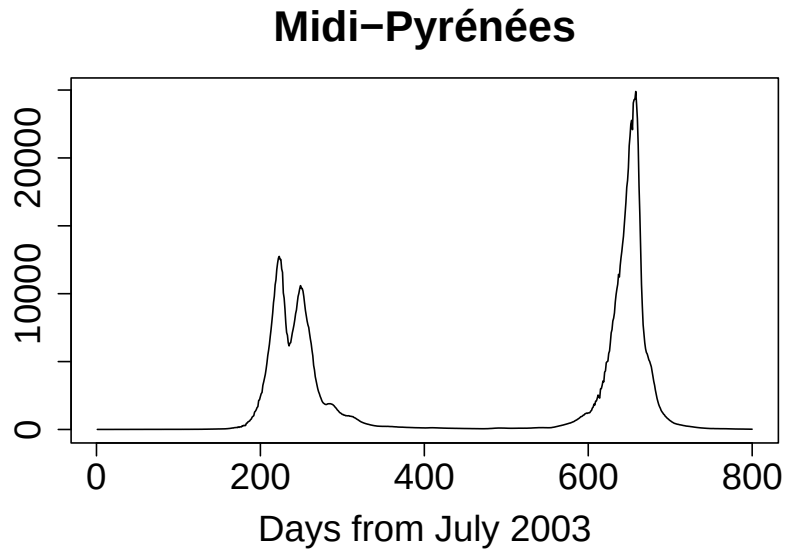


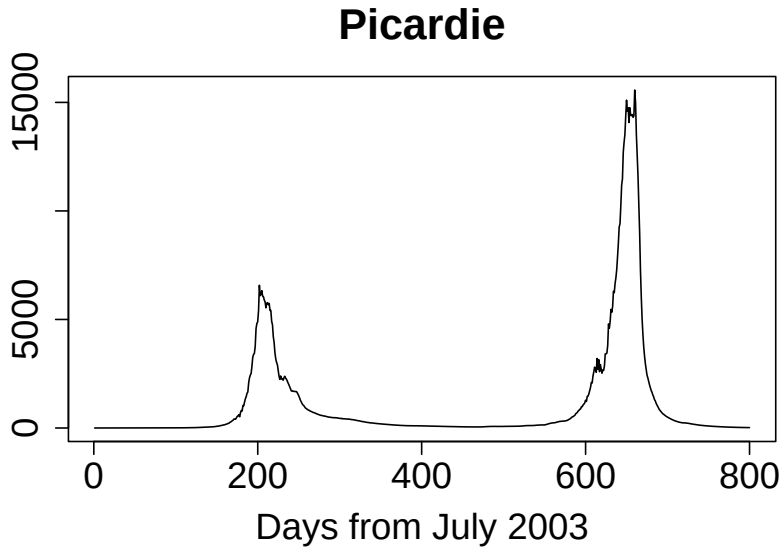


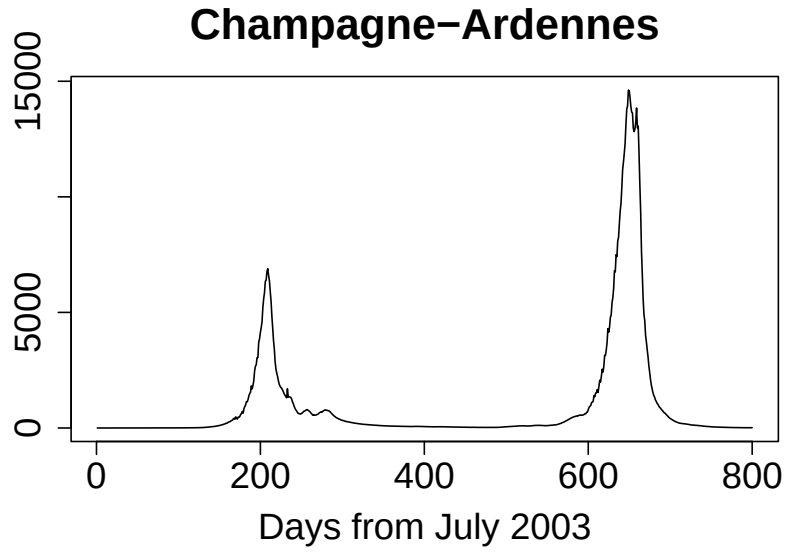
Assimilation of 2003-2005 data

Estimation of the number of infected individuals



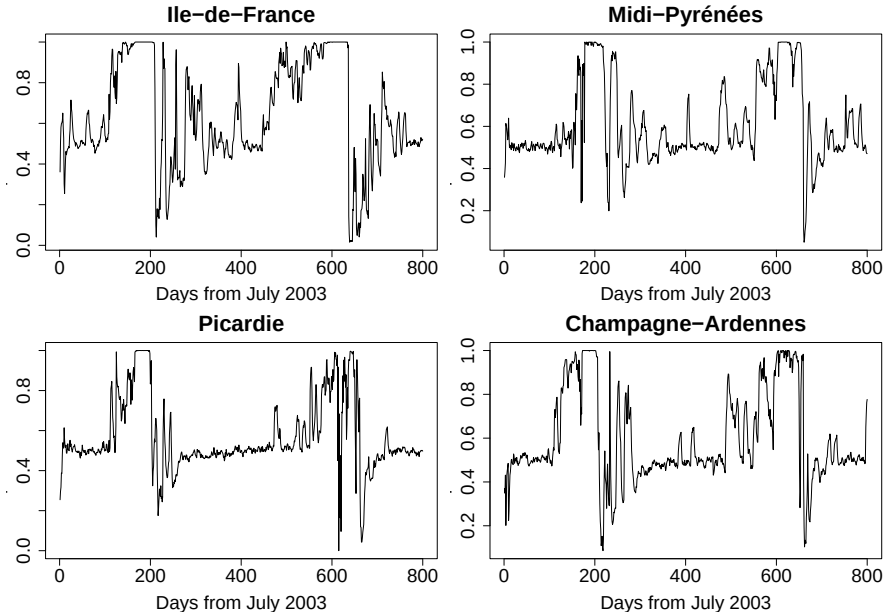


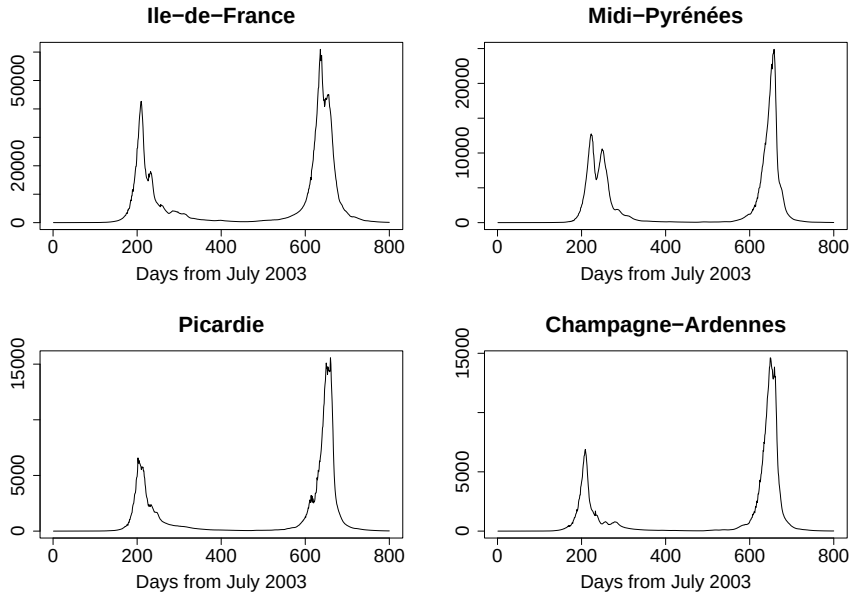




Assimilation of 2003-2005 data

Early detection of the epidemics...





