Centile estimation using GAMLSS

Mikis Stasinopoulos[1] Bob Rigby[1]

[1]STORM, London Metropolitan University

GAMLSS: Short course Campina Grande, Brazil, Jul 2013

gamlss

gamlss

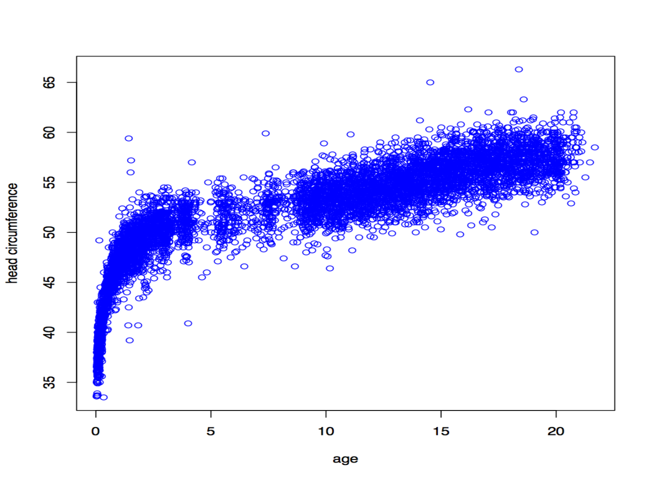
The Dutch boys data

`head` : the head head circumference of 7040 boys

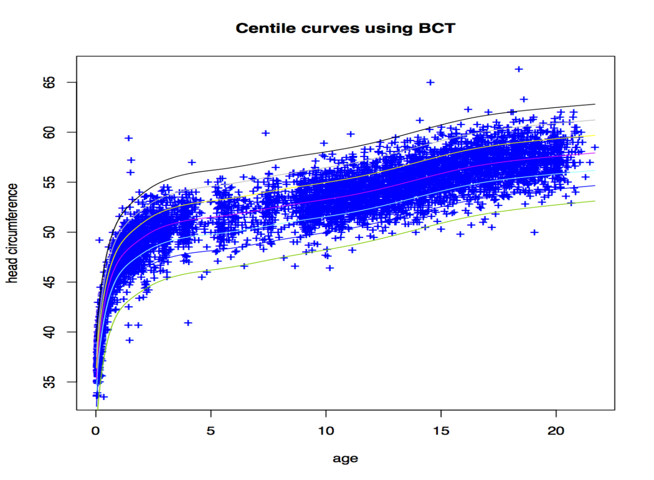
`age` : the age in years

Source: Buuren and Fredriks (2001)

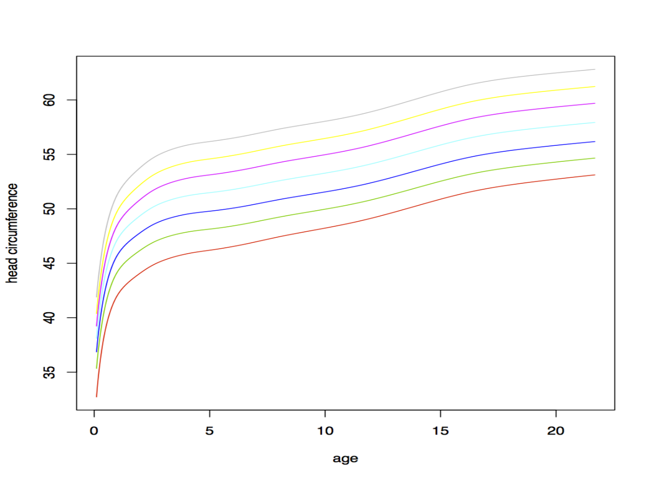
The Dutch boys data



The Dutch boys data with centiles



The Dutch boys data: centiles



The methods

i) the non parametric approach of quantile regression (Koenker, 2005; Koenker and Bassett, 1978)

ii) the parametric LMS approach of Cole (1988), Cole and Green (1992) and its extensions Rigby and Stasinopoulos (2004, 2006, 2007).

Centiles: the model

$Y \sim f_Y(y|\mu, \sigma, \nu, \tau)$ where $f_Y()$ is any distribution

$Y =$head circumference and $X = AGE^{\xi}$

$$\begin{aligned} \mu &= s(x, df_\mu) \\ \log(\sigma) &= s(x, df_\sigma) \\ \nu &= s(x, df_\nu) \\ \log(\tau) &= s(x, df_\tau) \end{aligned}$$

The LMS method and extensions

Let Y be a random variable with range $Y > 0$ defined through the transformed variable Z given by:

$$
\begin{aligned}
Z &= \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], \quad \text{if} \quad \nu \neq 0 \\
&= \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right), \quad \text{if} \quad \nu = 0.
\end{aligned}
$$

1. if $Z \sim N(0,1)$ then $Y \sim BCCG(\mu, \sigma, \nu)$ = LMS method
2. if $Z \sim t_\tau$ then $Y \sim BCT(\mu, \sigma, \nu, \tau)$ = LMST method
3. if $Z \sim PE(0,1,\tau)$ then $Y \sim BCPE(\mu, \sigma, \nu, \tau)$ = LMSP method adopted by WHO

