HousePrice

Lan Hu

July 13, 2019

This project uses spatial models to predict house prices

```
library(GISTools); ##reading shapefiles
```

```
## Loading required package: maptools
## Loading required package: sp
## Checking rgeos availability: TRUE
## Loading required package: RColorBrewer
## Loading required package: MASS
## Loading required package: rgeos
## Warning: package 'rgeos' was built under R version 3.5.2
## rgeos version: 0.4-2, (SVN revision 581)
##  GEOS runtime version: 3.6.1-CAPI-1.10.1
##  Linking to sp version: 1.3-1
##  Polygon checking: TRUE
```

```
library(spdep); #creating spatial weight matrix and eigenvectors
```

```
## Loading required package: Matrix
## Loading required package: spData
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
```

```
library(nlme); #multilevel model
library(MuMIn) #computing rsquared for multilevel model
```

```
## Warning: package 'MuMIn' was built under R version 3.5.3
```

```
library(knitr)
```

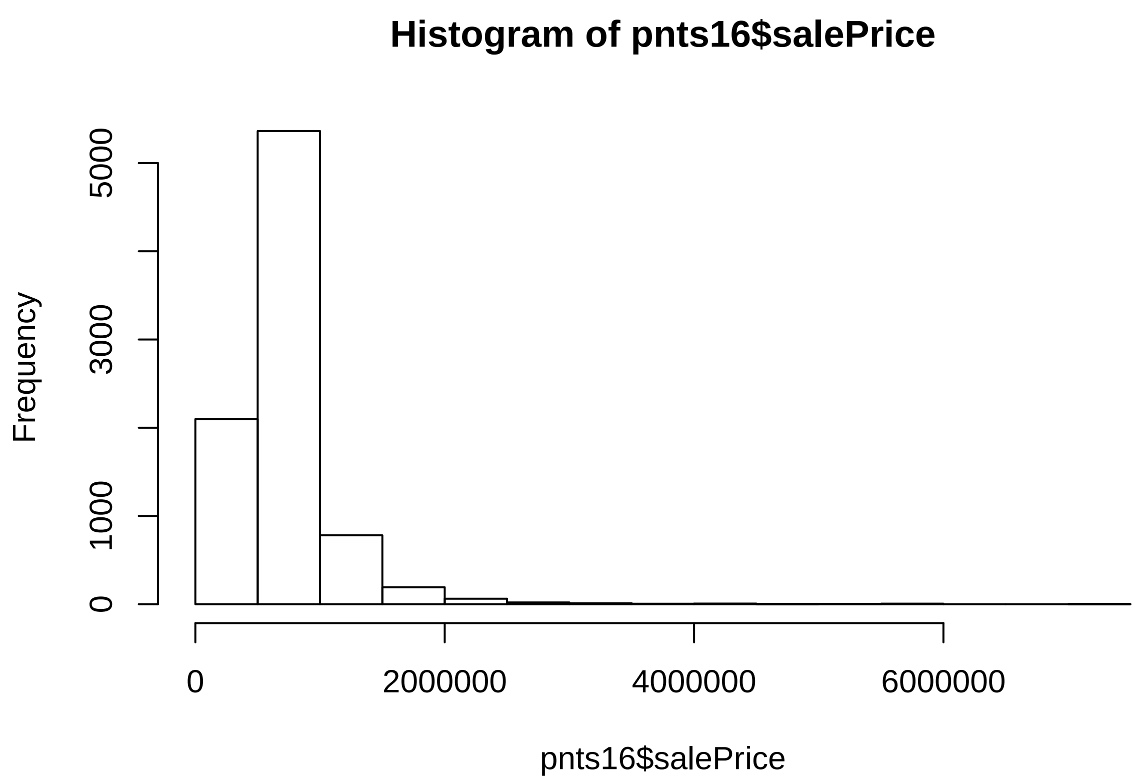
```
## Warning: package 'knitr' was built under R version 3.5.3
```

```
library(tinytex)

##block group effects
#read block group shapefile
bg<- readShapePoly("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1\\hp_project1\\
```

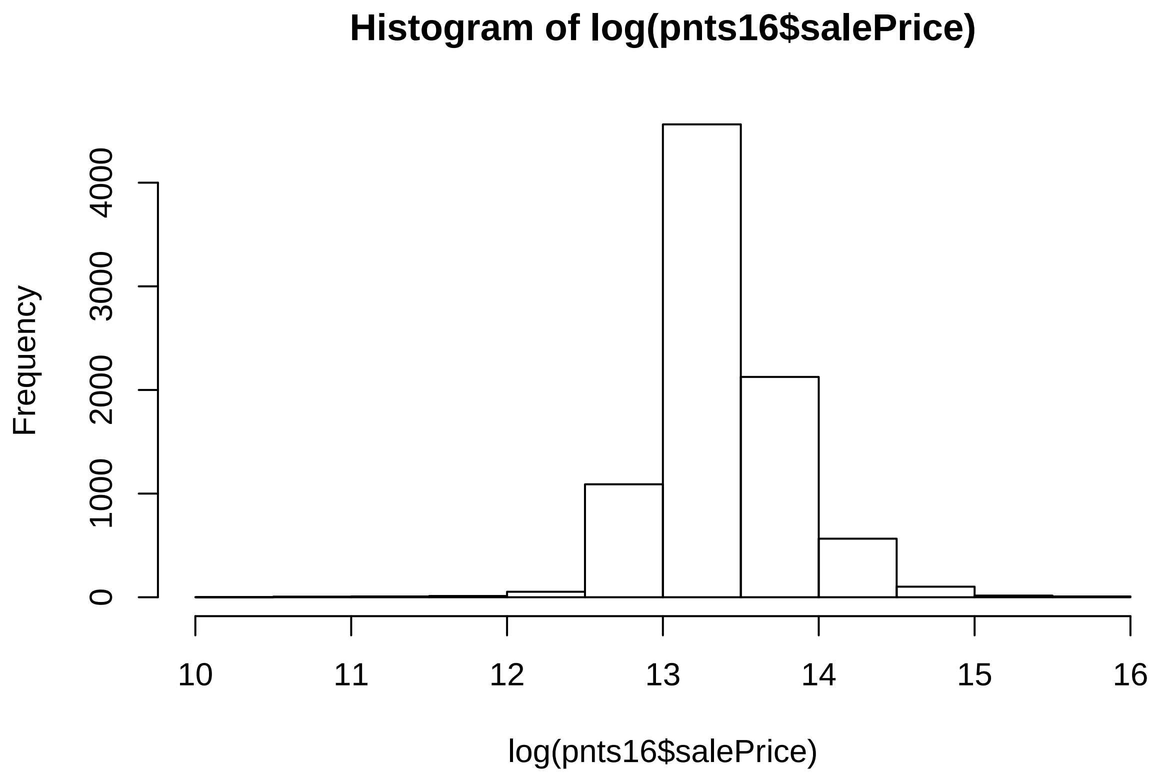
```
## Warning: readShapePoly is deprecated; use rgdal::readOGR or sf::st_read
```

```
#create spatial weight matrix
bg.nb <- poly2nb(bg)
bg.n <- length(bg.nb)
bg.listw <- nb2listw(bg.nb, style="W")
bg.listb <- nb2listw(bg.nb, style="B")
```

