R. Bivand (2019) Progress in the R ecosystem for open source spatial analysis software: Comments

Ghislain Geniaux, INRA Avignon

23/05/2019

The recent evolution from *sp* to *sf* is really a great job for applied spatial econometrics in which we need to control the workflow of spatial:

1. data preparation,
2. exploration,
3. modeling,
4. results visualization and web publishing.

Now, regarding data preparation, data exploration and web publishing, they now can be done more easily and more quickly with R since *sf package* is available and used in numerous other R packages.

I have some comments and suggestions to improve the third point for practitioners, i.e. spatial modelling, which corresponds to the *spatialreg* package currently in your presentation.

It was mentioned in your conclusion, but I will say it differently:

How to create more bridges between developments in machine learning (especially boosting) and spatial econometrics, with regard to variable selections and cross-validation techniques ?

Some suggestions :

- BLUP (from Goulard et al. 2017) for all spatial regression functions proposed in *spatialreg*.
- Integrated tools for systematic use of k folds cross validation adapted to spatial data.
- the addition of support for non-linear relationships in spatial regression functions.

would be a good first step.

- the addition of support for variable selections, taking into account possible non-linear relationships and variable interactions, based on boosting techniques

would be a very nice second step.

spatial econometrics and non linear model

Now, I will insist a little more on non linear relationships in spatial regressions.

Consider this very simple hedonic price model in which price is a function of non-linear form of LandArea and Dist2MainRoad variables:

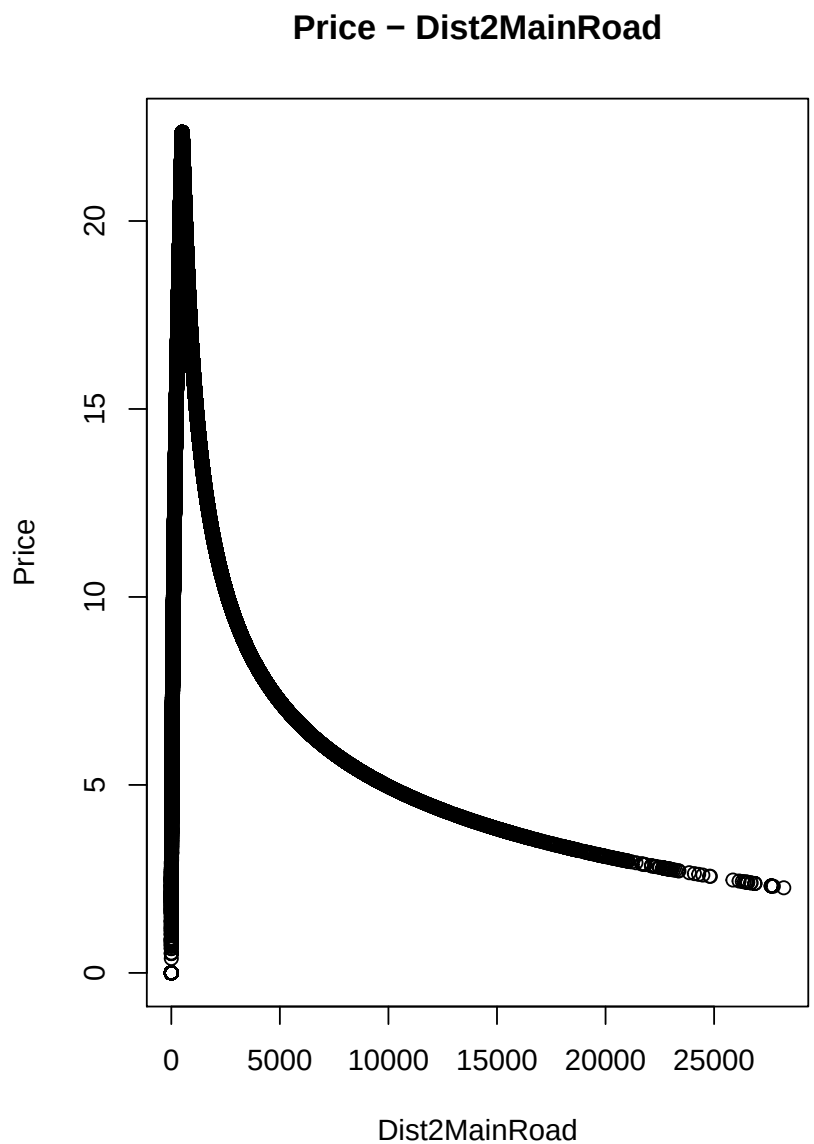
$$housePrice = HouseArea + f(LandArea) + g(Dist2MainRoad) + \epsilon$$