Centiles: estimation of the smoothing parameters

We need to select the five values $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$

- by trial and error
- minimize the generalized Akaike information criterion, GAIC($\sharp$)
- minimize the the validation global deviance VGD
- using local selection criteria, i.e. CV, ML

Diagnostics should be used in all above cases

Centiles: GAIC($\sharp$)

For different values of $\sharp$ minimize GAIC($\sharp$) to find estimates for $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$

- $\sharp = 2$, AIC
- $\sharp = 3$, our preference
- $\sharp = 3.84$, χ^2
- $\sharp = log(n)$, BIC, SBC

The gamlss functions `find.hyper()` can be used

Centiles: Validation Global Deviance

Divide the data into two components 60% and 40%

Use the first component for training (fitting)
Use the second component for estimating the parameters $df_\mu, df_\sigma, df_\nu, df_\tau, \xi$

The gamlss functions `VDG()` in combination of the R function `optim()` can be used

gamlss

Centiles: Local selection criteria

- GCV
- GAIC
- ML
- EM

The gamlss smoothing functions `pb()` uses ML as default
The gamlss smoothing functions `ga()` is an interface for the `gam()` function of Simon Wood and therefore uses GCV.

GAMLSS functions for centile estimation

`centiles()` to plot centile curves against an x-variable.

`centiles.com()` to compare centiles curves for more than one `object`.

`centiles.split()` as for `centiles()`, but splits the plot at specified values of x.

`centiles.fan()` fan plot as in `centiles()`.

`centiles.pred()` to predict and plot centile curves for new x-values.

`fitted.plot()` to plot fitted values for all the parameters against an x-variable

gamlss

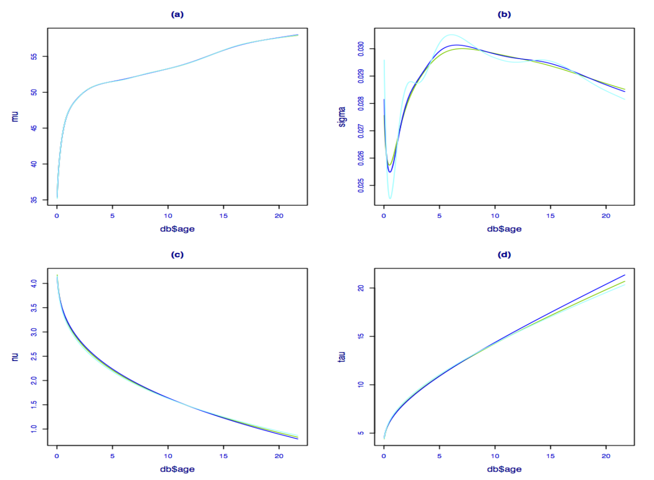
The head circumference data: results

Method	df_μ	df_σ	df_ν	df_τ	ξ	$GAIC(3)$
GAIC(3)	12.3	5.7	2	2	0.33	26811.7
VGD	15.8	8.1	2	2	0.28	26815.6
local ML	13.0	4.9	2	2	0.30	26811.1

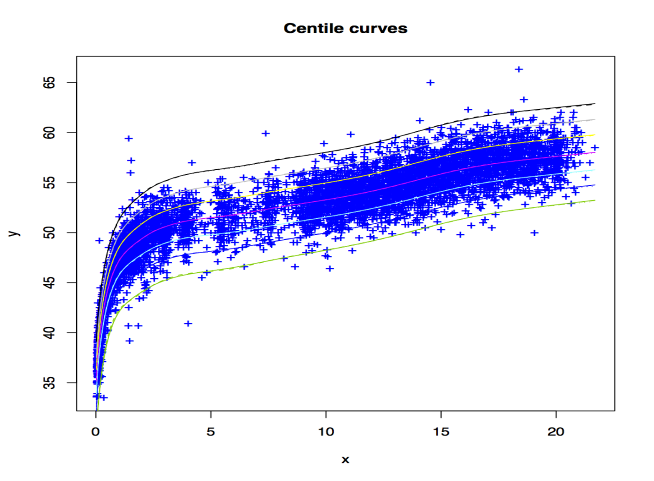
Table : The estimated degrees of freedom

