Using the SemiPar Package

NICHOLAS J. SALKOWSKI
Division of Biostatistics, School of Public Health,
University of Minnesota, Minneapolis, MN 55455, USA
salk0008@umn.edu

May 15, 2008

1 Introduction

The SemiPar package (Wand, 2006) was created to complement *Semiparametric Regression* (Ruppert et al., 2003). This package provides a convenient way to fit splines to data using R (R Development Core Team, 2007). The *SemiPar 1.0 Users' Manual* (Ganguli and Wand, 2005) provides details about syntax and options that are not available through the documentation in R.

For univariate smoothing, SemiPar's default basis function is radial: $\mid x - \kappa_k \mid^p$, where κ_k is the location of the k^{th} knot, and p is the degree. The default degree for the radial basis is 3, so the default function is:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k \mid x - \kappa_k \mid^3 . \tag{1}$$

Alternatively, users can choose a polynomial basis function: $(x - \kappa_k)_+^p$, where p is the degree of the polynomial. The function is thus:

$$f(x) = \sum_{j=0}^{p} \beta_j x^j + \sum_{k=1}^{K} u_k (x - \kappa_k)_+^p . \tag{2}$$

The default degree for the polynomial basis is 1.

For bivariate smoothing, SemiPar's basis function is: $\|x - \kappa_k\|^2 \log \|x - \kappa_k\|$, where x and κ_k are now two-dimensional vectors. The complete function is:

$$f(x) = \beta_0 + \beta_1^T x + \sum_{k=1}^{K} u_k \|x - \kappa_k\|^2 \log \|x - \kappa_k\| . \tag{3}$$

The SemiPar package has algorithms for selecting knots, if knots are not provided by the user. The smoothness of the fit can also be left to SemiPar, but users can control smoothness two ways. In a univariate fit, the smoothing parameter is the ratio of the smoothing variance to the error variance. Large values of the smoothing parameter ($\sigma_\epsilon^2/\sigma_u^2$) produce smoother functions. Alternatively, users can specify the degrees of freedom for the fit. The more degrees of freedom in the fit, the less smooth the function is. The details of using the SemiPar package are best shown through examples. First, the Global Mean Surface Temperature data set is used to demonstrate basic syntax and options in SemiPar. The Pig Jawbone data set is used to show more complex models, and the Slovenia Cancer data is used to demonstrate bivariate smoothing.

2 Global Mean Surface Temperature Example

The Global Mean Surface Temperature (GMST) data set consists of temperature deviations for the years 1881 through 2005. The errors are almost certain to have some autocorrelation, but this feature will be ignored at present. The default cubic spline can be fit to this data set using the following commands:

```
library(SemiPar)
## Get GMST data ##
gmst <- read.table("http://www.biostat.umn.edu/~hodges/GMST.txt",header=T)
## Fit ##
attach(gmst)
fit.gmst <- spm(temp.dev ~ f(Year))
```

In this case the `spm` function of the SemiPar package is used to fit the semiparametric model where temperature deviation is a function of the year. The `summary` function displays basic information regarding the semiparametric fit. In this case, `summary(fit.gmst)` displays:

```
Summary for non-linear components:
           df spar knots
f(Year) 7.247 32.5    30
Note this includes 1 df for the intercept.
```

Note that, by default, 30 knots were selected and 7.247 degrees of freedom were used in the fit, including 2 degrees of freedom for the intercept and slope term in Equation (1). Figure 1 was produced using the following commands:

```
plot(fit.gmst,ylab="Temperature Deviation",
main="Radial Cubic Fit: 7.247 df")
points(gmst)
```

The `df` option allows the user to specify the degrees of freedom for the fit. The degrees of freedom specified include the degrees of freedom associated with any fixed effects in the function. Increasing the degrees of freedom reduces the smoothness of the function:

```
fit.gmst <- spm(temp.dev ~ f(Year, df=15))
```

Reducing the degrees of freedom increases the smoothness of the function:

```
fit.gmst <- spm(temp.dev ~ f(Year, df=3.5))
```

Figure 2 shows the effect of changing the degrees of freedom of the fit. The `spar` option can be used to set the smoothing parameter value instead of specifying the degrees of freedom:

```
fit.gmst <- spm(temp.dev ~ f(Year,spar=10))
```

A smoothing parameter of 10 yields a fit with 14.96 degrees of freedom, so the resulting function is very much like the function produced by specifying 15 degrees of freedom. The degree of the basis function can be changed using the `degree` option, although only odd degrees are allowed for the radial basis. To fit a fifth degree radial spline:

```
fit.gmst <- spm(temp.dev ~ f(Year,degree=5))
```

Figure 3 shows that the resulting fit is quite similar to the fit produced using the default cubic basis. In order to use a polynomial basis, the option `basis="trunc.poly"` must be used:

```
fit.gmst <- spm(temp.dev ~ f(Year,basis="trunc.poly"))
```

Figure 4 shows the results from a first and second degree polynomial fits. For the global mean surface temperature data, the choice of basis function appears to have little impact on the resulting spline fit. Changing the amount of smoothing has an obvious impact on the fit.

3 Pig Jaw Bone Example

Now, consider a more complex data set. The Pig Jaw Bone data has two outcomes: elastic modulus (EM) and hardness. The data come from four pigs (Dino, Mario, Mika, and Paulo), two rest times (1 month and 4 months, 2 pigs for each rest time), the type of anatomy measured (cortical or trabecular), the line (1-9) along which EM or hardness was measured, and the distance along the line where each measurement was made. Lines 1-4 were in cortical bone, and lines 5-9 were in trabecular bone. Consider fitting models for EM and hardness as a function of the line:

```
fit.em <- spm(EM ~ f(Line))
fit.h <- spm(Hardness ~ f(Line))
```

SemiPar's algorithm fails to fit these models, and a somewhat cryptic error is produced:

```
Error in solve.default(sqrt.Omega, t(new.cols)) :
  Lapack routine dgesv: system is exactly singular
```

The SemiPar function `default.knots()` generates a vector of knots. Applying the `default.knots()` function to the Line vector produces only one knot. SemiPar allows the user to specify knots via the `knots` option. The `spm` function produces errors if the equals sign is used within one of its arguments. That is, `knots=seq(from=2,to=8,by=2)` produces an error, but `knots=seq(2,8,2)` is error free. Alternatively, a vector of knots can be created outside the `spm` function:

```
Line.knots <- seq(from=2,to=8, by=2)
fit.em <- spm(EM ~ f(Line,knots=Line.knots))
fit.h <- spm(Hardness ~ f(Line,knots=Line.knots))
```

Now, there are no errors produced.

SemiPar can fit models with more than one function. Consider fitting models where EM and hardness are functions of both the line and the distance. Simply add a function of distance:

```
fit.em <- spm(EM ~ f(Line,knots=Line.knots) + f(distance))
fit.h <- spm(Hardness ~ f(Line,knots=Line.knots) + f(distance))
```

The summary of the model for EM indicates that 3.094 degrees of freedom for function of line and 2.894 degrees of freedom were used for the function of distance. The summary of the model for hardness, however, shows that only one degree of freedom was used for each of the functions of line and distance. Figure 5 displays the results. The functions in the model for hardness are linear. This suggests that a linear regression model would be sufficient to model hardness as a function of line and distance.