Conclusion

- The particle filter is an efficient tool for the early detection of a change in epidemic state.
- Improvements are needed in the underlying SIR model and its parameterization.
- An important side-product of the system is that it provides an estimate of the total number of infected people at a given time for a given region.
- As a scenario simulator it can also provide an error estimate in the assessment of the severeness of an epidemy.

Perspectives I

Inclusion of climate parameters. Alternative filters

Knowing that a virus is likely to be already present at the beginning of a winter period:

- there is a need to characterize the meteorological configurations leading to an outbreak on the basis of the 23 years of available weekly *Sentinelles* data,
- relevant climatic parameters could be included into the epidemics' early detection system.

As the dimensionality of the system grows, the particle filter is in danger of becoming impracticable. Alternatives can be:

- the Ensemble Kalman Filter,
- the Marginalized Particle Filter (for mixed linear/non-linear state space models).

Perspectives II

Other diseases

The system can easily be adapted to handle other diseases than influenza:

- gastro-enteritis, chickenpox
  (also monitored by *Sentinelles*)
- bacterial meningitis
  (sub-sahelian zone: *meningitis belt*; China)
- . . .

Influenza epidemics

1) Short term: data assimilation by particle filtering

2) Long term (mortality data):

- extreme value analysis of epidemics

Extreme value analysis of US P&I mortality data

under consideration of demographic effects

Huey Chyi LEE, Hans WACKERNAGEL

Equipe de Géostatistique — MINES ParisTech

(with input from: Fabrice CARRAT, Magali LEMAITRE, Mark WILSON)

Motivation

- **Epidemiology** is intimately linked to **demography**.
- Studying epidemics at the scale of several decades requires a detailed analysis of the evolution of the age distribution in the same period.
- E.g. above 20 years old the P&I mortality increases exponentially with age.

Outline

Demography: exploratory analysis

- We consider the four most populated states of the United States.
- Population is subdivided into 19 age-groups.
- We consider the evolution between the 10-year censuses since 1970.
- We finally compute the average age for people 65 and older in each year.

Population: 19 age groups

00 = 0 years

01 = 1-4 years

02 = 5-9 years

03 = 10-14 years

04 = 15-19 years

...

