GAMLSS e R

Gaussian Markov random field models in GAMLSS
gamlss.spatial

Fernanda De Bastiani[1]

[1]Departamento de Estatística, Universidade Federal de Pernambuco

GAMLSS was developed by Rigby and Stasinopoulos (2005)

R Ladies - UFLA

Lavras, May 2019

The Munich rent data of 1999

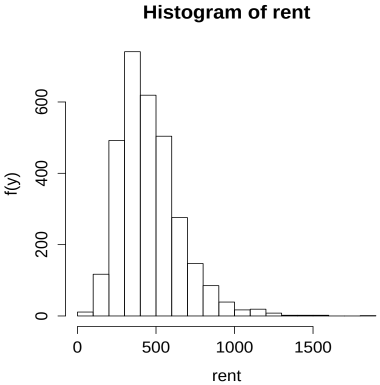
- `rent` : the rent per month (in Euro),
- `area` : living area in square meters,
- `yearc` : year of construction,
- `location` : quality of location: average, 1, good, 2, and top, 3,
- `bath` : quality of bathroom: standard, 0, or premium, 1,
- `kitchen` : quality of kitchen: standard, 0, or premium, 1,
- `cheating` : central heating: with 1, 0,
- `district` : district in Munich (spatial).

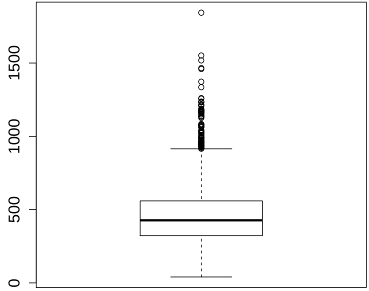
Source: Munich rental guide 1999

3082 observations

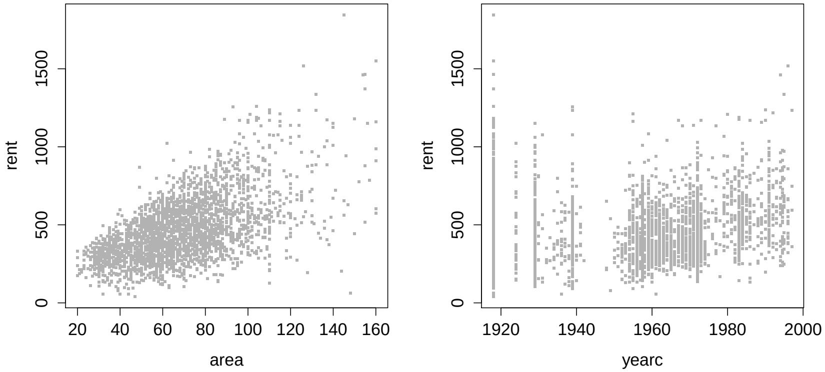
411 districts

Rent data: response

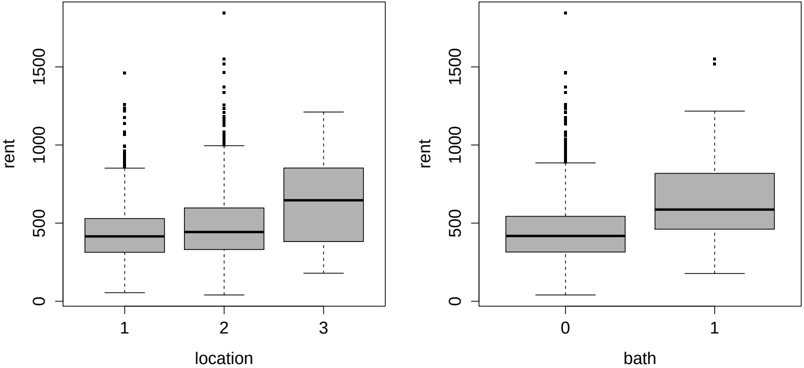




Rent data: area and year



Rent data: location and bath



Rent data: kitchen and central heating

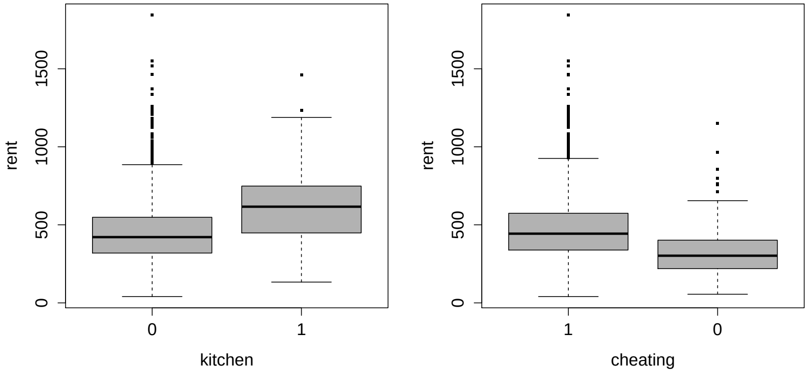
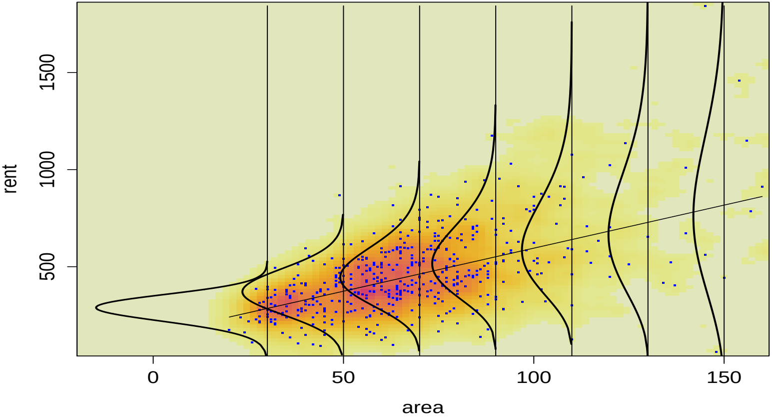


Figure: Geographical information of Munich rent data.