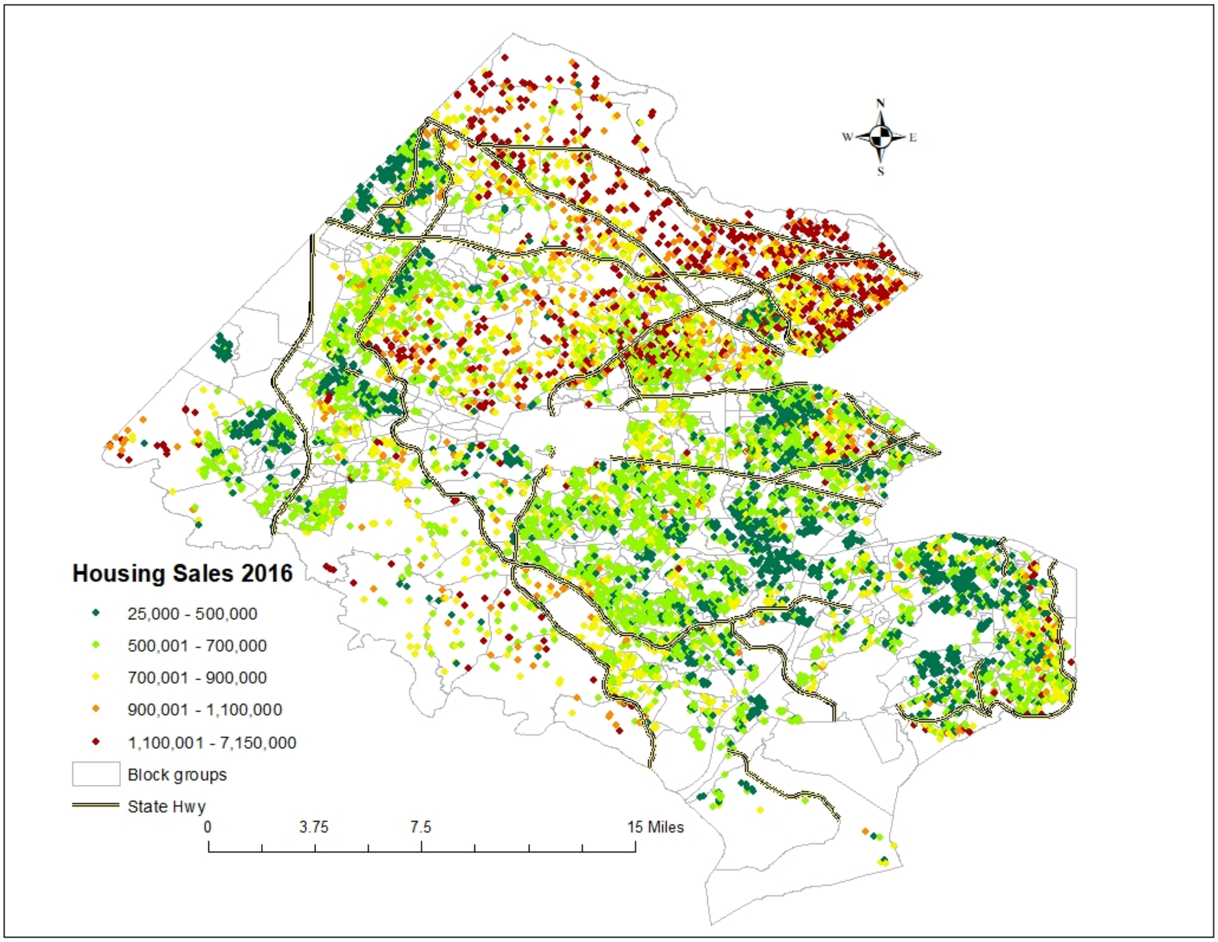
```r
#compute eigenvectors
bg.M <- diag(bg.n) - matrix(1,bg.n,bg.n)/bg.n
bg.B <- listw2mat(bg.listb)
bg.MBM <- bg.M %*% bg.B %*% bg.M
bg.eig <- eigen(bg.MBM, symmetric=T)
bg.EV <- as.data.frame(bg.eig$vectors[,bg.eig$values/bg.eig$values[1]>0.25])
colnames(bg.EV) <- paste("EV", 1:NCOL(bg.EV), sep="")
bg.EV.neg <- as.data.frame(bg.eig$vectors[,bg.eig$values/bg.eig$values[bg.n]>0.25])
colnames(bg.EV.neg) <- paste("V", (bg.n - NCOL(bg.EV.neg) + 1):bg.n, sep="")
bg.EV.both <- cbind(bg.EV, bg.EV.neg)

###regression analysis
#read house sale prices data in 2016
pnts16<- read.csv("C:\\Users\\lxh152030\\Desktop\\test\\houseP.csv")

#examine house prices distribution
hist(pnts16$salePrice)
```




```r
hist(log(pnts16$salePrice))
```




```
##Geographic distrution of house prices in 2016 created with ArcMap
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\housePric
```




```r
#summaries of variables
summary(pnts16[, 1:29])
```

```
##      GEOID                  X                 Y            Land_Area
##  Min.   :5.106e+11   Min.   :-77.53   Min.   :38.64   Min.   :0.002262
##  1st Qu.:5.106e+11   1st Qu.:-77.33   1st Qu.:38.78   1st Qu.:0.010220
##  Median :5.106e+11   Median :-77.25   Median :38.85   Median :0.012462
##  Mean   :5.106e+11   Mean   :-77.25   Mean   :38.85   Mean   :0.023222
##  3rd Qu.:5.106e+11   3rd Qu.:-77.18   3rd Qu.:38.91   3rd Qu.:0.020450
##  Max.   :5.106e+11   Max.   :-77.04   Max.   :39.05   Max.   :1.073318
##    salePrice          noStories        Bedroom        Full_Baths
##  Min.   :  25000   Min.   :1.000   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 505000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.: 2.000
##  Median : 625000   Median :2.000   Median :4.000   Median : 3.000
##  Mean   : 718539   Mean   :1.611   Mean   :4.038   Mean   : 2.815
##  3rd Qu.: 791875   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.: 3.000
##  Max.   :7150000   Max.   :3.000   Max.   :8.000   Max.   :12.000
##    Half_Baths        Fireplaces      Year_Built      liveArea
##  Min.   :0.0000   Min.   :0.00   Min.   :1764   Min.   : 0.032
##  1st Qu.:0.0000   1st Qu.:1.00   1st Qu.:1961   1st Qu.: 1.464
##  Median :1.0000   Median :1.00   Median :1976   Median : 2.076
##  Mean   :0.6894   Mean   :1.23   Mean   :1976   Mean   : 2.344
##  3rd Qu.:1.0000   3rd Qu.:2.00   3rd Qu.:1988   3rd Qu.: 2.877
```

```
##  Max.   :5.0000   Max.   :9.00   Max.   :2017   Max.   :14.165
##      disSch             disMall            disHwy
##  Min.   :0.0002287   Min.   :0.00000   Min.   :1.030e-06
##  1st Qu.:0.0143307   1st Qu.:0.03142   1st Qu.:3.911e-03
##  Median :0.0260121   Median :0.05054   Median :7.714e-03
##  Mean   :0.0296946   Mean   :0.05878   Mean   :9.752e-03
##  3rd Qu.:0.0417418   3rd Qu.:0.08078   3rd Qu.:1.326e-02
##  Max.   :0.1232210   Max.   :0.17122   Max.   :5.270e-02
##       yPop              white           hispanic           income
##  Min.   :0.06084   Min.   :0.2295   Min.   :0.00000   Min.   : 23220
##  1st Qu.:0.24782   1st Qu.:0.6136   1st Qu.:0.04106   1st Qu.:124688
##  Median :0.28571   Median :0.7319   Median :0.07700   Median :153631
##  Mean   :0.28068   Mean   :0.7049   Mean   :0.11587   Mean   :154170
##  3rd Qu.:0.31449   3rd Qu.:0.8118   3rd Qu.:0.15818   3rd Qu.:185385
##  Max.   :0.83979   Max.   :0.9951   Max.   :0.87797   Max.   :248357
##     homeVal          migration           medAge         Bgyears
##  Min.   : 165400   Min.   :0.00000   Min.   :20.1   Min.   :1948
##  1st Qu.: 482400   1st Qu.:0.04023   1st Qu.:39.3   1st Qu.:1967
##  Median : 606000   Median :0.06088   Median :42.3   Median :1977
##  Mean   : 648431   Mean   :0.06869   Mean   :42.6   Mean   :1976
##  3rd Qu.: 741500   3rd Qu.:0.08896   3rd Qu.:46.1   3rd Qu.:1985
##  Max.   :1750000   Max.   :0.24978   Max.   :68.5   Max.   :2006
##     season        sale_log         yrsOld           BGyrs
##  fall  :2197   Min.   :10.13   Min.   :  0.00   Min.   :11.00
##  spring:1713   1st Qu.:13.13   1st Qu.: 29.00   1st Qu.:32.00
##  summer:2991   Median :13.35   Median : 41.00   Median :40.00
##  winter:1649   Mean   :13.39   Mean   : 40.58   Mean   :40.83
##                3rd Qu.:13.58   3rd Qu.: 56.00   3rd Qu.:50.00
##                Max.   :15.78   Max.   :253.00   Max.   :69.00
##    income_log     homeVal_log
##  Min.   :10.05   Min.   :12.02
##  1st Qu.:11.73   1st Qu.:13.09
##  Median :11.94   Median :13.31
##  Mean   :11.90   Mean   :13.33
##  3rd Qu.:12.13   3rd Qu.:13.52
##  Max.   :12.42   Max.   :14.38
```

```r
attach(pnts16)
##linear model
#log-transformed house prices furnish the y variables
#two levels variables: individual house and block group levels
#individual level: Land_Area,liveArea, noStories, Full_Baths, Half_Baths,
                   #Fireplaces, Bedroom, yrsOld,
#                  disHwy, disSch, disMall, season
#block group level: yPop, white, hispanic, income_log, homeVal_log,
                    #migration, medAge, BGyrs

lm1<- lm(sale_log~Land_Area+liveArea+noStories+Full_Baths+Half_Baths+Fireplaces+Bedroom+
                  yrsOld+disHwy+disSch+disMall +season+yPop+white+hispanic+income_log+
                  homeVal_log+migration+medAge+BGyrs)
summary(lm1)
```

```
##
## Call:
## lm(formula = sale_log ~ Land_Area + liveArea + noStories + Full_Baths +
```

```
##     Half_Baths + Fireplaces + Bedroom + yrsOld + disHwy + disSch +
##     disMall + season + yPop + white + hispanic + income_log +
##     homeVal_log + migration + medAge + BGyrs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81416 -0.08416  0.00512  0.09494  1.89288
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.0238780  0.1659507  48.351  < 2e-16 ***
## Land_Area     0.6560294  0.0693159   9.464  < 2e-16 ***
## liveArea      0.1266217  0.0044272  28.601  < 2e-16 ***
## noStories    -0.0009091  0.0065072  -0.140  0.88890
## Full_Baths    0.0581451  0.0038756  15.003  < 2e-16 ***
## Half_Baths    0.0242631  0.0053706   4.518 6.33e-06 ***
## Fireplaces    0.0412574  0.0035906  11.490  < 2e-16 ***
## Bedroom       0.0171250  0.0037080   4.618 3.92e-06 ***
## yrsOld       -0.0026945  0.0001818 -14.821  < 2e-16 ***
## disHwy       -1.6902561  0.3227821  -5.237 1.68e-07 ***
## disSch       -1.9923976  0.1479469 -13.467  < 2e-16 ***
## disMall      -0.4419824  0.0801777  -5.513 3.64e-08 ***
## seasonspring  0.0055141  0.0071062   0.776  0.43780
## seasonsummer  0.0173597  0.0061969   2.801  0.00510 **
## seasonwinter -0.0235137  0.0071853  -3.272  0.00107 **
## yPop         -0.2846200  0.0572530  -4.971 6.78e-07 ***
## white         0.0667832  0.0221672   3.013  0.00260 **
## hispanic     -0.1715118  0.0285519  -6.007 1.97e-09 ***
## income_log   -0.0671711  0.0136211  -4.931 8.32e-07 ***
## homeVal_log   0.4349034  0.0137499  31.630  < 2e-16 ***
## migration    -0.2667152  0.0588025  -4.536 5.82e-06 ***
## medAge       -0.0027493  0.0006849  -4.014 6.02e-05 ***
## BGyrs         0.0037084  0.0002500  14.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2202 on 8527 degrees of freedom
## Multiple R-squared:  0.7243, Adjusted R-squared:  0.7236
## F-statistic:  1018 on 22 and 8527 DF,  p-value: < 2.2e-16
```

```r
AIC(lm1)
```

```
## [1] -1588.693
```

```r
##multilevel model
re.ini<- lme(sale_log~Land_Area+liveArea+noStories+Full_Baths+Half_Baths+Fireplaces+
                      Bedroom+yrsOld+disHwy+disSch+disMall +season+yPop+white+hispanic+
                      income_log+homeVal_log+migration+medAge+BGyrs, random = ~1|GEOID,
                      method = "ML")
summary(re.ini)$tTable
```

