Universität Innsbruck

Structured Additive Regression Models: An R Interface to BayesX

Nikolaus Umlauf, Daniel Adler, Thomas Kneib, Stefan Lang, Achim Zeileis

http://eeecon.uibk.ac.at/~umlauf/

Overview

- Introduction
- Structured Additive Regression Models (STAR)
- The main model fitting function
- More components of the interface
- Available additive terms
- Illustration
- Why does it always rain on me?
- References

Introduction: What is BayesX?

The free software **BayesX** is a standalone program comprising powerful tools for Bayesian and mixed model based inference in complex semiparametric regression models with structured additive predictor (STAR).

- Generalized additive models (GAM).
- Generalized additive mixed models (GAMM).
- Generalized geoadditive mixed models (GGAMM).
- Dynamic models.
- Varying coefficient models (VCM).
- Geographically weighted regression.

BayesX is written in C++ and utilizes numerically efficient (sparse) matrix architectures.

Introduction: What is BayesX?

In **BayesX**, estimation of regression parameters is based on three inferential concepts:

1. Full Bayesian inference via MCMC.
2. Inference via a mixed model representation.
3. Penalized likelihood including variable selection.

BayesX provides functionality for the following types of responses:

- Univariate exponential family.
- Categorical responses with unordered responses.
- Categorical responses with ordered responses.
- Continuous time survival models.
- Continuous time multi-state models.

Introduction: The R interface

Problems: BayesX only provides limited functionality for

- handling/manipulating data sets,
- handling/manipulating geographical maps,
- exploring/visualizing estimation results.

Introduction: The R interface

Now: Interface package **R2BayesX** for

- specifying/estimating STAR models with **BayesX** directly from R,
- standard methods and extractor functions for **BayesX** fitted model objects, e.g. producing high level graphics of estimated effects, model diagnostic plots, summary statistics and more.

In addition:

- Run already existing **BayesX** input program files from R.
- Automatically import **BayesX** output files into R.

To install the package directly within R type:

```
install.packages("R2BayesX")
```

STAR models

Distributional and structural assumptions, given covariates and parameters, are based on generalized linear models with

$$E(y|\mathbf{x}, \mathbf{z}, \gamma, \boldsymbol{\theta}) = h^{-1}(\eta)$$

and structured additive predictor

$$\eta = f_1(\mathbf{z}) + \ldots + f_p(\mathbf{z}) + \mathbf{x}'\gamma$$

- $\mathbf{x}'\gamma$ parametric part of the predictor.
- $\mathbf{z}$ represents a generic vector of all nonlinear modeled covariates, e.g. may include continuous covariates, time scales, location or unit or cluster indexes.
- The vector $\boldsymbol{\theta}$ comprises all parameters of the functions $f_1, \ldots, f_p$.
- f_j one-/two-/higher-dimensional, not necessarily continuous functions.

STAR models: Modeling the functions f_j

The functions f_j are possibly smooth functions comprising effects (and combinations) as e.g. given by:

- Nonlinear effects of continuous covariates.
- Two-dimensional surfaces.
- Spatially correlated effects.
- Varying coefficients.
- Spatially varying effects.
- Random intercepts.
- Random slopes.

STAR models: General form

- Vector of function evaluations $\mathbf{f}_j = (f_j(\mathbf{z}_1), \ldots, f_j(\mathbf{z}_n))$ of the $i = 1, \ldots, n$ observations can be written in matrix notation

$$\mathbf{f}_j = \mathbf{Z}_j \boldsymbol{\beta}_j,$$

  with $\mathbf{Z}_j$ as the design matrix, where $\boldsymbol{\beta}_j$ are unknown regression coefficients.
- Form of $\mathbf{Z}_j$ only depends on the functional type chosen.
- Penalized least squares:

$$\text{PLS}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = ||\mathbf{y} - \boldsymbol{\eta}||^2 + \lambda_1 \boldsymbol{\beta}_1' \mathbf{K}_1 \boldsymbol{\beta}_1 + \ldots + \lambda_p \boldsymbol{\beta}_p' \mathbf{K}_p \boldsymbol{\beta}_p$$

STAR models: General form

- Prior for β in the corresponding Bayesian approach

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j'\mathbf{K}_j\beta_j\right),$$

  τ_j^2 variance parameter, governs the smoothness of f_j.
- Structure of $\mathbf{K}_j$ also depends on the type of covariates and on assumptions about smoothness of $\mathbf{f}_j$.
- The variance parameter τ_j^2 is equivalent to the inverse smoothing parameter in a frequentist approach. Utilizing mixed model technology, restricted maximum likelihood (REML) forms a basis for determination. From a Bayesian perspective, this yields empirical Bayes/posterior mode estimates for the STAR models.

The main model fitting function

The arguments of the main model fitting function are

```
bayesx(formula, data, weights = NULL, subset = NULL,
  offset = NULL, na.action = na.fail, contrasts = NULL,
  family = "gaussian", method = "MCMC",
  control = bayesx.control(...), ...)
```

Families:

`"binomial"`, `"binomialprobit"`, `"gamma"`, `"gaussian"`,
`"multinomial"`, `"poisson"`, `"cox"`, `"cumprobit"`, `"multistate"`,
`"binomialcomploglog"`, `"cumlogit"`, `"multinomialcatsp"`,
`"multinomialprobit"`, `"seqlogit"`, `"seqprobit"`.

Methods:

`"MCMC"`, `"REML"`, `"STEP"`.

Note: `family` objects are currently not supported.

More components of the interface

Internally, function `bayesx()` calls the following functions:

1. `parse.bayesx.input()`
2. `write.bayesx.input()`
3. `run.bayesx()`
4. `read.bayesx.output()`

These functions are operating independently and may also be called by the R user.

The functionality is especially helpful for already existing **BayesX** program and output files.

Moreover, function `read.bayesx.output()` also returns objects of class `"bayesx"`.

Available additive terms

The main model term constructor function is function `sx()`, with arguments:

```
sx(x, z = NULL, bs = "ps", by = NA, ...)
```

`sx()` is simply an interface to function `s()` from package **mgcv**.

Basis/term types:
`"rw1"`, `"rw2"`, `"season"`, `"ps"` (`"psplinerw1"`, `"psplinerw2"`), `"te"` (`"pspline2dimrw1"`), `"kr"` (`"kriging"`), `"gk"` (`"geokriging"`), `"gs"` (`"geospline"`), `"mrf"` (`"spatial"`), `"bl"` (`"baseline"`), `"factor"`, `"ridge"`, `"lasso"`, `"nigmix"`, `"re"` (`"ra"`, `"random"`).