The Munich rent data: area



What we need for modelling the above data?

We need

- flexible distributions for the response variable,
- to be able to deal with heterogeneity in the data,
- to be able to model skewness (and kurtosis?),
- We need modelling all the parameters of the distributions,
- flexible functions to model the relationship between the parameter of the distribution and the explanatory variables.

Historical development

Important events in the creation of the GAMLSS models

- Linear model (Gauss, 1809) Go to LM
- Generalized Linear Models (Nelder and Wedderburn, 1972) Go to GLM
- Generalized Additive Models (Hastie and Tibshirani, 1990) Go to GAM
- Generalized Additive Models for Location Scale and Shape (GAMLSS) (Rigby and Stasinopoulos, 2005).

Generalized Additive Model for Location Scale and Shape

Generalized additive model for location scale and shape Rigby and Stasinopoulos (2005)

$$\mathbf{y} \sim D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$$

$$\begin{aligned}
g_\mu(\boldsymbol{\mu}) &= \mathbf{X}_\mu \boldsymbol{\beta}_\mu + h_{1,\mu}(\mathbf{x}_{1,\mu}) + \ldots + h_{k,\mu}(\mathbf{x}_{k,\mu}) \\
g_\sigma(\boldsymbol{\sigma}) &= \mathbf{X}_\sigma \boldsymbol{\beta}_\sigma + h_{1,\sigma}(\mathbf{x}_{1,\sigma}) + \ldots + h_{k,\sigma}(\mathbf{x}_{k,\sigma}) \\
g_\nu(\boldsymbol{\nu}) &= \mathbf{X}_\nu \boldsymbol{\beta}_\nu + h_{1,\nu}(\mathbf{x}_{1,\nu}) + \ldots + h_{k,\nu}(\mathbf{x}_{k,\nu}) \\
g_\tau(\boldsymbol{\tau}) &= \mathbf{X}_\tau \boldsymbol{\beta}_\tau + h_{1,\tau}(\mathbf{x}_{1,\tau}) + \ldots + h_{k,\tau}(\mathbf{x}_{k,\tau})
\end{aligned}$$

where $D(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$ can be any distribution and where $h_j(\mathbf{x}_j)$ are smooth functions of the X's.

Random effects form

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}$$

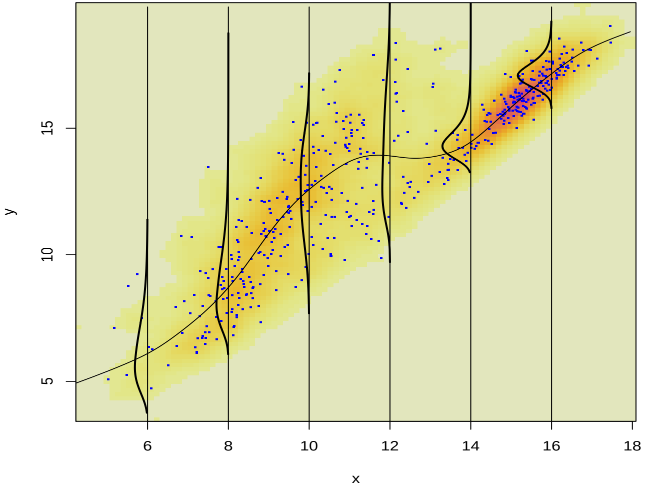
$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}.$$

where $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \lambda_{jk}^{-1}\mathbf{G}_{jk}^{-1})$.

GAMLSS assumptions



What is GAMLSS?

GAMLSS: are semi-parametric regression type models.

- regression type: we have many explanatory variables $\mathbf{X}$ and one response variable $\mathbf{y}$ and we believe that $\mathbf{X} \rightarrow \mathbf{y}$,
- parametric: a parametric distribution assumption for the response variable,
- semi: the parameters of the distribution, as functions of explanatory variables, may involve non-parametric smoothing functions,
- GAMLSS philosophy: try different models.

GAMLSS is a generalization of GLM and GAM models.

gamlss

GAMLSS: Distributions

There are around 100 discrete, continuous continuous, and implemented as `gamlss.family` in the R including highly skew and kurtotic distributions shapes,

- creating a new distribution is relatively easy,
- truncating truncated an existing distribution,
- using a censored version of an existing distribution,
- mixing mixture different distributions to create a new finite mixture distribution,
- discretise discretise continuous distributions,
- log or logit any continuous distribution in $(-\infty, \infty)$.

- Historically the distributions commonly used for modelling continuous and discrete count response variables were the normal and Poisson distributions respectively.
- In practice, for a continuous response variable, the shape of its distribution can be highly positively or negatively skewed and/or highly platykurtic or leptokurtic.
- Continuous variables can also have different ranges from that of the normal distribution, i.e. $(-\infty, \infty)$.
- Also in practice a discrete count response variable can have a distribution which is overdispersed or underdispersed and/or leptokurtic and/or have an excess or reduced incidence of zero values.
- ...

