Exploratory Variography

Manoradhan Murugesan

11/13/2016

Overview Dataset description

Variogram is a function describing the degree of spatial dependence of a spatial random field or stochastic process in terms of the variance of the difference between field values at two locations across realizations of the field. Seattle's 98103 zip code house prices dataset is used in this exploratory variography analysis to study the trends in price/sqft.

Summary of findings

The variogram plot reveals an upward sloping trend in data, which is a violation of the stationarity of variance. Outliers were identified from the variogram cloud plot, revealing that other factors such as a view of the lake in this zip code and building amenities and age differences are at play here leading to extreme differences in rates per sqft between properties located close to each other. The upward sloping trend was moderated by the transformation, log(RATEPSQFT)~long+lat. Since distances larger than 0.015 have large fluctuations and a downward slope, the cutoff was chosen as 0.015. The directional variograms dont reveal significant differences in scale (sill and nuggets) between angles, however, spatial correlation is the weakest in 135° direction and strongest in 0°. Smaller bin sizes result in more granular plots which dont flatten, and this however must be taken in the light of reduced number of points per bin and the scale of y-axis. This could be due to structural differences and age differences between properties which get scaled up due to reduced number of points per bin.

Analysis

The data was loaded for analysis and the summary reveals the variable of interest, RATEPSQFT, and the location information.

```
library(foreign)
library(sp)
library(gstat)

kng <- read.dbf("98103.dbf")

summary(kng)
```

```
##    RATEPSQFT              id                          date
##  Min.   :0.001283   Min.   :9.100e+07   20140623T000000:  7
##  1st Qu.:0.002379   1st Qu.:1.972e+09   20140714T000000:  7
##  Median :0.002777   Median :4.083e+09   20140507T000000:  6
##  Mean   :0.002895   Mean   :4.704e+09   20140508T000000:  6
##  3rd Qu.:0.003363   3rd Qu.:7.354e+09   20140709T000000:  6
##  Max.   :0.007067   Max.   :9.551e+09   20140716T000000:  6
##                                         (Other)        :564
##      price            bedrooms       bathrooms       sqft__ving
##  Min.   : 238000   Min.   : 1.00   Min.   :0.000   Min.   : 390
##  1st Qu.: 432125   1st Qu.: 2.00   1st Qu.:1.000   1st Qu.:1242
```

```
##  Median : 550000   Median : 3.00   Median :2.000   Median :1505
##  Mean   : 584919   Mean   : 3.06   Mean   :1.663   Mean   :1651
##  3rd Qu.: 695000   3rd Qu.: 3.00   3rd Qu.:2.000   3rd Qu.:1960
##  Max.   :1695000   Max.   :33.00   Max.   :4.000   Max.   :4360
##
##     sqft_lot        floors        waterfront      view
##  Min.   : 651   Min.   :1.000   Min.   :0    Min.   :0.0000
##  1st Qu.:1563   1st Qu.:1.000   1st Qu.:0    1st Qu.:0.0000
##  Median :3500   Median :1.000   Median :0    Median :0.0000
##  Mean   :3482   Mean   :1.703   Mean   :0    Mean   :0.1445
##  3rd Qu.:4800   3rd Qu.:3.000   3rd Qu.:0    3rd Qu.:0.0000
##  Max.   :9450   Max.   :3.000   Max.   :0    Max.   :4.0000
##
##    condition         grade         sqft_above     sqft__ment
##  Min.   :1.000   Min.   : 5.00   Min.   : 390   Min.   :   0.0
##  1st Qu.:3.000   1st Qu.: 7.00   1st Qu.:1070   1st Qu.:   0.0
##  Median :3.000   Median : 7.00   Median :1365   Median :   0.0
##  Mean   :3.483   Mean   : 7.41   Mean   :1405   Mean   : 245.8
##  3rd Qu.:4.000   3rd Qu.: 8.00   3rd Qu.:1603   3rd Qu.: 480.0
##  Max.   :5.000   Max.   :11.00   Max.   :3920   Max.   :1540.0
##
##     yr_built      yr_re_ated        zipcode           lat
##  Min.   :1900   Min.   :   0.0   Min.   :98103   Min.   :47.65
##  1st Qu.:1917   1st Qu.:   0.0   1st Qu.:98103   1st Qu.:47.66
##  Median :1934   Median :   0.0   Median :98103   Median :47.68
##  Mean   :1953   Mean   : 109.3   Mean   :98103   Mean   :47.68
##  3rd Qu.:2005   3rd Qu.:   0.0   3rd Qu.:98103   3rd Qu.:47.69
##  Max.   :2015   Max.   :2014.0   Max.   :98103   Max.   :47.70
##
##       long          sqft__ng15     sqft_lot15
##  Min.   :-122.4   Min.   : 690   Min.   :1026
##  1st Qu.:-122.3   1st Qu.:1330   1st Qu.:1610
##  Median :-122.3   Median :1500   Median :3850
##  Mean   :-122.3   Mean   :1524   Mean   :3472
##  3rd Qu.:-122.3   3rd Qu.:1700   3rd Qu.:4560
##  Max.   :-122.3   Max.   :2660   Max.   :8431
##
```