gamlss

The head circumference data: `fitted.plot()`

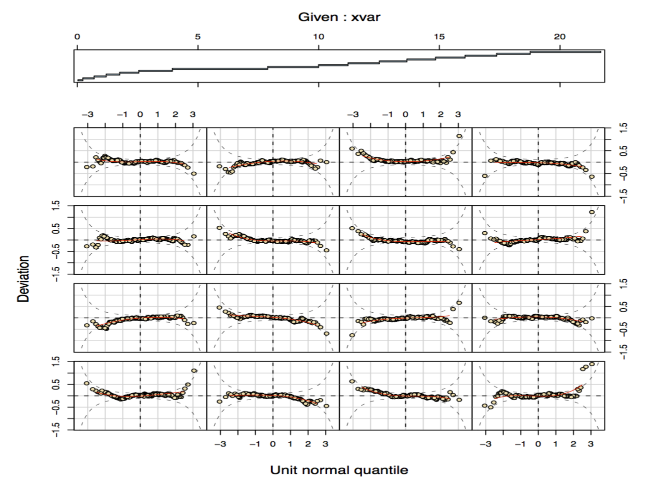


The head circumference data: comparison of the centiles, `centiles.com()`



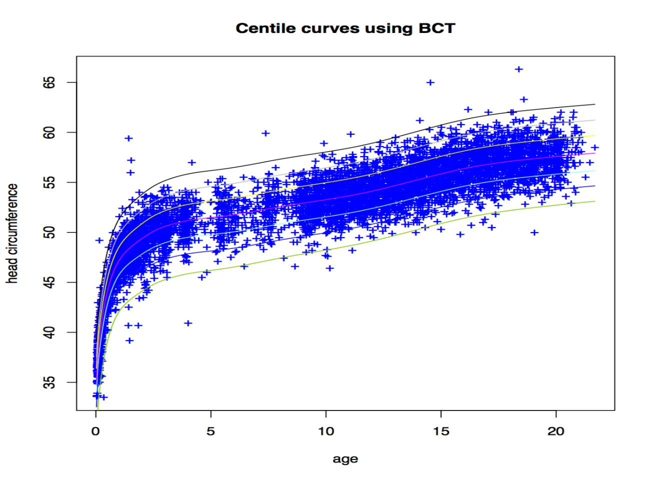
gamlss

The head circumference data: The worm plots, `wp()`

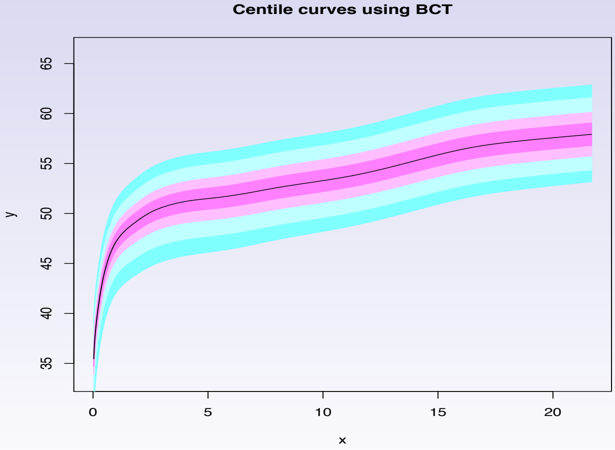


gamlss

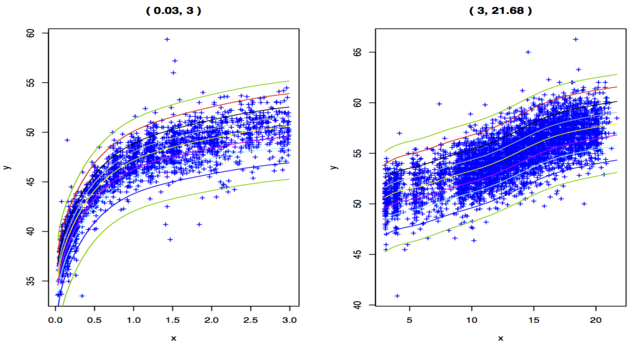
The head circumference data: `centiles()`



The head circumference data: `centiles.fan()`



The head circumference data: centiles.split()

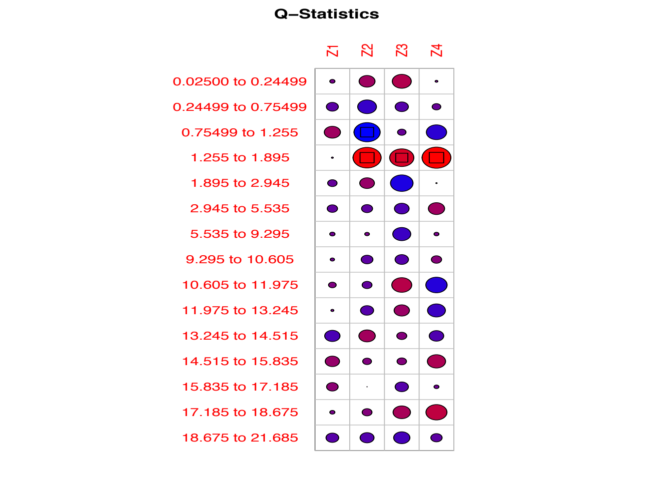


gamlss

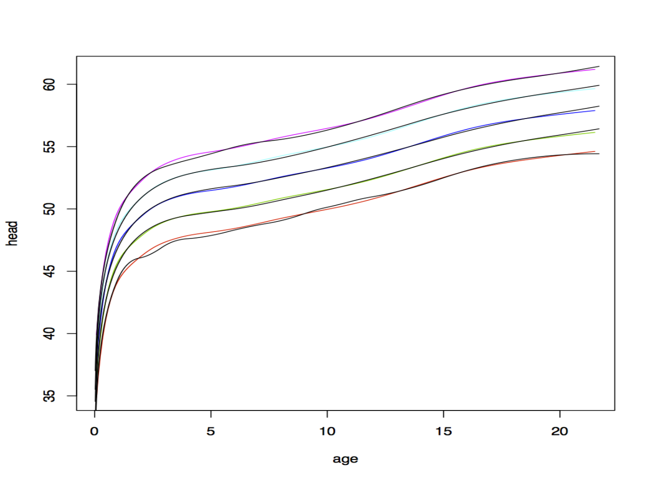
The head circumference data: The Q statistics