Additive terms

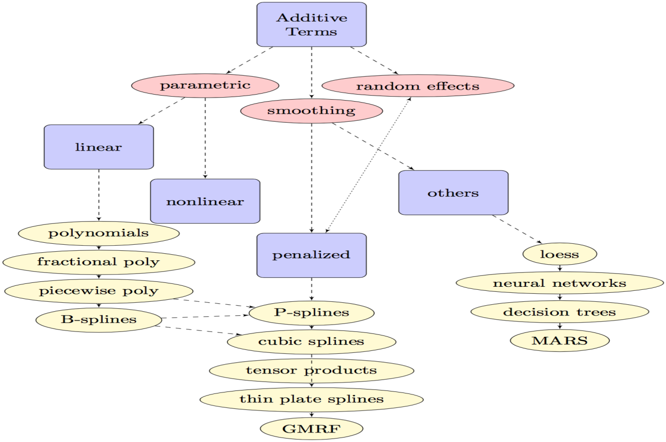


Figure: Diagram showing the different additive terms that can be fitted within GAMLSS.

GAMLSS: R implementation

GAMLSS is implemented in series of packages in R

gamlss	the original package
gamlss.dist	for distributions
gamlss.data	for data
gamlss.demo	for demos
gamlss.nl	for non-linear terms
gamlss.add	for extra additive terms
gamlss.tr	for truncated distributions
gamlss.cens	for censored (left, right or interval) response variables
gamlss.mx	for finite mixtures and random effects
gamlss.countKinf	for count response variables
gamlss.spatial	for Markov Random fields

gamlss

- Fitting the parametric model requires only estimates for the 'betas' $\boldsymbol{\beta}$.

- Fitting the random effects GAMLSS model requires estimates for the 'betas' $\boldsymbol{\beta}$, the 'gammas' $\boldsymbol{\gamma}$, and also the 'lambdas' $\boldsymbol{\lambda}$, where:

$$\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{11}^{\top}, \ldots, \boldsymbol{\lambda}_{1J_1}^{\top}, \boldsymbol{\lambda}_{21}^{\top}, \ldots, \boldsymbol{\lambda}_{4J_4}^{\top})^{\top} \ .$$

- Within **gamlss**, the parametric GAMLSS model is fitted by maximum likelihood estimation with respect to $\boldsymbol{\beta}$, whereas

- the more general random effects model is fitted by maximum penalized likelihood estimation (or equivalently posterior mode or maximum a posteriori (MAP), estimation with respect to $\boldsymbol{\beta}$ and γ for fixed $\boldsymbol{\lambda}$.

For more details, see Chapter 3 of Stasinopoulos et al (2017).

Estimation

The log-likelihood function for the GAMLSS model under the assumption that observations of the response variable are independent is given by

$$\ell = \sum_{i=1}^{n} \log f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i),$$

where $f_Y(.)$ represents the probability (density) function of the response variable. The respective penalized log-likelihood function is given by

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^{4} \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^\top \mathbf{G}_{jk} \gamma_{jk}.$$

The two basic algorithms for fitting the parametric model with respect to $\boldsymbol{\beta}$, and the nonparametric model with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for fixed $\boldsymbol{\lambda}$, implementted in **gamlss** in **R** are:

- **RS** and
- **CG**.

Both use an iteratively reweighted (penalized) least squares algorithm (Rigby and Stasinopoulos, 2013 and Stasinopoulos et al, 2017).

Markov random field (MRF)

- A Markov random field (MRF) is a set of random variables where a local defined assumption is used to determine their joint (or global) distribution, (Banerjee et al, 2014);
- Their local behaviour is described through Markov properties based on conditional independence assumptions;
- Those Markovian assumptions can be presented as an undirected graph $\mathcal{G}$, where each vertex represents an areal unit and each edge connects two areal units and represents a neighbouring relationship, Rue and Held (2005).

Gaussian Markov random field

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph that consists of vertices $\mathcal{V} = (1, 2, \ldots, q)$, and a set of edges $\mathcal{E}$, where a typical edge is (m, t), $m, t \in \mathcal{V}$.

A random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^\top$ is called a Gaussian MRF (i.e. GMRF) with respect to the graph $\mathcal{G}$, with mean $\boldsymbol{\mu}$ and precision matrix $\lambda\mathbf{G}$, if and only if its density has the form

$$\pi(\boldsymbol{\gamma}) \propto \exp\left[-\frac{1}{2}\lambda(\boldsymbol{\gamma} - \boldsymbol{\mu})^\top \mathbf{G}(\boldsymbol{\gamma} - \boldsymbol{\mu})\right] \tag{1}$$

and

$$G_{mt} \neq 0 \Longleftrightarrow (m, t) \in \mathcal{E} \text{ for } m \neq t,$$

where G_{mt} is the element of matrix $\mathbf{G}$ for row m and column t (Rue and Held, 2005).

- The conditional autoregressive (CAR) model (Besag, 1974) is a GMRF model of the form given in Equation (1) where **G** is a non-singular matrix.