Available additive terms

Additional options within "`...`" and `xt` for each basis/term type and method may be looked up using function `bayes.term.options()`, e.g.

```
R> bayesx.term.options(bs = "ps", method = "MCMC")

possible options for 'bs = "ps"':

     degree: the degree of the B-spline basis functions.
           Default: integer, 'degree = 3'.

      knots: number of inner knots.
           Default: integer, 'knots = 20'.

      order: only if 'bs = "ps"', the order of the
           difference penalty.
           Default: integer, 'order = 2'.
        .
        .
        .
```

Illustration

Following Kandala, Lang, Klasen and Fahrmeir (2001), the task is to model `stunting` of newborn children on the following covariates:

Variable	Description
`stunting`	Standardized Z-score for stunting.
`mbmi`	Body mass index of the mother.
`agechild`	Age of the child in months.
`district`	District where the mother lives.
`memployment`	Is the mother employed?
`meducation`	Mother's educational status.
`urban`	Is the domicile in an urban region?
`gender`	Gender of the child.

The predictor of the STAR model is given by

$$\begin{aligned}\eta \quad = \quad & \gamma_0 + \gamma_1 \texttt{memploymentyes} + \gamma_2 \texttt{urbanno} + \gamma_3 \texttt{genderfemale} + \\ & \gamma_4 \texttt{meducationno} + \gamma_5 \texttt{meducationprimary} + \\ & f_1(\texttt{mbmi}) + f_2(\texttt{agechild}) + f_{str}(\texttt{district}) + f_{unstr}(\texttt{district})\end{aligned}$$

Illustration

The formula is set with

```
R> f <- stunting ~ memployment + urban + gender + meducation +
+   sx(mbmi) + sx(agechild) +
+   sx(district, bs = "mrf", map = ZambiaBnd) +
+   sx(district, bs = "re")
```

The model is then fitted using MCMC by calling

```
R> set.seed(321)
R> zm <- bayesx(f, family = "gaussian", method = "MCMC",
+   data = ZambiaNutrition, iterations = 12000, burnin = 2000,
+   step = 10)
```

Model summary

```
R> summary(zm)
```

Illustration

```
Call:
bayesx(formula = f, data = ZambiaNutrition, family = "gaussian",
    method = "MCMC", iterations = 12000, burnin = 2000, step = 10)

Fixed effects estimation results:

Parametric Coefficients:
                     Mean      Sd    2.5%     50%   97.5%
(Intercept)        0.0991  0.0475  0.0046  0.1018  0.1863
memploymentno     -0.0084  0.0135 -0.0359 -0.0084  0.0170
urbanno           -0.0895  0.0217 -0.1306 -0.0893 -0.0450
genderfemale       0.0582  0.0133  0.0320  0.0578  0.0850
meducationno      -0.1722  0.0269 -0.2248 -0.1719 -0.1163
meducationprimary -0.0611  0.0262 -0.1115 -0.0614 -0.0091

Smooth terms variances:
                Mean     Sd   2.5%    50%  97.5%    Min    Max
sx(agechild) 0.0062 0.0060 0.0014 0.0042 0.0233 0.0007 0.0570
sx(district) 0.0360 0.0191 0.0094 0.0325 0.0813 0.0025 0.1784
sx(mbmi)     0.0019 0.0028 0.0003 0.0011 0.0081 0.0002 0.0468
```

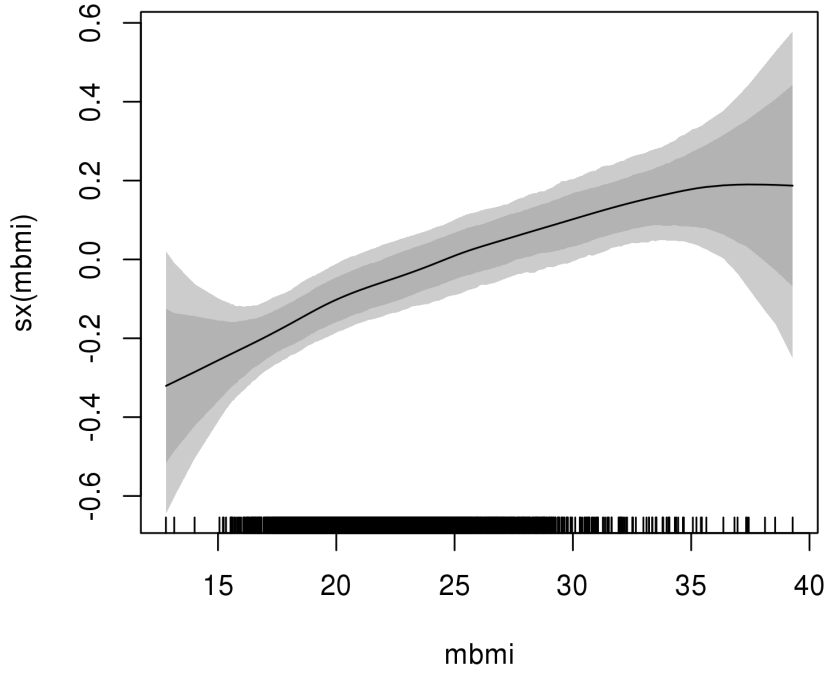
Illustration

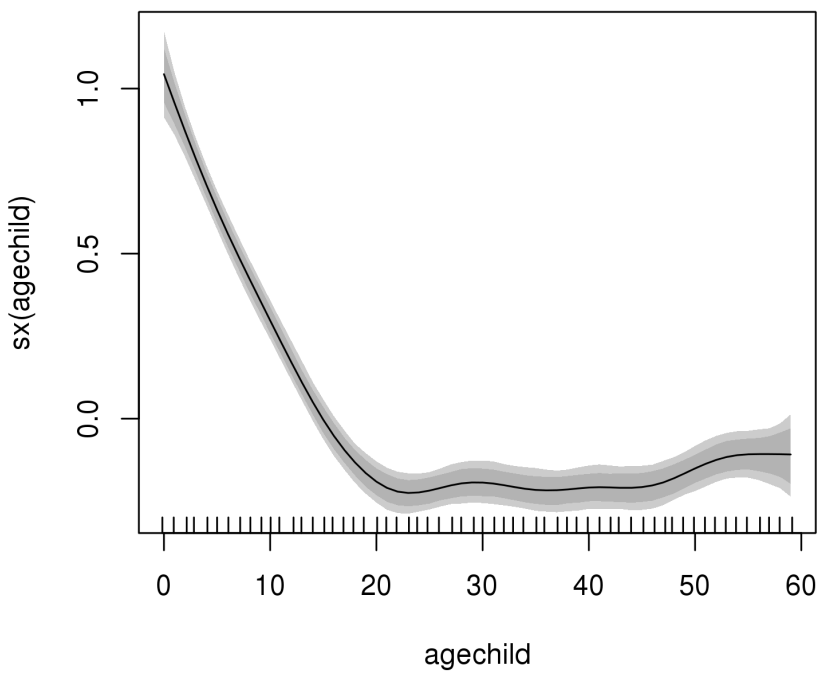
```
Random effects variances:
                Mean     Sd   2.5%    50%  97.5%    Min    Max
sx(district) 0.0076 0.0064 0.0008 0.0062 0.0226 0.0003 0.0701

Scale estimate:
         Mean     Sd   2.5%    50% 97.5%
Sigma2 0.8023 0.0163 0.7721 0.8017 0.836

N = 4847  burnin = 2000  DIC = 4899.506  pd = 50.41262
method = MCMC  family = gaussian  iterations = 12000  step = 10
```