The spm function can also include fixed effects in the model. Consider adding the indicator for a four month rest time into the model for EM:

```
fit.em <- spm(EM ~ f(Line,knots=Line.knots) + f(distance) + rest4)
```

Now the summary for the model includes fixed effects:

```
Summary for linear components:
           coef     se ratio p-value
intercept 7.959 0.7235 11.00       0
rest4     1.303 0.1292 10.08       0
Summary for non-linear components:
               df   spar knots
f(Line)     2.489   19.7     4
f(distance) 2.558 3614.0    24
```

Figure 6 shows that the functions of line and distance when rest time is included are smoother than when rest time is not included in the model. This is also indicated by the decrease in the degrees of freedom for the functions.

The spm function has the capability to include a random intercept in a model using random and group arguments. Consider adding a random effect for pig to the previous model for EM:

```
fit.em <- spm(EM ~ f(Line,knots=Line.knots) + f(distance) + rest4,
random=~1, group=Animal)
```

Unfortunately, this produces an error:

```
Error in qr.default(G) : NA/NaN/Inf in foreign function call (arg 1)
```

Looking at Figure 6, the function of distance might be approximated reasonably well by a quadratic function. If the function of distance is replaced with fixed effects for distance and squared distance, then spm can fit the model:

```
fit.em <- spm(EM ~  f(Line,knots=Line.knots) + distance +I(distance^2)+ rest4,
random=~1, group=Animal)
```

The summary for the fit indicates that the fixed effects are both significant, 4.535 degrees of freedom were used for the function of line, and 0.1198 degrees of freedom were used for the random intercept.

```
Summary for linear components:
                    coef        se  ratio p-value
intercept      7.058e+00 3.607e+00  1.956  0.0505
distance       3.203e-03 5.732e-04  5.587  0.0000
I(distance^2) -1.092e-06 3.929e-07 -2.780  0.0054
rest4          1.349e+00 1.358e-01  9.930  0.0000
Summary for non-linear components:
           df  spar knots
f(Line) 4.539 3.664     4
Summary for random intercept component:
                     df
random intercept 0.1228
```

It is possible to treat line and distance as a bivariate predictor. The spm function produces errors when a random effect for animal is included in the model along with a bivariate function of line and distance. SemiPar's bivariate smoothing capabilities can be better demonstrated using the Slovenian stomach cancer data.

4 Slovenian Stomach Cancer Example

Data analyzed by Reich et al. (2006) was acquired from Jim Hodges. The outcome is the count of stomach cancer cases observed for 194 municipalities in Slovenia from 1995 to 2001 (O). Each municipality has an expected number of cases (E). Each municipality is also characterized as either urban or rural (Urban). Also, each municipality is categorized into one of five ordered socioeconomic categories. These categories were then rescaled and centered (SEc), and then were treated as a continuous variable. Municipality centroids are not available, but the two-dimensional midpoint of each municipality is available. The two-dimensional midpoint is defined as the point whose east-west coordinate is the mean of the east-west coordinates of the easternmost and westernmost points in the municipality, and whose north-south coordinate is the mean of the north-south coordinates of the northernmost and southernmost points of the municipality. These midpoints were then rescaled and approximately centered (X1c, X2c).

The observed stomach cancer cases can be modeled through Poisson regression using the `glm` function. The expected number of cases is an offset in the model. If SEc is used as a predictor:

```
fit.glm.poisson <- glm(O ~ SEc + offset(log(E)),family=poisson(link="log"))
```

fits the poisson regression model. The command `summary(fit.glm.poisson)` displays:

```
Call:
glm(formula = O ~ SEc + offset(log(E)), family = poisson(link = "log"))
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-5.27479  -0.98524  -0.02356   0.75129   3.54262
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.15643    0.01839   8.504  < 2e-16 ***
SEc         -0.13683    0.01966  -6.959 3.42e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 414.22  on 193  degrees of freedom
Residual deviance: 367.46  on 192  degrees of freedom
AIC: 1152.9
Number of Fisher Scoring iterations: 4
```

In this model, SEc is a significant predictor of the stomach cancer incidence rate. The `spm` function can fit the same model using the statement:

```
fit.spm.poisson <- spm(O ~ SEc + offset(log(E)),family="poisson")
```

The summary information is more limited in this case:

```
Summary for linear components:

             coef      se  ratio p-value

intercept  0.1564 0.01839  8.504       0

SEc       -0.1368 0.01966 -6.959       0
```

Yet, the fit and conclusions are the same. This model treats every municipality as independent. Reich et al. (2006) used a conditional autoregressive model to account for spatial correlation. Now, consider a modelling spatial correlations through a bivariate smoothing function:

$$\log(\mathrm{E}\left[y_O\right]) = \log(y_E) + \beta_0 + \beta_{SEc}x_{SEc} + \beta_1^T x + \sum_{k=1}^{K} u_k \left\|x - \kappa_k\right\|^2 \log\left\|x - \kappa_k\right\|$$

where y_O is the observed number of cases, y_E is the expected number of cases, x_{SEc} is the centered socioeconomic status, x is the coordinates of the municipality, κ are the knot locations, and β_0, β_{SEc}, β_1 are fixed predictors. SemiPar can fit this model:

```
fit.spm <- spm(O ~ SEc + offset(log(E)) +

f(X1c,X2c,knots=Slovenia.knots), family="poisson")
```

When the user does not supply a matrix of knot locations, SemiPar generates a set of knots, displays them, and prompts the user to save the knots to a file. This way, the knots do not need to be generated every time a new model is fit. R can then read the file to produce an object (Slovenia.knots) that can be used in the `spm` function. The summary of the fit displayed by `summary(fit.spm)` is :

```
Summary for linear components:

            coef      se  ratio p-value

intercept 2.7830 0.02371 117.40       0

SEc       0.6445 0.02822  22.84       0
```

```
Summary for non-linear components:

            df   spar knots

f(X1c,X2c)   2 158100    48
```

The fixed effect for SEc is still significant, but with the opposite sign. Also, only 2.000003 degrees of freedom were used for the smoothing function. Since there is a fixed effect for each direction (i.e., X1c and X2c), a fit with just over 2 degrees of freedom should be similar to the fit from a poisson regression model with fixed effects for X1c and X2c. This model can be fit with the `glm` function and yields estimates $\hat{\beta}_{SEc} = -0.05207$, $SE(\hat{\beta}_{SEc}) = -0.02822$, $\hat{\beta}_{X1c} = 0.18057$, $SE(\hat{\beta}_{X1c}) = 0.04856$, and $\hat{\beta}_{X2c} = 0.02064$, $SE(\hat{\beta}_{X2c}) = 0.06648$. Strangely, the standard error for $\hat{\beta}_{SEc}$ is the same for the two models, but the parameter estimates are clearly different. In the Poisson regression model, the coefficient for SEc is negative, but not significant.

The `adf` option can be used to change the amount of bivariate smoothing. SemiPar cannot guarantee to get the degrees of freedom of a bivariate smoothing function to be exactly what the user chooses, but can generally approximate the degrees of freedom reasonably well. The statement:

```
fit.spm.df2.001 <- spm(O ~ SEc + offset(log(E)) +

f(X1c,X2c, knots=Slovenia.knots,adf=2.001),

family="poisson")
```

generates a model with 2.001 degrees of freedom for bivariate smoothing, and estimates $\hat{\beta}_{SEc} = -0.05207$, $SE(\hat{\beta}_{SEc}) = 0.02822$. This fit is nearly identical to the Poisson fit including fixed effects for the two directions. The estimated coefficients for the fixed effects for X1c and X2c were 0.18058 and 0.02064, respectively. (These coefficients can be found in `fit.spm.df2.001$fit$coef$fixed`.) Since the model with the default degrees of freedom (2.000003) does not appear to be similar to the model with 2.001 or the poisson model including direction effects, it is likely that some error occurred with the default fit.