- The CAR model can also be specified by:

$$\gamma_i | \boldsymbol{\gamma}_{-i} \sim N\left(\sum_j \alpha_{ij}\gamma_j, k_i\right),$$

  where $\boldsymbol{\gamma}_{-i} = (\gamma_1, \gamma_2, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_q)$ and $\alpha_{ii} = 0$, $\alpha_{ij} = -G_{ij}/G_{ii}(i \neq j)$ and $k_i = 1/(\lambda G_{ii})$ for $i = 1, 2, \ldots, q$.

- The joint distribution for γ is of the form given in Equation (1) with $\boldsymbol{\mu} = \mathbf{0}$, provided $\alpha_{ij}k_j = \alpha_{ji}k_i$ for all i and j to ensure that matrix **G** is symmetric (Besag and Krooperberg, 1985).

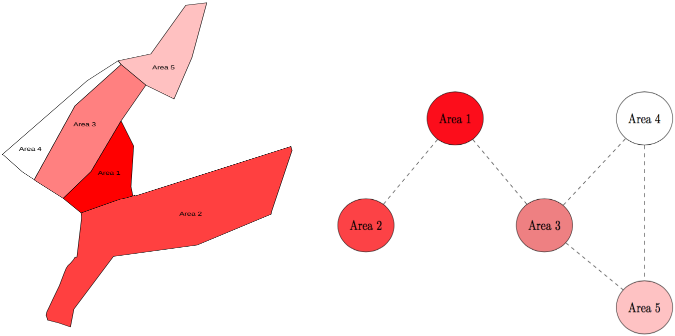


Figure: Subset regions of `rent99` data (left) and diagram showing the relationship between five regions (right).

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}, \quad \mathbf{D}_w = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

$$\text{and} \quad \mathbf{G} = \begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 3 & -1 & -1 \\ -1 & 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{bmatrix}.$$

$\mathbf{G} = \mathbf{D}_w - \mathbf{W}$

When the **G** matrix is singular the GMRF model can be represented by the intrinsic autoregressive model (IAR). To incorporate IAR models within the GAMLSS model

- set **Z** to be an index matrix defining which observation belongs to which area,
- let γ be the vector of q spatial random effects and
- assume $\gamma \sim N_q(0, \lambda^{-1}\mathbf{G}^{-1})$, where $\mathbf{G}^{-1}$ is the (generalized) inverse of a $q \times q$ matrix, **G**.

- A non-zero value in matrix **G** implies a connection between the two corresponding regions in the graph $\mathcal{G}$ (they are connected neighbours).
- The zero value in matrix **G** implies no connection between the two regions in the graph $\mathcal{G}$ and hence
- that the corresponding spatial random effects γ_m and γ_t for the two regions are conditionally independent (given the other spatial random effects γ_r for all r not equal to m or t).

Model selection Strategy

Let $\mathcal{M} = \{\mathcal{D}, \mathcal{L}, \mathcal{T}, \boldsymbol{\lambda}\}$ represent a GAMLSS model.

- $\mathcal{D}$ specifies the distribution of the response variable,
- $\mathcal{L}$ specifies the set of link functions for the distribution parameters μ, σ, ν and τ,
- $\mathcal{T}$ specifies the terms appearing in the predictors for μ, σ, ν and τ,
- $\boldsymbol{\lambda}$ specifies the smoothing hyperparameters which determine the amount of smoothing of continuous explanatory variables (`area` and `yearc`) and of the spatial effect (`district`).

Box-Cox Cole and Green - BCCG

Suitable for positively or negatively skew data.

Let $Y > 0$ be a positive random variable having a Box-Cox Cole and Green distribution, denoted by $BCCG(\mu, \sigma, \nu)$, defined through the transformed random variable Z given by

$$f_Y(y) = \frac{y^{\nu-1} \exp\left(-\frac{1}{2}z^2\right)}{\mu^\nu \sigma \sqrt{2\pi}\Phi\left(\frac{1}{\sigma|\nu|}\right)},$$

for $y > 0$ and Z is assumed to follow a truncated standard normal distribution.

Rent data analysis: model selection

$$BCCGo(\mu, \sigma, \nu)$$

1. forward GAIC selection procedure to select an appropriate model for μ,
2. forward selection procedure to select an appropriate model for σ,
3. forward selection procedure to select an appropriate model for ν,
4. backward selection procedure to select an appropriate model for σ,
5. backward selection procedure to select an appropriate model for μ.

R implementation

The function `gmrf()` accepts three different ways to pass the geographical information:

i) `polys`, is a **R** list comprising the region label followed by coordinates of points in two columns in matrix form defining the boundary for each area,

ii) `neighbour`, is a **R** list comprising each region label followed by its neighbouring region labels,

iii) `precision`, is a **R** matrix containing the **G** matrix.

gamlss.spatial package in **R**!