Illustration

Plotting of specific terms

```
R> plot(zm, term = c("sx(mbmi)", "sx(agechild)"))
```

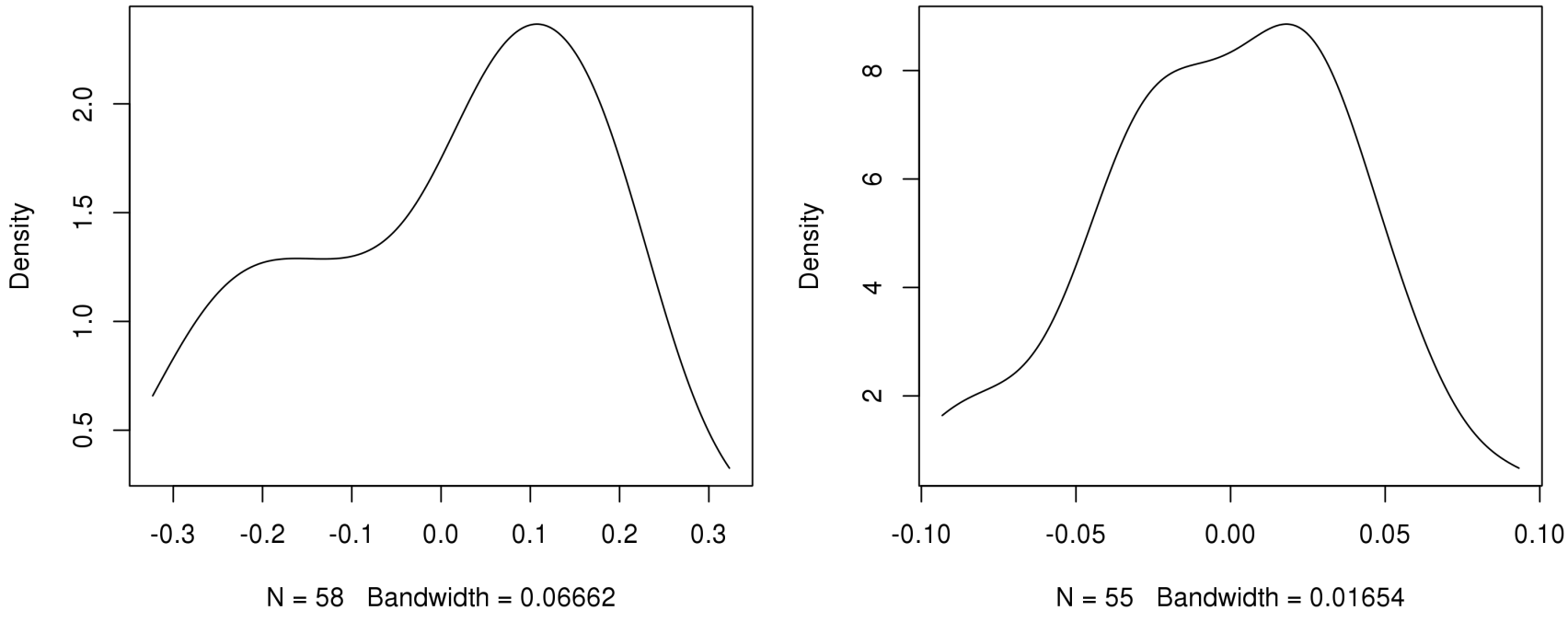




Illustration

Spatial effects, kernel density estimates

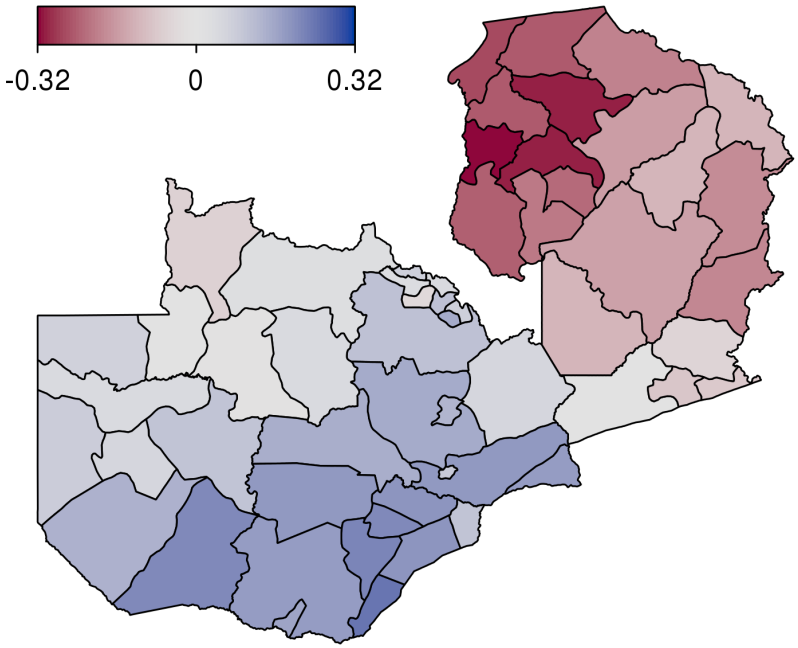
```
R> plot(zm, term = c("sx(district):mrf", "sx(district):re"))
```

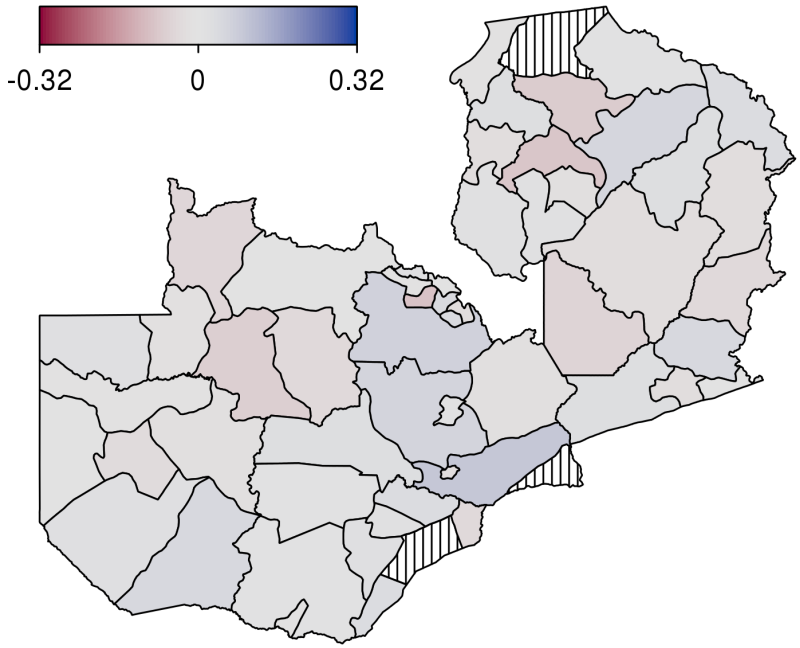


Illustration

Spatial effects, map effect plots

```
R> plot(zm, term = "sx(district):mrf", map = ZambiaBnd)
R> range <- c(-0.32, 0.32)
R> plot(zm, term = "sx(district):re", map = ZambiaBnd,
+   range = range, lrange = range)
```