We convert this to a SpatialPointsDataFrame with the coordinates function.

```
coordinates(kng) <- ~long + lat
summary(kng)
```

```
## Object of class SpatialPointsDataFrame
## Coordinates:
##            min       max
## long -122.3640 -122.3290
## lat    47.6485   47.7011
## Is projected: NA
## proj4string : [NA]
## Number of points: 602
## Data attributes:
##    RATEPSQFT              id                        date
```

```
##  Min.   :0.001283   Min.   :9.100e+07   20140623T000000:  7
##  1st Qu.:0.002379   1st Qu.:1.972e+09   20140714T000000:  7
##  Median :0.002777   Median :4.083e+09   20140507T000000:  6
##  Mean   :0.002895   Mean   :4.704e+09   20140508T000000:  6
##  3rd Qu.:0.003363   3rd Qu.:7.354e+09   20140709T000000:  6
##  Max.   :0.007067   Max.   :9.551e+09   20140716T000000:  6
##                                         (Other)        :564
##      price            bedrooms       bathrooms       sqft__ving
##  Min.   : 238000   Min.   : 1.00   Min.   :0.000   Min.   : 390
##  1st Qu.: 432125   1st Qu.: 2.00   1st Qu.:1.000   1st Qu.:1242
##  Median : 550000   Median : 3.00   Median :2.000   Median :1505
##  Mean   : 584919   Mean   : 3.06   Mean   :1.663   Mean   :1651
##  3rd Qu.: 695000   3rd Qu.: 3.00   3rd Qu.:2.000   3rd Qu.:1960
##  Max.   :1695000   Max.   :33.00   Max.   :4.000   Max.   :4360
##
##     sqft_lot        floors        waterfront      view
##  Min.   : 651   Min.   :1.000   Min.   :0    Min.   :0.0000
##  1st Qu.:1563   1st Qu.:1.000   1st Qu.:0    1st Qu.:0.0000
##  Median :3500   Median :1.000   Median :0    Median :0.0000
##  Mean   :3482   Mean   :1.703   Mean   :0    Mean   :0.1445
##  3rd Qu.:4800   3rd Qu.:3.000   3rd Qu.:0    3rd Qu.:0.0000
##  Max.   :9450   Max.   :3.000   Max.   :0    Max.   :4.0000
##
##    condition         grade         sqft_above     sqft__ment
##  Min.   :1.000   Min.   : 5.00   Min.   : 390   Min.   :   0.0
##  1st Qu.:3.000   1st Qu.: 7.00   1st Qu.:1070   1st Qu.:   0.0
##  Median :3.000   Median : 7.00   Median :1365   Median :   0.0
##  Mean   :3.483   Mean   : 7.41   Mean   :1405   Mean   : 245.8
##  3rd Qu.:4.000   3rd Qu.: 8.00   3rd Qu.:1603   3rd Qu.: 480.0
##  Max.   :5.000   Max.   :11.00   Max.   :3920   Max.   :1540.0
##
##     yr_built      yr_re_ated        zipcode        sqft__ng15
##  Min.   :1900   Min.   :   0.0   Min.   :98103   Min.   : 690
##  1st Qu.:1917   1st Qu.:   0.0   1st Qu.:98103   1st Qu.:1330
##  Median :1934   Median :   0.0   Median :98103   Median :1500
##  Mean   :1953   Mean   : 109.3   Mean   :98103   Mean   :1524
##  3rd Qu.:2005   3rd Qu.:   0.0   3rd Qu.:98103   3rd Qu.:1700
##  Max.   :2015   Max.   :2014.0   Max.   :98103   Max.   :2660
##
##    sqft_lot15
##  Min.   :1026
##  1st Qu.:1610
##  Median :3850
##  Mean   :3472
##  3rd Qu.:4560
##  Max.   :8431
##
```