Now, consider increasing the degrees of freedom for bivariate smoothing. With the default knot selection, up to 50 degrees of freedom can be assigned to the bivariate smoothing function, since there are 48 knots and 2 fixed effects direction. Figure 7 shows the effect of increasing the smoothing degrees of freedom on the estimate for the SEc fixed effect, its standard error, and the p-value for the test of the hypothesis that the SEc fixed effect is equal to zero. As the bivariate smoothing degrees of freedom increase, both the estimate of the SEc fixed effect and its standard error increase. It is troubling that the estimate of the SEc fixed effect changes sign depending on the degrees of freedom for bivariate smoothing. Meaningful interpretation of the fixed effect is difficult, since its sign depends on the level of spatial smoothing. Perhaps it is fortunate that the fixed effect for SEc is not significant when spatial effects are included in the model. Reich et al. (2006) found a similar effect. When a conditional autoregressive model for the spatial correlation was included, the absolute value of $\hat{\beta}_{SEc}$ decreased, and centered socioeconomic status was not longer statistically significant.

This effect on the fixed effect is similar to the effect described in Reich et al. (2006). When the bivariate smoothing is allocated about 26 degrees of freedom, the estimates of both the fixed effect for SEc (-0.022) and its standard error (0.034) are near the values under a CAR model (Reich et al., 2006). This suggests that the spatial variation in stomach cancer can be described similarly by a CAR model orbivariate smoothing with default knots and about 26 degrees of freedom.

The `plot` function in SemiPar can be used to generate a plot of the bivariate smoothing function. If the user does not supply a boundary matrix, SemiPar prompts the user to draw a boundary and provides an opportunity to save the boundary data to a file. R can then read the file to produce and object (sb) that can be used with the `plot()` function to define the boundary. Figure 8 shows the bivariate smoothing functions with approximately 2.001 and 26 degrees of freedom. Note that the function with 26 degrees of freedom is less smooth. It clearly has more peaks and valleys, and its minimum and maximum values are farther apart.

The first panel was produced with the following statement:

```
plot(fit.spm.df2.001,
bdry=sb,plot.it=FALSE,plot.image=TRUE,
image.xlim=c(-1.5,1.5),image.ylim=c(-1.5,1.5),leg.loc=c(-1,-0.9),
leg.dim=c(2,0.3),image.main="2.001 df",image.grid.size = c(128,128))
```

The first argument is the model object. The `bdry` option chooses the boundary for the plot. The `plot.it` option selects whether to plot the fixed effects. If there were more than one fixed effect, a logical vector could be supplied to select which fixed effects to plot. The `plot.image` option selects whether to plot the bivariate smoothing function. The `image.grid.size` option controls the number of pixels in the plot. The remaining options relate the axes, legend, and title.

Looking at the locations of the knots associated with the largest absolute random effects can provide some insight into the bivariate smoothing function. Figure 9 shows the ordered absolute random effects for a model with about 5 degrees of freedom for bivariate smooothing. The locations of the knots associated with the largest seven absolute estimated random effects is shown, as well. The knots are located at the ends of the longest axis of Slovenia (i.e., the southwest-northeast axis) and in central Slovenia. In Slovenia, the stomach cancer rates are higher than expected in the northeast and lower than expected in the southwest. Also, socioeconomic status tends to be higher in the southwest and lower in the northeast. So, the locations of the knots with the largest absolute estimated random effects appears to be closely related to the underlying pattern of disease.

Figures 10 and 11 show similar information for models with about 10 and 20 degrees of freedom for bivariate smoothing. The locations of the knots associated with the highest absolute estimated random effects are similar in each case. As the degrees of freedom increase, however, the largest absolute estimated random effects increase.

5 Summary

The SemiPar package provides a convenient and easy way to fit spline models using R. It produces useful plots which include confidence bands for univariate spline functions. Unfortunately, its summary output is limited. Documentation of its fitted model objects is essentially nonexistent, making interpreting the results difficult. When the `spm` function fails, errors may be uninformative. The SemiPar can also handle Poisson regression models with the log link and logistic regression models.

One of the strengths of the package is for bivariate smoothing. Fitting models with different degrees of smoothing is simple. This is quite positive, since selection of a default level of smoothing may lead to errors. Analysis of the Slovenian stomach cancer data using bivariate smoothing yields results similar to CAR models.

References

Ganguli, B. and Wand, M. (2005). SemiPar 1.0 Users Manual.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich, B., Hodges, J., and Zadnik, V. (2006). Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models. *Biometrics*, 62(4):1197–1206.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Wand, M. (2006). *SemiPar: Semiparametic Regression*. R package version 1.0-2.