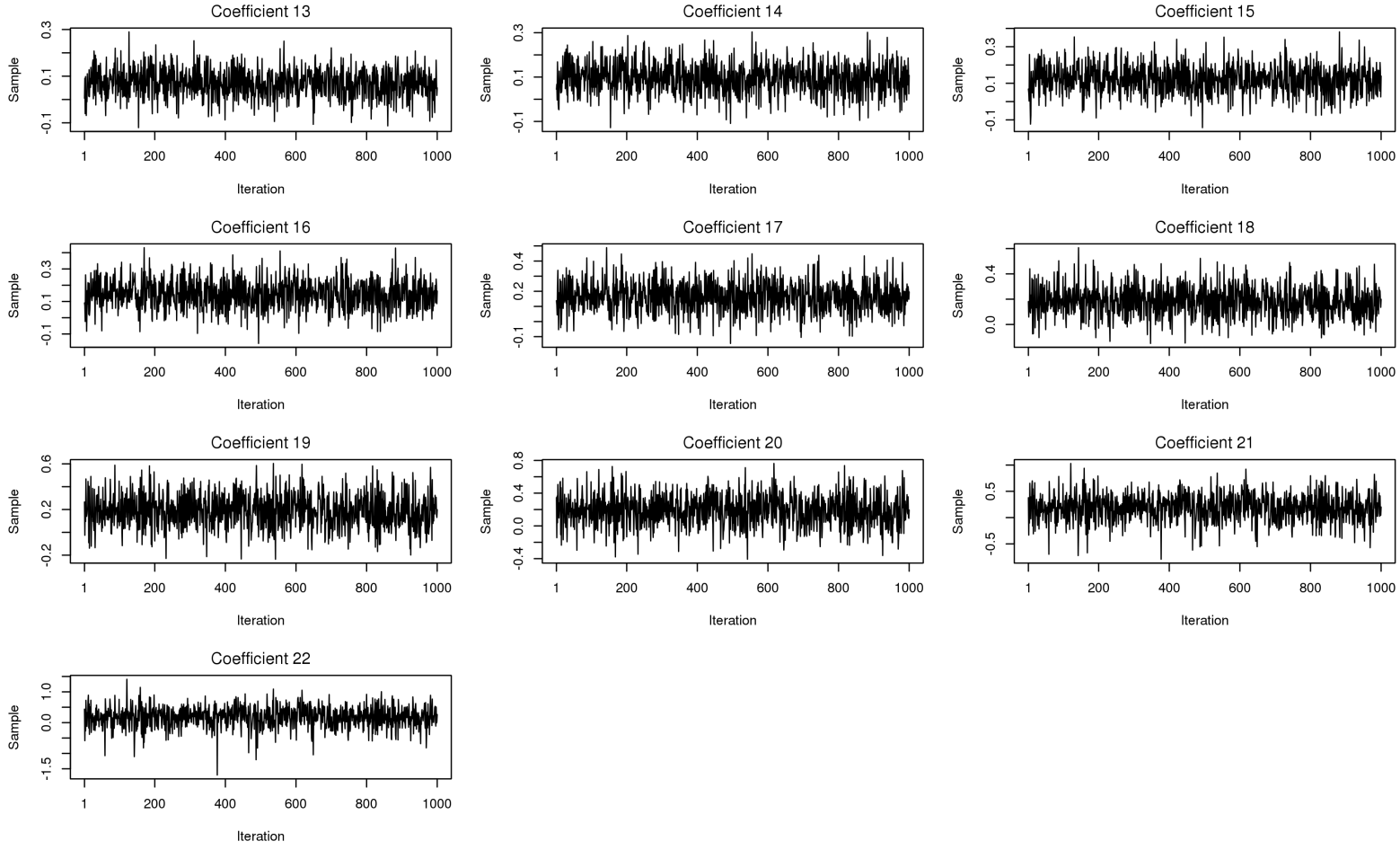




Illustration

Diagnostic plots, sampling paths

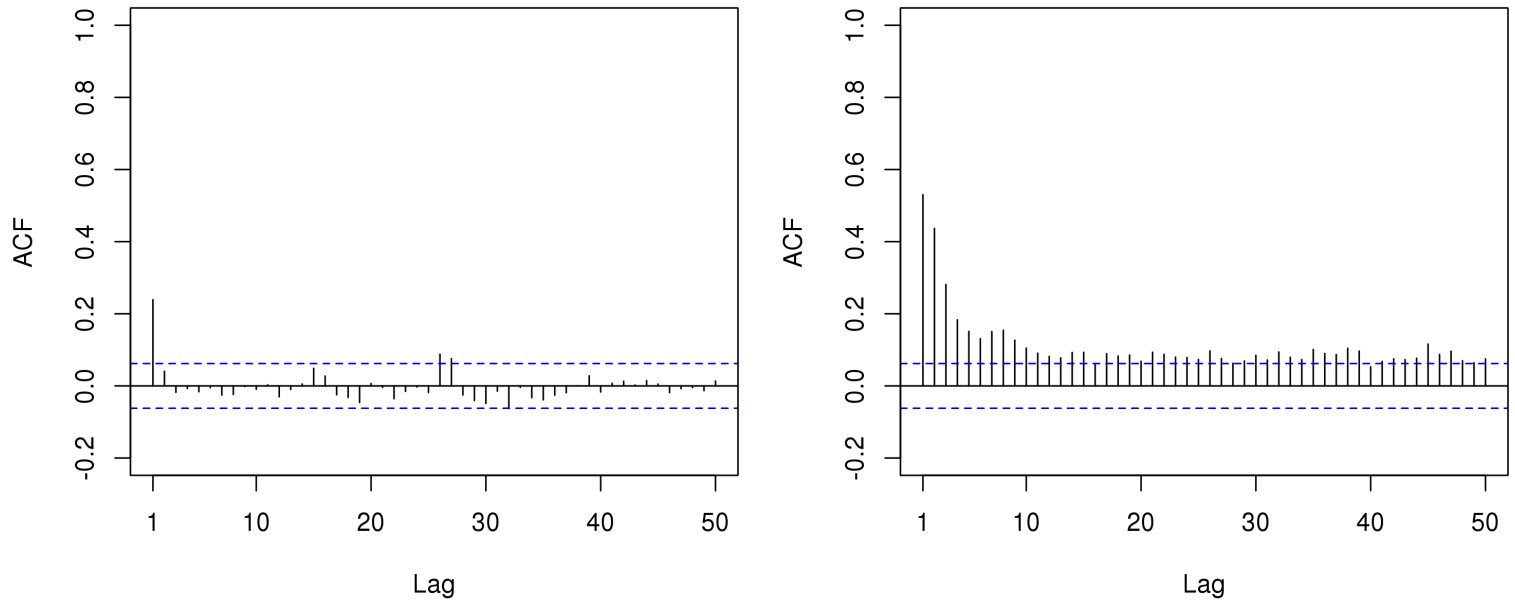
```
R> plot(zm, term = "sx(mbmi)", which = "coef-samples")
```



Illustration

Diagnostic plots, autocorrelation functions and maximum autocorrelation of parameters

```
R> plot(zm, term = "sx(mbmi)", which = "var-samples", acf = TRUE)
R> plot(zm, which = "max-acf")
```



Further inspection through extractor function `samples()`, e.g. with package **coda** is possible.