Rent data analysis: fitted model

$$
\begin{aligned}
Y \sim & BCCGo(\hat{\mu}, \hat{\sigma}, \hat{\nu}), \\
\log(\hat{\mu}) = & 6.06 + s_{11}(year) + s_{12}(area) + s(district) \\
& + 0.079(\text{if location=good}) + 0.211(\text{if location=top}) \\
& - (0.255 - 0.0038 nyearc)(\text{if no central heating}) \\
& + (0.146 - 0.0034 nyearc + 0.0023 narea)(\text{if kitchen=premium}) \\
& + 0.067(\text{if bath=1, premium}), \\
\log(\hat{\sigma}) = & 11.811 + s_{21}(year) + 0.0016 area \\
& + 0.231(\text{if no central heating}), \\
\hat{\nu} = & -12.377 + s_{31}(year) + s_{32}(area) \\
& + 2.381(\text{if kitchen=premium}),
\end{aligned}
$$

Rent data analysis: fitted mu model

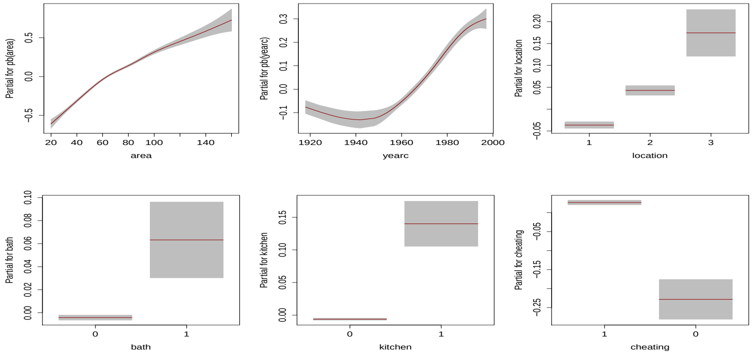


Figure: Terms plots for μ.

good location: 8.2% increase; premium location: 23.5% increase; premium bathroom: 6.9% increase.

Rent data analysis: fitted mu model linear interactions

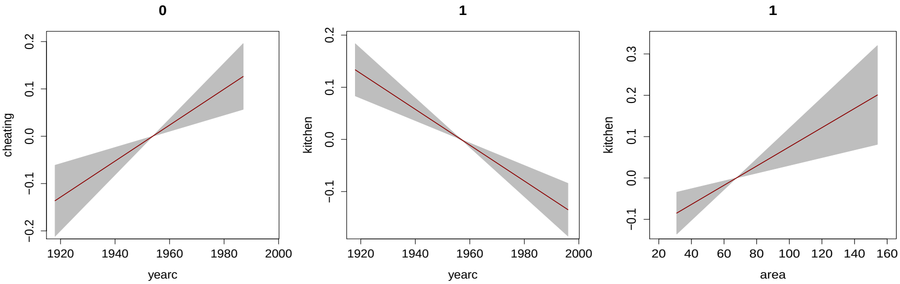


Figure: Term plots of the interactions for μ.

no central heating: 22.5% decrease (depends on `yearc`);
premium kitchen: 15.6% increase (depends on `yearc` and `area`);

Rent data analysis: fitted mu model spatial factor

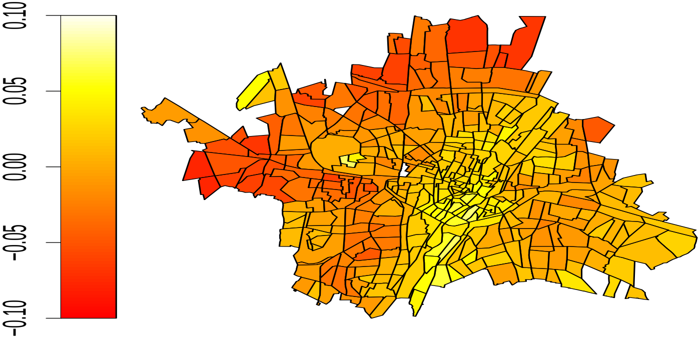


Figure: The fitted spatial effect for μ for the chosen model with spatial effect.

Best district: 10.5% higher fitted median rent;
Worst district: 9.5% lower fitted median rent.

Rent data analysis: fitted sigma model

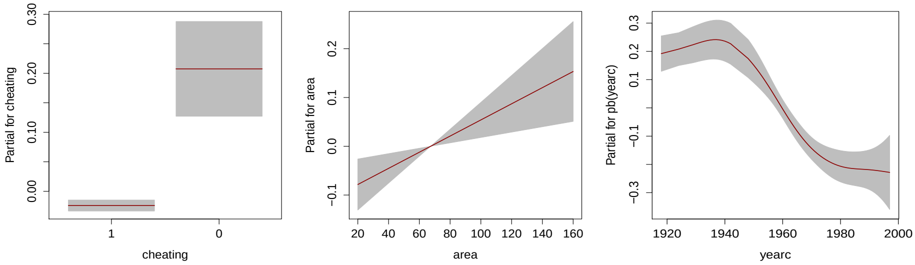


Figure: Term plots for σ.

No central heating: increase 26.1%

Rent data analysis: fitted nu model

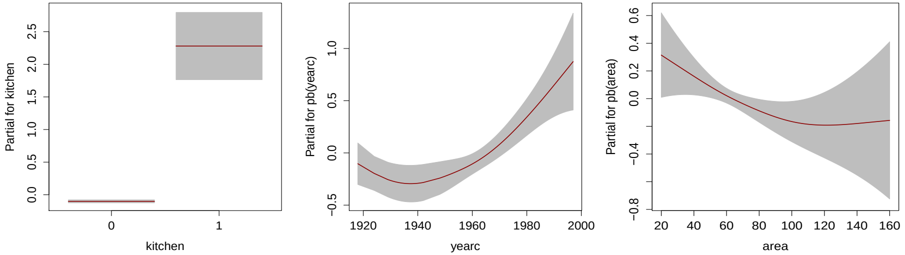


Figure: Term plots for ν.