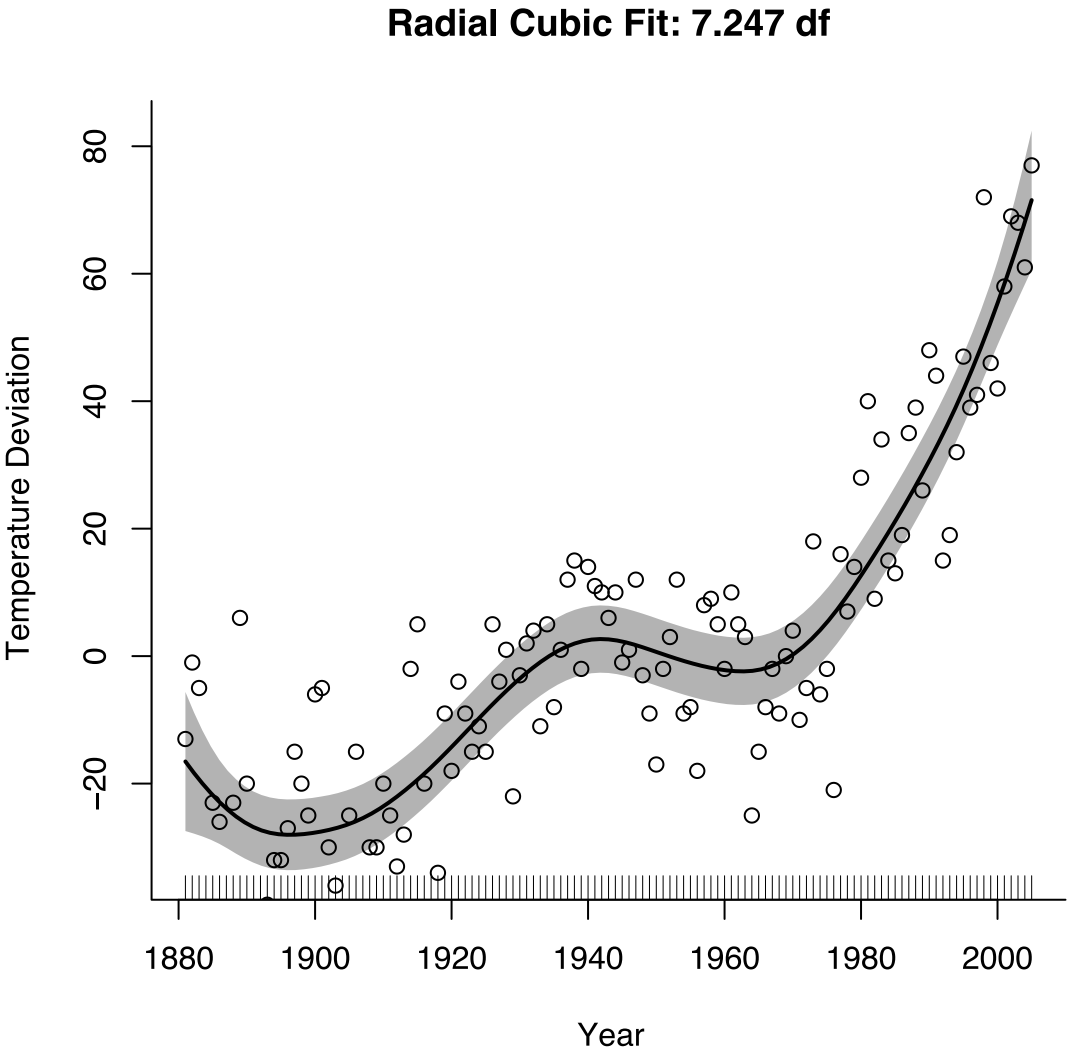


Figure 1: Default radial cubic fit for the GMST data.

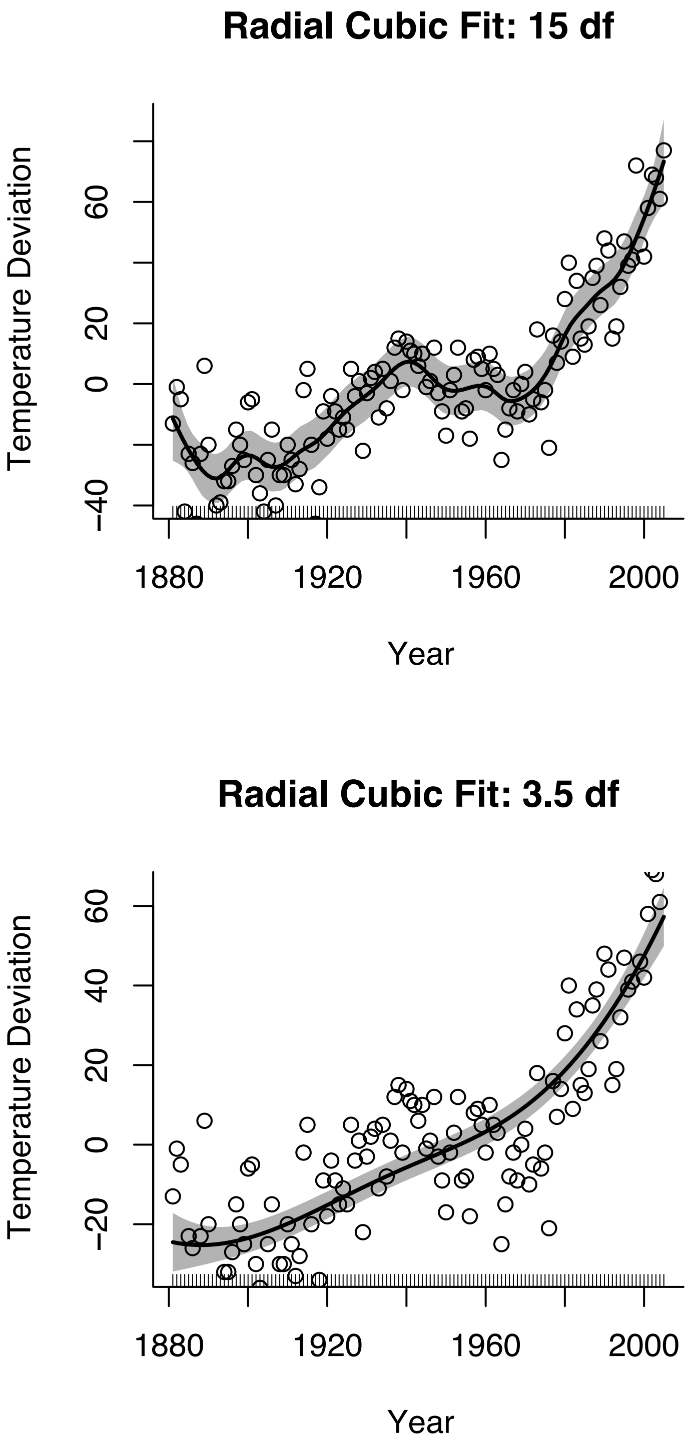


Figure 2: Changing the degrees of freedom changes the smoothness of the fit.

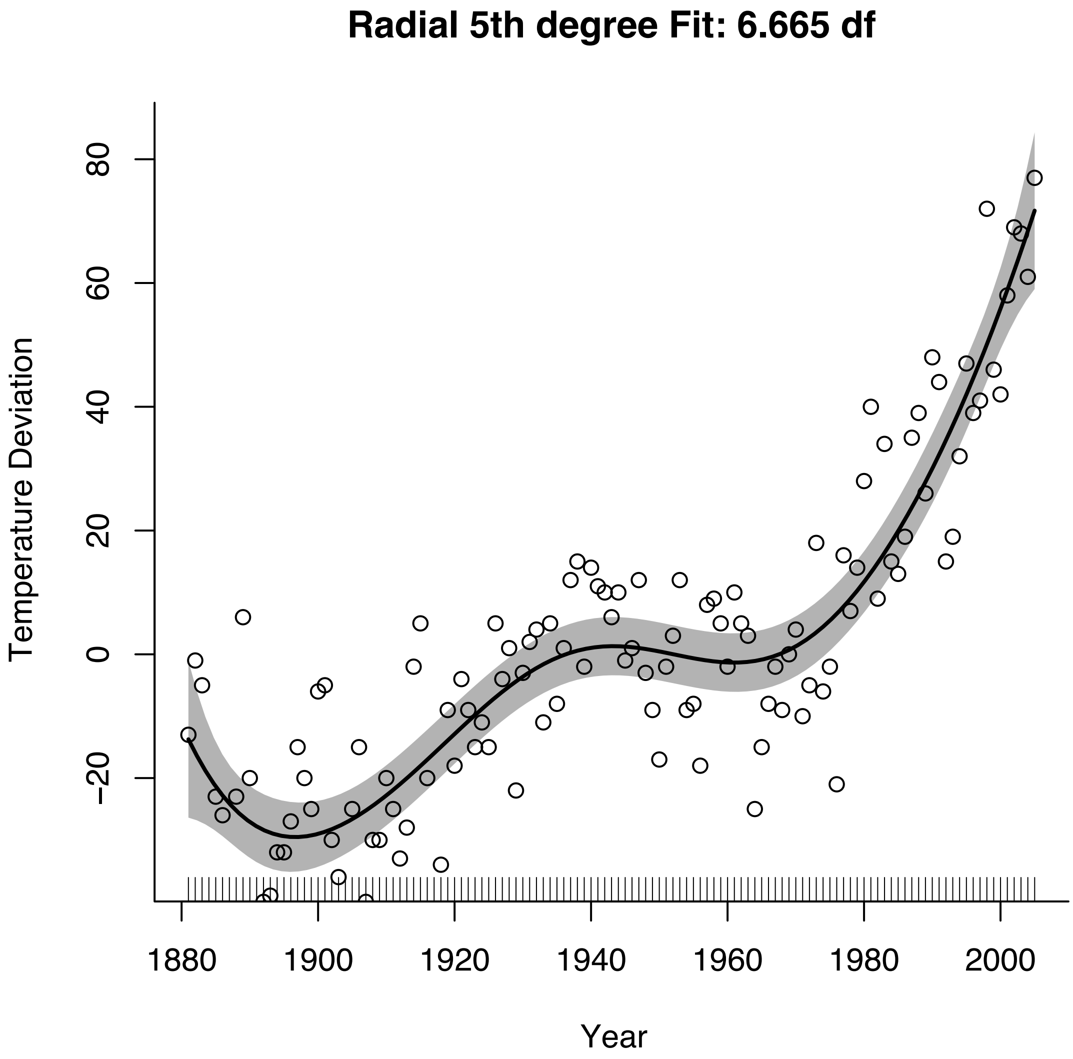
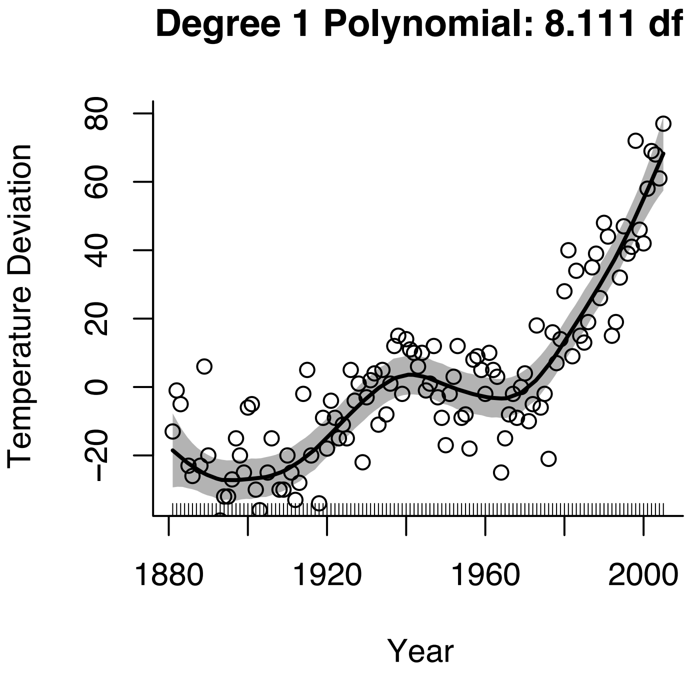