Illustration

Inspecting the log-file of the **BayesX** binary

```
R> bayesx_logfile(zm)
```

```
> bayesreg b
> map ZambiaBnd
> ZambiaBnd.infile using /tmp/Rtmpa3Z6WF/bayesx/ZambiaBnd.bnd
NOTE: 57 regions read from file /tmp/Rtmpa3Z6WF/bayesx/ZambiaBnd.bnd
> dataset d
> d.infile using /tmp/Rtmpa3Z6WF/bayesx/bayesx.estim.data.raw
NOTE: 14 variables with 4847 observations read from file
/tmp/Rtmpa3Z6WF/bayesx/bayesx.estim.data.raw

> b.outfile = /tmp/Rtmpa3Z6WF/bayesx/bayesx.estim
> b.regress stunting = mbmi(psplinerw2,nrknots=20,degree=3) +
    agechild(psplinerw2,nrknots=20,degree=3) +
    district(spatial,map=ZambiaBnd) + district(random) + memploymentyes +
    urbanno + genderfemale + meducationno + meducationprimary,
    family=gaussian iterations=12000 burnin=2000 step=10
    setseed=2052766222 predict using d
      .
      .
      .
```

Why does it always rain on me?

Is the weather “bad” more frequently on weekends than on other days of the week?

Meteorological literature does report some evidence for such human-induced weekly cycles.

To contribute to this discussion, we apply a modern data-driven approach using STAR models to a newly available high-quality data set for Austria.

Daily precipitation observations over 60 years for a rather dense net of meteorological stations.

Data are homogenized to adjust for effects, e.g. caused through changes in the data collection process or measurement technology.

Why does it always rain on me?

Data are taken from the HOMSTART project

`http://www.zamg.ac.at/forschung/klimatologie/klimawandel/homstart/`

Freely available online for research purposes.

Precipitation is measured in millimeters in a standardized reservoir with a resolution of 0.1mm, transformed to four ordered categories to capture censoring and threshold effects (especially at zero precipitation)

Altogether the data set consists of almost 1,120,000 observations.

Variable	Description
`cat`	Rain intensity: none (≤ 0) (56%), low $(0, 1)$ (11%), medium $[1, 5)$ (16%) and high (≥ 5) (17%).
`long`, `lat`	The longitude and latitude coordinate of the meteorological station.
`time`	The daily calendar date of the measurement, from 1948 to 2009.
`id`	A station id, overall 57 stations across Austria.

Why does it always rain on me?

We apply a threshold model with cumulative probit link given by

$$\Phi^{-1}\left\{P(\texttt{rain}_{it} \leq r)\right\} = \eta_{it}^{(r)},$$

with rain intensity categories $r = (\texttt{none}, \texttt{low}, \texttt{medium})$, stations $i = 1, \ldots, 57$ and time $t = 1, \ldots, 22645$.

Probabilities for the individual categories can then be obtained by taking differences of the cumulative probabilites

$$P(\texttt{rain}_{it} = \texttt{high}) = 1 - P(\texttt{rain}_{it} \leq \texttt{medium}).$$

Why does it always rain on me?

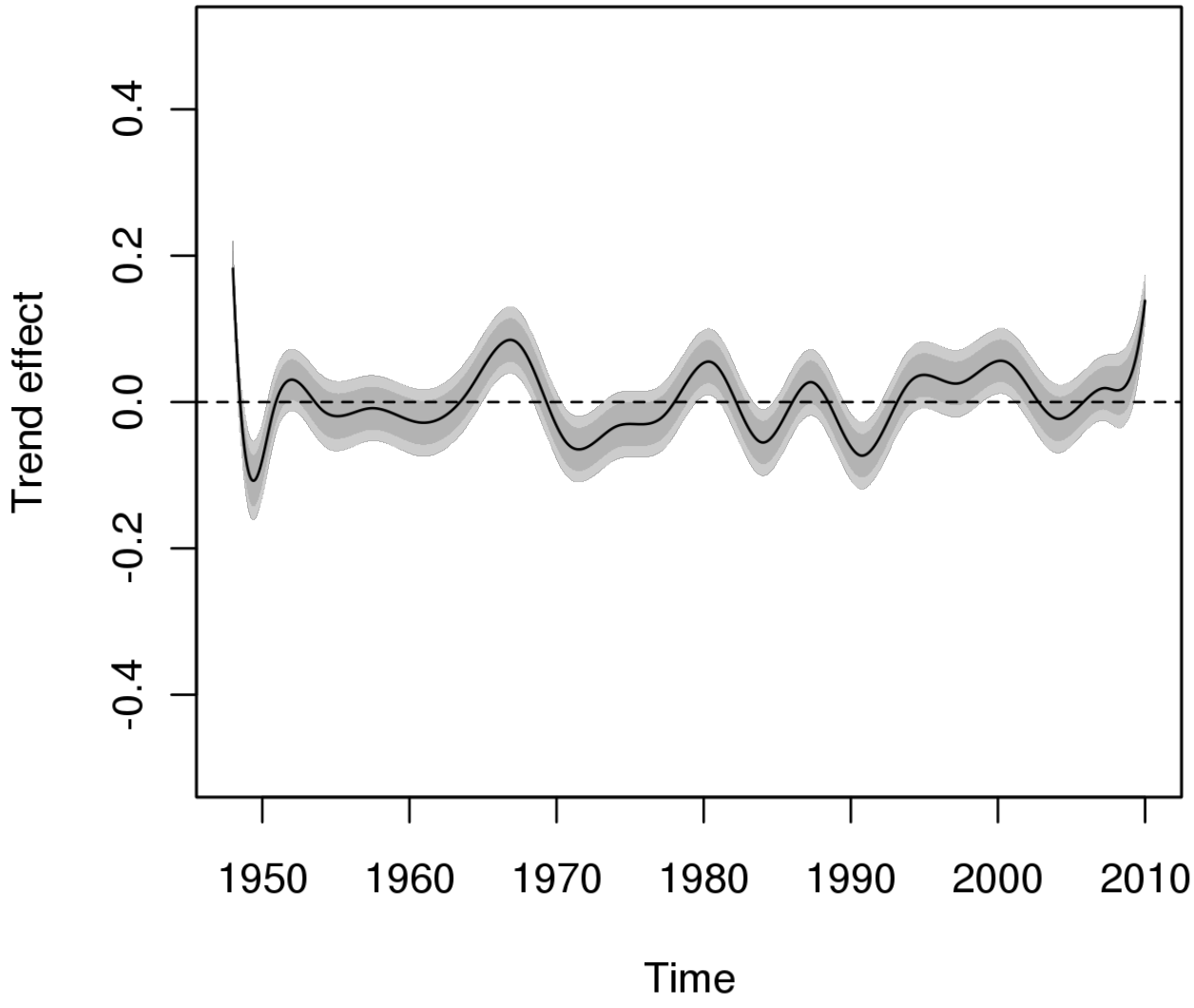
The STAR predictor is represented by

$$\begin{aligned}\eta_{it}^{(r)} &= \xi_r - \{f_{\text{kr}}(\texttt{long}_i, \texttt{lat}_i) + f_{\text{ps}}(t) + \\ &\quad \alpha_{i,1} \cdot \cos(2\pi \cdot t + \phi_{i,1}) + \alpha_{i,2} \cdot \cos(4\pi \cdot t + \phi_{i,2}) \\ &\quad + \omega_i \cdot I_{\texttt{weekend}}(t)\}\,,\end{aligned}$$

