Trabalho: Estatística Espacial
Prof.: Jony Arrais Pinto Junior
Aluno: Willian Teixeira Silva de Sousa
Filial: FGV RJ 2 - Barra da Tijuca
Turma: MRJ21973-TBABD-7
Data da Entrega: 12/ago/2020

Questão 1

Objetivo: Avaliar os três passos a seguir descritos como plano para compreender o comportamento espacial do número de veículos furtados nos bairros da cidade de Vitória. Nos casos incorretos ou incompletos, corrigir ou complementar as conclusões apresentadas.

Passo I: Criação de um mapa coroplético para a variável número de veículos furtados nos bairros, associando tonalidades de vermelho escuro para locais com maiores números de furtos e vermelho claro para locais com menores números.

Passo II: Criação de uma matriz de vizinhança W, considerando um critério.

Passo III: Cálculo do Índice de Moran Global (I = 0.61) com p-valor de 0.008.

Entende-se por “compreender o comportamento espacial” como o objetivo de identificar a existência de padrões espaciais nos valores observados. Como os bairros de uma região podem ter interrelações de acordo com suas interações econômicas/ecológicas, é possível que um valor de um bairro afete o valor de outro. Logo, é preciso investigar se há alguma tendência em que bairros próximos apresente valores parecidos com bairros distantes.

A criação do mapa coroplético do Passo I é útil, pois sombreia cada bairro com uma intensidade de cor proporcional ao número de veículos furtados. No entanto, é preciso observar se a distribuição desses números por bairro possa afetar a visualização das variações de alguma forma. Dependendo desta distribuição, pode-se tomar decisões sobre o número de classes/intervalos nesta escala além dos valores em cada intervalo.

É preciso observar também se há grandes variações na área dos bairros. Tais condições podem afetar a interpretação dos padrões, logo precisam ser consideradas pelo cientista de dados.

Segundo o enunciado, o cientista concluiu que “os bairros ao norte da cidade estavam mais escuros”, logo haveria “um menor número de furtos de carros nessa área”. Tal conclusão não faz sentido de acordo com a descrição da escala utilizada no mapa, onde a tonalidade mais escura estaria relacionada a “locais com maiores números de furtos”.

Ao executar os Passos II e III, o cientista também teria chegado à conclusão que “não haveria um padrão espacial para o número de veículos furtados”. A matriz de vizinhança mencionada define quais bairros são vizinhos entre si e define pesos para este relacionamento. Os critérios para essa definição de pesos podem ser de contiguidade (a partir da contagem de bordas e/ou vértices em comum) ou em relação às distâncias entre os centroides dos bairros.

A partir desta matriz de vizinhança, o Índice de Moran Global pode ser calculado para compreender a magnitude da autocorrelação espacial, ou seja, se localizações próximas tendem apresentar comportamentos similares – no caso, o número de veículos furtados. Fala-se em índice global, pois abrange a dependência da região analisada como um todo. Já o índice local é utilizado quando se quer entender essa relação de maneira mais detalhada.

A interpretação do Índice de Moran Global é feita de maneira que se o índice é positivo, regiões com valores similares se agrupam. Já se o índice é negativo, as regiões vizinhas entre si possuem diferenças consideráveis. Dessa forma, com o índice global de 0.61, o cientista errou ao dizer que não haveria um padrão espacial, uma vez que esse valor indica a existência de aglomerações nos números de cada bairro com validade estatística, já que o p-valor é relativamente baixo (0.8%). O índice positivo indica que bairros com número alto de veículos furtados são próximos de bairros com número semelhante, da mesma forma que também indica que bairros com baixo número de furtos estão próximos de outros bairros com números parecidos. Logo, o dado indica uma tendência espacial de aglomeração.

