Spatio-temporally varying coefficient modeling using the spmoran package

Daisuke Murakami

2024/09/24

Contents

1 Introduction **1**

2 Regression with residual spatio-temporal dependence **2**

3 Regression with spatio-temporally varying coefficients (STVCs) **3**

4 Implemented extentions **8**

5 Reference **8**

1 Introduction

As of version 0.3.0, the spmoran package provides functions for modeling spatio-temporally varying coefficients (STVCs) and residual spatio-temporal dependence. The functions estimate purely spatial, purely temporal, and spatio-temporal interaction patterns for each coefficient and residuals in a computationally efficiently manner (see Murakami et al., 2024).

This tutorial demonstrates how to implement spatio-temporal models using the Lucas housing price dataset with the sample size of 25,357:

```
require(spData)
data(house)
dat0  <- st_as_sf(house)
dat   <- data.frame(st_coordinates(dat0), dat0)
```

The explained variable (y) is the logarithm of the housing price. Explanatory variables (x) assuming STVCs are "lotsize" (lot size) and "TLA" (total floor area) while those (xconst) assuming constant coefficients are "rooms" (number of rooms) and "beds" (number of beds):

```
y       <- log( dat[,"price"] )
x     <- dat[,c("lotsize","TLA")]
xconst<- dat[,c("rooms","beds")]
```

Spatial coordinates and one or two temporal coordinates can be considered. We consider the following temporal coordinates: year of building (yrbuilt) and year of sale (syear):

```
coords  <- dat[,c("X","Y")]
byear   <- house$yrbuilt
#is(house$syear)  # check the format of syear
```

```r
syear   <- as.numeric(as.character(house$syear)) # factor -> numeric
coords_z<- cbind(byear,syear)
```

As yrbuilt is in a factor format, it is converted to a numeric format.

2 Regression with residual spatio-temporal dependence

This section considers the explained variable $y(s_i, \mathbf{t}_i)$ observed at site s_i at times $\mathbf{t}_i = \{t_{1,i}, t_{2,i}\}$. For example, $t_{1,i}$ and $t_{2,i}$ may be month and hour respectively. $y(s_i, \mathbf{t}_i)$ is assumed to obey the following model:

$$y(s_i, \mathbf{t}_i) = \sum_{k=1}^{K} x_k(s_i, \mathbf{t}_i) b_k + \beta_0(s_i, \mathbf{t}_i) + e(s_i, \mathbf{t}_i), \qquad e(s_i, \mathbf{t}_i) \sim N(0, \sigma^2),$$

where $\sum_{k=1}^{K} x_{i,k} b_k$ represents a regression term ($x_{i,k}$: k-th explanatory variate; b_k: k-th regression coefficient) and ϵ_i represents noise.

The spatio-temporal process $\beta_0(s_i, \mathbf{t}_i)$ is required to eliminate residual spatio-temporal dependence and estimate/infer regression coefficients b_k appropriately. It is specified as

$$\beta_0(s_i, \mathbf{t}_i) = \tau_s f(s_i; \alpha_s) + \sum_{m=1}^{M} \tau_{t,m} f(t_{m,i}; \alpha_{t,m}) + \sum_{m=1}^{M} \tau_{st,m} f(s_i, t_{m,i}; \alpha_{st,m}),$$

where (i) $f(s_i; \alpha_s)$ is a (low rank) spatial Gaussian process, (ii) $f(t_{m,i}; \alpha_{t,m})$ is the m-th temporal process, and (iii) $f(s_i, t_{m,i}; \alpha_{st,m})$ is the m-th spatio-temporal interaction process. $M = 2$ in our case. The variance parameters $\tau_s, \tau_{t,m}, \tau_{st,m}$ estimate the amount of variations explained by each process. For example, if $\tau_s = 0$, $\beta_0(s_i, \mathbf{t}_i)$ has no (pure) spatial variation. As τ_s increases, spatial variation gets strong. The scale/smoothness of each process is estimated by another parameters $\alpha_s, \alpha_{t,m}, \alpha_{st,m}$.