where ξ_r is the category specific threshold, functions $f_{\text{kr}}(\cdot,\cdot)$ and $f_{\text{ps}}(\cdot)$ are penalized terms, while the remaining parameters are classical parametric terms.

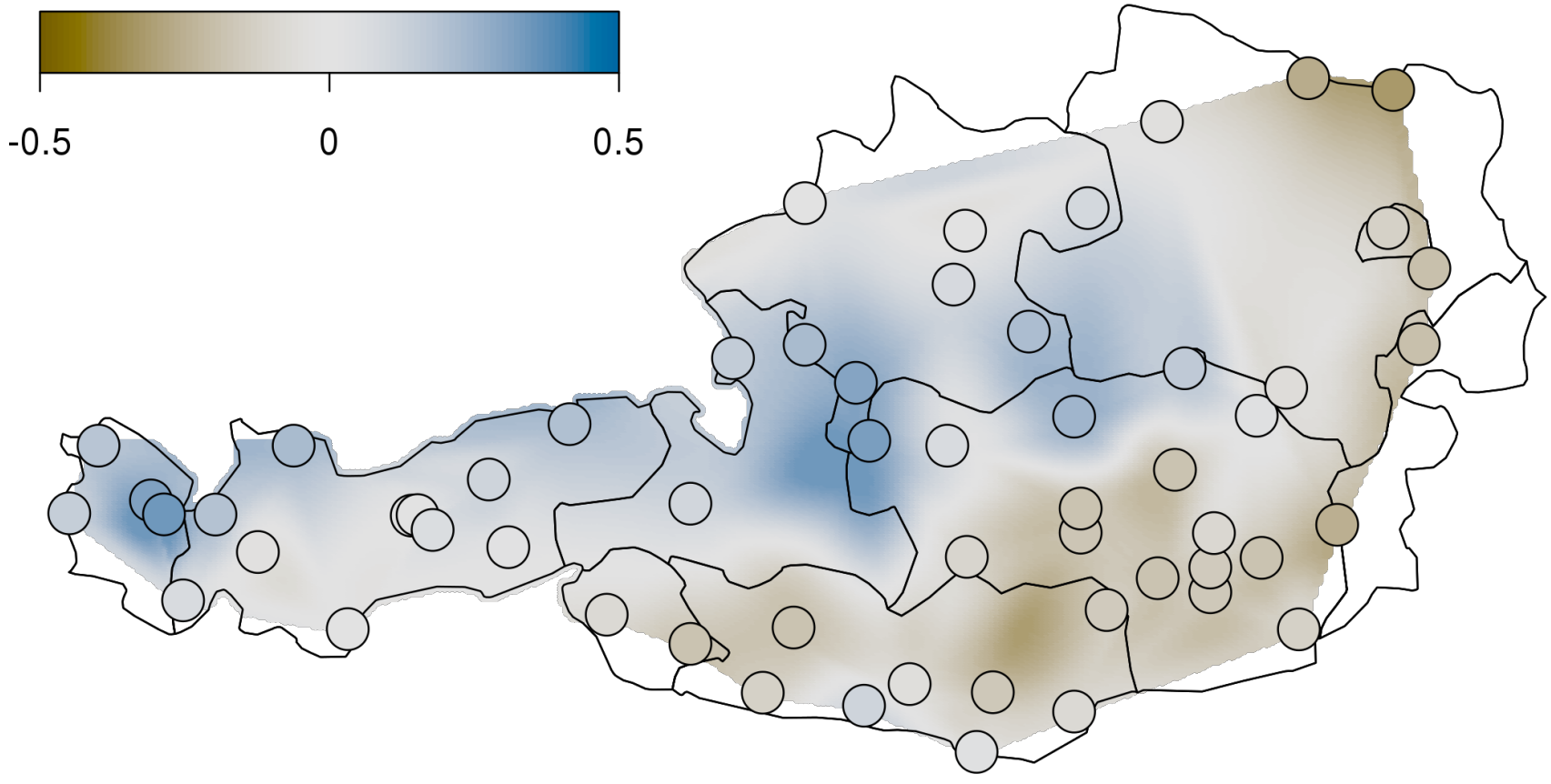
Why does it always rain on me?

Estimated nonlinear time trend across years $\hat{f}_{\text{ps}}(t)$.