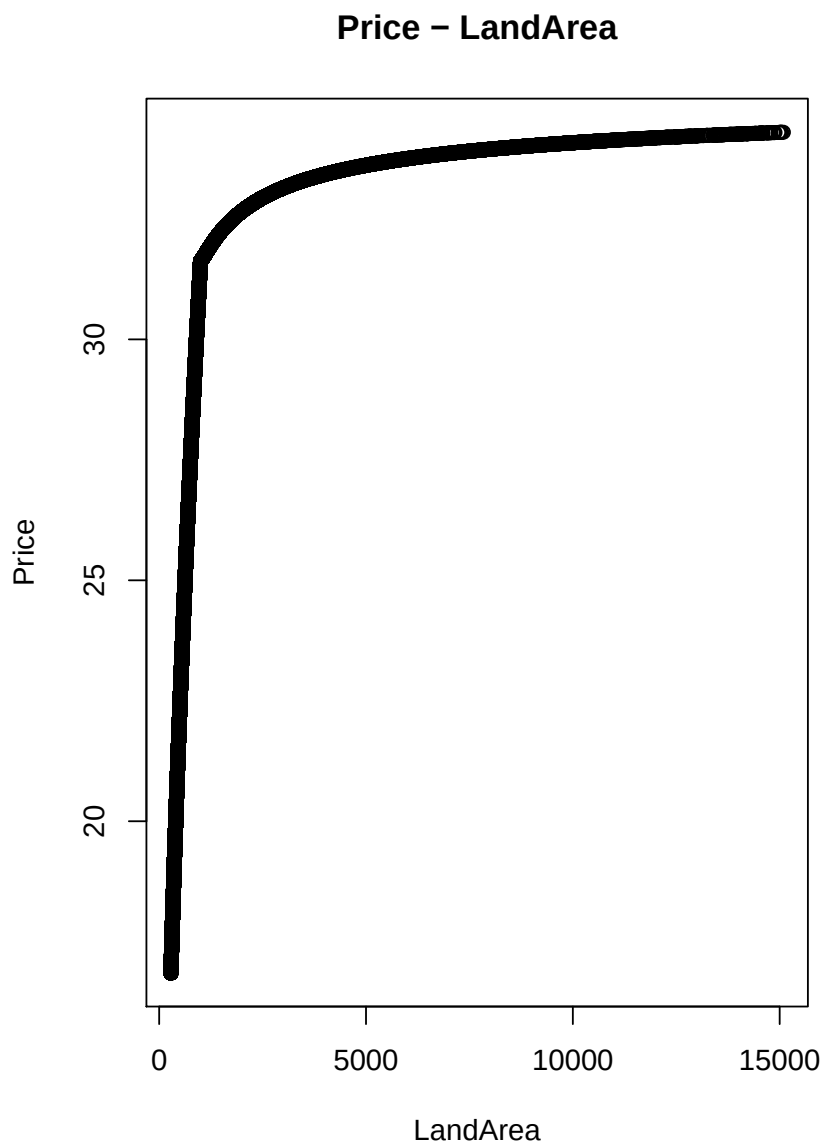
Distance to main road is generally non linear and non monotone because between 0 to 500 meters it's desamenity and after 500 it's an amenity by reducing time access to others locations. Land Area is also often non linear and log transformation could be insufficient.

hyp 1: Suppose we have these realtionship with price.

```
knitr::asis_output("\\tiny");par(mfrow=c(1,2));
```

```
plot(sort(HousePriceData_sf$Dist2MainRoad),sapply(sort(HousePriceData_sf$Dist2MainRoad),fs1),main='Price -
plot(sort(HousePriceData_sf$LandArea),sapply(sort(HousePriceData_sf$LandArea),fs2),main='Price - LandArea'
```