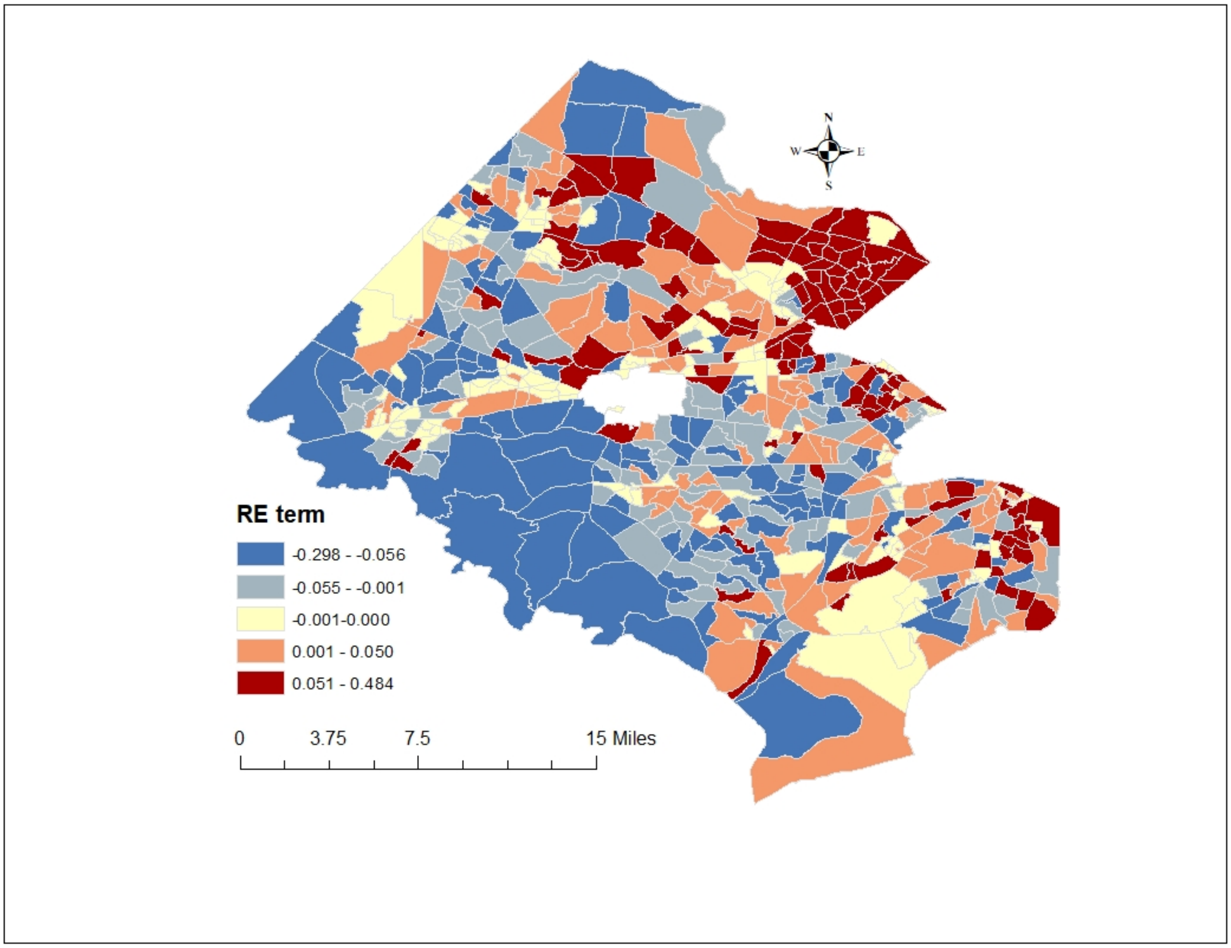
```
##                      Value    Std.Error   DF     t-value       p-value
## (Intercept)   8.6813433266 0.3330192249 8044  26.0685951 6.899404e-144
## Land_Area     1.1276093069 0.0737068154 8044  15.2985759  4.211700e-52
## liveArea      0.1169023883 0.0043067752 8044  27.1438336 2.533300e-155
## noStories    -0.0005580893 0.0064423216 8044  -0.0866286  9.309689e-01
```

```
## Full_Baths    0.0543753543 0.0036615768 8044  14.8502562  3.099468e-49
## Half_Baths    0.0332657434 0.0050771993 8044   6.5519869  6.026970e-11
## Fireplaces    0.0346254713 0.0034590713 8044  10.0100484  1.887843e-23
## Bedroom       0.0182259737 0.0035268579 8044   5.1677653  2.426036e-07
## yrsOld       -0.0029256752 0.0001794654 8044 -16.3021666  8.293743e-59
## disHwy       -0.2616878440 0.5110217154 8044  -0.5120875  6.086038e-01
## disSch       -1.7077291368 0.2854836849 8044  -5.9818800  2.300085e-09
## disMall      -0.4474888570 0.1719169476 8044  -2.6029363  9.259877e-03
## seasonspring  0.0047538862 0.0065794835 8044   0.7225318  4.699886e-01
## seasonsummer  0.0216587214 0.0057510716 8044   3.7660323  1.670451e-04
## seasonwinter -0.0198834521 0.0066635971 8044  -2.9838917  2.854624e-03
## yPop         -0.3185743880 0.1114717483  483  -2.8578935  4.448809e-03
## white         0.0774622928 0.0472100571  483   1.6408007  1.014898e-01
## hispanic     -0.1569721160 0.0578877614  483  -2.7116633  6.933304e-03
## income_log   -0.0475063048 0.0277238133  483  -1.7135559  8.725195e-02
## homeVal_log   0.3699902496 0.0265522910  483  13.9344002  2.448627e-37
## migration    -0.2360935169 0.1267206689  483  -1.8631019  6.305468e-02
## medAge       -0.0025181625 0.0014178166  483  -1.7760848  7.634841e-02
## BGyrs         0.0030836506 0.0005083468  483   6.0660374  2.649219e-09
```

```r
AIC(re.ini)
```

```
## [1] -2473.492
```

```r
#(intra-block group spatial effects)
ranef.ini.re<-unlist(ranef(re.ini)) #random effects
##display intra-block group moderate spatial autocorrelation, z-score of Moran's I: 13.94
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\re_ini2.J
```




```
##multilevel ESF model
source("C:\\Users\\lxh152030\\Box\\Researches\\SAE\\lungCancer\\fl\\county\\stepAICc.R")
EV<- pnts16[,30:368] #eigenvectors

#a stepwise procedure is used to selected eigenvectors
#it takes time to run the stepwise function
#re.ini<- lme(sale_log~Land_Area+liveArea+noStories+Full_Baths+Half_Baths+Fireplaces+
#                        Bedroom+yrsOld+disHwy+disSch+disMall +season+yPop+white+hispanic+
#                        income_log+homeVal_log+migration+medAge+BGyrs, random = ~1|GEOID,
#                        method = "ML")

#re.full<- lme(sale_log~Land_Area+liveArea+noStories+Full_Baths+Half_Baths+Fireplaces+
#                        Bedroom+yrsOld+disHwy+disSch+disMall +season+yPop+white+hispanic+
#                        income_log+homeVal_log+migration+medAge+BGyrs+., random = ~1|GEOID,
#                        method = "ML")
#re.esf<- stepAICc(re.ini, scope=list(upper=re.full), direction="forward", trace = 1)

#79 eigenvectors selected
re.esf<- lme(sale_log~Land_Area+liveArea+noStories+Full_Baths+Half_Baths+Fireplaces+
                        Bedroom+yrsOld+disHwy+disSch+disMall+season+yPop+white+hispanic+
                        income_log+homeVal_log+migration+medAge+BGyrs+EV5 + EV3 + EV14 +
                        EV7 + EV23 + EV1 + EV10 + EV102 + EV13 + EV6 + EV47 + EV37 +
                        EV12 + EV30 + EV15 + EV53 +V385 + V383 + EV29 + V402 + EV44 +
```