The processes (i)-(iii) are given by a weighted sum of the (i) spatial, (ii) temporal, and (iii) spatio-temporal Moran eigenvectors, which are basis functions. The eigenvectors are extracted by the meigen function for small-to-moderate samples (e.g., N<3000). For larger samples, the meigen_f function performs an approximation to exract these computationally efficiently. The commands are as follows:

```r
### For small to moderate samples
#meig   <- meigen(coords=coords, coords_z=cbind(byear,syear),interact=TRUE)

### For large samples
meig    <- meigen_f(coords=coords, coords_z=cbind(byear,syear),interact=TRUE)
```

The space-temporal interaction term (iii) is considered if interact=TRUE, whereas ignored for fast computation if interact=FALSE. An argument coords_z specifies the temporal coordinates which are byear and syear this time. If coords_z = byear, only byear is considered. Any one-dimensional coordinate (including non-temporal one) is available for coords_z as long as the same column does not appear in x and xconst.

The Gaussian regression model with residual spatio-temporal dependence is estimated as follows:

```r
res   <- resf(y = y, x = cbind(x, xconst), meig = meig)
res
```

```
## Call:
## resf(y = y, x = cbind(x, xconst), meig = meig)
##
## ----Coefficients------------------------------
##                  Estimate           SE    t_value      p_value
## (Intercept) 1.010834e+01 1.307557e-02 773.070937 0.000000e+00
```

```
## lotsize     1.108241e-06 8.817751e-08  12.568295 0.000000e+00
## TLA         3.959125e-04 6.057748e-06  65.356382 0.000000e+00
## rooms       1.295219e-02 3.008271e-03   4.305527 1.672197e-05
## beds        2.711555e-02 4.426194e-03   6.126154 9.136549e-10
##
## ----Variance parameter------------------------
##
## Spatial effects (residuals):
##                                    (Intercept)
## random SD: Spatial                0.0429151079
## --Moran.I/max(Moran.I)            0.3727802661
## random SD: coords_z[,1]           0.4127041056
## random SD: coords_z[,2]           0.0958532443
## random SD: Spatial x coords_z[,1] 0.0000150781
## random SD: Spatial x coords_z[,2] 0.0000000000
##
## ----Error statistics-------------------------
##                      stat
## resid SE        0.2952891
## adjR2(cond)     0.8501454
## rlogLik     -5701.3333214
## AIC         11426.6666428
## BIC         11524.3563640
##
## NULL model: lm( y ~ x )
##    (r)loglik: -22738.27 ( AIC: 45488.55,  BIC: 45537.39 )
##
## Note: AIC and BIC are based on the restricted/marginal likelihood.
##       Use method="ml" for comparison of models with different fixed effects (x)
```

Based on the t-values, TLA has the strongest positive influence. The other explanatory variables are also positively statistically significant. The error statistics confirm the high accuracy of our model.

Based on the variance estimates, "random SD: coords_z[,1]" has largest variance. It means that housing price has a strong temporal dependence according to the year of building (coords_z[,1]). The non-zero variances of "random SD: Spatial" and "random SD: coords_z[,2]" show that housing price has spatial dependence and temporal dependence over sales years (coords_z[,2]). A space-time interaction term "random SD: Spatial x coords_z[,1]" is also estimated to be influential.

3 Regression with spatio-temporally varying coefficients (STVCs)

The Gaussian STVC model implemented in this package is formulated as

$$y(s_i, \mathbf{t}_i) = \sum_{k=1}^{K} x_k(s_i, \mathbf{t}_i)\beta_k(s_i, \mathbf{t}_i) + \beta_0(s_i, \mathbf{t}_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

$$\beta_k(s_i, \mathbf{t}_i) = b_k + \tau_{s,k} f(s_i; \alpha_{s,k}) + \sum_{m=1}^{M} \tau_{t,m,k} f(t_{m,i}; \alpha_{t,m,k}) + \sum_{m=1}^{M} \tau_{st,m,k} f(s_i, t_{m,i}; \alpha_{st,m,k}),$$

Just like the geographically and temporally weighted regression (GTWR) model, which is widely used these days, our STVC model estimates STVCs. Major advantages of our model over the conventional GTWR model are as follows:

- It estimates the spatial, temporal, and spatiotemporal scale/smoothness of the coefficients using the $\alpha_{s,k}, \alpha_{t,k}, \alpha_{st,k}$ parameters, while the classical GTWR assumes only one bandwidth/scale parameter across STVCs.
- The STVCs are easily replaced with constant coefficients by assuming $\beta_{i,k} = b_k$.
- This model is faster and available for large samples as demonstrated in this tutorial.