	Z1	Z2	Z3	Z4	Agostino	N
0.02500 to 0.24499	0.09	0.85	1.21	0.02	1.47	477
0.24499 to 0.75499	-0.48	-1.18	-0.58	-0.25	0.40	473
0.75499 to 1.255	0.88	-2.27	-0.24	-1.37	1.94	467
1.255 to 1.895	0.01	2.62	1.96	2.76	11.47	460
1.895 to 2.945	-0.29	0.72	-1.70	-0.01	2.90	473
2.945 to 5.535	-0.35	-0.40	-0.72	0.88	1.30	466
....						
15.835 to 17.185	0.45	0.00	-0.59	-0.08	0.35	466
17.185 to 18.675	0.09	0.32	1.01	1.48	3.23	472
18.675 to 21.685	-0.54	-0.65	-0.88	-0.42	0.96	467
TOTAL Q stats	2.85	16.90	15.43	18.04	33.46	7040
df for Q stats	1.95	12.01	12.35	13.00	25.35	0
p-val for Q stats	0.23	0.15	0.24	0.16	0.13	0

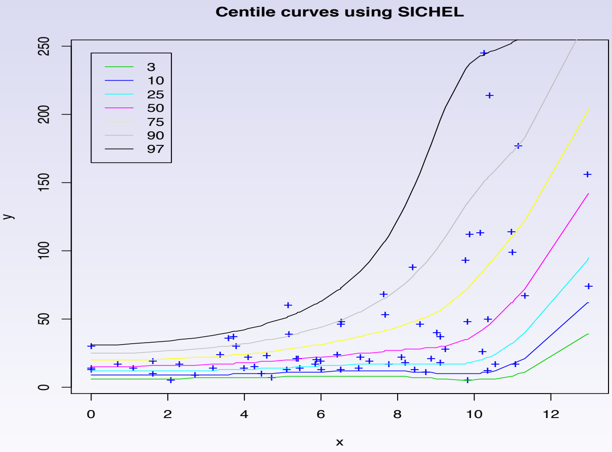
The head circumference data: `Q.stats()`



Comparing GAMLSS and QR: fitted centiles



Discrete distributions: fitted centiles



Further work

- model selection model selection
- time series models time series models
- peak oil peak oil
- fitting the tail of the distribution tails

gamlss

Conclusions: GAMLSS

- Unified framework for univariate regression type of models
- Allows any distribution for the response variable Y
- Models all the parameters of the distribution of Y
- Allows a variety of additive terms in the models for the distribution parameters
- The fitted algorithm is modular, where different components can be added easily
- Models can be fitted easily and fast
- Explanatory tool to find appropriate set of models
- It deals with overdispersion, skewness and kurtosis

gamlss

Rigby, R. A. and Stasinopoulos D. M. (2004). Smooth centile curves for skew and kurtotic data modelled ..., *Statistics in Medicine*, **23**, pp 3053-3076.

Rigby, R. A. and Stasinopoulos D. M. (2005). Generalized Additive Models for Location, Scale and Shape, (with discussion). *Appl. Statist.*, **54**, pp 507-554.

Rigby, R. A. and Stasinopoulos D. M. (2006). Using the Box-Cox t distribution in GAMLSS ... *Statistical Modelling*, **6**, pp 209-229.

Stasinopoulos D. M. Rigby R.A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R, *Journal of Statistical Software*, Vol. **23**, Issue 7.

Rigby R. A., Stasinopoulos M.D. and Akantziliotou C. (2008) A framework for modelling overdispersed count data, including ... *CSDA*, 53, pp 381-393.)

gamlss

End

END

for more information see

www.gamlss.org

gamlss

Time series and GAMLSS

What is available:

- Modelling μ
  - ARIMA (for normal errors)
  - Structural time series models (for normal errors)
  - GARMA and others (for exponential family)
  - Hidden Markov Chain models
- Modelling σ
  - GARCH and its by product (for normal and long tails)
  - Stochastic volatility models
  - Markov-Switching Multifractal models

What is suitable for GAMLSS? go back next page

gamlss

GARMA

What is our experience?

GARMA

- We have expand the GARMA model to accept any `gamlss.family` distributions
- Modelling μ only
- Not clear how lags should define in the case of ν and τ
- No explicit stochastic model

go back next page

gamlss

Hidden Markov Chain

go back · next page

A Latent-State Generalized Additive Model for Location, Scale and Shape for Multivariate Pollutant Concentrations

Antonello Maruotti
Universitá di Roma Tre, Roma, Italy.
Walter Zucchini
Institut für Statistik und Ökonometrie, Göttingen, Germany.

Summary. XXX YYY ZZZ

Keywords: GAMLSS, latent-state models, multivariate time series, pollutant concentrations.

1. Introduction

The model developed in this paper was motivated by the analysis of a daily multivariate time series of pollutant concentrations taken at a monitoring station in Rome, Italy. The objective is to quantify the relationship between pollutant levels and the meteorological factors that affect them.