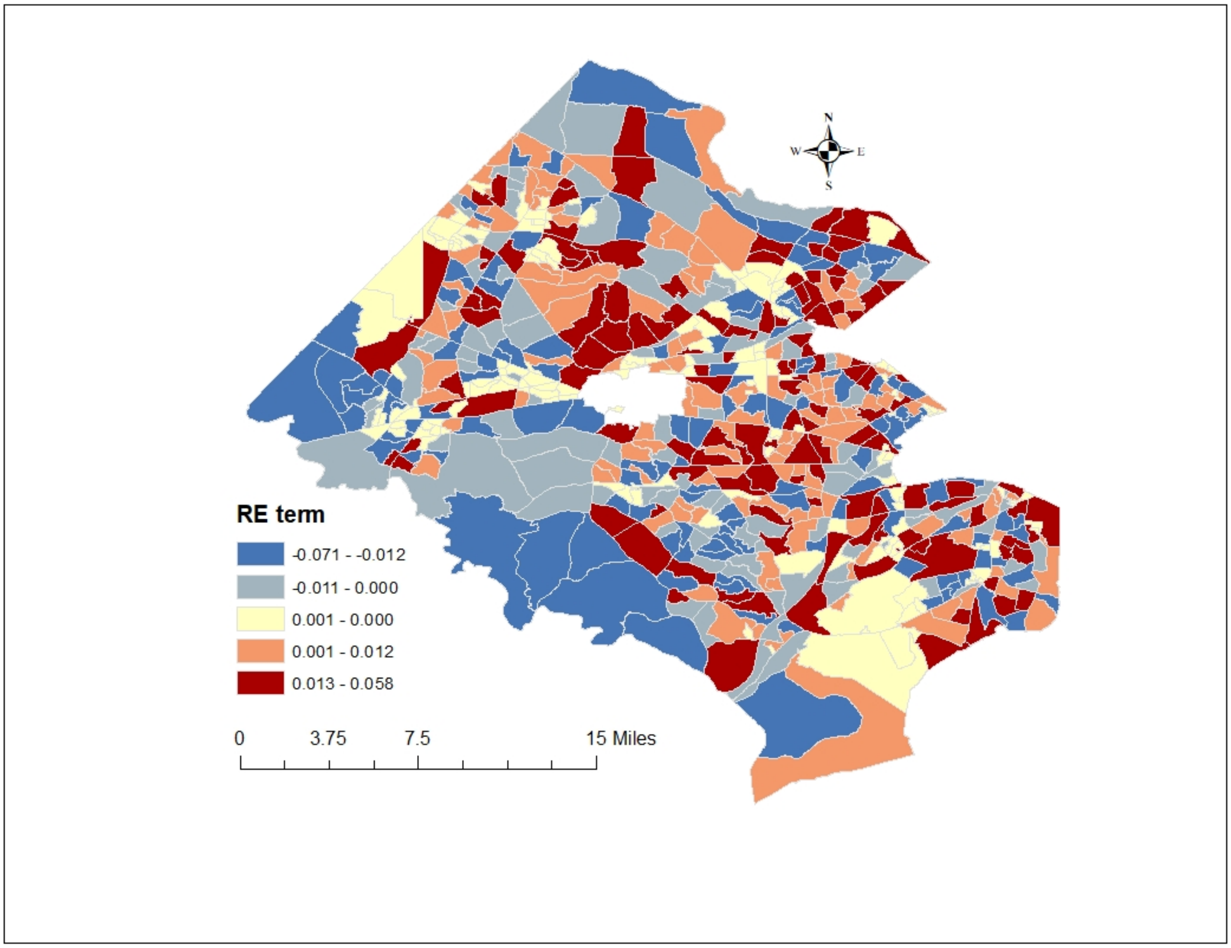
```r
                        EV26 + EV24 + V273 + EV90 +EV48 + V287 + EV17 + EV79 + EV60 +
                        V314 + V300 + V297 + V362 + EV93 + V399 + V429 + EV51 + V386 +
                        EV107 + EV11 + V450 + EV95 + EV9 + EV42 +  EV20 + EV22 + V491 +
                        EV82 + EV75 + EV21 +EV2 + V404 + V290 + EV19 + EV109 + V353 +
                        V378 + V405 + V301 + V288 + V462 + V270 + EV41 + V274 + V382 +
                        V310 + V272 + V327 +V345 + EV4 + EV33 + EV28 + EV16 + EV8 +
                        EV62 + EV18 + EV106 + EV81, data=EV, random = ~1|GEOID,
                        method = "ML")
summary(re.esf)$tTable[1:23,]
```

```
##                       Value    Std.Error   DF      t-value       p-value
## (Intercept)   1.046079e+01 0.2281223626 8044  45.8560613  0.000000e+00
## Land_Area     1.129826e+00 0.0696754256 8044  16.2155587  3.258853e-58
## liveArea      1.143159e-01 0.0042081871 8044  27.1651274 1.490540e-155
## noStories    -1.880415e-03 0.0062449685 8044  -0.3011089  7.633393e-01
## Full_Baths    5.411062e-02 0.0036062457 8044  15.0046964  3.256259e-50
## Half_Baths    3.657335e-02 0.0050099332 8044   7.3001676  3.148392e-13
## Fireplaces    3.408628e-02 0.0033875774 8044  10.0621391  1.120848e-23
## Bedroom       1.857621e-02 0.0034747755 8044   5.3460178  9.237794e-08
## yrsOld       -3.052233e-03 0.0001752292 8044 -17.4185180  9.911602e-67
## disHwy        1.007712e+00 0.4310951089 8044   2.3375628  1.943433e-02
## disSch       -1.204140e+00 0.2642365211 8044  -4.5570541  5.264149e-06
## disMall      -5.164842e-01 0.1944877778 8044  -2.6556128  7.932052e-03
## seasonspring  2.793953e-03 0.0065338146 8044   0.4276144  6.689433e-01
## seasonsummer  2.014150e-02 0.0057052033 8044   3.5303741  4.172939e-04
## seasonwinter -2.360094e-02 0.0066142164 8044  -3.5682143  3.615150e-04
## yPop         -8.913917e-02 0.0670635425  404  -1.3291748  1.845404e-01
## white         4.044940e-02 0.0286331522  404   1.4126773  1.585205e-01
## hispanic     -1.283193e-01 0.0349367444  404  -3.6729031  2.720699e-04
## income_log    5.596232e-02 0.0169743541  404   3.2968750  1.064141e-03
## homeVal_log   1.366249e-01 0.0191285522  404   7.1424602  4.299746e-12
## migration    -1.158589e-01 0.0717052356  404  -1.6157660  1.069255e-01
## medAge        1.217417e-03 0.0008109600  404   1.5012041  1.340839e-01
## BGyrs        -8.626184e-05 0.0003619852  404  -0.2383022  8.117676e-01
```