Figure 3: Radial 5th degree fit for the GMST data.



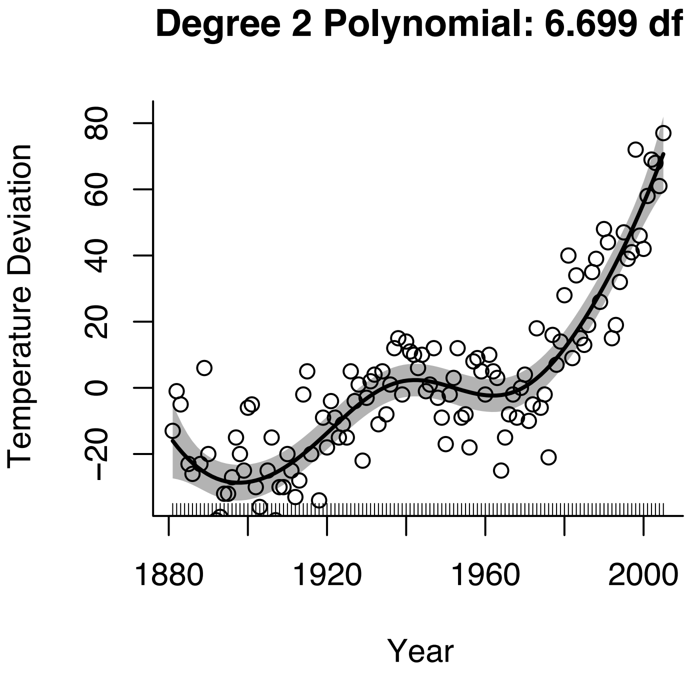


Figure 4: Polynomial fits. 1st degree (top) and 2nd degree (bottom).

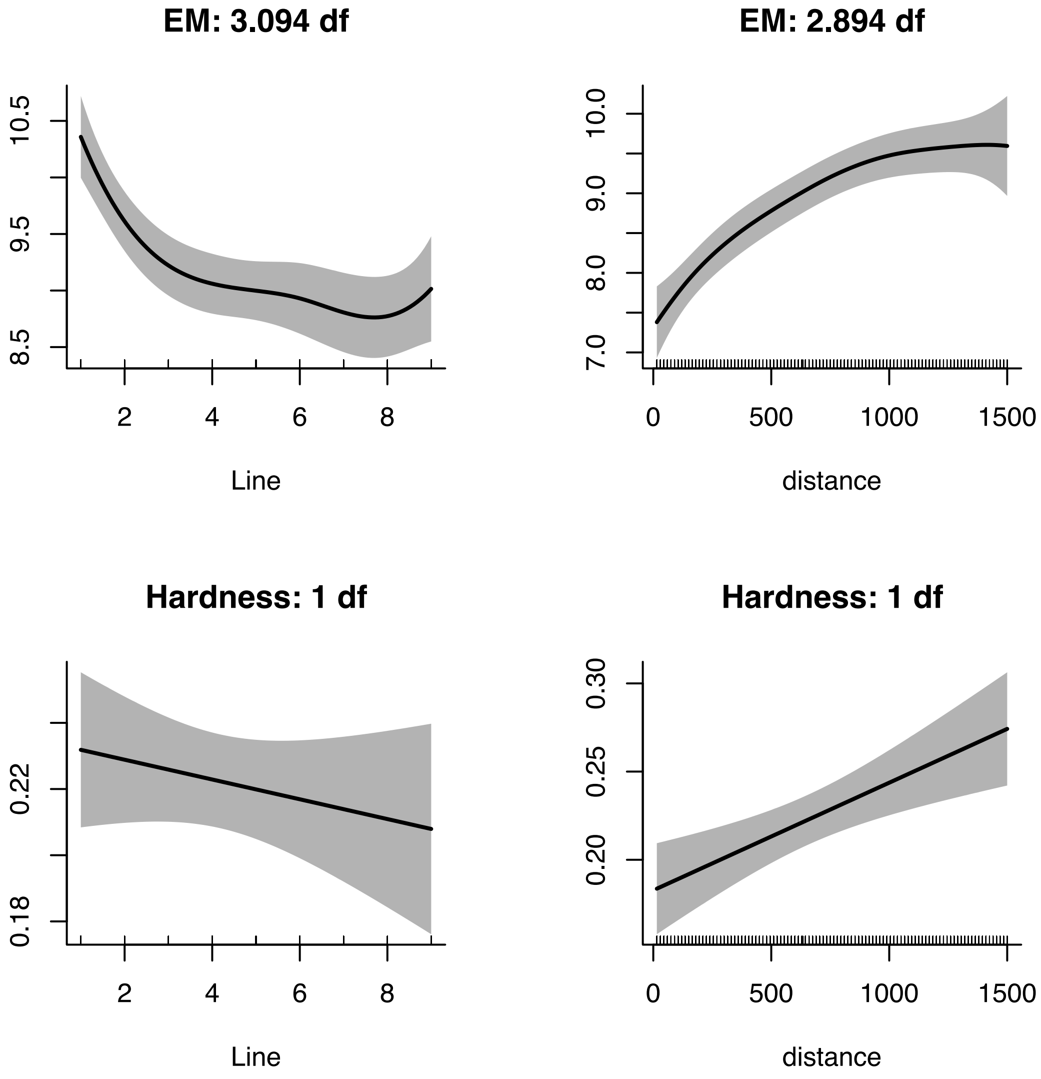


Figure 5: EM and hardness as functions of line and distance.