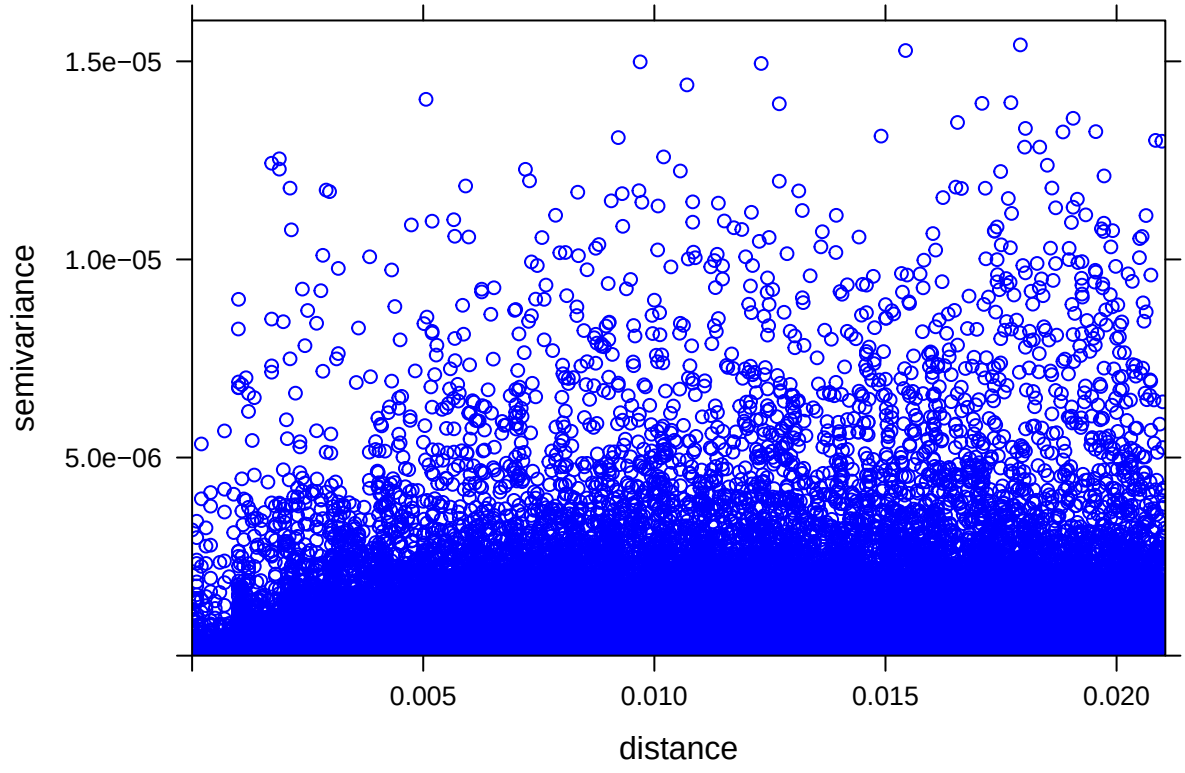
We then create a variogram cloud and plot it to identify outliers.

```
vcloud <- variogram(RATEPSQFT ~ 1, data = kng, cloud = TRUE)
summary(vcloud)
```