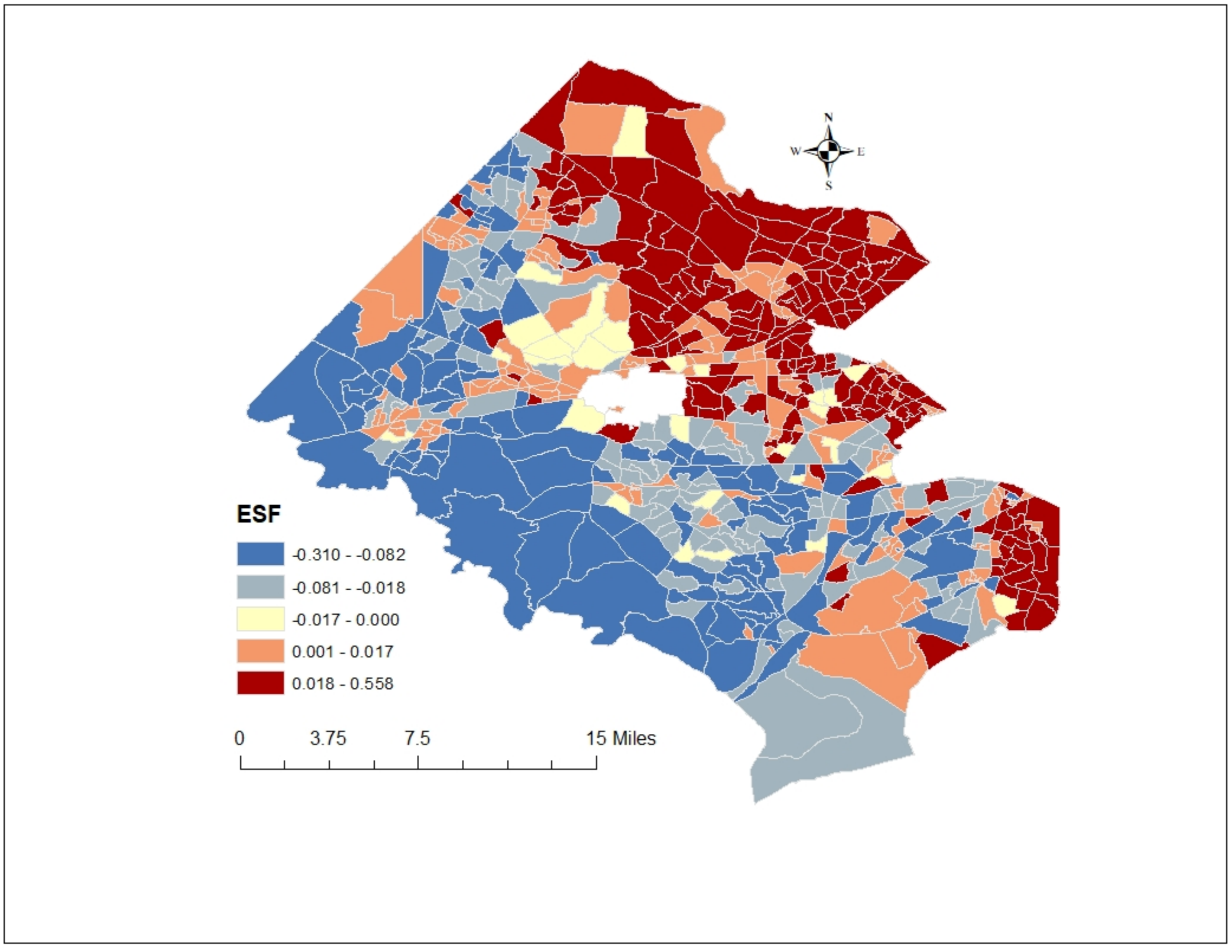
```r
AIC(re.esf)
```

```
## [1] -2961.103
```

```r
#(intra-block group spatial effects)
ranef.ini.re<-unlist(ranef(re.esf))
#intra-block group spatial autocorrelation reduces to z-score of Moran's I: 2.34
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\re_esf2.J
```




```
#(inter-block group spatial effects) a linear combination of the selected eigenvectors
esf.co<- as.matrix(re.esf$coefficients$fixed)[24:102]
eig.sel<- as.matrix(names(re.esf$coefficients$fixed))[-c(1:23)]
esf.sel<- as.matrix(pnts16[, eig.sel])%*%as.matrix(esf.co)
#strong inter-block group spatial autocorrelation, z-score of Moran's I: 32.16(<0.001)
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\ESF2.JPG")
```




```
detach(pnts16)

###2017 house price prediction
pre17<- read.csv("C:\\Users\\lxh152030\\Desktop\\test\\pre2017.csv")
pre17$income_log<- log(pre17$income)
pre17$homeVal_log<- log(pre17$homeVal)

#create the model matrix
ModelM<- model.matrix(lm(sale_log~Land_Area+liveArea+noStories+FullBaths+HalfBaths+
                                  Fireplaces+Bedroom_Co+yrsOld+disHwy+disSch+disMall+
                                  season+yPop+white+hispanic+income_log+
                                  homeVal_log+migration+medAge+BGyrs,data = pre17))
#prediction with hedonic estimates
co.lm<- coefficients(lm1)
pre.lm<-  exp(ModelM%*% as.matrix(co.lm))

##prediction with multilevel model estimates
coe.re<- as.matrix(re.ini$coefficients$fixed)
pre.re<-  exp(ModelM %*%coe.re+ pre17$re_ini)

##prediction with multilevel ESF model estimates
ModelM.esf<- model.matrix(lm(sale_log~Land_Area+liveArea+noStories+FullBaths+
                                      HalfBaths+Fireplaces+Bedroom_Co+yrsOld+
```

```
                                            disHwy+disSch+disMall +season+yPop+white+
                                            hispanic+income_log+homeVal_log+migration+
                                            medAge+BGyrs +EV5 + EV3 + EV14 + EV7 + EV23 +
                                            EV1 + EV10 + EV102 + EV13 + EV6 + EV47 + EV37 +
                                            EV12 + EV30 + EV15 + EV53 +V385 + V383 +
                                            EV29 + V402 + EV44 + EV26 + EV24 + V273 +
                                            EV90 +EV48 + V287 + EV17 + EV79 + EV60 + V314 +
                                            V300 + V297 + V362 + EV93 + V399 + V429 + EV51 +
                                            V386 + EV107 + EV11 + V450 + EV95 + EV9 + EV42 +
                                            EV20 + EV22 + V491 + EV82 + EV75 + EV21 +EV2 +
                                            V404 + V290 + EV19 + EV109 + V353 + V378 + V405 +
                                            V301 + V288 + V462 + V270 +EV41 + V274 + V382 +
                                            V310 + V272 + V327 +V345 + EV4 + EV33 + EV28 +
                                            EV16 + EV8 + EV62 + EV18 + EV106 + EV81, data = pre17))

coe.esf<- as.matrix(re.esf$coefficients$fixed)
pre.esf<-  exp(ModelM.esf%*% as.matrix(coe.esf)+ pre17$re_esf3)

##prediction analysis
pre_pre<- cbind(pre17$Recent_S_1, pre.lm, pre.re, pre.esf)
pre_pre<- pre_pre/1000

#correlations between observed prices and predicted prices
cor(pre_pre[,1], pre_pre[,2])
```

```
## [1] 0.8448542
```