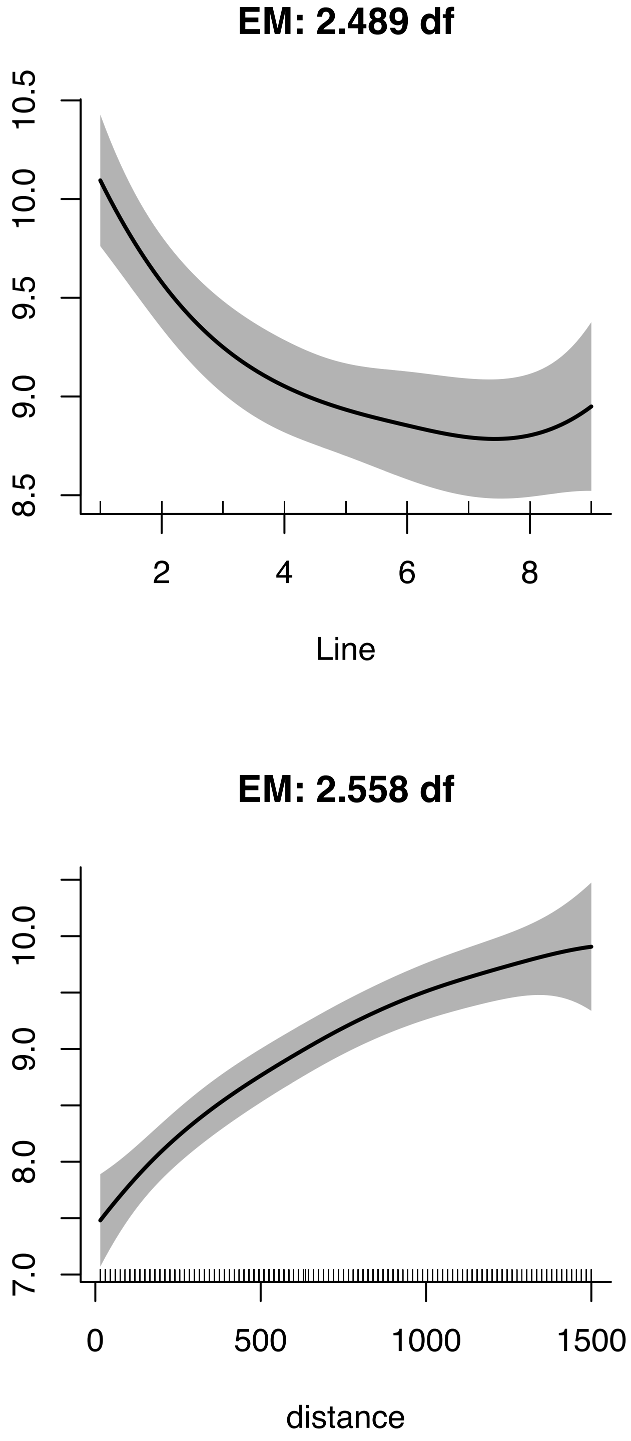


Figure 6: EM as a function of line and distance when a fixed effect for rest time is included.

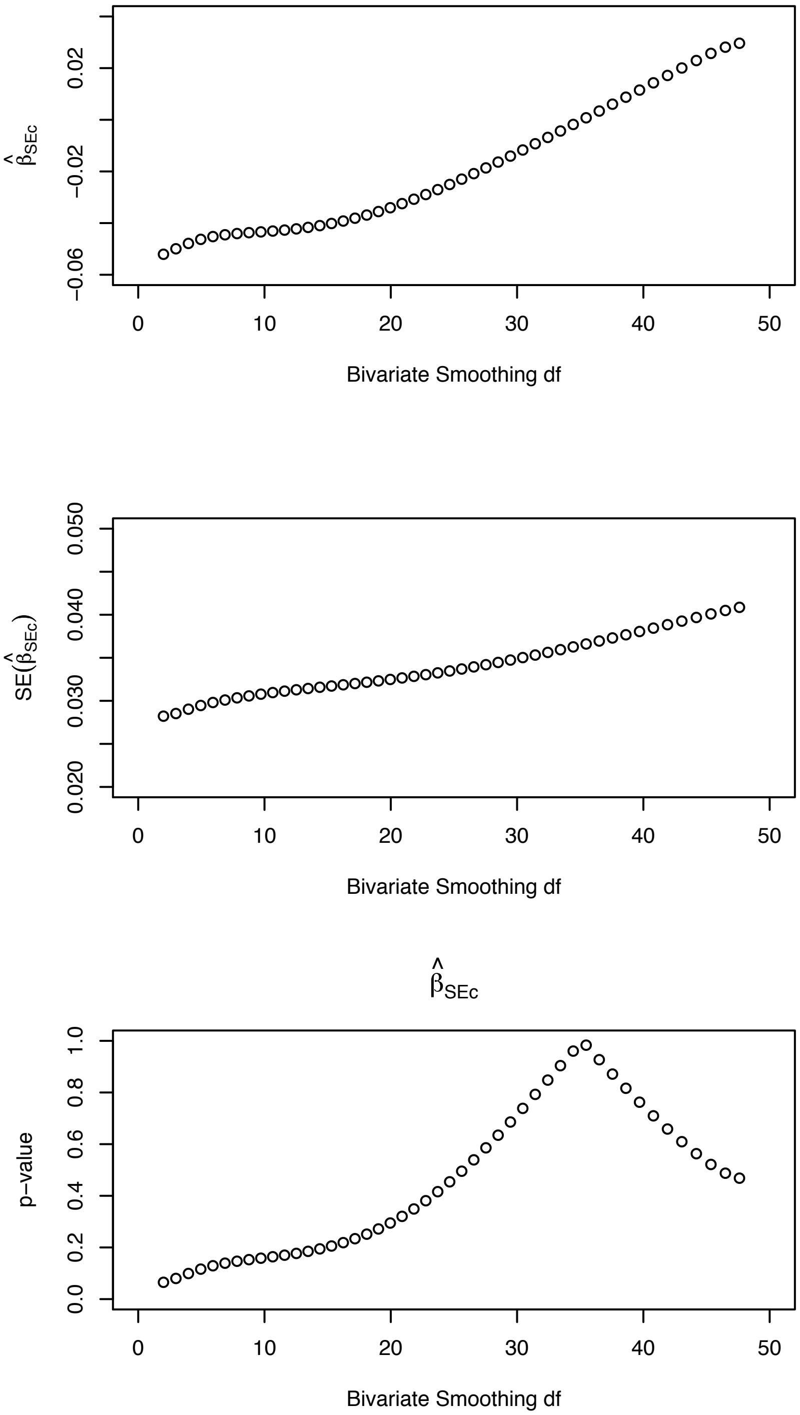


Figure 7: $\hat{\beta}_{SEc}$, the standard error of $\hat{\beta}_{SEc}$, and the p-value for the test of the hypothesis that $\hat{\beta}_{SEc} = 0$ as a function of degrees of freedom for bivariate smoothing.