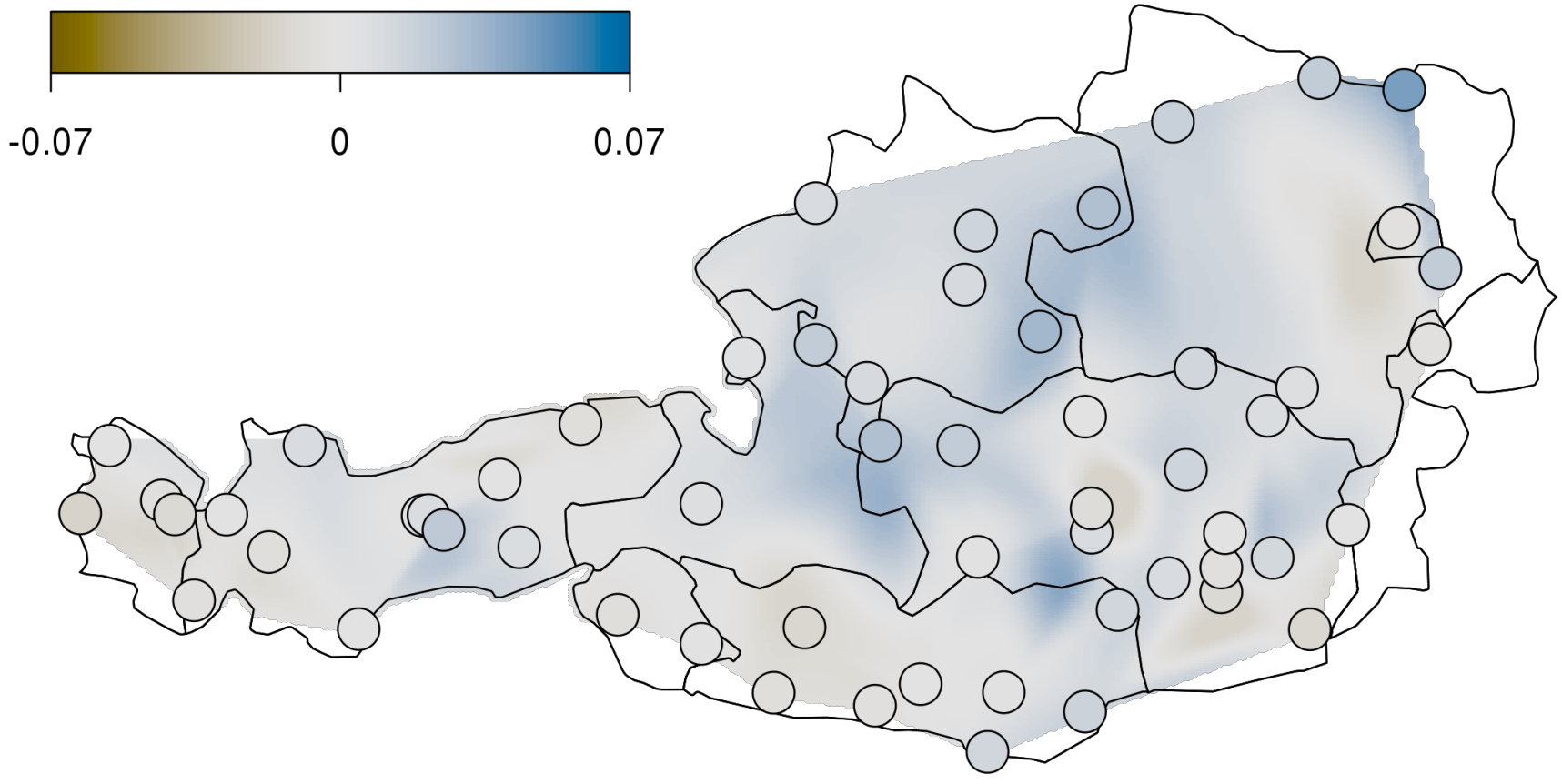
Why does it always rain on me?

Estimated spatially-correlated effect $\hat{f}_{\text{kr}}(\texttt{long}_i, \texttt{lat}_i)$.



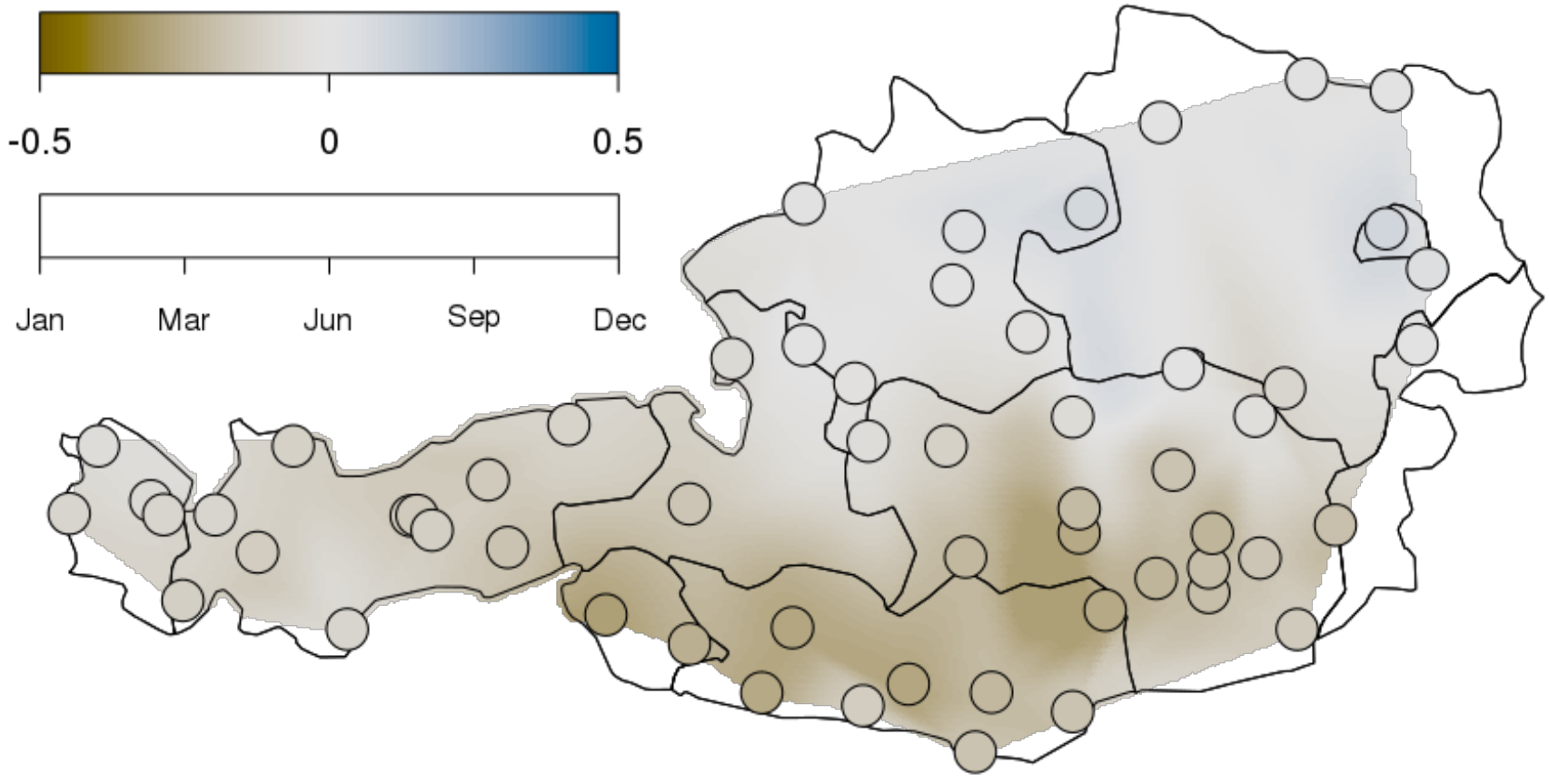
Why does it always rain on me?

Spatial variation of weekend effect $\hat{\omega}_i$.