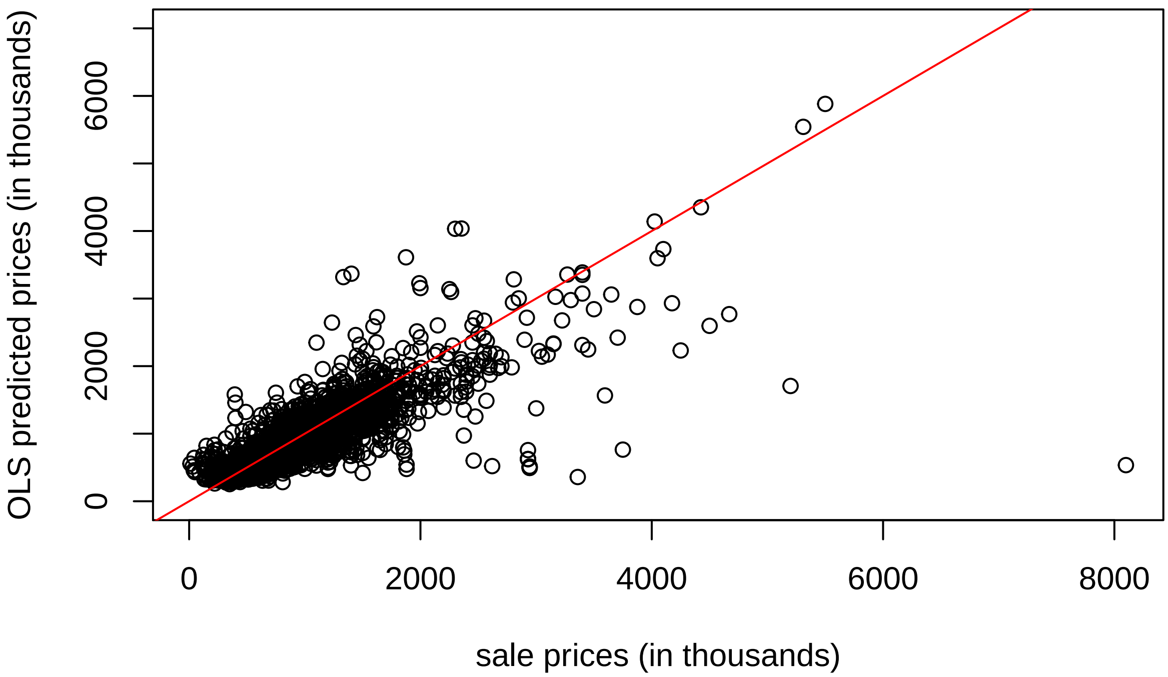
```
cor(pre_pre[,1], pre_pre[,3])
```

```
## [1] 0.8668865
```