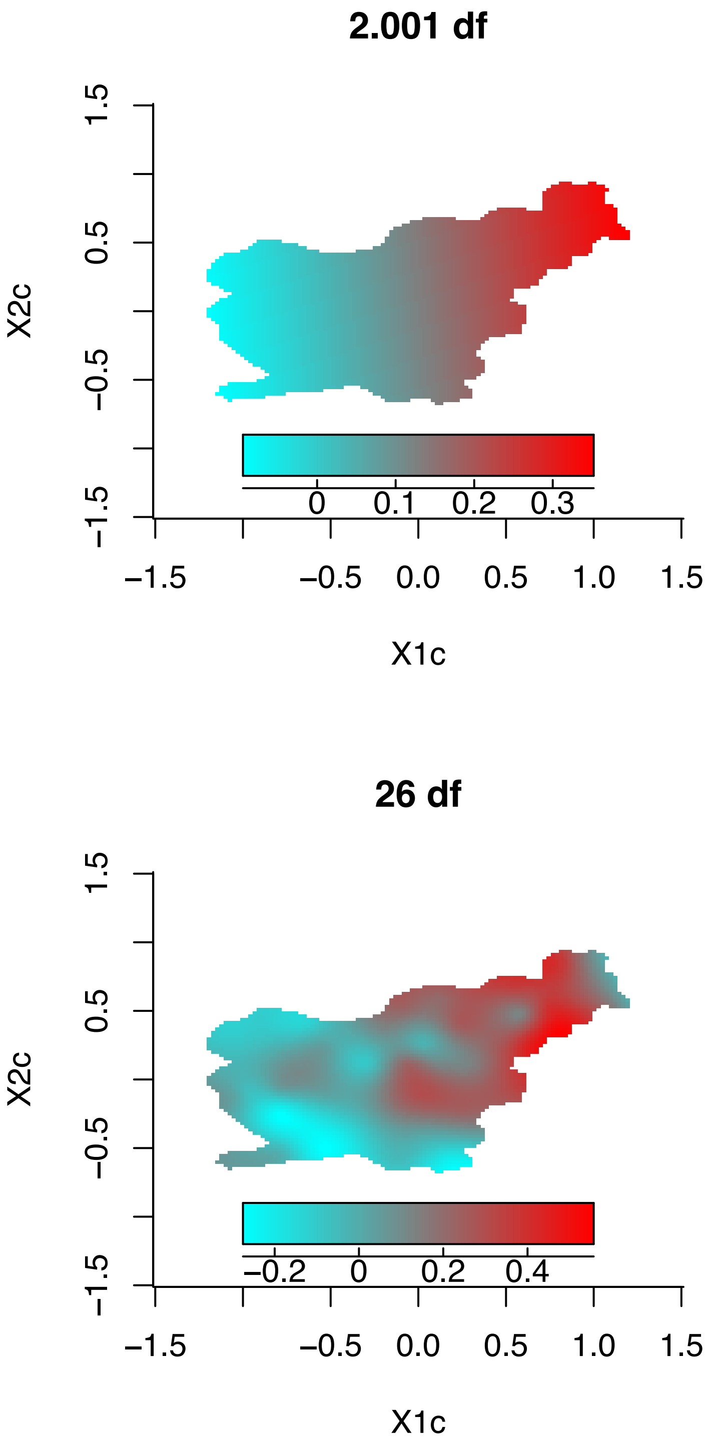


Figure 8: Plots of the bivariate smoothing function with 2.001 and 26 degrees of freedom.

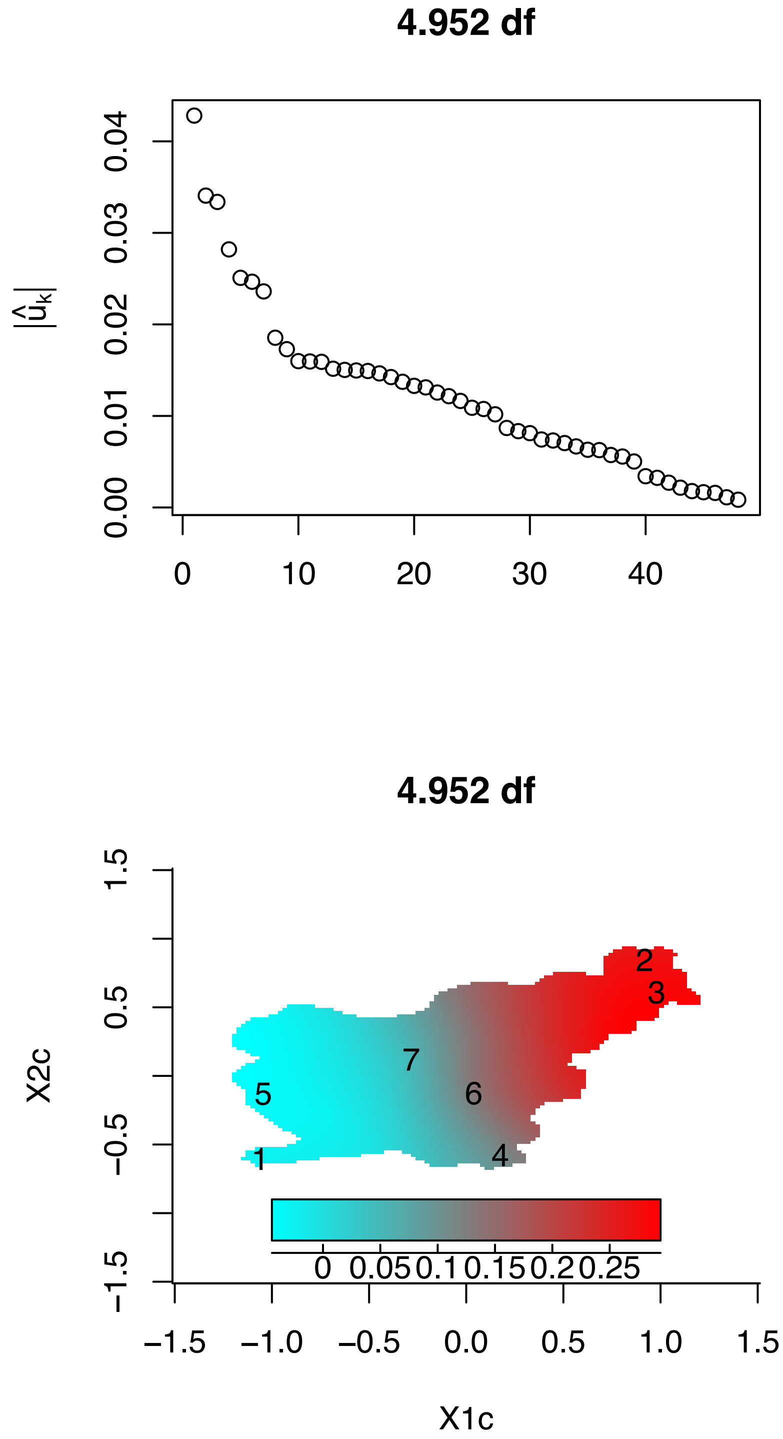


Figure 9: Ordered absolute estimated random effects with 4.952 degrees of freedom (top). Location of the knots associated with the 7 largest absolute estimated random effects.