14 = 65-69 years

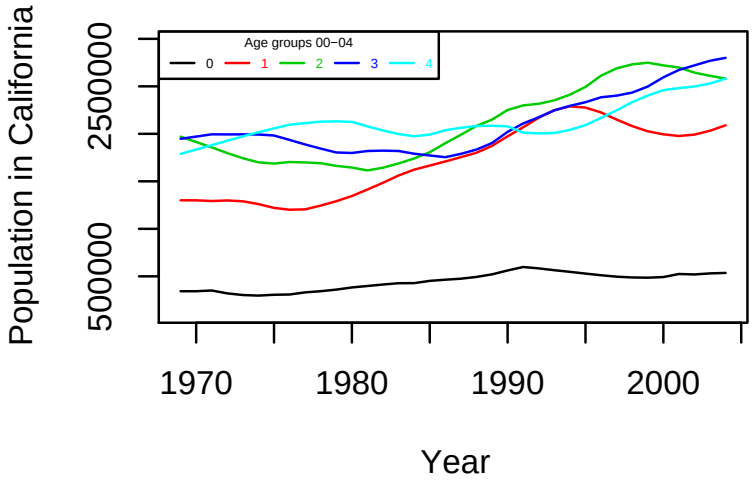
15 = 70-74 years

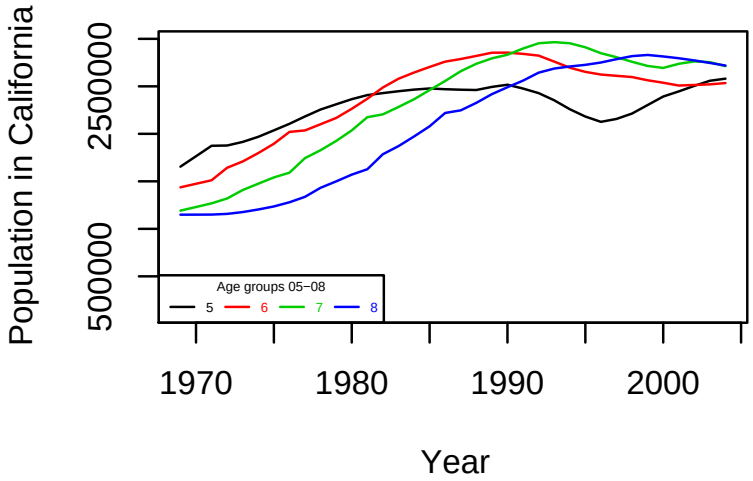
16 = 75-79 years

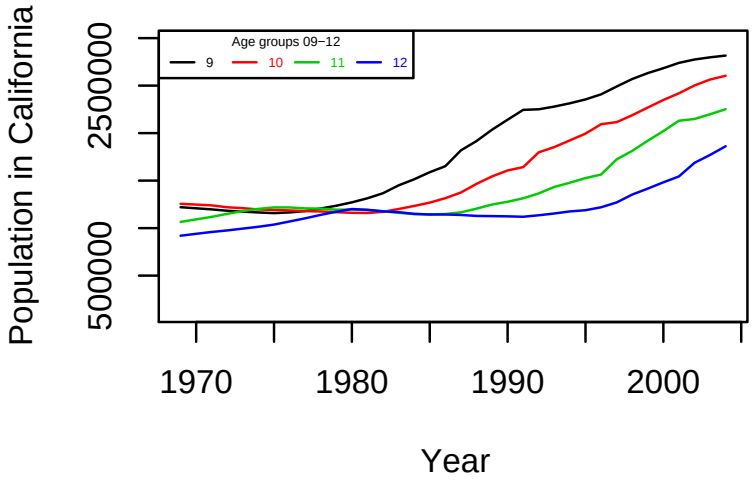
17 = 80-84 years

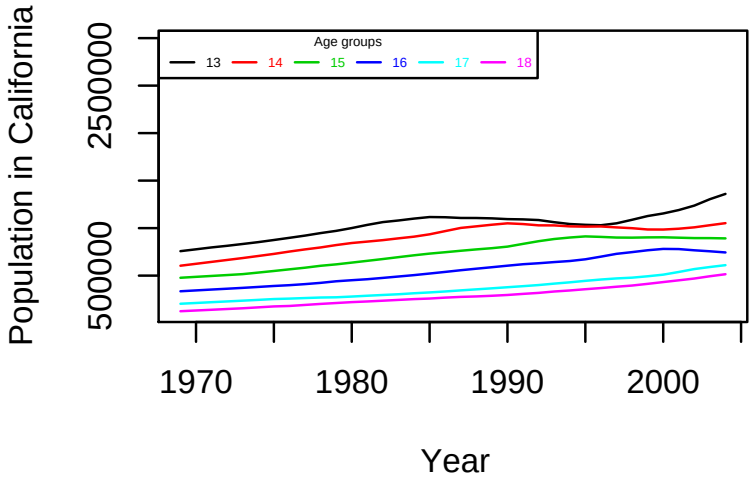
18 = 85 years and older

California population: age-group evolution



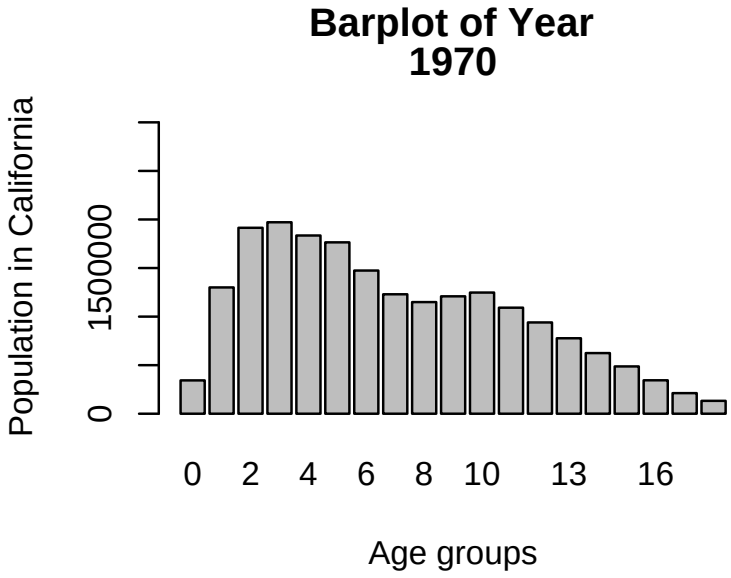


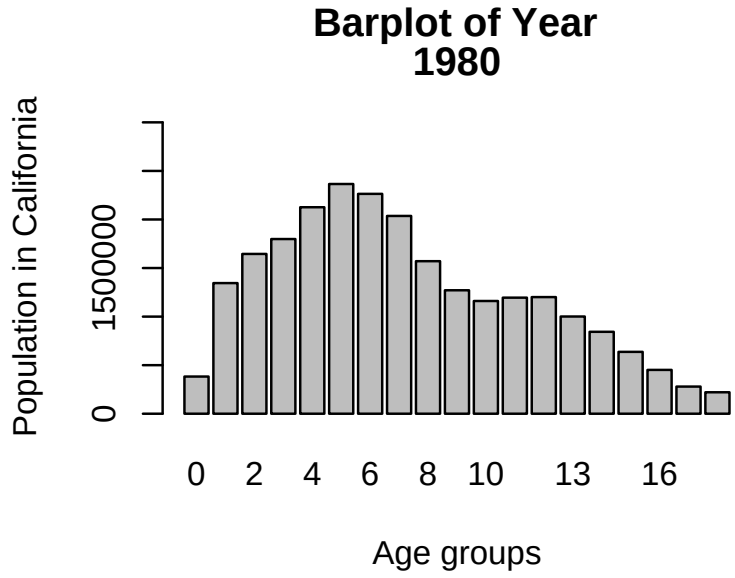


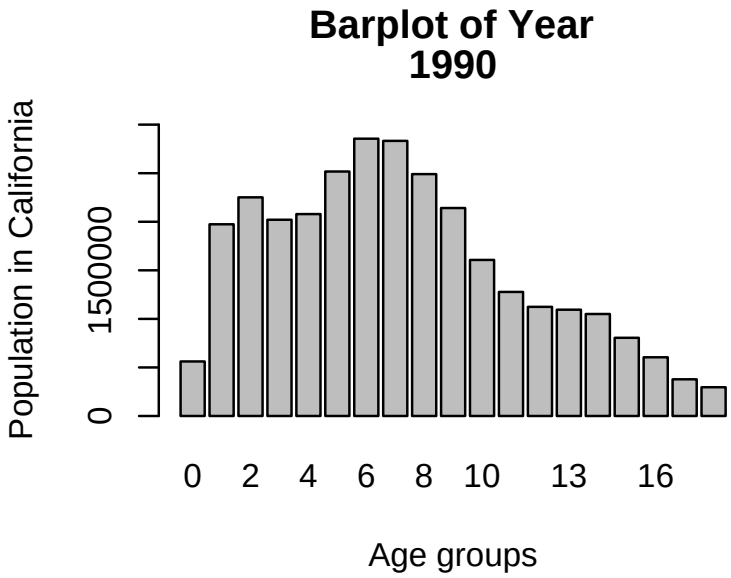


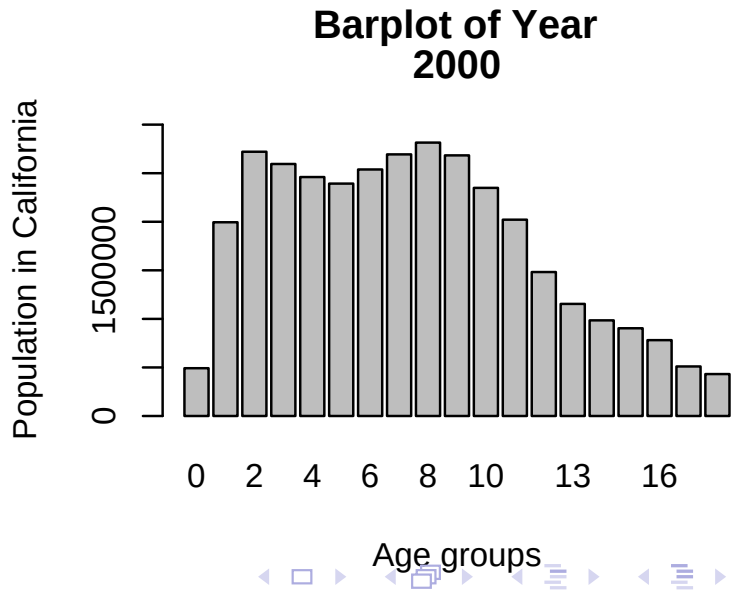
California population by age-groups

Evolution over 4 decades





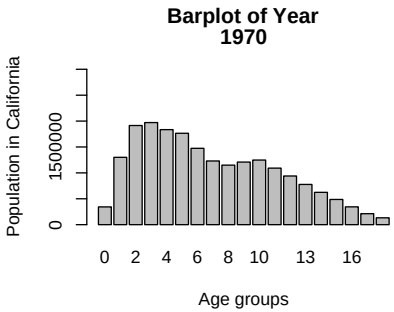


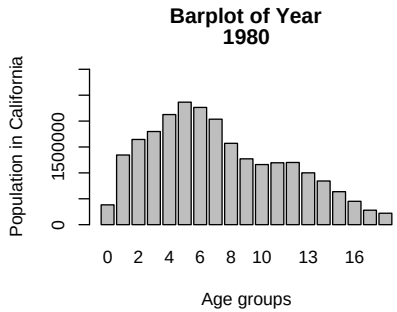


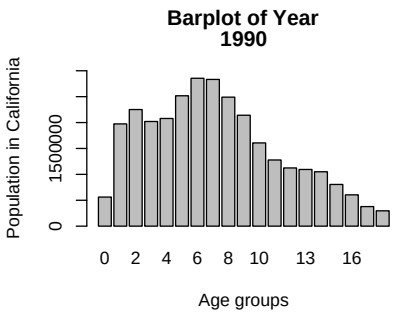
Comparing California and Texas

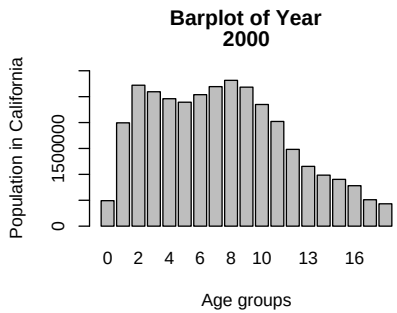
Population evolution over 4 decades

California