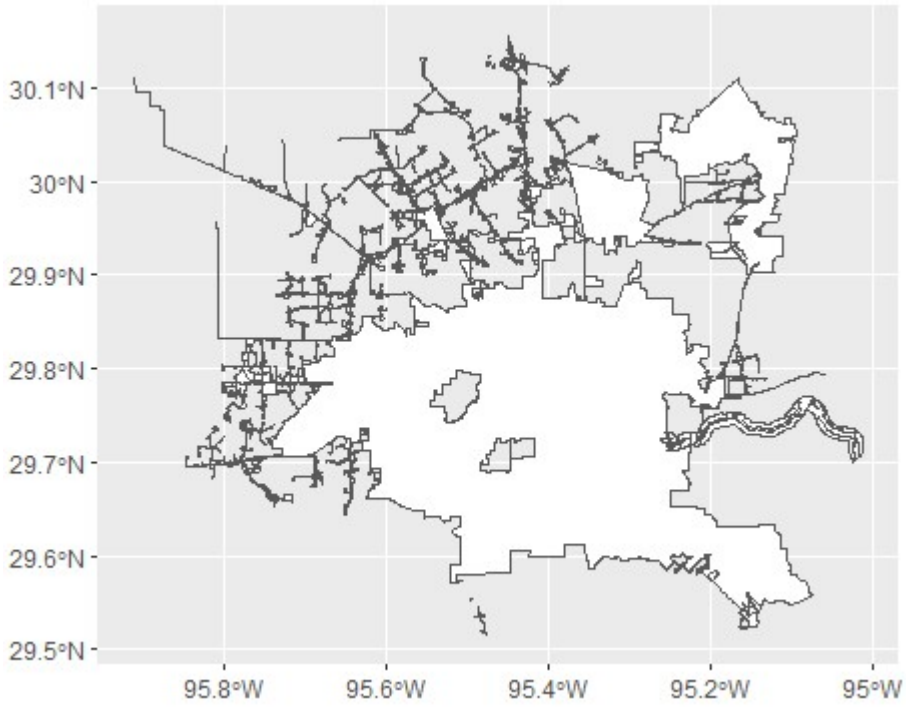
Questão 2

Objetivo: Fazer uma análise exploratória das localizações das ocorrências de crimes de acordo com seus tipos.

A cidade de Houston, no Texas (EUA), é o cenário da análise proposta. O arquivo sf (shape file) da localidade foi disponibilizado e o sistema de coordenadas presente é o WGS 84 (CRS 4326).

```
> Houston <- read_sf("Houston_City_Limit.shp")
> st_crs(Houston)
Coordinate Reference System:
  User input: WGS 84
  wkt:
GEOGCRS["WGS 84",
    DATUM["World Geodetic System 1984",
        ELLIPSOID["WGS 84",6378137,298.257223563,
            LENGTHUNIT["metre",1]]],
    PRIMEM["Greenwich",0,
        ANGLEUNIT["degree",0.0174532925199433]],
    CS[ellipsoidal,2],
        AXIS["latitude",north,
            ORDER[1],
            ANGLEUNIT["degree",0.0174532925199433]],
        AXIS["longitude",east,
            ORDER[2],
            ANGLEUNIT["degree",0.0174532925199433]],
    ID["EPSG",4326]]
```

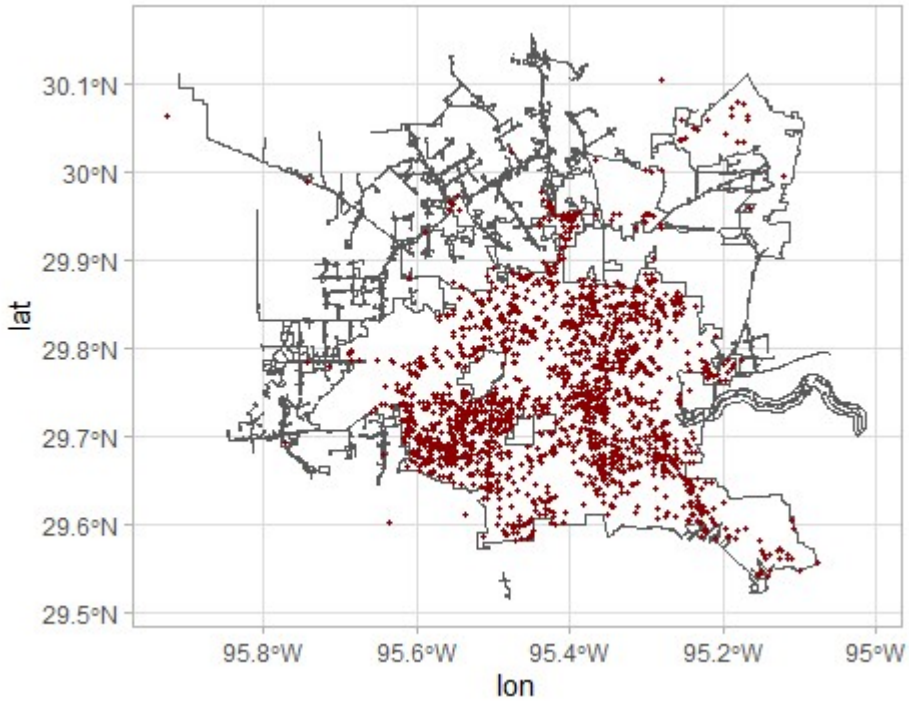
Uma simples visualização com o shape file de Houston resulta no gráfico abaixo:



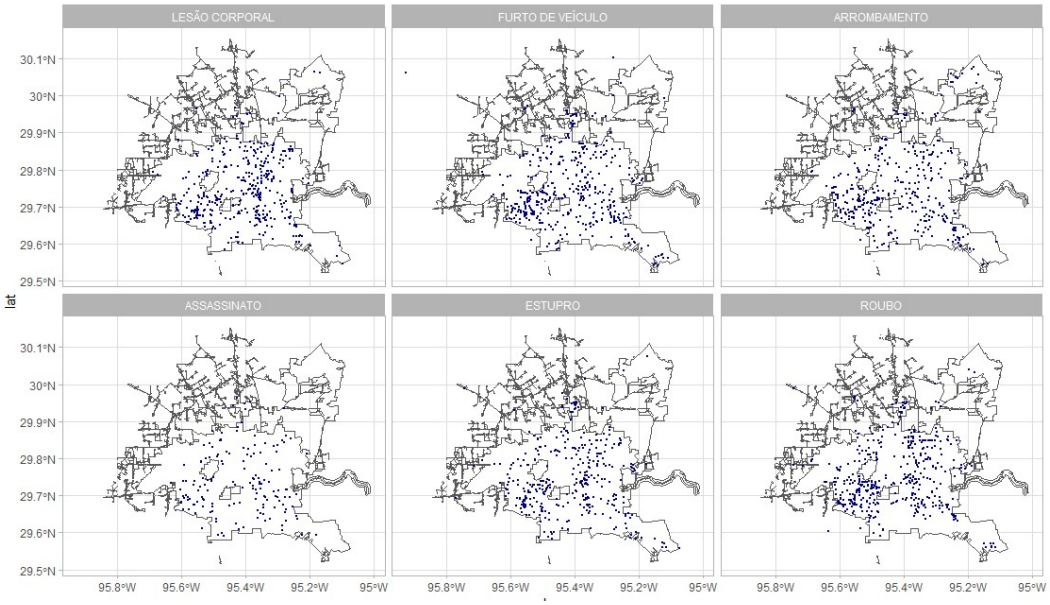
Além do shape file da cidade, foi disponibilizado também uma tabela contendo dados criminalísticos. A coluna “offense” representa seis tipos diferentes de crimes observados em Houston.

```
> crimeHouston <- read_csv("Base Houston.csv")
> head(crimeHouston)
# A tibble: 6 x 17
  time  date   hour premise offense beat  block street type  suffix number month
  <chr> <chr> <dbl> <chr>   <chr>   <chr> <chr> <chr>  <chr> <chr>   <dbl> <chr>
1 01/0~ 01/0~     0 18A     murder  1,50~ 9600~ marli~ ln    -           1 janu~
2 05/0~ 01/0~    15 20A     murder  6B60  1300~ greens pkwy  -           1 janu~
3 06/0~ 01/0~    22 18A     murder  19G50 1020~ bisso~ st    -           1 janu~
4 07/0~ 01/0~    18 20R     murder  6B40  1000~ marjo~ NA    -           1 janu~
5 07/0~ 01/0~    18 20R     murder  8C50  8200~ parker rd    -           1 janu~
6 07/0~ 01/0~     1 13R     murder  10H50 2700~ canfi~ st    -           1 janu~
# ... with 5 more variables: day <chr>, location <chr>, address <chr>, lon <dbl>,
#   lat <dbl>
> unique(crimeHouston$offense)
[1] "murder"             "aggravated assault" "rape"
[4] "burglary"           "auto theft"         "robbery"
```

A análise gráfica das ocorrências dos delitos em relação à localidade registrada pode levar em conta o total ou o número por tipo de ofensa:



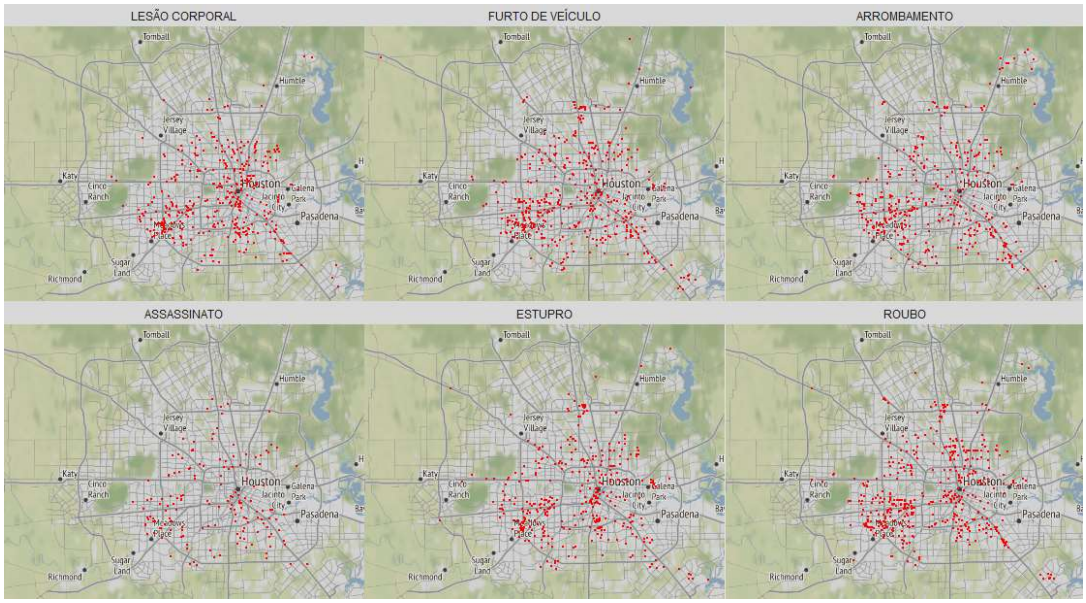
Todos os delitos registrados em Houston



Delitos registrados em Houston por tipo de crime

A mesma visualização executada pela função qmplot oferece mais detalhes do mapa:

```
>qmplot(x=lon,y=lat, data=crimeHouston, colour=I('red'), size=I(0.3), darken=0.1) +
  facet_wrap(~offense,
             labeller = labeller(offense = c("aggravated assault"="LESÃO CORPORAL",
                                             "rape" = "ESTUPRO",
                                             "robbery" = "ROUBO",
                                             "burglary" = "ARROMBAMENTO",
                                             "murder" = "ASSASSINATO",
                                             "auto theft" = "FURTO DE VEÍCULO")))
```