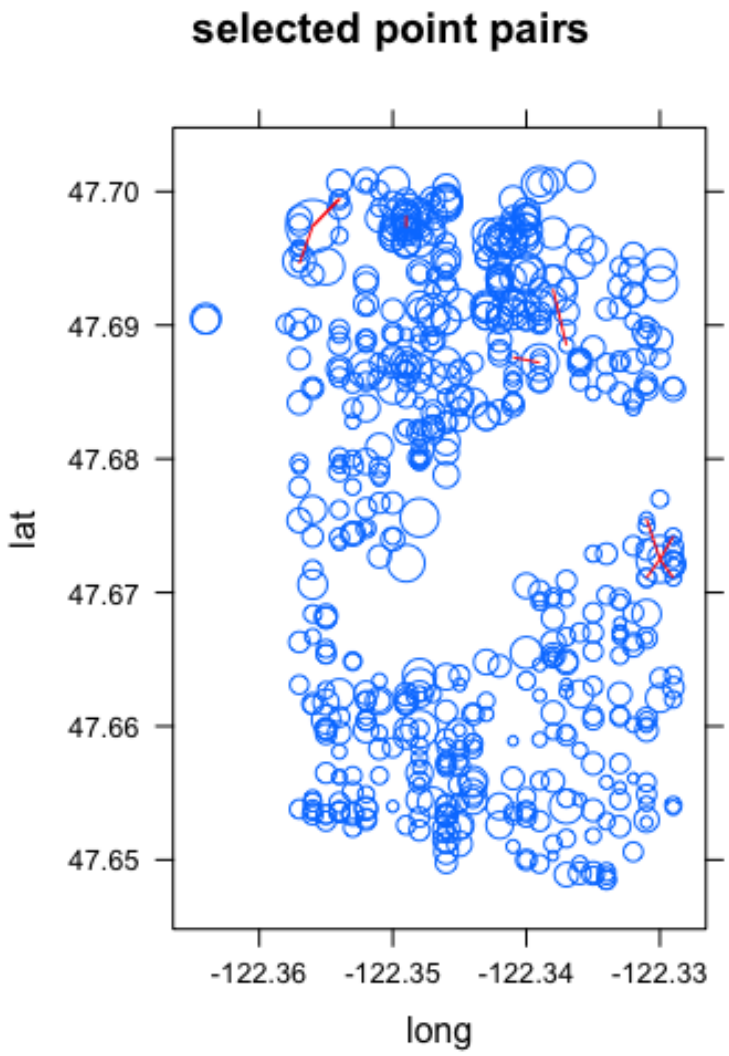
```
##        np                dist              gamma              dir.hor
```

```
##  Min.   :1.000e+00   Min.   :0.000000   Min.   :0.000e+00   Min.   :0
##  1st Qu.:3.393e+11   1st Qu.:0.007632   1st Qu.:4.369e-08   1st Qu.:0
##  Median :7.387e+11   Median :0.011942   Median :2.042e-07   Median :0
##  Mean   :8.481e+11   Mean   :0.011850   Mean   :5.525e-07   Mean   :0
##  3rd Qu.:1.280e+12   3rd Qu.:0.016386   3rd Qu.:6.347e-07   3rd Qu.:0
##  Max.   :2.573e+12   Max.   :0.021054   Max.   :1.542e-05   Max.   :0
##     dir.ver     id
##  Min.   :0   var1:88719
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