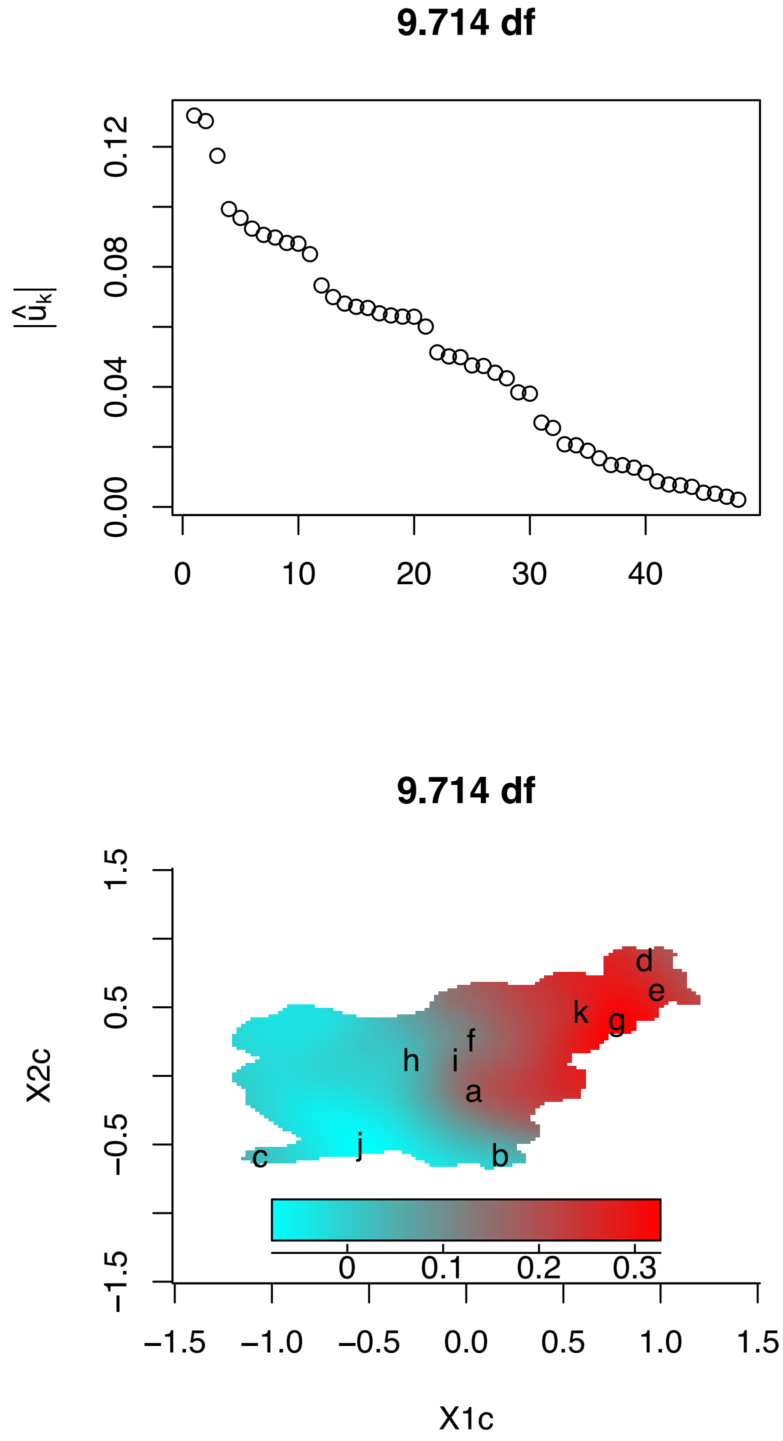


Figure 10: Ordered absolute estimated random effects with 9.714 degrees of freedom (top). Location of the knots associated with the 11 largest absolute estimated random effects.

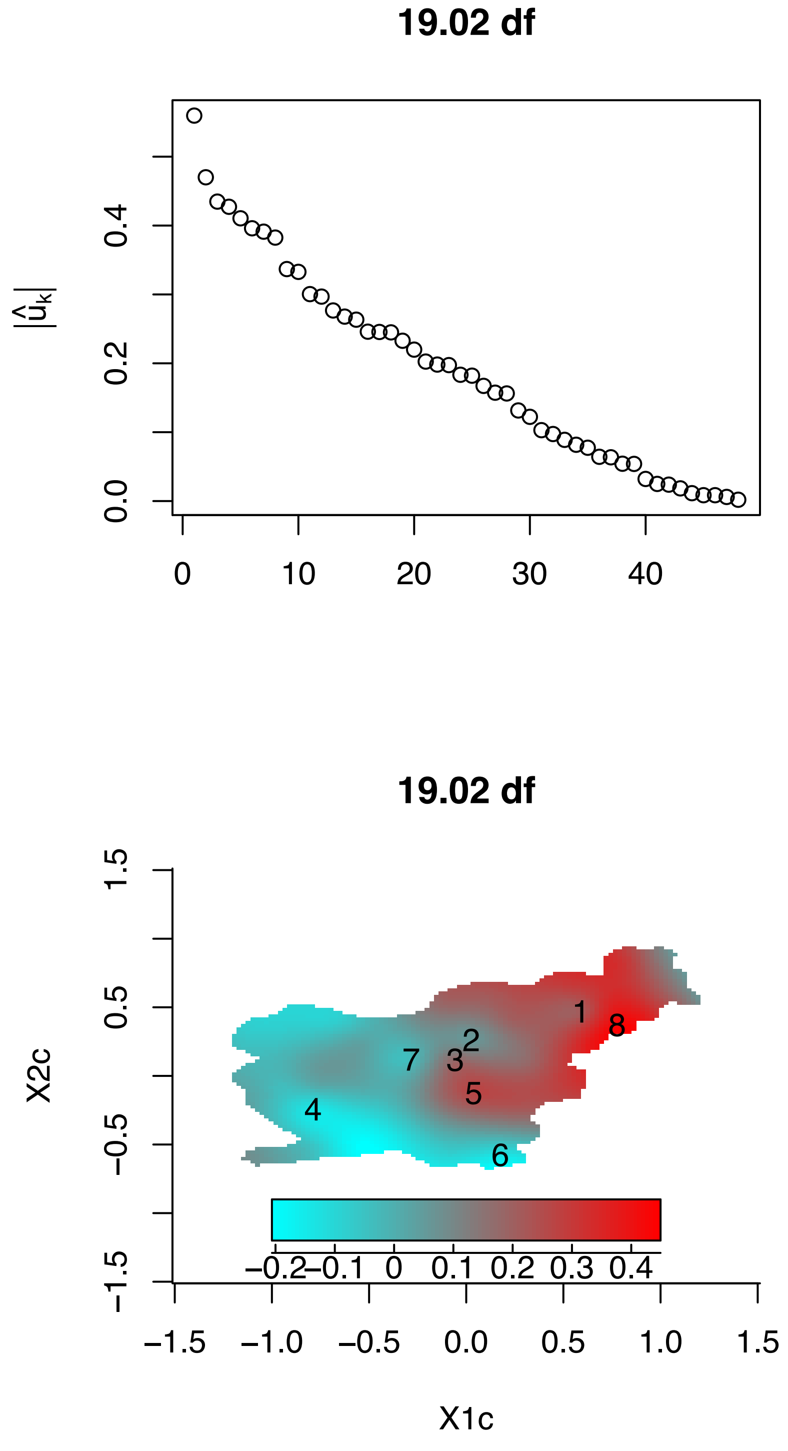


Figure 11: Ordered absolute estimated random effects with 19.02 degrees of freedom (top). Location of the knots associated with the 8 largest absolute estimated random effects.