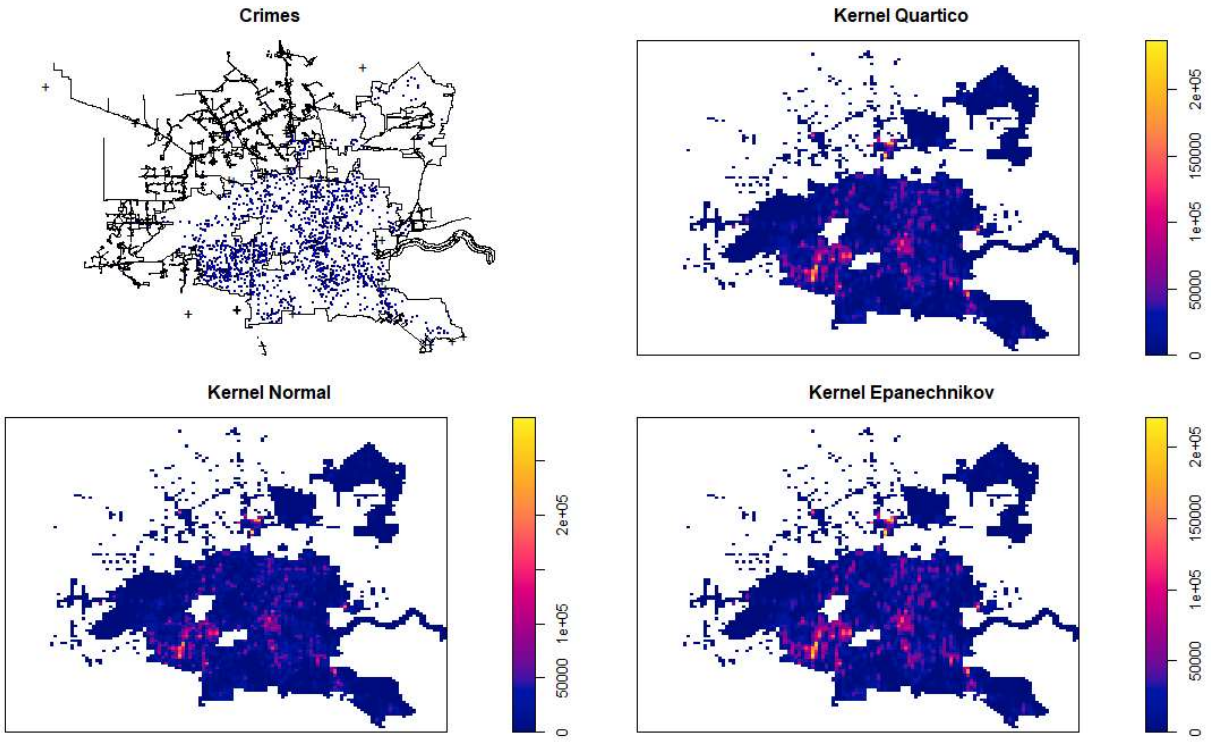
A partir de uma análise visual destes gráficos, apesar de ocorrências de crime acontecerem por toda a cidade, há uma aparente concentração no centro da cidade de Houston e outra perto da localidade identificada no mapa como Meadows Place.

No entanto, esta análise pode não ser eficaz pois estes gráficos não ilustram possíveis superposições de ocorrências/pontos, logo não é indicado para avaliar a densidade dos dados no mapa.

Visualizar a intensidade dos padrões de pontos (efeito de primeira ordem) permite tirar melhores conclusões sobre os fenômenos observados.

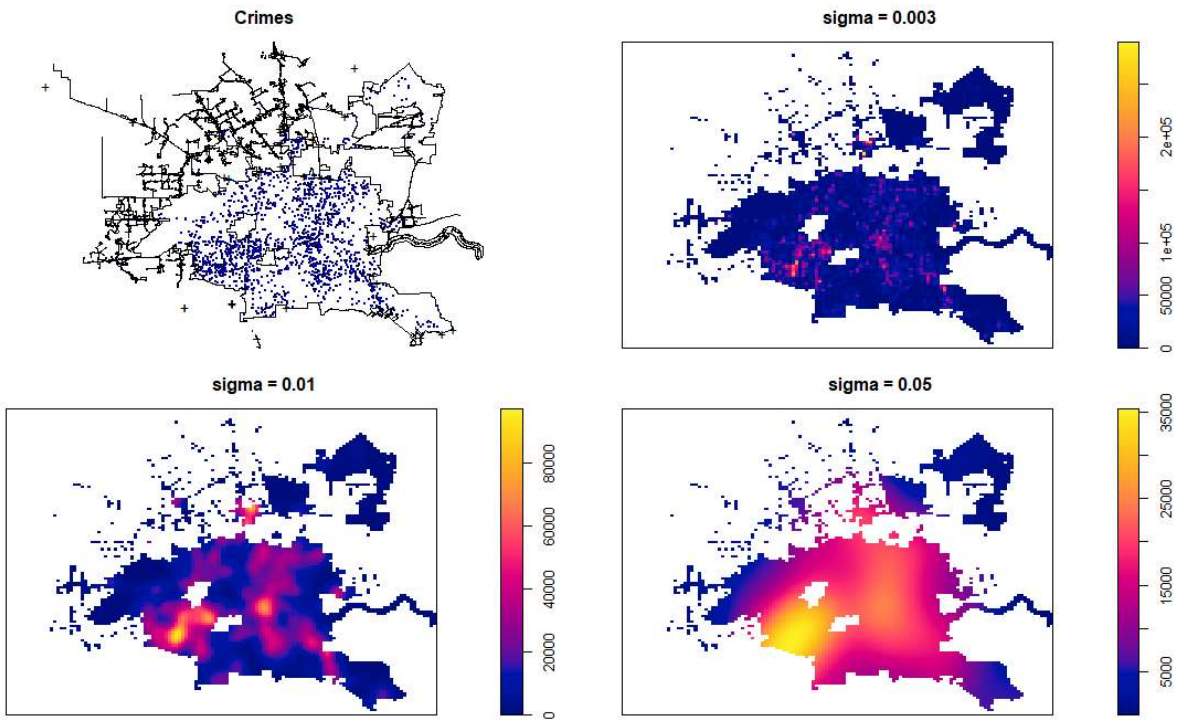
O código abaixo define uma janela de observação ("owin") a partir do shape file para criar o padrão dos pontos no plano ("ppp") utilizando os dados criminalísticos de Houston. Além disso, para o cálculo da densidade, o código também estima um raio ótimo para a aplicação dos diferentes possíveis kernels (Quartico, Normal e Epanechnikov):

```
>HoustonShp2 <- readShapePoly("Houston_City_Limit.shp")
>HoustonOWin <- as.owin(HoustonShp2)

>Houstonppp = ppp( x = crimeHouston$lon,
             y = crimeHouston$lat,
             window = HoustonOWin)

>raio.est = bw.diggle(Houstonppp)
>raio.est
#Sigma otimizado = 0.003

>Houstonkde.g1 = density.ppp(x = Houstonppp,
                        sigma = 0.003,
                        kernel="quartic")
>Houstonkde.g2 = density.ppp(x = Houstonppp,
                        sigma = 0.003,
                        kernel="gaussian")
>Houstonkde.g3 = density.ppp(x = Houstonppp,
                        sigma = 0.003,
                        kernel="epanechnikov")
```