The STVC model is implemented as follows:

```
res     <- resf_vc(y=y,x=x,xconst=xconst,meig=meig )
```

```
## [1] "-------  Iteration 1  -------"
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9876.126"
## [1] "-------  Iteration 2  -------"
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9819.75"
## [1] "-------  Iteration 3  -------"
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9817.187"
## [1] "-------  Iteration 4  -------"
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9817.162"
## [1] "-------  Iteration 5  -------"
```

```
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9817.161"
## [1] "-------  Iteration 6  -------"
## [1] "1/9"
## [1] "2/9"
## [1] "3/9"
## [1] "4/9"
## [1] "5/9"
## [1] "6/9"
## [1] "7/9"
## [1] "8/9"
## [1] "9/9"
## [1] "BIC: 9817.161"
## [1] "-------  Iteration (interaction term) 1  -------"
## [1] "BIC: 9518.243"
## [1] "-------  Iteration (interaction term) 2  -------"
```

```
res
```

```
## Call:
## resf_vc(y = y, x = x, xconst = xconst, meig = meig)
##
## ----Spatio-temporally varying coefficients on x (summary)----
##
## Coefficient estimates:
##   (Intercept)        lotsize                 TLA
##  Min.   : 7.586   Min.   :-1.005e-05   Min.   :-3.699e-05
##  1st Qu.: 9.931   1st Qu.: 3.498e-06   1st Qu.: 3.173e-04
##  Median :10.415   Median : 8.237e-06   Median : 3.874e-04
##  Mean   :10.267   Mean   : 1.443e-05   Mean   : 3.971e-04
##  3rd Qu.:10.762   3rd Qu.: 2.153e-05   3rd Qu.: 4.727e-04
##  Max.   :11.944   Max.   : 6.905e-05   Max.   : 8.634e-04
##
## Statistical significance (adjusted by Benjamini-Yekutieli method):
##                         Intercept lotsize   TLA
## Not significant                 0   14323   279
## Significant (10% level)         0    1004    83
## Significant ( 5% level)         0    1725   222
## Significant ( 1% level)     25357    8305 24773
##
## ----Constant coefficients on xconst----------------------------
##         Estimate          SE  t_value      p_value
## rooms 0.00991385 0.002871660 3.452306 0.0005567428
## beds  0.01653253 0.004270333 3.871485 0.0001084511
##
## ----Variance parameters----------------------------------
##
```

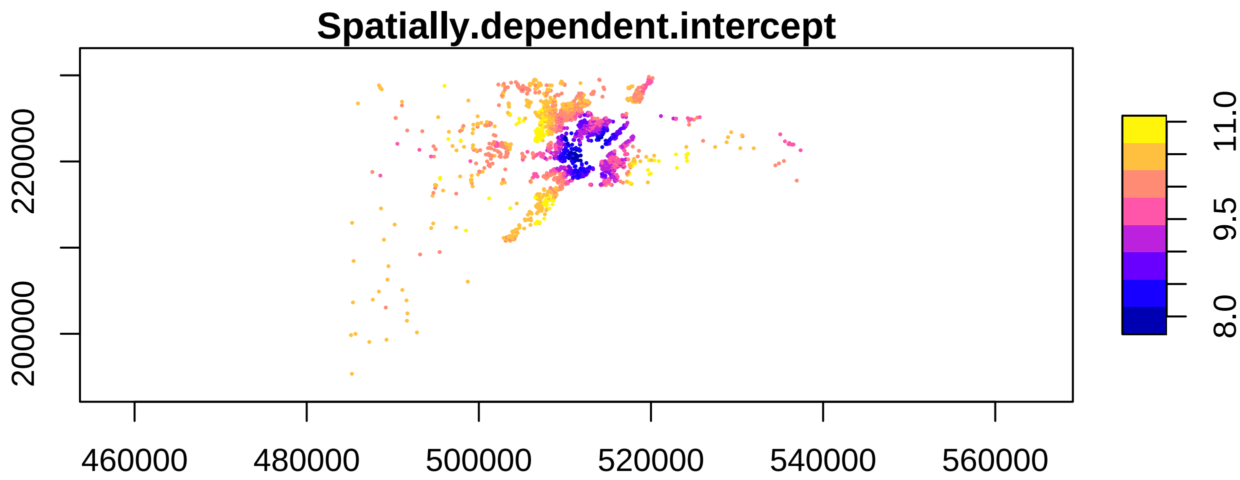
```
## Spatio-temporal effects (coefficients on x):
##                                (Intercept)      lotsize          TLA
## random SD: Spatial             5.687896e-02 1.656749e-06 1.559795e-05
## --Moran.I/max(Moran.I)         3.726172e-01 4.042222e-01 6.394207e-02
## random SD: coords_z[,1]        4.395052e-01 2.065568e-06 1.122794e-04
## random SD: coords_z[,2]        9.642453e-02 0.000000e+00 0.000000e+00
## random SD: Spatial x coords_z[,1] 1.573200e-09 0.000000e+00 0.000000e+00
## random SD: Spatial x coords_z[,2] 0.000000e+00 0.000000e+00 0.000000e+00
##
## ----Estimated probability distribution of y--------------
##                 Estimates
## skewness                0
## excess kurtosis         0
##
## ----Error statistics-------------------------------------
##                     stat
## resid SE        0.2762681
## adjR2(cond)     0.8687879
## rlogLik     -4647.5725502
## AIC          9335.1451003
## BIC          9497.9613024
##
## NULL model: lm( y ~ x + xconst )
##    (r)loglik: -22738.27 ( AIC: 45488.55,  BIC: 45537.39 )
##
## Note: AIC and BIC are based on the restricted/marginal likelihood.
##       Use method="ml" for comparison of models with different fixed effects (x and xconst)
```