```
plot(vcloud, pch = 1, col = "blue")
```

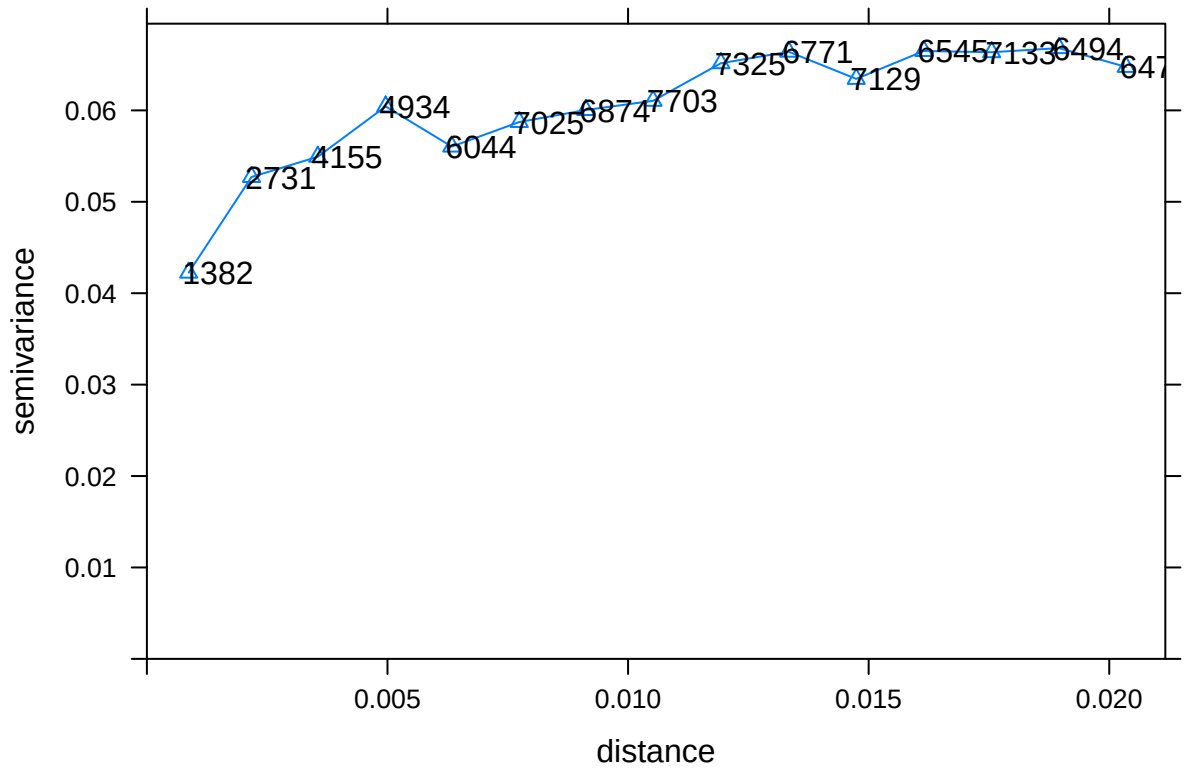



```
plot(plot(vcloud, identify = T), kng, cex = 3 * kng$RATEPSQFT/max(kng$RATEPSQFT),
    pch = 1)
```



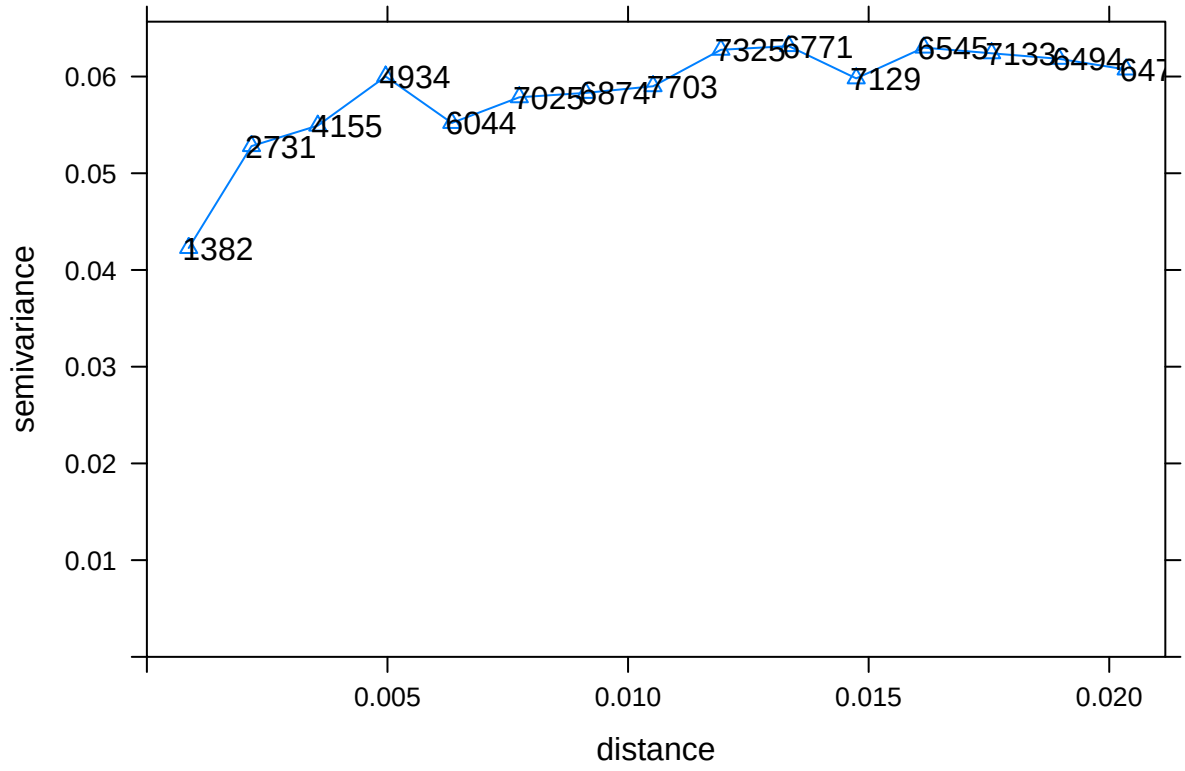
We then plot an empirical variogram plot and examine the number of points in each bin.