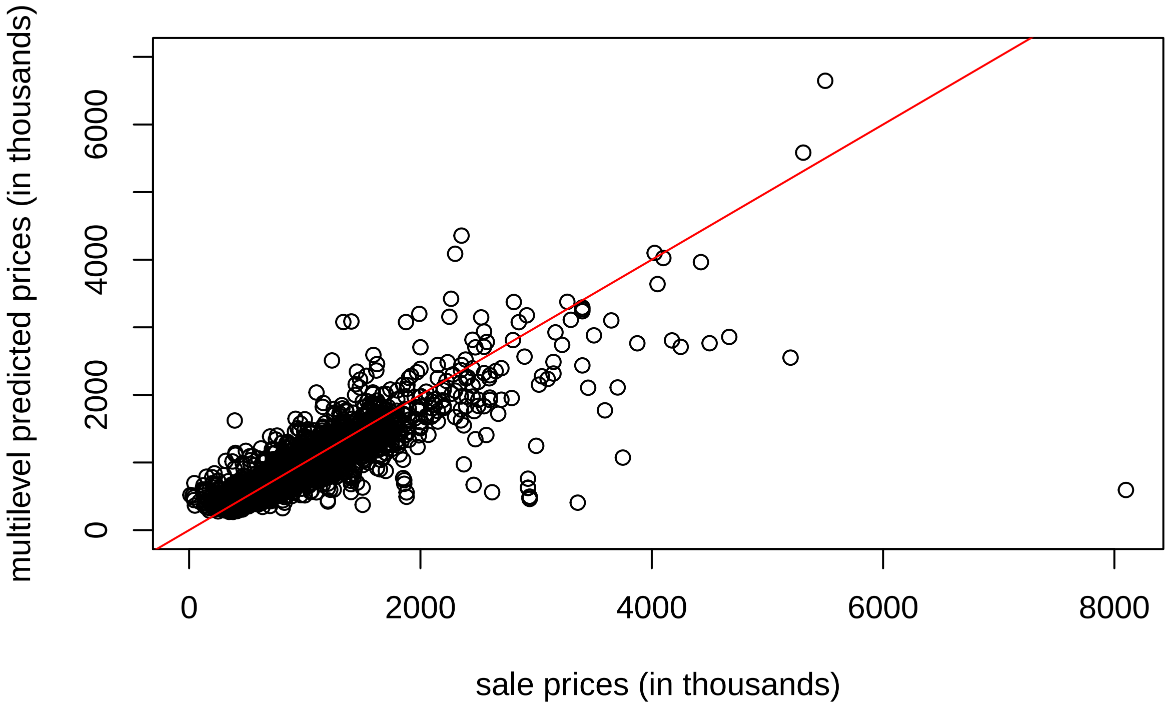
```
cor(pre_pre[,1], pre_pre[,4])
```

```
## [1] 0.8679157
```

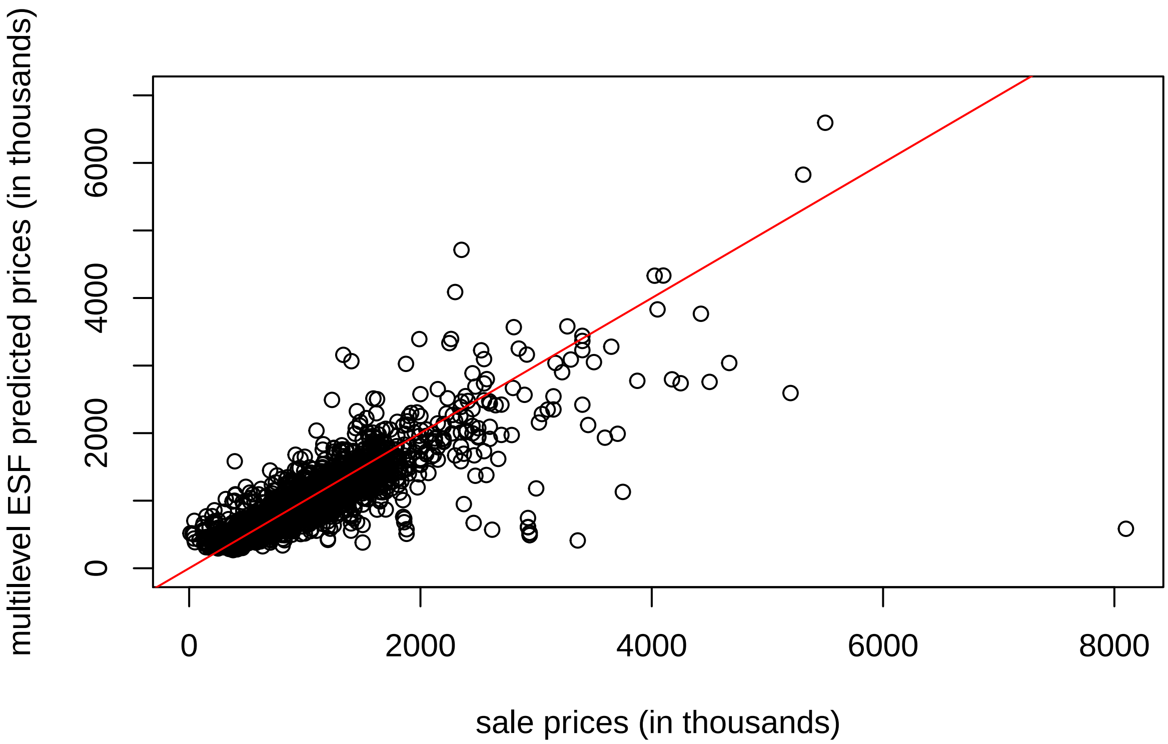
```
options(scipen=999)
#scatterplots of observed and predicted prices
plot(pre_pre[,1], pre_pre[,2], ylim=c(0, 7000), xlab="sale prices (in thousands)",
     ylab="OLS predicted prices (in thousands)")
abline(0,1, col = 2)
```