The code assumes assumed STVCs on the explanatory variables in x while constant coefficients on those in xconst. The STVCs on lotsize and TLA tend to be positive, meaning that larger residences are more expensive. Based on their statistical significance, TLA is more influential than lotsize. The constant coefficients on rooms and beds suggest their positive influence on housing price. All these results are intuitively reasonable. The error statistics confirm the better accuracy of the STVC model over the non-STVC model implemented in the previous section.

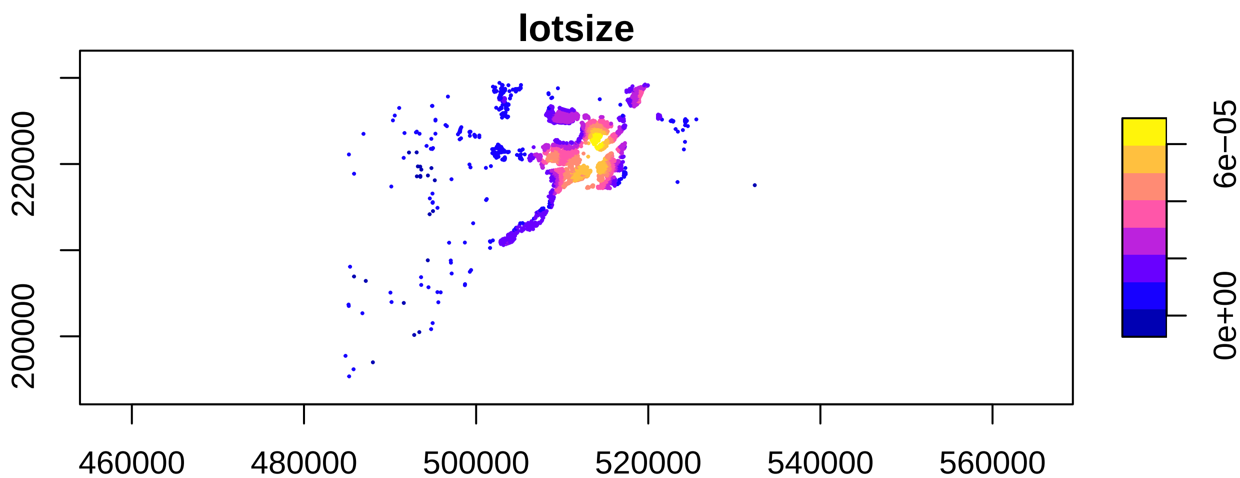
The resf_vc function optimizes the combination of (i), (ii), (iii) by STVC, by minimizing the Bayesian information criterion. The selected elements among (i)-(iii) have non-zero variance whereas those who are not-selected have zero variance. The estimated variance parameters are summarized in a table above. Regarding the varying intercept, "random SD: coords_z[,1]" has the largest variance, while "random SD: Spatial" and "random SD: coords_z[,2]" have relatively small variances. These results that suggest the strong temporal dependence over coords_z[,1] is consistent with the result in the previous section. Regarding lotsize and TLA, their estimated STVCs become [constant coefficient]+[pure spatial process (random SD: Spatial)] + [pure temporal process (random SD) according to coords_z[,1]]. In other words, no space-time interaction pattern is detected from both these coefficients (i.e., "random SD: Spatial x coords_z[,1]" and "random SD: Spatial x coords_z[,2]" took zero values).

