`varycoef`: Modeling Spatially Varying Coefficients

Jakob A. Dambon[1, 2], Fabio Sigrist[2], Reinhard Furrer[1, 3]

[1]University of Zurich,
Department of Mathematics

[2]Lucerne University of Applied Sciences and Arts,
Institute of Financial Services Zug

[3]University of Zurich,
Department of Computational Science

Introduction

```r
library(sp)
data(meuse)
head(meuse[, 1:8]) # first columns
```

```
##        x      y cadmium copper lead zinc  elev       dist
## 1 181072 333611    11.7     85  299 1022 7.909 0.00135803
## 2 181025 333558     8.6     81  277 1141 6.983 0.01222430
## 3 181165 333537     6.5     68  199  640 7.800 0.10302900
## 4 181298 333484     2.6     81  116  257 7.655 0.19009400
## 5 181307 333330     2.8     48  117  269 7.480 0.27709000
## 6 181390 333260     3.0     61  137  281 7.791 0.36406700
```

Introduction: Linear Regression

Linear Model:

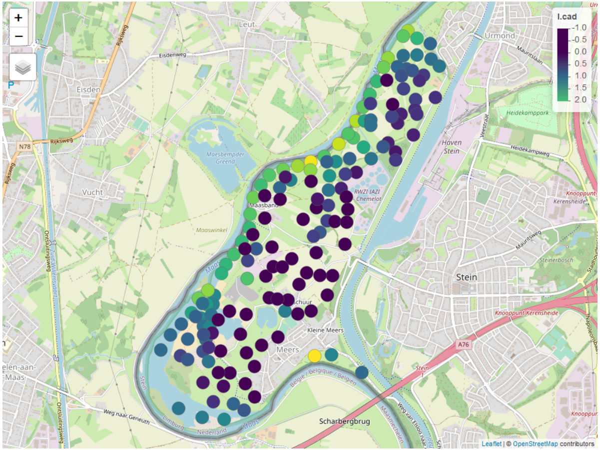
$$y_i = \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i$$

Fitting a simple Linear Model:

```
linmod <- lm(log(cadmium)~elev+dist, data = meuse)
coef(linmod)
```

```
## (Intercept)        elev        dist
##   5.5038907  -0.5298496  -2.5681280
```

Introduction: Spatial Structure



Introduction: Spatial Statistics

Geo-Statistical Model:

$$y_i = \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + Z(\mathbf{s}_i)$$

where $\mathbf{Z}(\mathbf{s})$ depends on location $\mathbf{s}$

underlying the **First Law of Geography** by Tobler (1970):

"Everything is related to everything else, but near things are more related than distant things."

Introduction: Spatial Statistics

Geo-Statistical Model:

$$y_i = \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + Z(\mathbf{s}_i)$$

where $\mathbf{Z}(\mathbf{s})$ depends on location $\mathbf{s}$

underlying the **First Law of Geography** by Tobler (1970):

“Everything is related to everything else, but near things are more related than distant things.”

Introduction: Spatial Statistics

Modeling $\mathbf{Z}(\mathbf{s})$:

- Splines `{akima}`
- (Bayesian) Gaussian processes `{geoR, gstat, spBayes}`

Spatially Varying Coefficient Models

Idea: Each coefficient has a spatial structure.

$$y_i = \beta_1(\mathbf{s}_i)x_i^{(1)} + \ldots + \beta_p(\mathbf{s}_i)x_i^{(p)} + \varepsilon_i$$

Pros:

- Not a black-box:

  Given a location **s**, the SVC model reduces to a linear model.

- High flexibility

Cons:

- Computational intensive, depending on how SVCs are defined.
- Moderate size of data necessary

Spatially Varying Coefficient Models

Idea: Each coefficient has a spatial structure.

$$y_i = \beta_1(\mathbf{s}_i)x_i^{(1)} + \ldots + \beta_p(\mathbf{s}_i)x_i^{(p)} + \varepsilon_i$$

Pros:

- Not a black-box:

  Given a location **s**, the SVC model reduces to a linear model.