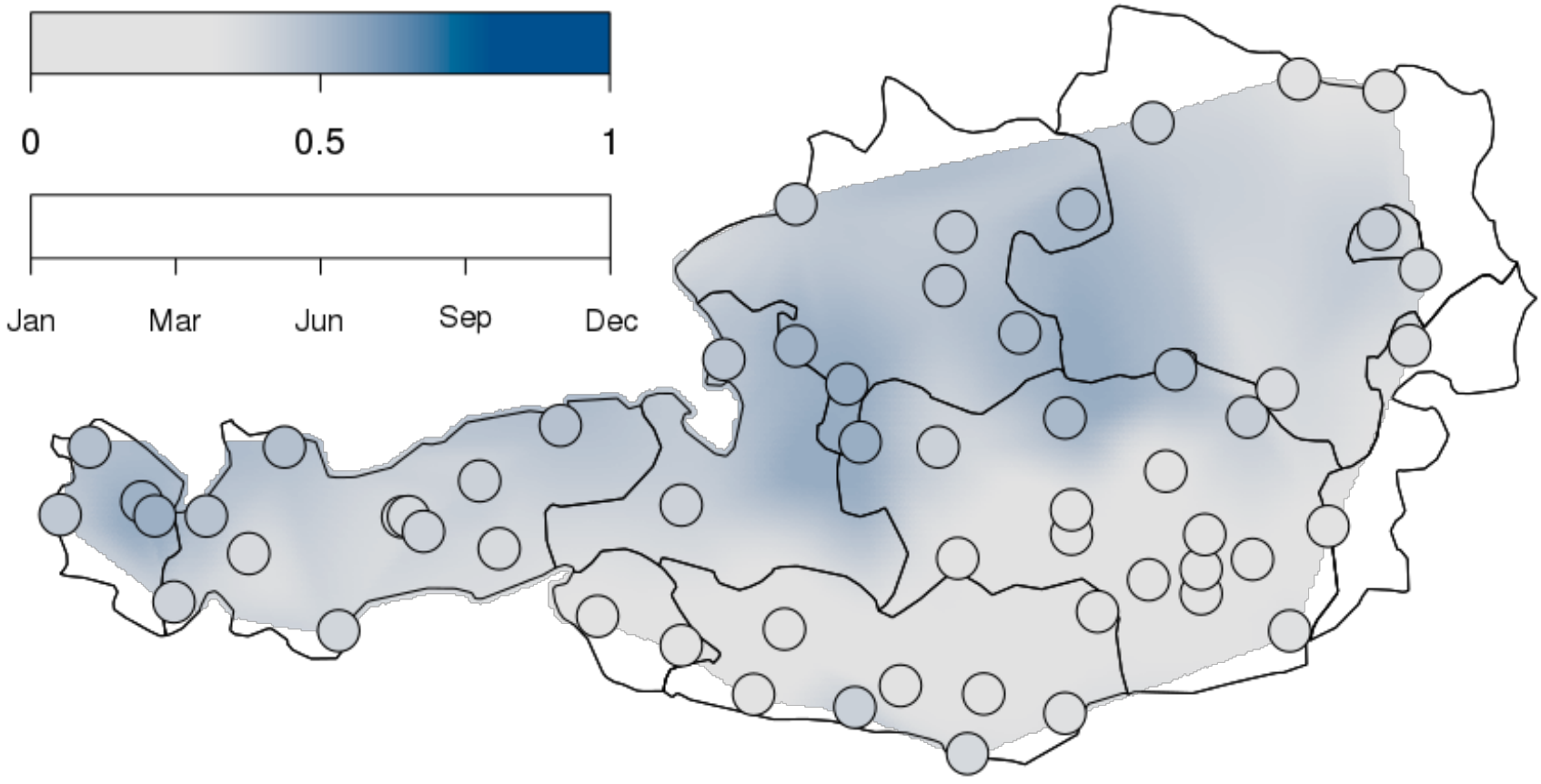
Why does it always rain on me?

Spatial variation of the seasonal effect.



Why does it always rain on me?

Spatial variation of fitted rain probabilities $1 - P(\texttt{rain}_{it} \leq \texttt{none})$.



Why does it always rain on me?

Fitted probabilities for all four categories.

	none: ≤ 0		low: $(0, 1)$		medium: $[1, 5)$		high: ≥ 5	
Location	Jan 1	Jul 1	Jan 1	Jul 1	Jan 1	Jul 1	Jan 1	Jul 1
Bregenz	51.9	40.3	11.4	11.5	17.3	19.6	19.5	28.6
Innsbruck (University)	61.9	41.6	10.5	11.5	14.4	19.4	13.2	27.4
Salzburg (Airport)	53.8	38.9	11.2	11.5	16.7	19.8	18.2	29.8
Hörsching	55.2	49.3	11.1	11.5	16.4	17.9	17.3	21.4
Klagenfurt	70.8	50.9	9.1	11.4	11.4	17.5	8.7	20.2
Graz (University)	58.2	51.0	10.9	11.4	15.5	17.5	15.4	20.1
Zwettl (Stift)	69.9	48.9	9.3	11.5	11.7	18.0	9.1	21.7
Vienna (Hohe Warte)	57.4	57.9	11.0	10.9	15.8	15.6	15.9	15.6
Eisenstadt	64.4	59.1	10.1	10.8	13.6	15.2	11.9	14.8

Why does it always rain on me?

Fitted probabilities for all four categories.

	none: ≤ 0		low: $(0, 1)$		medium: $[1, 5)$		high: ≥ 5	
Location	Jan 1	Jul 1	Jan 1	Jul 1	Jan 1	Jul 1	Jan 1	Jul 1
Bregenz	51.9	40.3	11.4	11.5	17.3	19.6	19.5	28.6
Innsbruck (University)	61.9	41.6	10.5	11.5	14.4	19.4	13.2	27.4
Salzburg (Airport)	53.8	38.9	11.2	11.5	16.7	19.8	18.2	29.8
Hörsching	55.2	49.3	11.1	11.5	16.4	17.9	17.3	21.4
Klagenfurt	70.8	50.9	9.1	11.4	11.4	17.5	8.7	20.2
Graz (University)	58.2	51.0	10.9	11.4	15.5	17.5	15.4	20.1
Zwettl (Stift)	69.9	48.9	9.3	11.5	11.7	18.0	9.1	21.7
Vienna (Hohe Warte)	57.4	57.9	11.0	10.9	15.8	15.6	15.9	15.6
Eisenstadt	64.4	59.1	10.1	10.8	13.6	15.2	11.9	14.8

References

Umlauf N, Kneib T, Lang S, Zeileis A (2012). "Structured additive regression models: An R interface to **BayesX**". *Technical report*, Department of Statistics, Universität Innsbruck.
URL `http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2012-10.pdf`

Umlauf N, Mayr G, Messner J, Zeileis A (2012). "Why does it always rain on me? A spatio-temporal analysis of precipitation in Austria". *Austrian Journal of Statistics,* **41**(1), 81–92.
URL `http://www.stat.tugraz.at/AJS/ausg121/121Zeileis.pdf`

Brezger A, Kneib T, Lang S (2005). "**BayesX**: Analyzing Bayesian Structured Additive Regression Models". *Journal of Statistical Software*, **14**(11), 1–22.
URL `http://www.jstatsoft.org/v14/i11/`

Belitz C, Brezger A, Kneib T, Lang S (2012). ***BayesX*** *– Software for Bayesian Inference in Structured Additive Regression. Models*. Version 2.1. URL `http://www.BayesX.org/`

Fahrmeir L, Kneib T, Lang S (2009). *Regression – Modelle, Methoden und Anwendungen*. 2nd edition. Springer, Berlin.

Kandala NB, Lang S, Klasen S, Fahrmeir L (2001). "Semiparametric analysis of the socio-demographic and spatial determinants of undernutrition in two African countries". *Research in Official Statistics,*, **1**, 81–100.