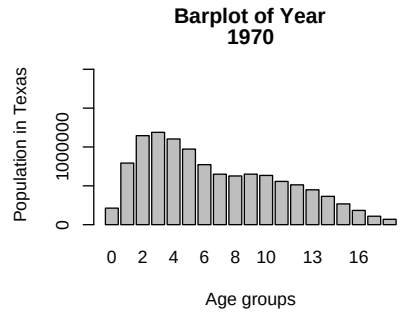


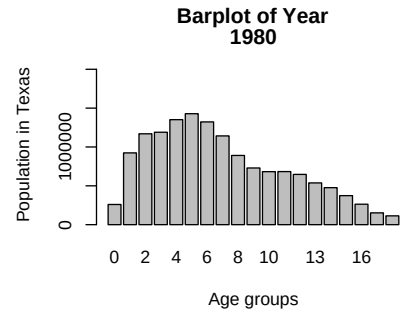


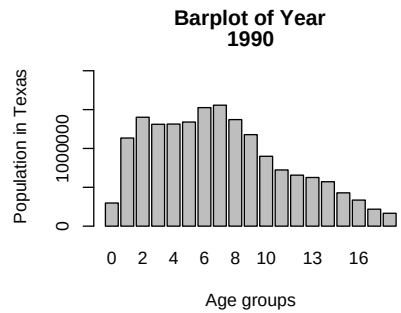


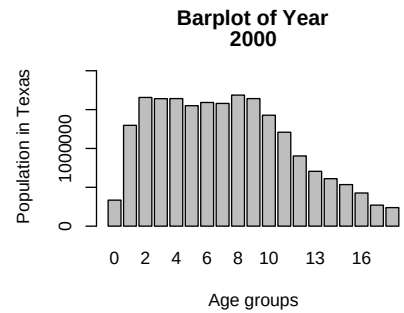


Texas





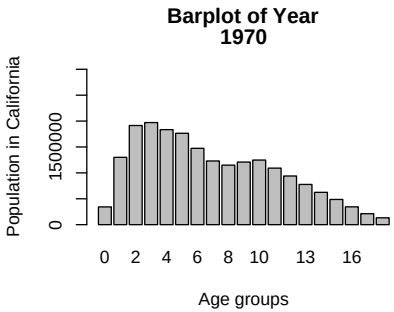


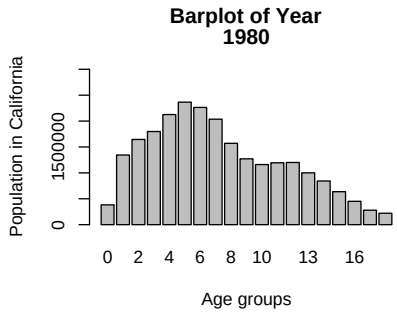


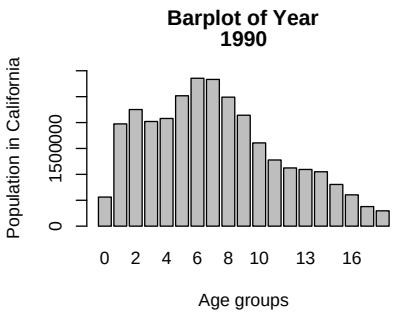
Comparing California and New York

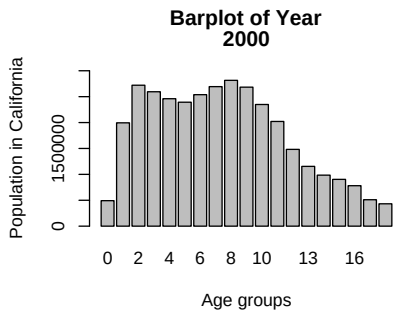
Population evolution over 4 decades

California

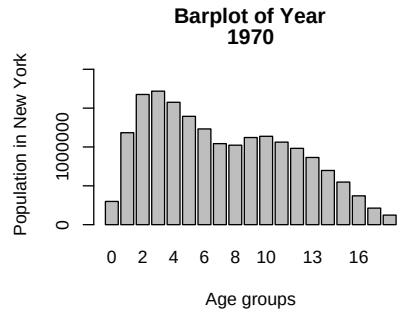


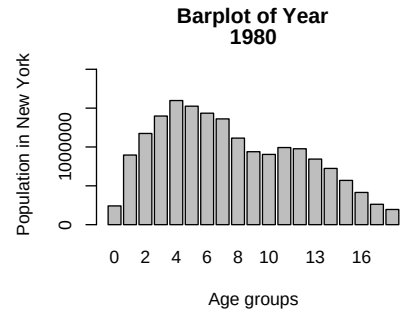


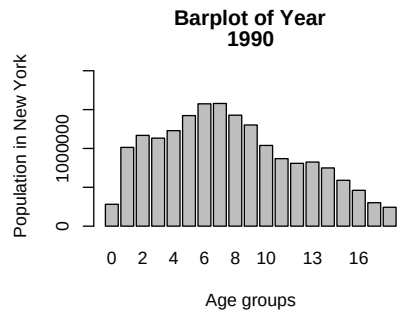


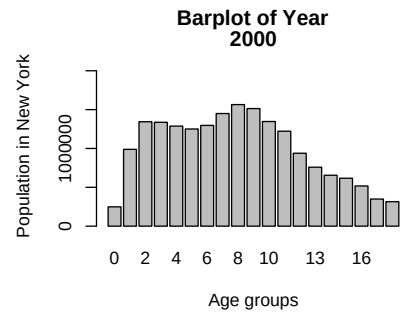


New York





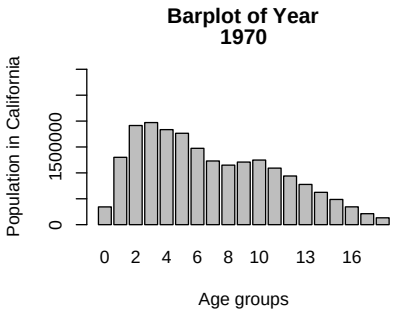


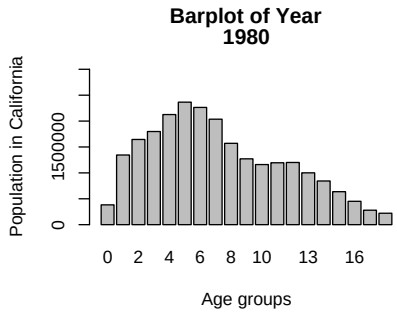


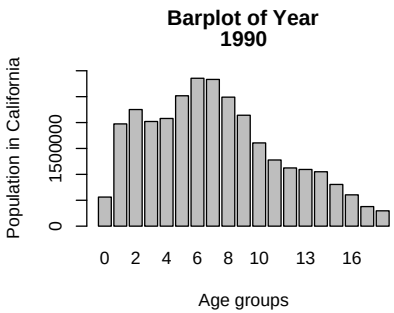
Comparison between California and Florida