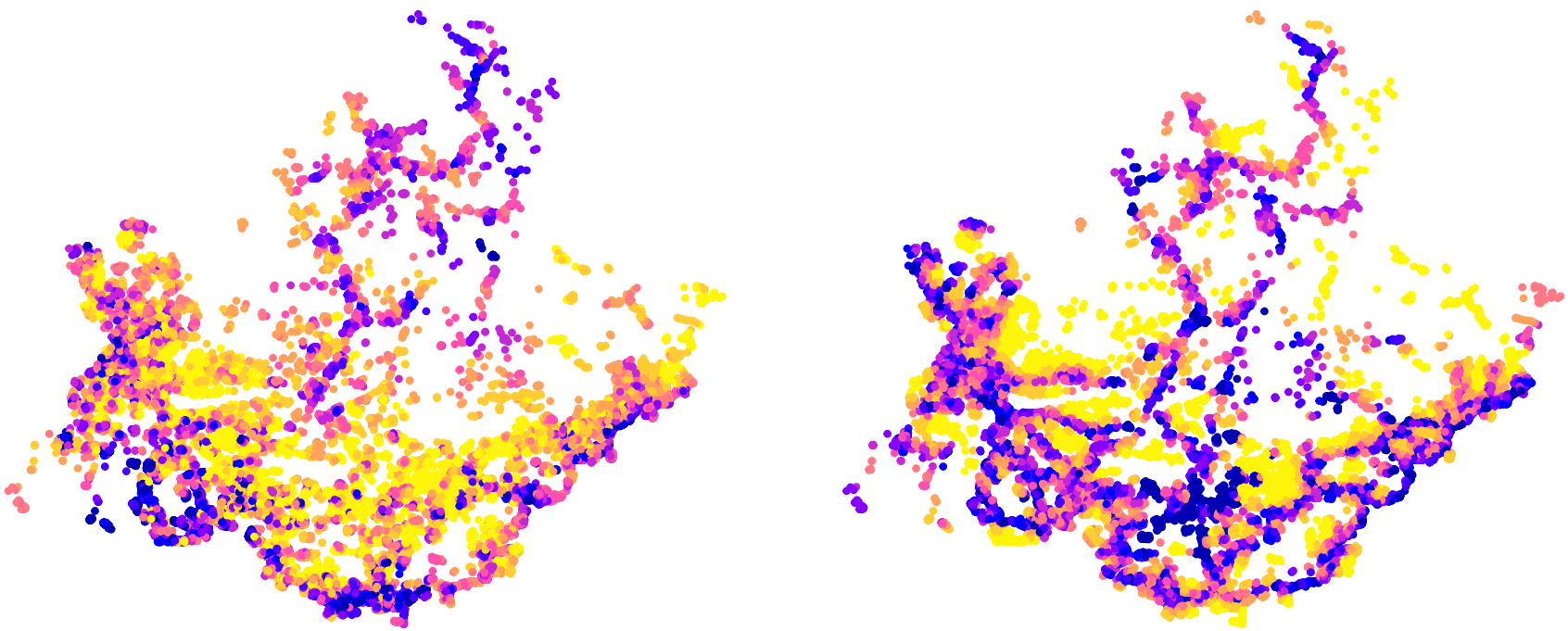


hyp 2: Suppose HouseArea and Dist2MainRoad are not randomly distributed in space (very weak hypothesis).

For example, the following slide illustrates the spatial patterns of LandArea and Dist2MainRoad for 76945 houses in Provence.

```
knitr::asis_output("\\tiny"); par(mfrow=c(1,2))
```

```
plot(HousePriceData_sf['LandArea_ks'],
     breaks='quantile',pch=19,cex=0.5, key.pos = NULL, reset = FALSE)
plot(HousePriceData_sf['Dist2MainRoad_ks'],
     breaks='quantile',pch=19,cex=0.5, key.pos = NULL, reset = FALSE)
```



Then simulate price using $f()$ and $g()$:

$$SimHousePrice = 130000 + 5000 * HouseArea$$

$$+f(Dist2MainRoad) + g(Landarea) + \epsilon$$

where ϵ follow uniform(0,250000). The True DGP has no spatial autocorrelation.