Premium kitchen: increase of 2.4%

Rent data analysis: fitted nu model

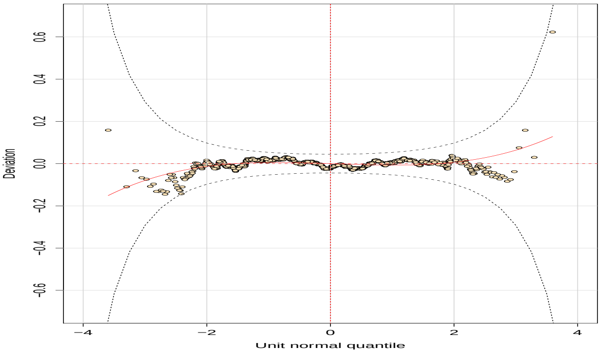


Figure: Worm plot of the residuals for the chosen final model.

Rent data analysis: fitted nu model

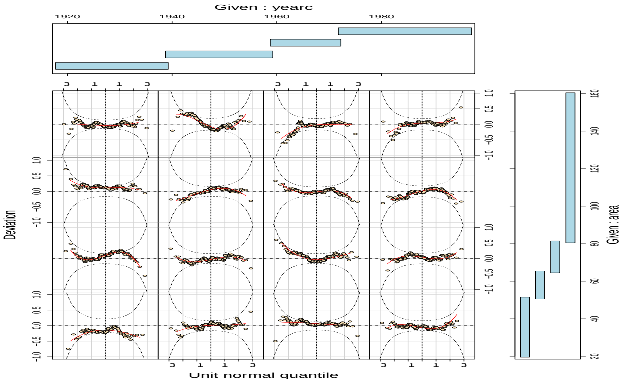


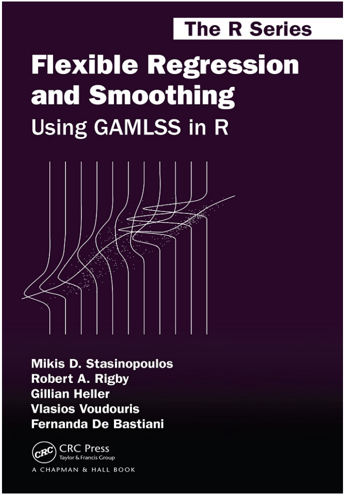
Figure: Worm plot of the residuals split by the yearc and area variables for the final model.

Conclusions

- GAMLSS provides a platform to **fit**, **compare** and **check** spatial models for the parameters of the distribution of a response variable which may be non exponential family;
- Can deal with continuous response variable distributions which are higly positively or negatively skewed and/or have high or low kurtosis, discrete count distributions that are overdispersed or have excess zeros, or mixed continuous distributions;
- The methodology presented can be applied to other data sets that have geographical information specifying the neighbours of each region.

gamlss

Information



Forthcoming Book

Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R

Robert A. Rigby , Mikis D. Stasinopoulos, Gillian Z. Heller and Fernanda De Bastiani

The book is about

- Statistical distributions and how they can be used in practice within the distributional regression modelling approach of the GAMLSS.
- It describes over 100 distributions (available in the **R** package **gamlss.dist**), their properties, limitations and their use in data applications.
- It follows an earlier book 'Flexible Regression and Smoothing: Using GAMLSS in R'.

The GAMLSS software is available from:

- The Comprehensive R Archive Network: `https://cran.r-project.org`.
- GitHub: `https://github.com/gamlss` (source code of **gamlss** in R and Java)

References

- S. Banerjee, B. P. Carlin and A.E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, Second Edition. Chapman and Hall/CRC, Boca Raton, 2014.
- J. Besag, *Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion)*, Journal of the Royal Statistical Society, Series B, 36, pp. 192–236, 1974.
- J. Besag and C. Kooperberg, *On conditional and intrinsic autoregressions*, Biometrika 82, pp. 733–746, 1995.
- F. De Bastiani, R.A. Rigby, D.M. Stasinopoulos, A.H.M.A. Cysneiros and M.A. Uribe-Opazo, *Gaussian Markov random field spatial models in GAMLSS*, Journal of Applied Statistics, 45 (1), pp. 168-186, 2018 .

gamlss

References

- R.A. Rigby and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape (with discussion)*, Applied. Statistics, 54, (2005), 507–554, 2005.
- R.A. Rigby , D. M. Stasinopoulos, G. Z. Heller and F. De Bastiani *Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R*, Chapman and Hall/CRC The R Series, (forthcoming).
- H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, Chapman and Hall, New York, (2005).
- D.M. Stasinopoulos, R.A. Rigby, G.Z. Heller, V. Voudouris and F. De Bastiani, *Flexible Regression and Smoothing: Using GAMLSS in R*, Chapman and Hall/CRC The R Series, 2017.

gamlss

For more GAMLSS

Muito obrigada!