Para efeito de visualização e análise gráfica, a decisão sobre o kernel impacta menos que a decisão do raio para compreender a densidade das ocorrências. Dessa forma, podem ser gerados novas imagens com raios um pouco maiores.

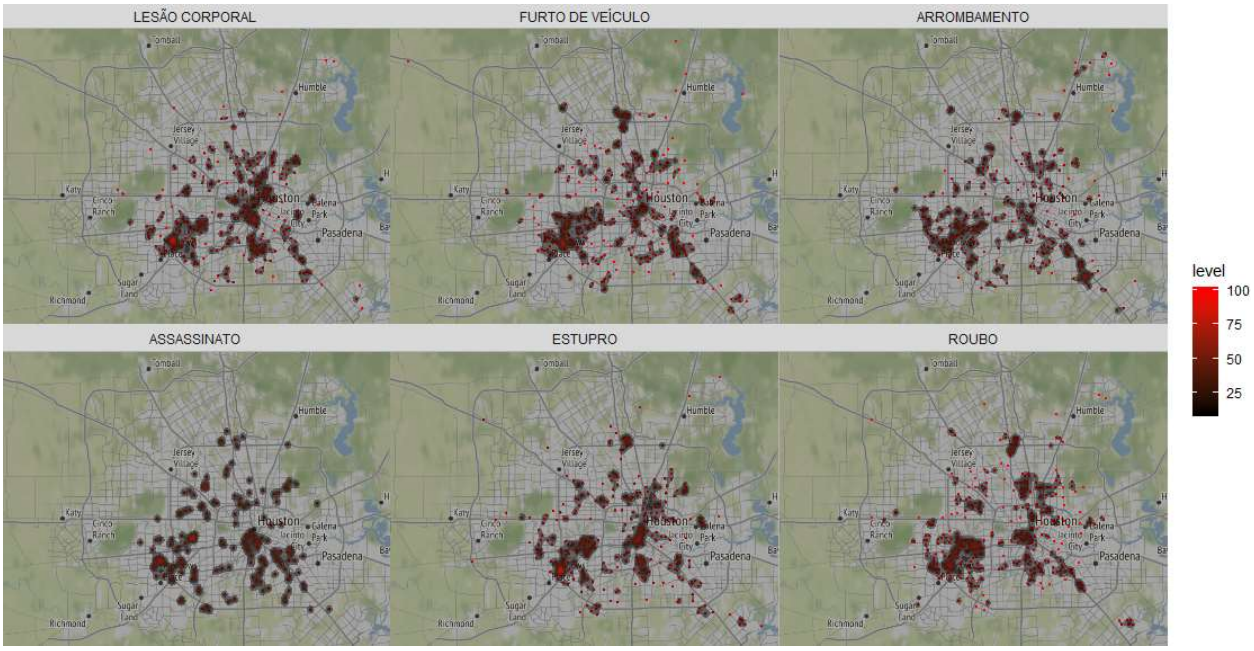
```
>Houstonkde.tau1 = density.ppp(x = Houstonppp,
                               sigma = 0.003,
                               kernel="gaussian")
>Houstonkde.tau2 = density.ppp(x = Houstonppp,
                               sigma = 0.01,
                               kernel="gaussian")
>Houstonkde.tau3 = density.ppp(x = Houstonppp,
                               sigma = 0.05,
                               kernel="gaussian")

>par(mfrow=c(2,2))
>par(mar=c(0.5,0.5,1.5,0.5))
>plot(Houstonppp,   pch=21,   cex=0.3,   bg="blue",   main="Crimes",
cex.main=0.5)
>plot(Houstonkde.tau1, main="sigma = 0.003", cex.main=0.5)
>plot(Houstonkde.tau2, main="sigma = 0.01")
>plot(Houstonkde.tau3, main="sigma = 0.05")
```