The study and monitoring of environmental phenomena often involves the analysis of large multivariate time series. The growing interest in air quality has led to a substantial literature. One branch of this is concerned with synthetic environmental indices designed to summarize a series of measurements (Barnett and Bown, 2002; Saisana et al., 2005; Bellini

Markov-Switching Multifractal Models

go back next page

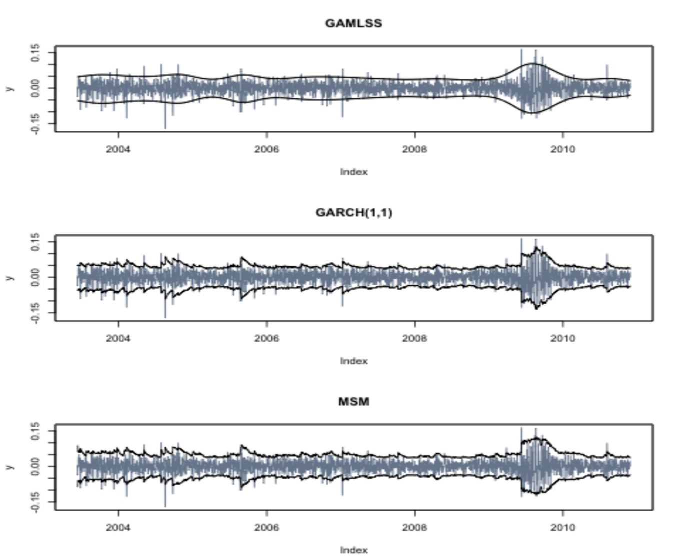
Markov-Switching Multifractal Models

- We have used them to model σ
- Very slow especially larger than 7 number of stages
- We have not find them to be an improvement compare to other models
- There is explicit stochastic model

go back next page

Oil returns

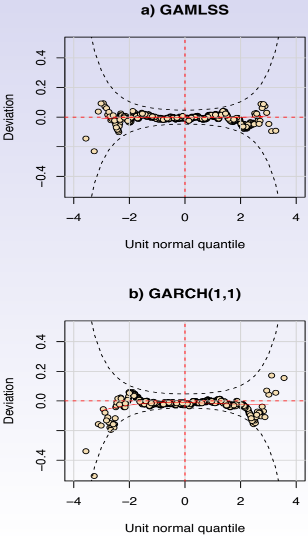
go back next page



gamlss

Oil returns

go back next page



Structural/Stochastic Volatility models

Structural/Stochastic Volatility models

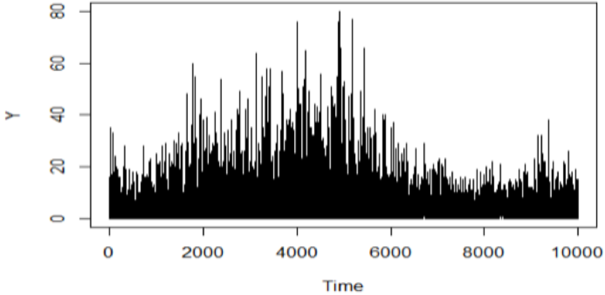
- A proper stochastic model exist
- This is the most promising approach at the moment
- It uses sparse matrices (rather than Kalman filter)
- There are difficulties to estimate the smoothing parameters but we have done good progress go back next page

gamlss

Simulated negative binomial type I

go back next page

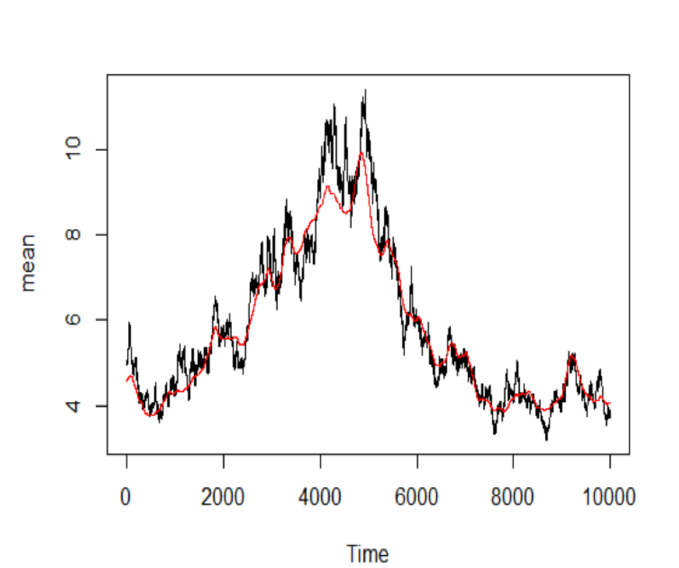
$$
\begin{aligned}
g_1(\mu_t) &= \eta_{\mu,t} = \eta_{\mu,t-1} + b_{\mu,t} \\
g_2(\sigma_t) &= \eta_{\sigma,t} = \eta_{\sigma,t-1} + b_{\sigma,t}
\end{aligned} \qquad (1
$$