- High flexibility

Cons:

- Computational intensive, depending on how SVCs are defined.
- Moderate size of data necessary

Existing Methodologies for Large Data

- Geographically Weighted Regression (non-model-based, Fotheringham et al., 2002) {`spgwr`, `GWmodel`, `gwrr`}
- Approximation via Gaussian Markov Random Fields (Lindgren et al., 2011) {`INLA`}
- **Maximum Likelihood Estimation** (Dambon et al., 2020) {`varycoef`}

Existing Methodologies for Large Data

- Geographically Weighted Regression (non-model-based, Fotheringham et al., 2002) {`spgwr`, `GWmodel`, `gwrr`}
- Approximation via Gaussian Markov Random Fields (Lindgren et al., 2011) {`INLA`}
- **Maximum Likelihood Estimation** (Dambon et al., 2020) {`varycoef`}

The R Package `varycoef`

- Models SVC as Gaussian processes on large data with
  - *Covariance tapering* (Furrer et al., 2006) `{spam}`: large n feasible ($n > 10^4$)
  - *Parallelized optimization* (Gerber and Furrer, 2019) `{optimPrallel}`: $p > 10$
- Fitted model of class `"SVC_mle"`

```
library(varycoef)
methods(class = "SVC_mle")
```

```
##  [1] coef            fitted          logLik          nobs
##  [5] plot            predict         print           residuals
##  [9] summary         SVC_mle_control
## see '?methods' for accessing help and source code
```

The R Package `varycoef`

- Models SVC as Gaussian processes on large data with
  - *Covariance tapering* (Furrer et al., 2006) `{spam}`: large n feasible ($n > 10^4$)
  - *Parallelized optimization* (Gerber and Furrer, 2019) `{optimPrallel}`: $p > 10$
- Fitted model of class `"SVC_mle"`

```
library(varycoef)
methods(class = "SVC_mle")
```

```
##  [1] coef            fitted          logLik          nobs
##  [5] plot            predict         print           residuals
##  [9] summary         SVC_mle_control
## see '?methods' for accessing help and source code
```

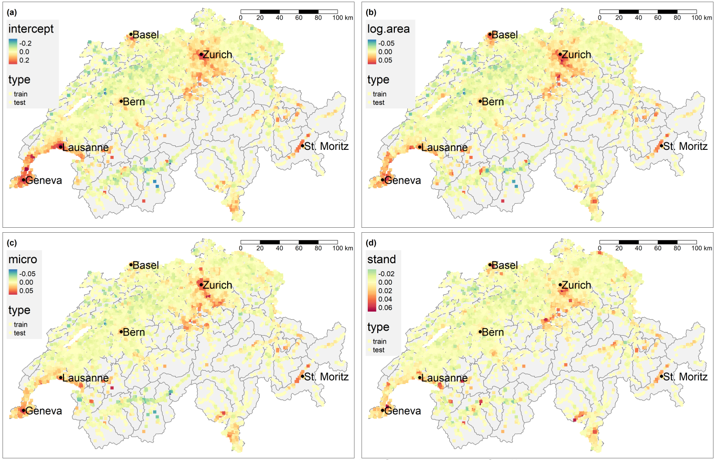
Modeling using `varycoef`

```r
X <- model.matrix(~1+elev+dist, data = meuse) # covariates
locs <- as.matrix(meuse[, 1:2])                # locations in CRS
## ----- PREPARING MLE -----
control <- SVC_mle_control(
  profileLik = TRUE,
  init = c(rep(c(0.4, 0.35), ncol(X)), 0.25)  # initial values
)
## ----- MLE -----
VC.fit <- SVC_mle(
  y = log(meuse$cadmium), X = X, locs = locs, control = control
)
```

Application: Real Estate Data

- Apartment price prediction (in Switzerland)
- ca. 15’000 observations for training
- 8 SVC modeled
- c.f. Dambon et al. (2020)

Application: Selection of Estimated SVCs



Future Work

- Adding Gaussian Process covariance functions
- Dependency Measures (`locs` argument):
  - Higher dimensional Domain (`ncol(locs)` > 2)
  - Time, Space-Time
  - Social Economic Status, etc.
- SVC Selection, similar to `{glmnet}`:
  - Which SVC / covariate do I need?

References I

Dambon, J. A., Sigrist, F., and Furrer, R. (2020). Maximum Likelihood Estimation of Spatially Varying Coefficient Models for Large Data with an Application to Real Estate Price Prediction. *ArXiv e-prints*.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, Chichester.

Furrer, R., Genton, M. G., and Nychka, D. W. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

References II

Gerber, F. and Furrer, R. (2019). `optimParallel`: An R Package Providing a Parallel Version of the L-BFGS-B Optimization Method. *The R Journal*, 11(1):352–358.

Lindgren, F. K., Rue, H., and Lindström, J. (2011). An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46(sup1):234–240.