Here are examples of plotting the estimated STVCs using the plot_s function:

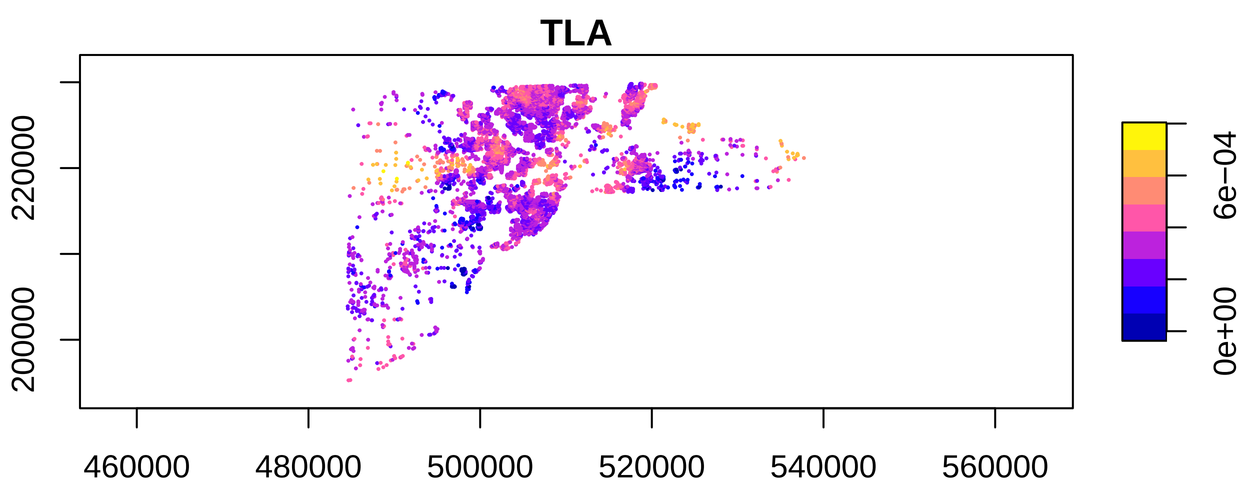
```r
# Varying intercept for byear <=1950 and syear==1998
plot_s(res,0, coords_z1_lim=c(-Inf, 1950),coords_z2_lim=1998,cex=0.2)
```




```
# 1st STVC (statistically significant at the 5% level) for byear <= 1950
plot_s(res,1, coords_z1_lim=c(-Inf, 1950), pmax=0.05,cex=0.2)
```




```
# 2nd STVC for byear >= 1951
plot_s(res,2, coords_z1_lim=c(1951,Inf),cex=0.2)
```



Note that the plot_s function is designed to quickly check the estimated coefficients. We recommend using sf, mapview or another package to make high quality maps.

4 Implemented extentions

If the arguments coords_z and interact are specified in the meigen/meigen_f function, a wide variety of spatio-temporal models can be estimate using the resf and resf_vc functions, by using the same code as spatial modelling (See codes and vignettes in https://github.com/dmuraka/spmoran/tree/master). Here are examples of models that are implemented:

- STVC model considering group effects (specify xgroup)
- Model with spatio-temporally and non-linearly varying coefficients (x_nvc=TRUE)
- Non-Gaussian spatio-temporal models (specify nongauss)
- Spatio-temporal prediction (specify coords_z0 in the meigen0 function, and substitute it in the predict0 function)

5 Reference

- Murakami, D., Shirota, S., Kajita, S., and Kajita, S. (2024) Fast spatio-temporally varying coefficient modeling with reluctant interaction selection. ArXiv.