hyp 3: lastly, suppose that a spatial econometrician use a linear model (very little hypothesis again since 95 % of papers published in spatial econometrics litterature use linear models)

As expected, the spatial autocorrelation tests for linear model indicate the presence of spatial autocorrelation and suggest a SARMA model.

```
knitr::asis_output("\\tiny")
```

```
model=lm('simPrice~HouseArea+LandArea+Dist2MainRoad'
         ,data=HousePriceData_sf)
W4=KNN(st_coordinates(HousePriceData_sf),4)
W30=KNN(st_coordinates(HousePriceData_sf),30)
moran.test(model$residuals, mat2listw(W4))
```

```
##
##  Moran I test under randomisation
##
## data:  model$residuals
## weights: mat2listw(W4)
##
## Moran I statistic standard deviate = 25.604, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      6.054305e-02     -1.299646e-05      5.593745e-06
```

```r
knitr::asis_output("\\tiny")
```

```r
lm.LMtests(model, mat2listw(W4),test=c('LMerr','LMlag'))
```

```
## Warning in lm.LMtests(model, mat2listw(W4), test = c("LMerr", "LMlag")):
## Spatial weights matrix not row standardized
```

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = "simPrice~HouseArea+LandArea+Dist2MainRoad",
## data = HousePriceData_sf)
## weights: mat2listw(W4)
##
## LMerr = 655.24, df = 1, p-value < 2.2e-16
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = "simPrice~HouseArea+LandArea+Dist2MainRoad",
## data = HousePriceData_sf)
## weights: mat2listw(W4)
##
## LMlag = 500.71, df = 1, p-value < 2.2e-16
```

```r
knitr::asis_output("\\tiny")
```

```r
lm.LMtests(model, mat2listw(W4),test=c('SARMA'))
```

```
## Warning in lm.LMtests(model, mat2listw(W4), test = c("SARMA")): Spatial
## weights matrix not row standardized
```

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = "simPrice~HouseArea+LandArea+Dist2MainRoad",
## data = HousePriceData_sf)
## weights: mat2listw(W4)
##
## SARMA = 864.22, df = 2, p-value < 2.2e-16
```

While true non-linear model and non linear model approximations using gam model with spline (mgcv package) or piecewise linear model using MARS model (earth package) provides residuals without spatial autocorrelation.

```
knitr::asis_output("\\tiny")
```

```
### model_NonLin_True
model_NonLin_True=lm('simPrice~HouseArea+g_LandArea+f_Dist2MainRoad',
                     data=HousePriceData_sf)
### model_gam
model_gam=gam(as.formula('simPrice~s(HouseArea)+s(Dist2Mainroad)+s(Landarea)'),
              data=as.data.frame(HousePriceData_sf))
### model_earth
model_earth=earth(as.formula('simPrice~HouseArea+Dist2Mainroad+Landarea'),
              data=as.data.frame(HousePriceData_sf))
```

```r
knitr::asis_output("\\tiny")
```

```r
##moran.test
moran.test(model_NonLin_True$residuals, mat2listw(W4))
```

```
##
##  Moran I test under randomisation
##
## data:  model_NonLin_True$residuals
## weights: mat2listw(W4)
##
## Moran I statistic standard deviate = 1.168, p-value = 0.1214
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      2.749489e-03     -1.299646e-05      5.593763e-06
```

```r
moran.test(model_gam$residuals, mat2listw(W4))
```

```
##
##  Moran I test under randomisation
##
## data:  model_gam$residuals
## weights: mat2listw(W4)
##
## Moran I statistic standard deviate = 1.1608, p-value = 0.1229
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      2.732354e-03     -1.299646e-05      5.593763e-06
```

```r
moran.test(model_earth$residuals, mat2listw(W4))
```

```
##
```

Here, it is simple because the true model is a non-linear model without spatial autocorrelation, and we can use the non-linear regression functions existing in R.

but if there is also a spatial dependence... and also spatial heterogeneity.

–> How to facilitate the use of autoregressive models with non-linear or discontinuous relationship: spline ? kernel smoothing ? boosting ? What is your plan for the newt two years ?