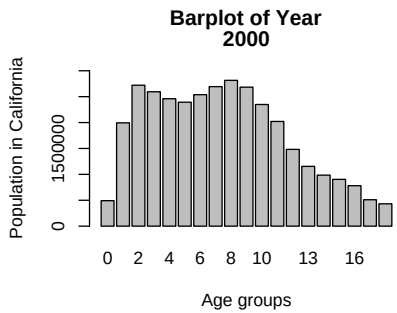
Population evolution over 4 decades

California

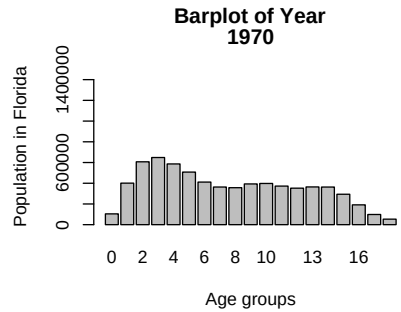


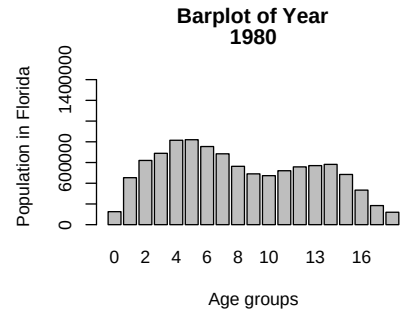


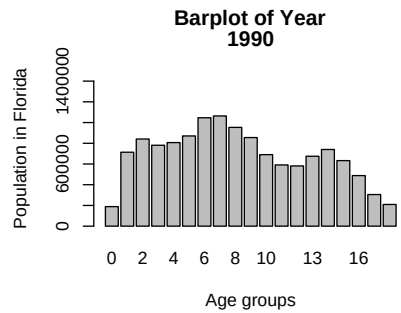


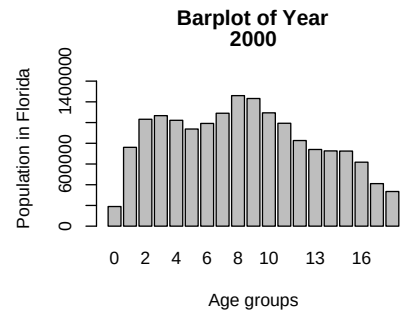


Florida

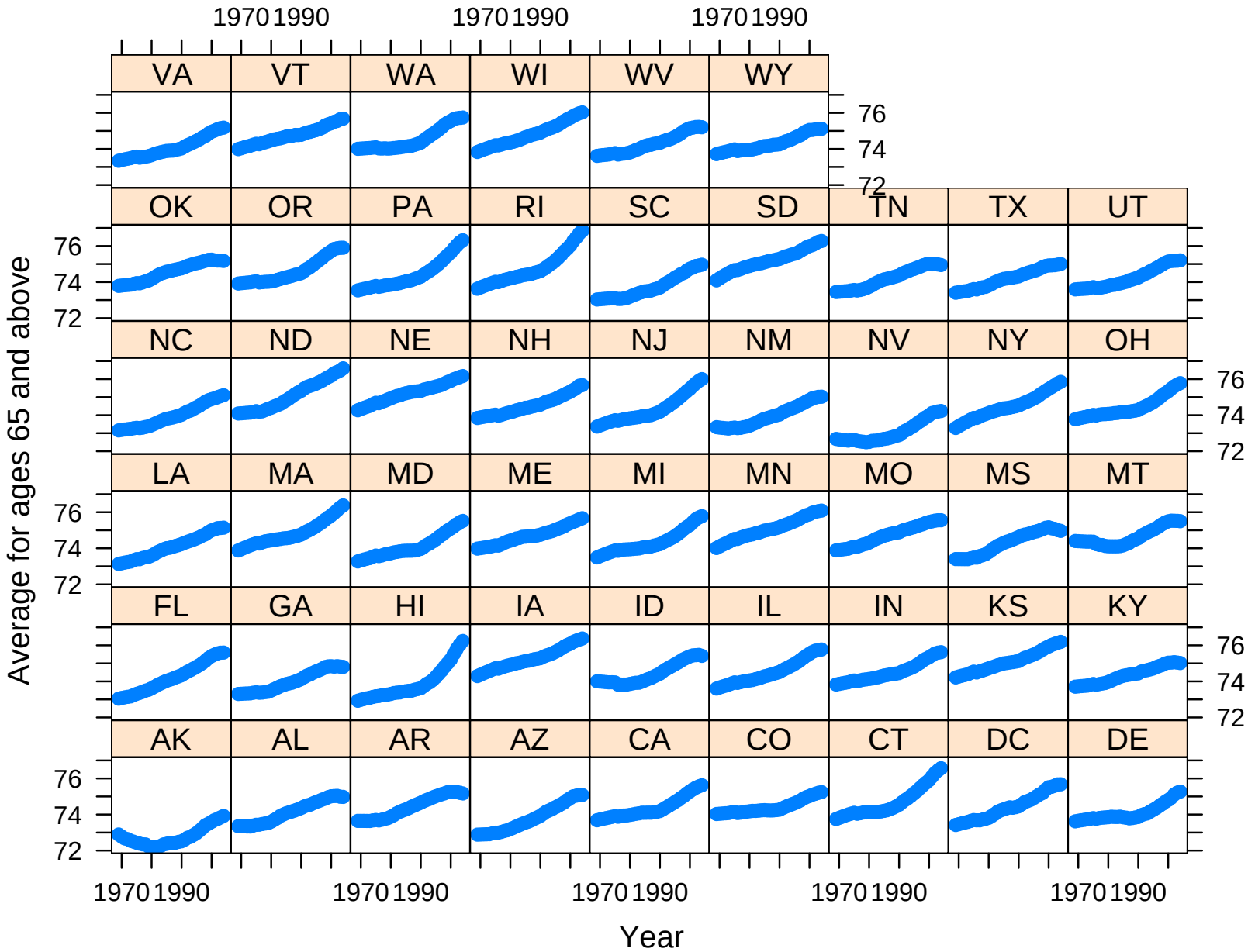








Average age for 65 and above



Mortality data

P&I proportion of total mortality

- Proportion of deaths due to Pneumonia & Influenza among the total mortality in the age group.
- The age-group analysed consists of people of age 65 and above.
- The block maxima of P&I mortality are taken in each July-June period.

- We need to take account of the effect of aging of the US population over the last decades.
- In the non-stationary model we will thus use the average-age above 65 as covariate.