```
v1 <- variogram(log(RATEPSQFT) ~ 1, kng)
plot(v1, type = "b", pch = 2, pl = T)
```



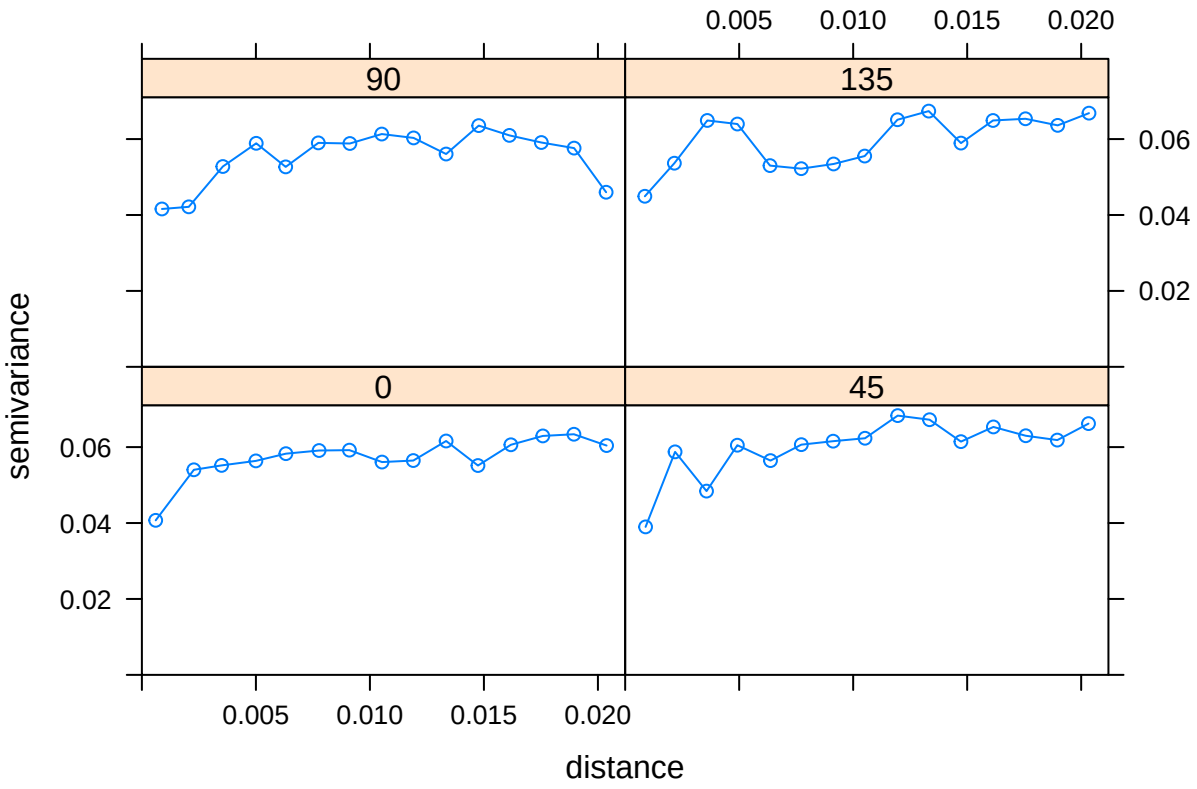
We see that there is an upward sloping trend, revealing the spatial dependence. We use a regression specification of RATEPSQFT in terms of the location coordinates to make the plot flatten out.