gamlss

Negative binomial fit for the mean

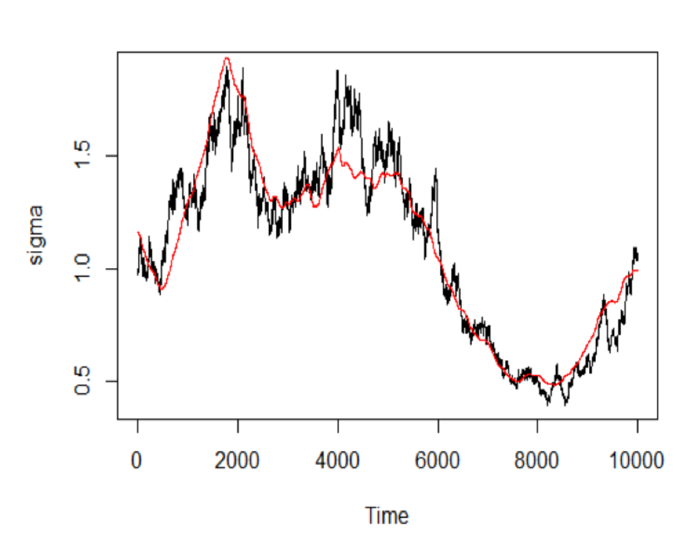
go back next page



gamlss

Negative binomial fit for sigma

go back next page



gamlss

GAMLSS and time series

go back

- GEST= GAMLSS+structural models
- The GEST approach looks very promising

Peak Oil

"*Peak Oil* is the point in time when the maximum rate of petroleum extraction is reached, after which the rate of production is expected to enter terminal decline", (Wikipedia).

go back next page

Peak Oil: scenarios FURTHER

go back next page

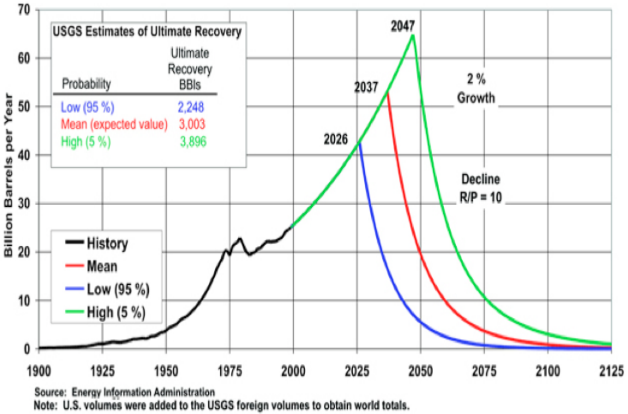


Figure 2: Discrete scenarios based on finite number of uncertainties[23]

gamlss

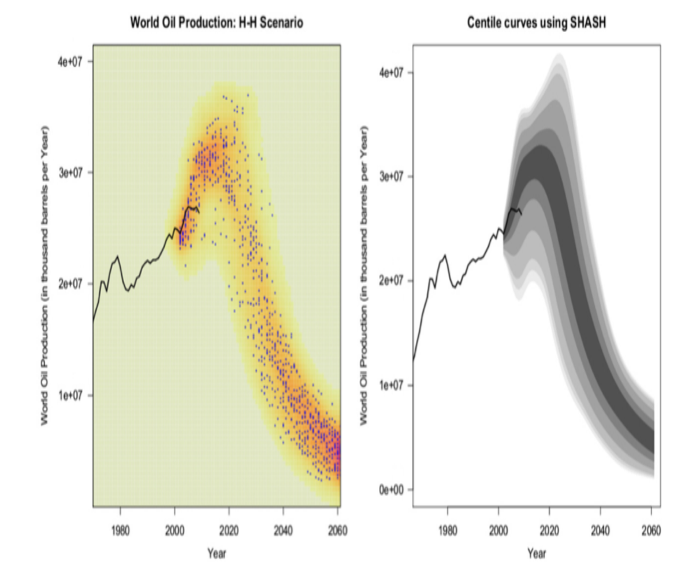
ACEGES models

- ACEGES Stands for Agent-based Computational Economics of the Global Energy System.
- It is an agent-based modelling simulation
- It tries to understand the outlook for oil production by accounting for uncertainties in resource estimates, demand growth, production growth and peak/decline point.
- People: Vlasis Voudouris, Ken'ichi Matsumoto, Bob Rigby, Paul Eilers, Michael Jefferson, John Sedgwick, Carlo Di Maio
- Interests: Oil, Gas, Electricity, Renewable energy