Mortality data

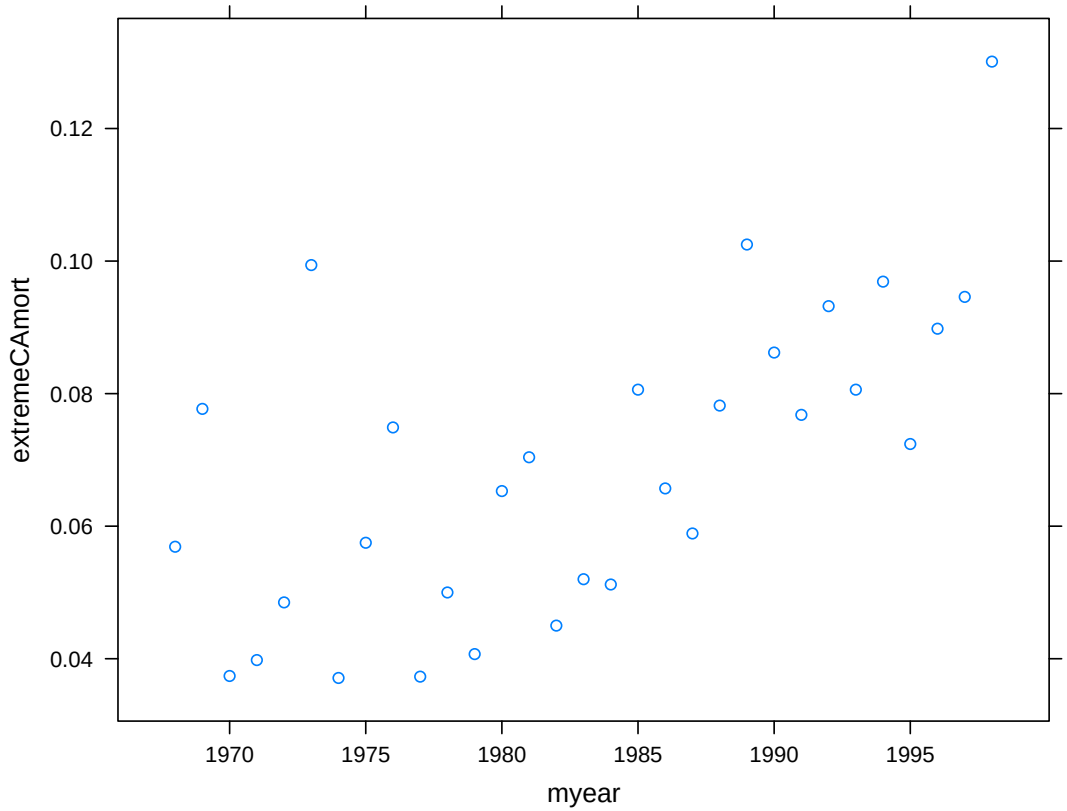
P&I proportion of total mortality

- Proportion of deaths due to Pneumonia & Influenza among the total mortality in the age group.
- The age-group analysed consists of people of age 65 and above.
- The block maxima of P&I mortality are taken in each July-June period.

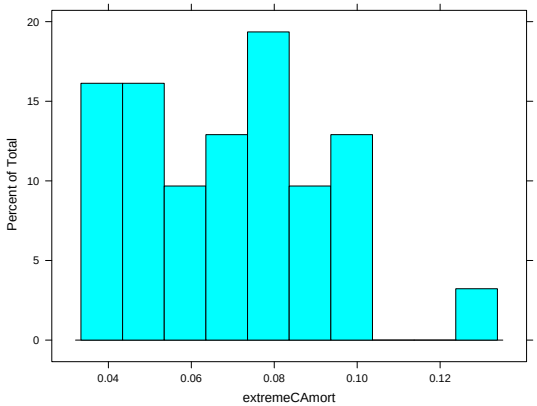
- We need to take account of the effect of aging of the US population over the last decades.
- In the non-stationary model we will thus use the average-age above 65 as covariate.

California: maxima of P&I mortality

Time series



Histogram



Motivation: extreme value analysis

- **Frequency and intensity of extremes**
  Return period:
  - how likely is the advent of an unusual weather event of a given type within the next decade/century?
  - what about a major epidemic?

- Return level:
  - what level could the event reach as compared to past events?

- **Extreme value analysis**
  A rationale to compute the probability and level of events within and beyond the range of past measurements.

Motivation: extreme value analysis

- **Frequency and intensity of extremes**
  Return period:
  - how likely is the advent of an unusual weather event of a given type within the next decade/century?
  - what about a major epidemic?

- Return level:
  - what level could the event reach as compared to past events?

- **Extreme value analysis**
  A rationale to compute the probability and level of events within and beyond the range of past measurements.

Motivation: extreme value analysis

- **Frequency and intensity of extremes**
  Return period:
  - how likely is the advent of an unusual weather event of a given type within the next decade/century?
  - what about a major epidemic?

- Return level:
  - what level could the event reach as compared to past events?

- **Extreme value analysis**
  A rationale to compute the probability and level of events within and beyond the range of past measurements.

Modeling extreme influenza epidemics

- As there is only one epidemic per year we use the block maxima approach.
- Non-stationarity in the location and scale parameters of the generalized extreme value distribution (GEV) will be considered.

Generalized Extreme Value distribution

The **generalized extreme value distribution** (family):

$$G(x) = \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)$$

defined on $\{x : 1+\xi(x-\mu)/\sigma > 0\}$.

- μ : location parameter
- σ : scale parameter
- ξ : shape parameter determines the rate of decay in the tail.

Fact

The case $\xi = 0$ corresponds to the Gumbel distribution (which has no shape parameter).

Domains of attraction

Gumbel domain ($\xi = 0$): The distribution of maxima of exponential, normal, lognormal, logistic, gamma distributed variables tends asymptotically to a **Gumbel distribution**.

Fréchet domain ($\xi > 0$): Asymptotic distribution for maxima of Pareto, Cauchy, t, F distributed variables...

Weibull domain ($\xi < 0$): Asymptotic distribution for maxima of uniform, beta, Burr distributed variables...

Extreme value paradigm

The extreme value paradigm consists in modeling the tails of a distribution F using asymptotically motivated distributions G of the maxima. It is actually not necessary to know F, the distribution of the complete data set.

Domains of attraction

Gumbel domain ($\xi = 0$): The distribution of maxima of exponential, normal, lognormal, logistic, gamma distributed variables tends asymptotically to a **Gumbel distribution**.

Fréchet domain ($\xi > 0$): Asymptotic distribution for maxima of Pareto, Cauchy, t, F distributed variables...

Weibull domain ($\xi < 0$): Asymptotic distribution for maxima of uniform, beta, Burr distributed variables...

Extreme value paradigm

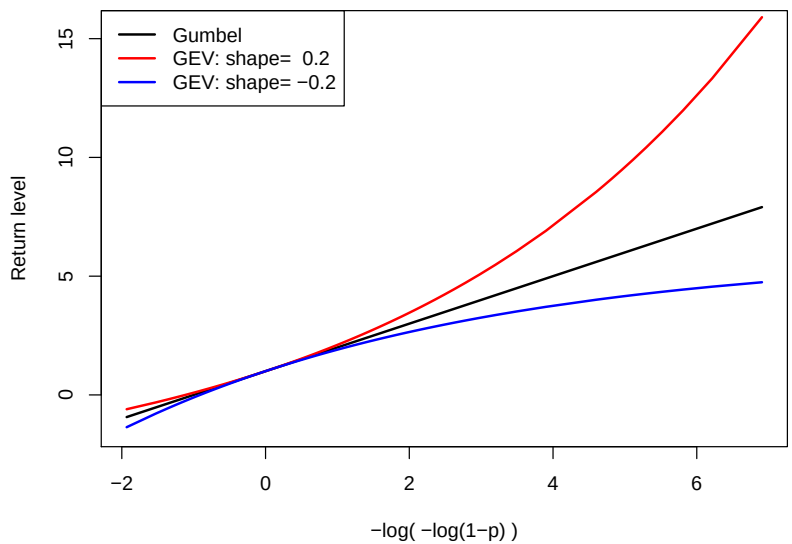
The extreme value paradigm consists in modeling the tails of a distribution F using asymptotically motivated distributions G of the maxima. It is actually not necessary to know F, the distribution of the complete data set.

Return level plot

- **Return period:**
  mean waiting time between two extreme events.
- **Return level:**
  the level associated with a return period.