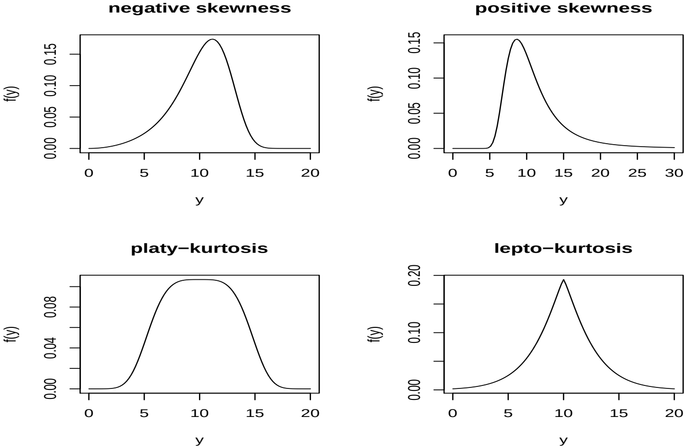
debastiani@de.ufpe.br

for more information see

www.gamlss.org

Continuous distributions: different shapes

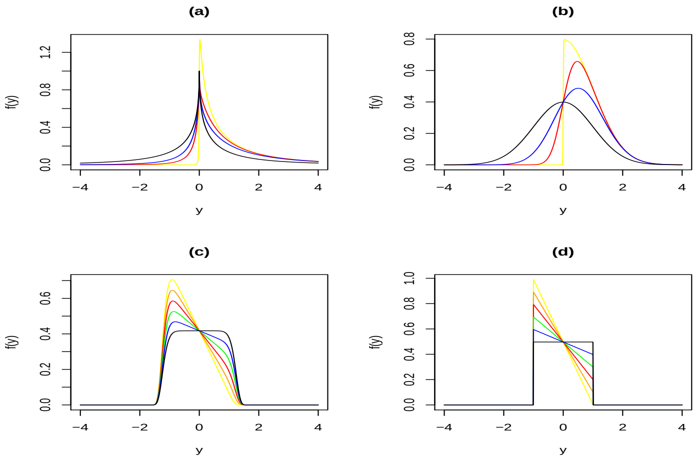
Go back Distributions



gamlss

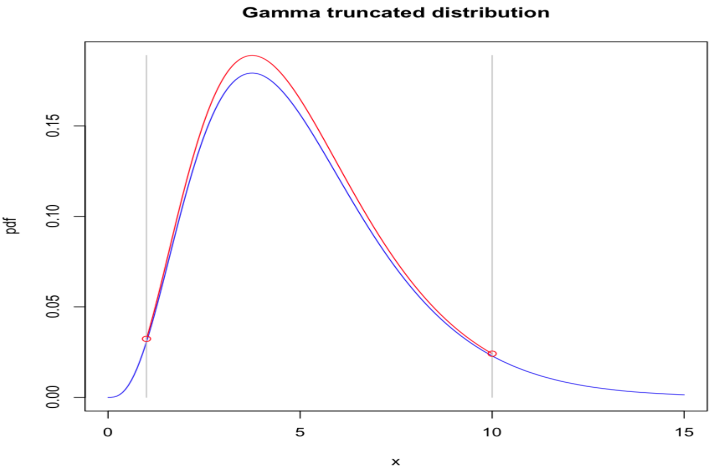
Continuous distributions: different types

Go back Distributions



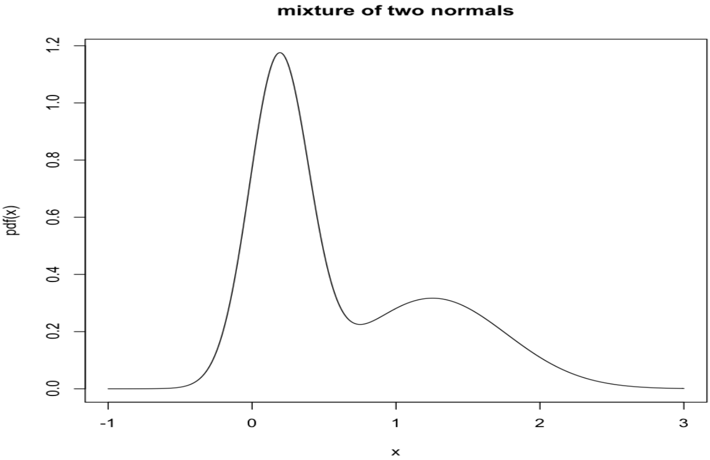
Continuous distributions: different types

Go back Distributions



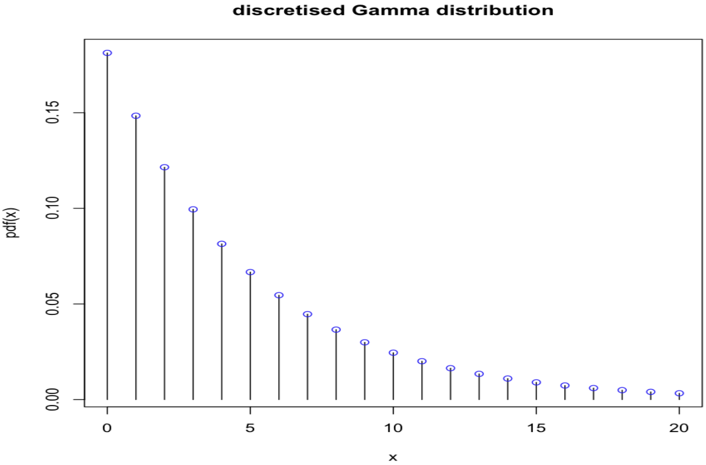
Continuous distributions: different types