```
v2 <- variogram(log(RATEPSQFT) ~ long + lat, kng)
plot(v2, type = "b", pch = 2, pl = T)
```



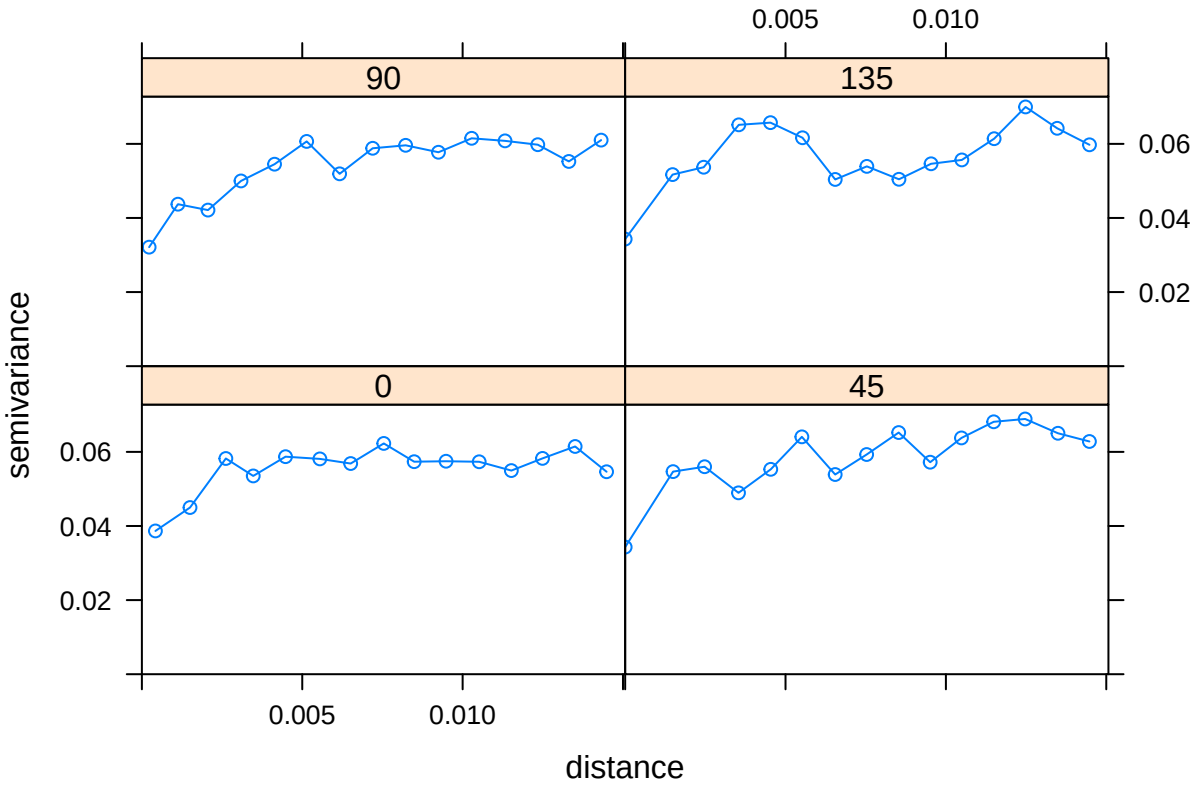
We then plot directional variograms to see if the observed semivariance trend is the same in all the four directions.