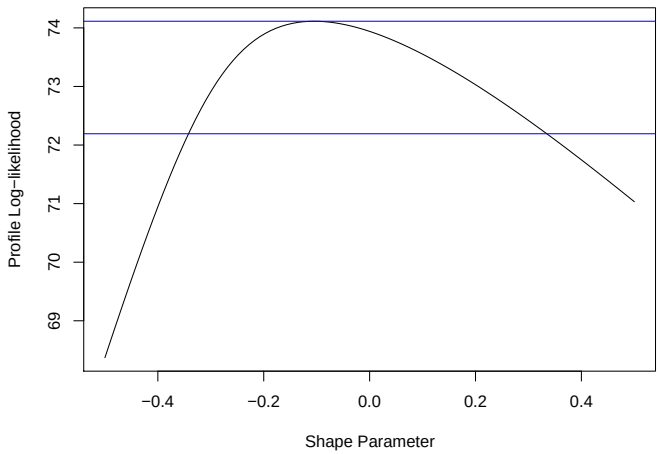
Example

A centenary flood is a flood expected to occur only once in a century.

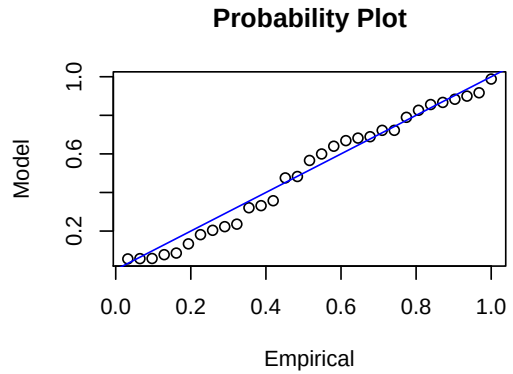


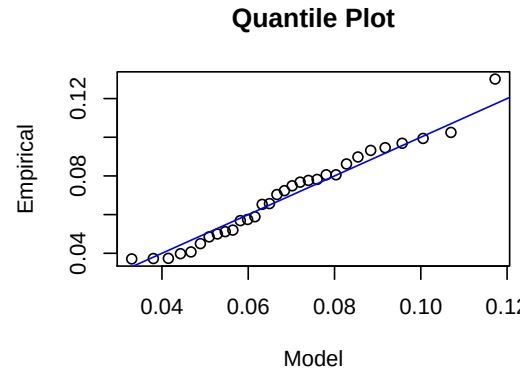
GEV fit to California P&I mortality

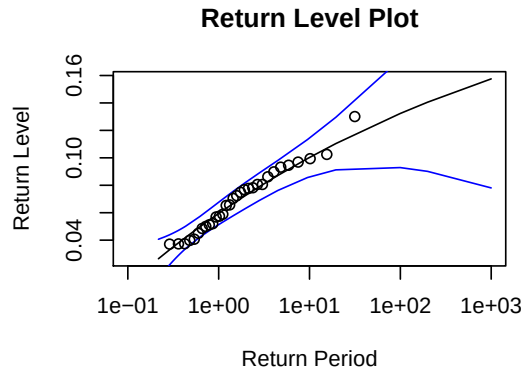
Profile likelihood for ξ

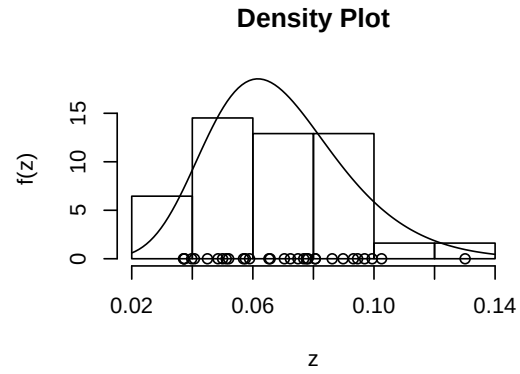


Diagnostic plots



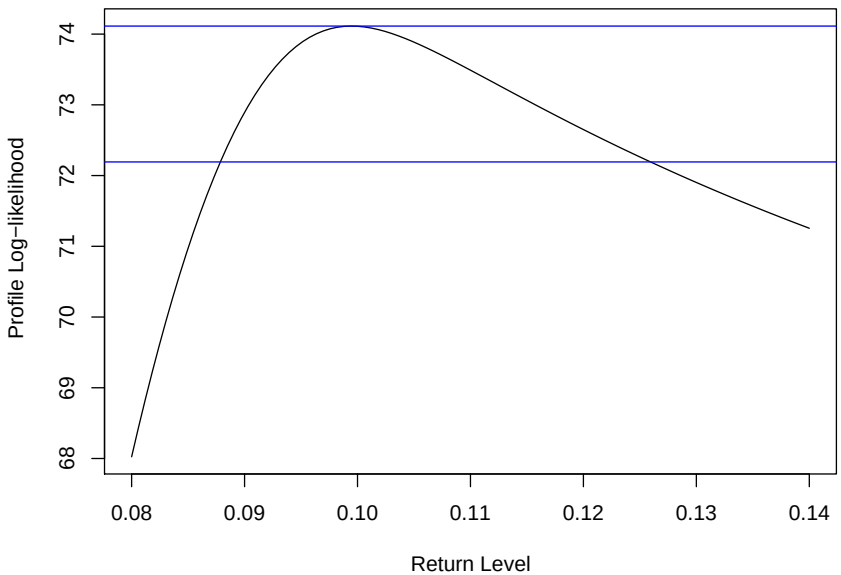




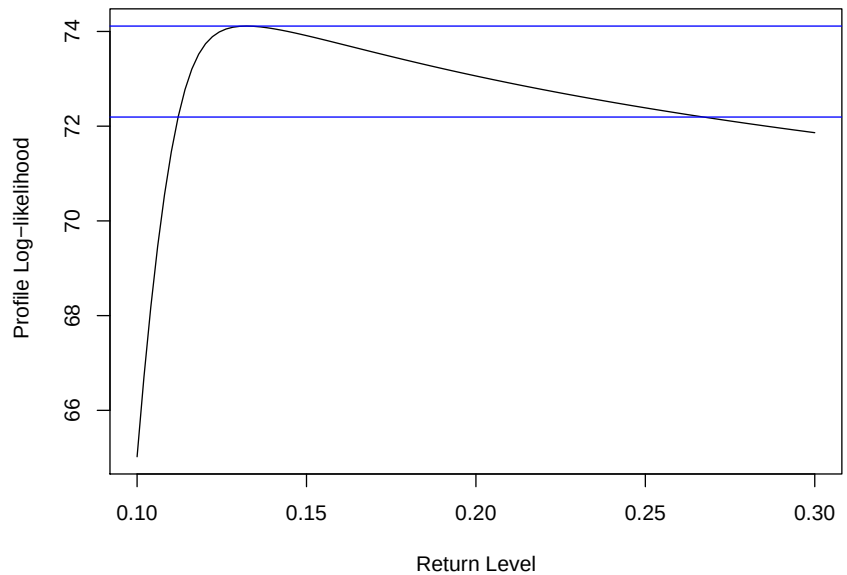


Profile likelihood for return levels

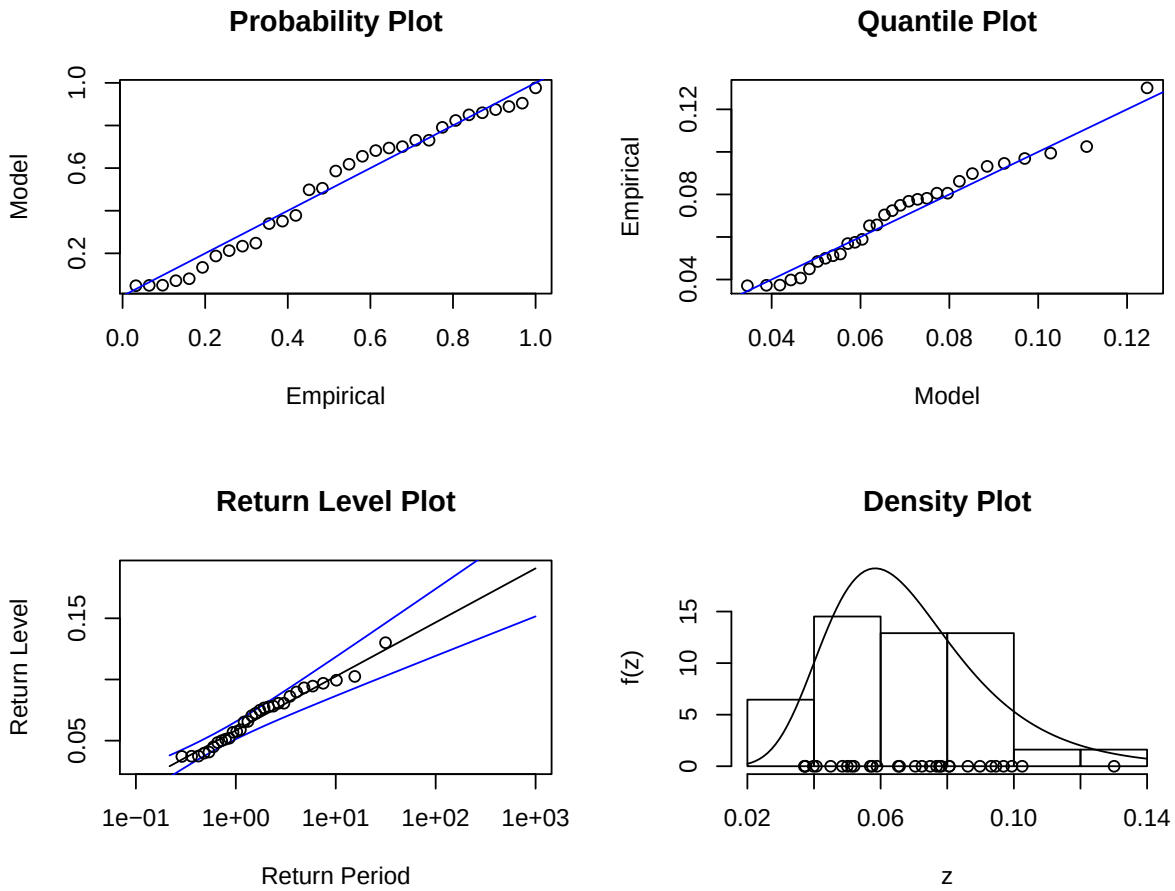
10-year return level



100-year return level



Gumbel model for California P&I mortality



- Considering $\xi = 0$ provides also a good fit.
- The confidence intervals for the return levels are much narrower ! (Parsimonious model: one parameter less!)

Non-stationary extreme value modeling

Introducing covariates: year as well as average-age above 65

Model 1: linear trend for **location parameter**.

Model 2: linear trend for **location parameter** and exponential trend for **scale parameter**.

Model 3: linear trend and linear dependence on average-age for **location parameter**.

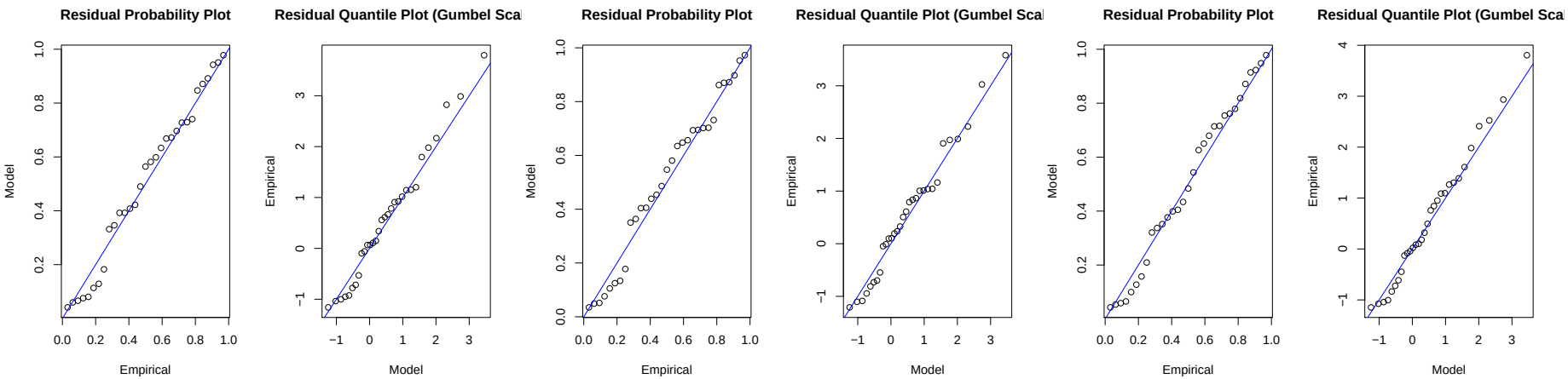
Summary table: California non-stationary fit

	Model 1: μ depends linearly on year	Model 2:σ depends exponentially on year	Model 3: Linear dependence of μ on year and average age
Calculated Maximum likelihood estimates	$\mu = 0.06 + 0.19 \cdot year$ σ=0.01 ξ=0.1	$\mu = 0.06 + 0.2 \cdot year$ $\sigma = \exp(-4.4 + 1.1 \cdot year)$ ξ=0.2	$\mu = -1.66 + 0.1 \cdot year - 0.02 \cdot age$ σ=0.01 ξ=0.08