go back next page

Peak Oil: scenario one

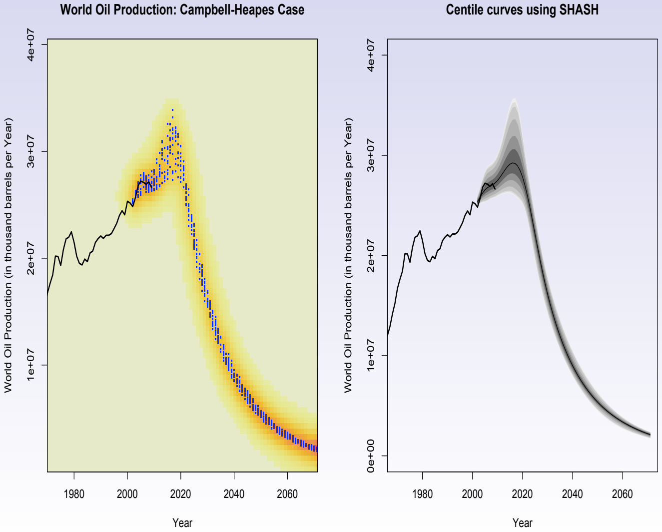
go back next page



gamlss

Peak Oil: scenario two

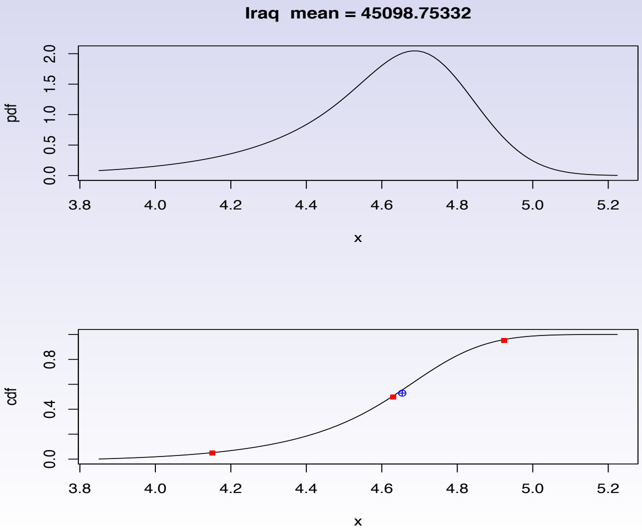
go back next page



gamlss

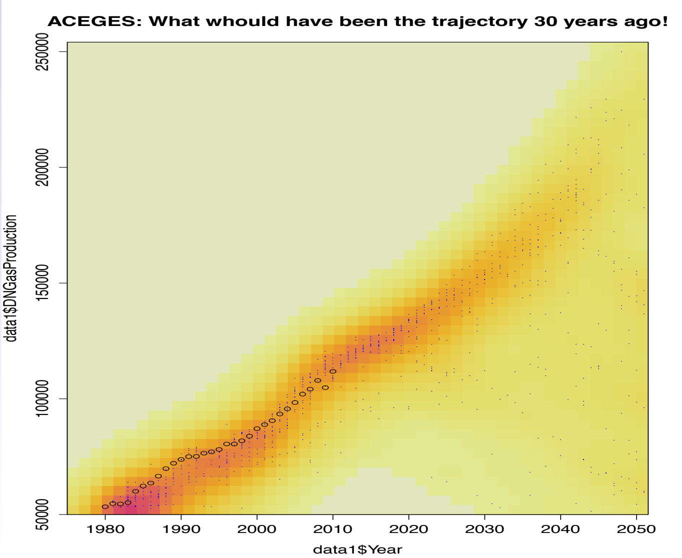
Peak Oil: estimated ultimate recovery EUR

go back
next page



Natural gas

go back next page



gamlss

Peak Oil

- GAMLSS is used as a by product here
- ACEGES: statistical improvements with the simulations
- Can we use statistical inference to make the system self-learning?

go back

Selection of explanatory variables

How to select explanatory variables?

	x_1	x_2	x_3	x_4	x_5	x_6
μ						
σ		✓				
ν						
τ						

- Strategy A
- Strategy B
- Other strategies?
- Boosting

strategy A

Strategy A:

- It starts with a forward stepwise selection using GAIC.
- Each x variables is set for selection first for μ then for σ, ν and τ
- then it does a backward elimination for ν, σ and μ.

	x_1	x_2	x_3	x_4	x_5	x_6
μ	×		×	×		×
σ			×	×		
ν	×		×			
τ				×		

strategy B

Strategy B: forward stepwise selection using GAIC in which an x variable is selected for all the parameters

Table : selecting explanatory variables

	x_1	x_2	x_3	x_4	x_5	x_6
μ	$\times$		$\times$			$\times$
σ	$\times$		$\times$			$\times$
ν	$\times$		$\times$			$\times$
τ	$\times$		$\times$			$\times$

go back next page

gamlss

Question one: boosting

GAMLSS for high-dimensional data – a flexible approach based on boosting

Andreas Mayr*[1], Nora Fenske[2], Benjamin Hofner[1], Thomas Kneib[3], Matthias Schmid[1]

[1] Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2] Institut für Statistik, Ludwig-Maximilians-Universität München, Germany
[3] Institut für Mathematik, Carl von Ossietzky Universität Oldenburg, Germany

go back

Fitting the tail of the distribtion

- Is the distribution fits well in both the centre and the tail of the distribution?
- Tails are important for VaR (value at risk) and ES (expected shortfall)
- Quantile and expectile regressions are less reliable in the tail of the distribution

go back next page

gamlss

Fitting the tail of the distribution: procedure

The procedure is:

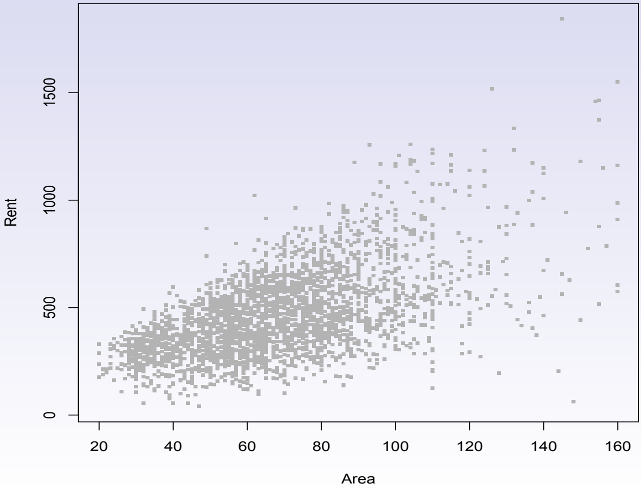
1. Find the α quantile by fitting an appropriate quantile or GAMLSS model from which we obtain the observations in the upper tail of the distribution.
2. Fit a suitable truncated distribution to the tail data where the truncation parameter is the fitted α quantile above.
3. Obtain the β quantile of the truncated distribution which corresponds to the $\tau = \alpha + \beta(1 - \alpha)$ quantile of the original data

go back next page

gamlss

Fitting the tail of the distribution: the data,

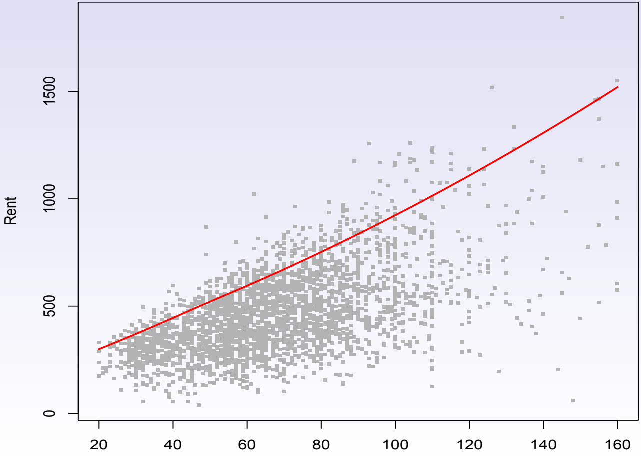
go back next page



Fitting the tail of the distribution: Fitted 90%

go back

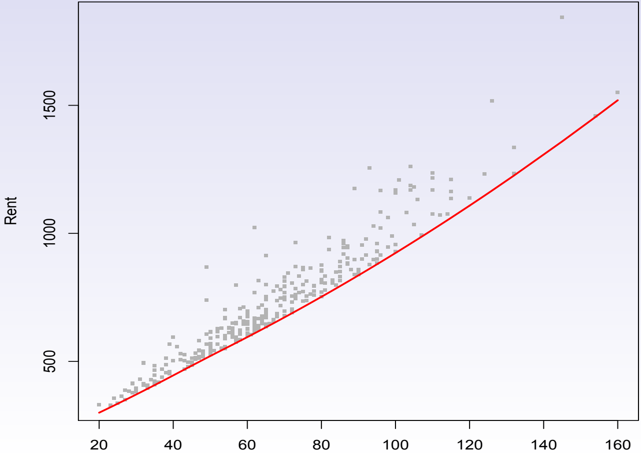
next page



Fitting the tail of the distribution: 90% data

go back

next page



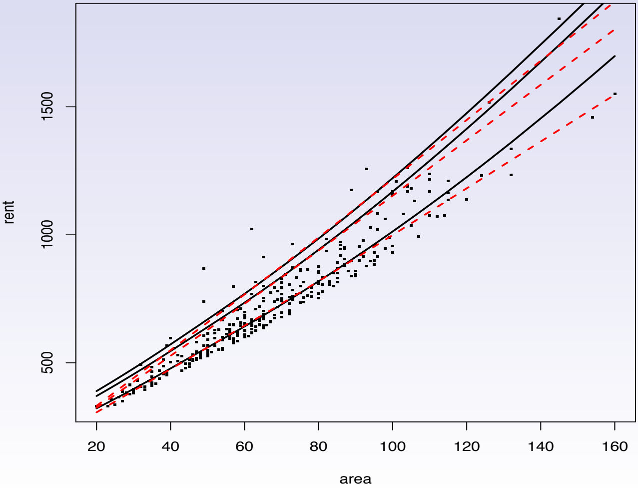
```
gen.trun(par=tail$fv,family="WEI3", name="tail",
            type="left", varying=TRUE)

gen.trun(par=tail$fv,family="GU", name="tail",
            type="left", varying=TRUE)
```

```
t1 <- gamlss(rent~pb(area), data=tail, family=GUtail,
                         n.cyc=500)
t2 <- gamlss(rent~pb(area) , data=tail, family=WEI3tail,
                       n.cyc=500)
t3 <- gamlss(rent~pb(area), sigma.fo=~pb(area), data=tail,
                 family=GUtail,     n.cyc=200)
t4 <- gamlss(rent~pb(area) , sigma.fo=~pb(area), data=tail,
                    family=WEI3tail, n.cyc=500)
```

```
AIC(t1,t2,t3,t4)
df      AIC
t1 3.000191 3212.143
t2 3.989311 3212.563
t3 4.000178 3213.096
t4 5.066734 3214.528
```

Fitting the tail of the distribution: 90% data

go back



gamlss