```
vgm.aniso <- variogram(log(RATEPSQFT) ~ long + lat, kng, alpha = c(0, 45, 90, 135))
plot(vgm.aniso, type = "b")
```



Since the plots tend to fluctuate more after 0.015 and since we are only concerned with the points that are closer to each other, we choose a safe cut-off distance of 0.015.

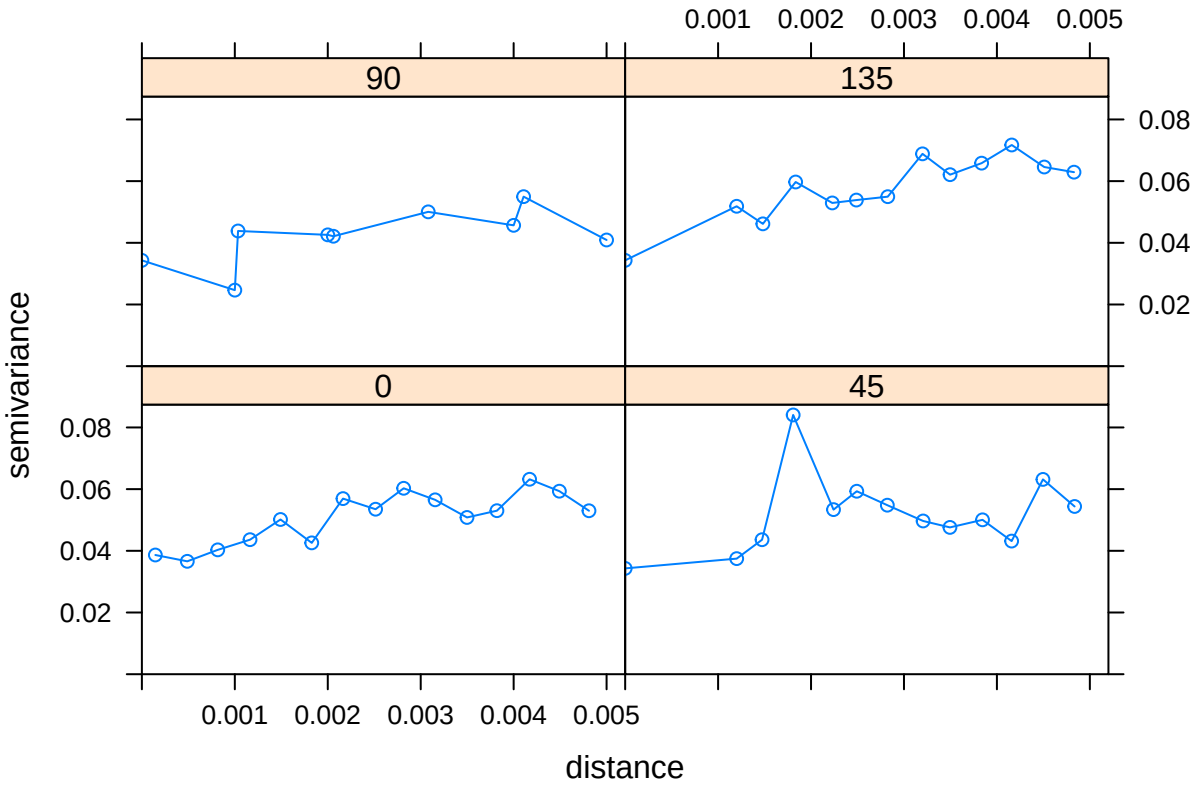
```
vgm.aniso <- variogram(log(RATEPSQFT) ~ long + lat, kng, alpha = c(0, 45, 90, 135),
    cutoff = 0.015)
plot(vgm.aniso, type = "b")
```



From the directional plots, we see that there is slight change in sill and nuggets between 0° and other directions. Also 135° direction has the least strong correlation while 0° has the strongest correlation.

Reducing bin size, we see more volatile plots, revealing more granular differences in the rates. This could be due to structural differences and age differences between properties which get scaled up due to reduced number of points per bin.

```
vgm.aniso <- variogram(log(RATEPSQFT) ~ long + lat, kng, alpha = c(0, 45, 90, 135),
    cutoff = 0.005)
plot(vgm.aniso, type = "b")
```


90
135
0
45
semivariance
distance
0.001
0.002
0.003
0.004
0.005
0.02
0.04
0.06
0.08