Negative log-likelihood	-84.9	-84.2	-85.4
Deviance from GEV model	D_1=2{84.9-74.1} =21.6	D_2=2{84.2-74.1} =20.2	D_3=2{85.4-74.1} =22.6

Residual diagnostic plots for California

The 3 non-stationary models

Model 1 Model 2 Model 3



Conclusion: non-stationary model

- **California P&I mortality** is best fitted with the non-stationary GEV Model 2:
  - location parameter μ depends linearly on time,
  - scale parameter σ depends exponentially on time.
- That model seems to provide better results than the inclusion of average-age ≥ 65 as covariate.

Perspectives

This *preliminary study* is now being improved for the following aspects:

- Use P&I mortality rate
  (instead of P&I proportion of total mortality).
- Consider several P&I mortality age-groups above 65.
- Duplicate study in France.
- Go back until 1945 to include the 1957 (asian flu) epidemic.
- Consider a spatially non-stationary model.

References

CAPPÉ, O., MOULINES, E., AND RYDEN, T.
Inference in Hidden Markov Models.
Springer, 2005.

COLES, S.
An Introduction to Statistical Modeling of Extreme Values.
Springer, London, 2001.

JÉGAT, C., CARRAT, F., LAJAUNIE, C., AND WACKERNAGEL, H.
Early detection and assessment of epidemics by particle filtering.
In *GeoENV VI – Geostatistics for Environmental Applications* (2008), A. Soares, M. J. Pereira, and R. Dimitrakopoulos, Eds., Springer, pp. 23–35.