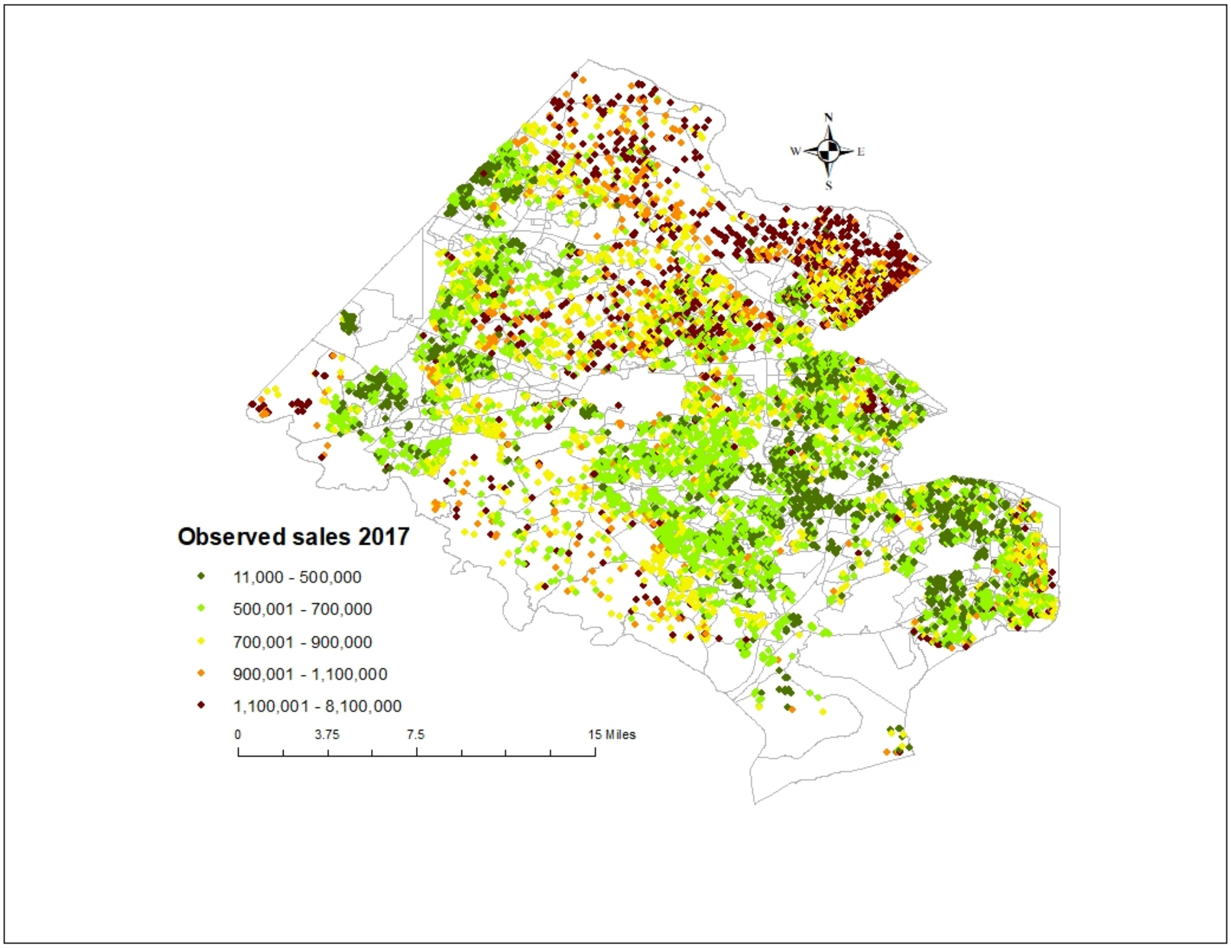

```
plot(pre_pre[,1], pre_pre[,3], ylim=c(0, 7000), xlab="sale prices (in thousands)",
     ylab="multilevel predicted prices (in thousands)")
abline(0,1, col = 2)
```



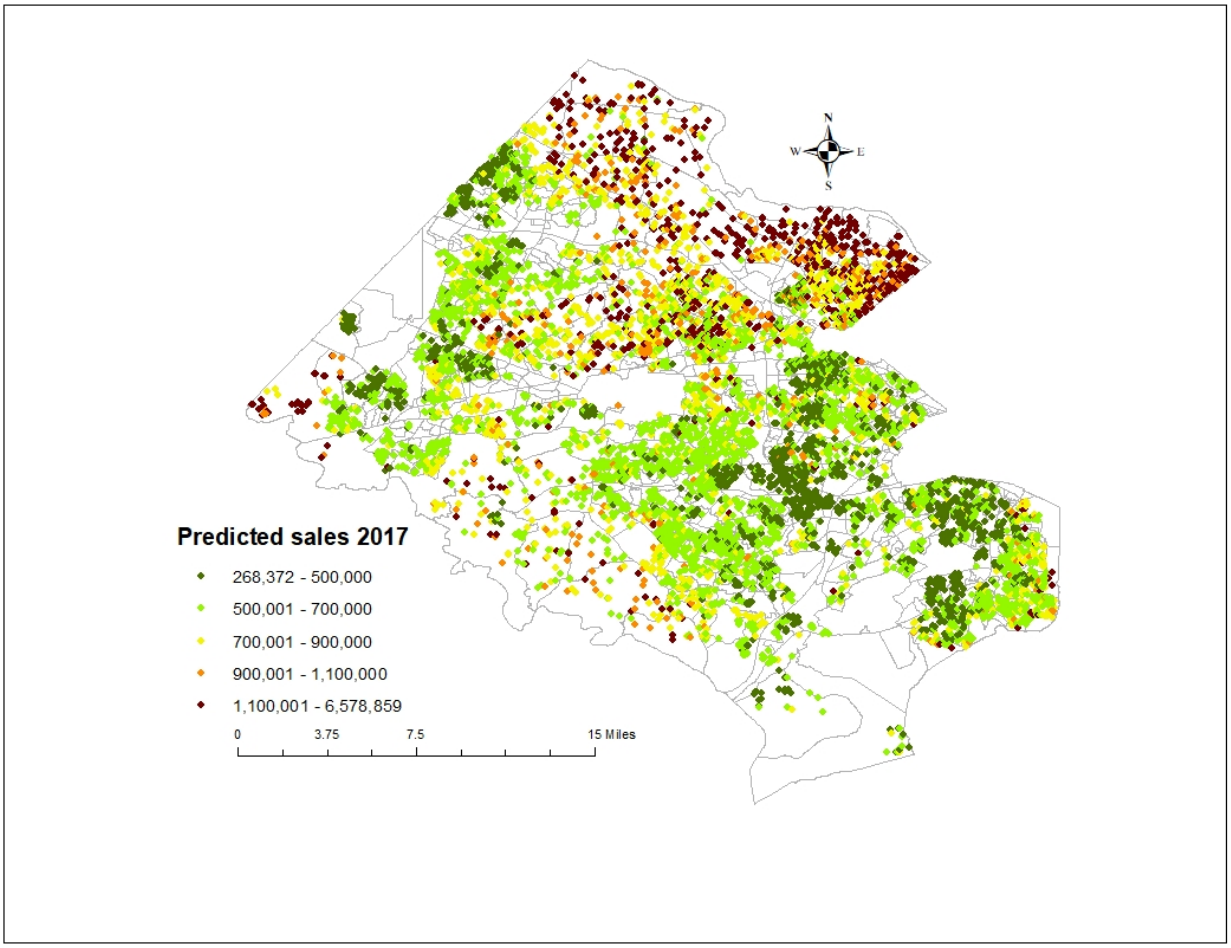

```r
plot(pre_pre[,1], pre_pre[,4], ylim=c(0, 7000), xlab="sale prices (in thousands)",
     ylab="multilevel ESF predicted prices (in thousands)")
abline(0,1, col = 2)
```




```
#prediciton maps
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\obs_sales
```




```
include_graphics("C:\\Users\\lxh152030\\Box\\Researches\\manuscripts\\housePrice\\pro1P\\new\\pred_sale
```


N
W
E
S
Predicted sales 2017
268,372 - 500,000
500,001 - 700,000
700,001 - 900,000
900,001 - 1,100,000
1,100,001 - 6,578,859
0
3.75
7.5
15 Miles