Go back Distributions

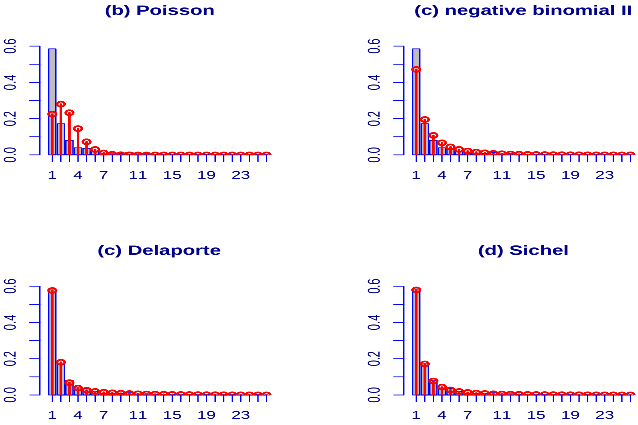


Continuous distributions: different types

Go back Distributions



The stylometric data, Go back to distributions



The linear model

Linear Model, Gauss

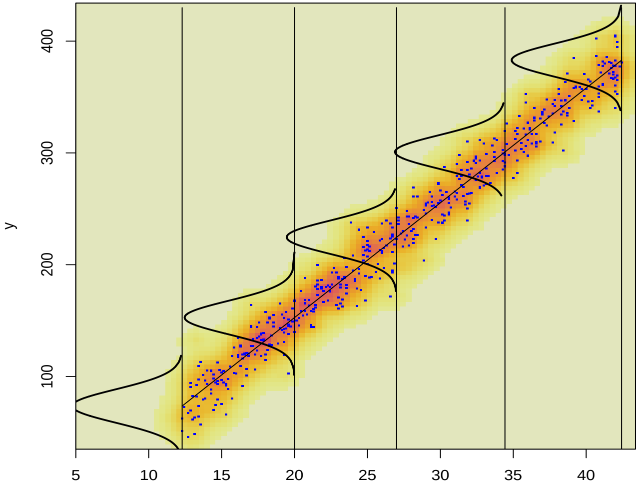
$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim NO(\mathbf{0}, \sigma^2\mathbf{I})$

The model can be also written as:

$\mathbf{y} \sim NO(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$

Go back

The linear model assumptions Go back



gamlss

The generalized linear model

Generalised Linear Model, Nelder and Wedderburn(1972)

$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ where $\mathbf{y} \sim \mathit{ExpFamily}(\boldsymbol{\mu}, \phi)$

where $g()$ is the link function

The exponential family

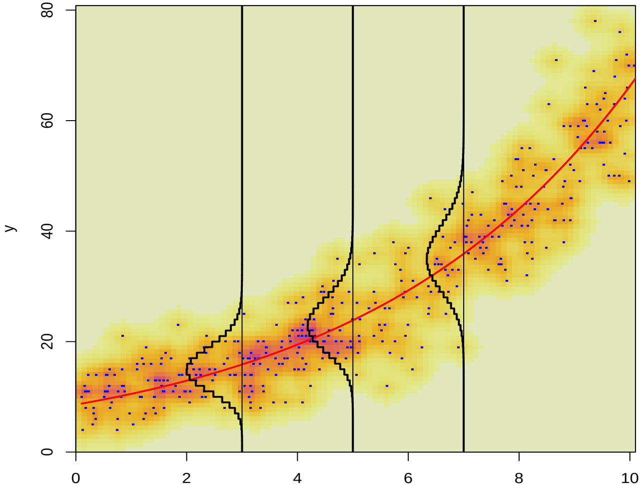
1. normal
2. Gamma
3. inverse Gaussian
4. Poisson
5. binomial

Go back 70-80 Next page

gamlss

The generalized linear model

Go back 70-80



gamlss

The generalized additive model

Generalized additive model Hastie and Tibshirani (1990)

$$\mathbf{y} \sim ExpFamily(\boldsymbol{\mu}, \phi)$$

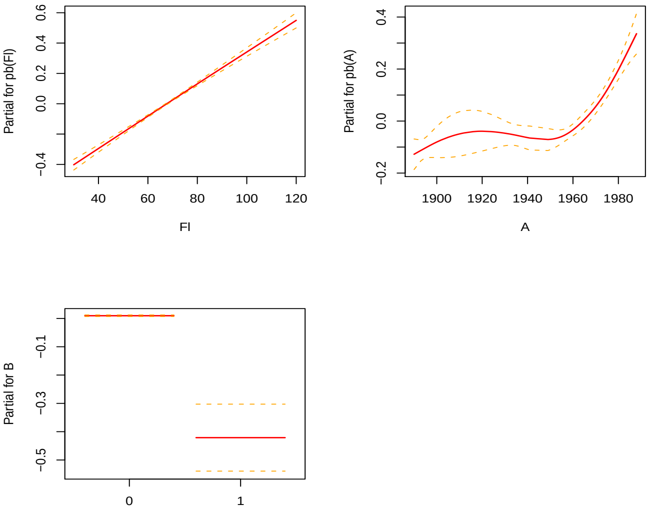
$$g(\boldsymbol{\mu}) = \mathbf{X}\beta + h_1(\mathbf{x}_1) + ... + h_k(\mathbf{x}_k)$$

where $h_j(\mathbf{x}_j)$ are smooth functions of the X's.

Go back Historical Next page

The generalized additive model

Go back Historical Next page



The generalized additive model

- the estimation is achieved by modified backfitting;
- modified backfitting: is a combination of IRLS for the linear part and penalised IRLS for the smoothing part;
- the smoothing parameters λ's are fixed.

Go back Historical