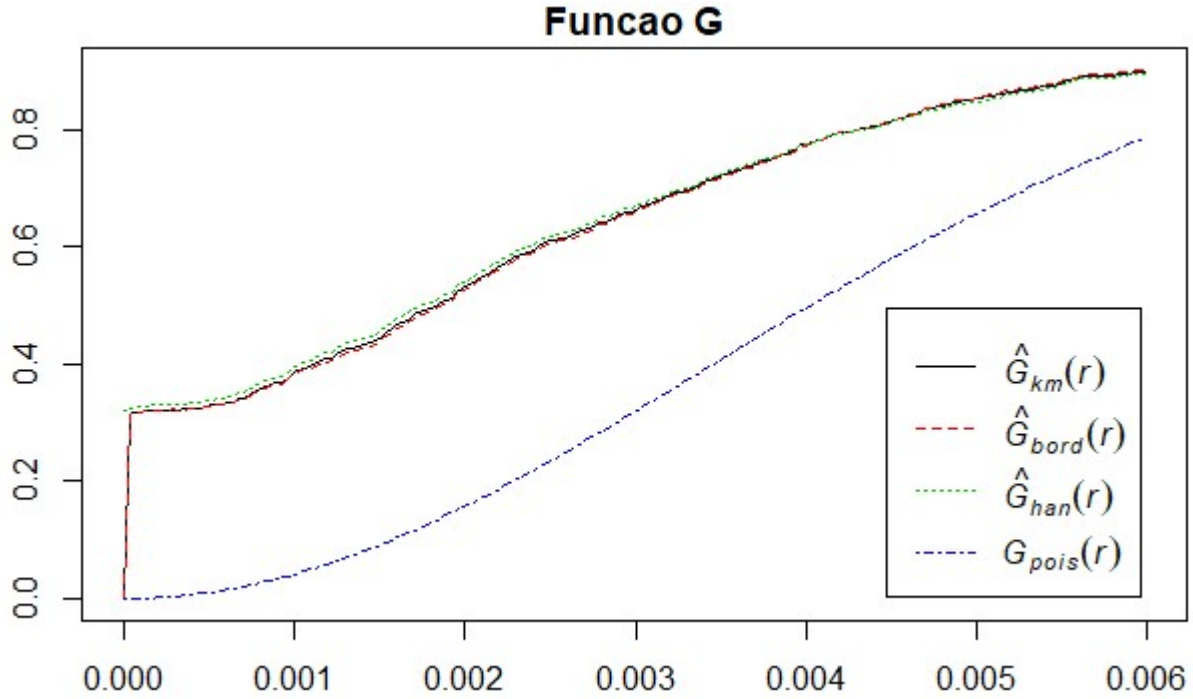
A imagem com raio 0.05 corrobora a conclusão visual original da existência de pelo menos dois grandes grupos associados às regiões próximas ao centro da cidade e de Meadows Place. Porém, a imagem com raio 0.01 parece ser a mais eficaz dadas as condições da visualização para identificar clusters de ocorrências com melhor resolução.

Uma forma alternativa de estimar a densidade é utilizando a função ggmap e fatiando os dados por tipos de crime:

```
>density_ggmap <- qmplot(x = lon,
                         y = lat,
                         data = crimeHouston,
                         colour = I('red'),
                         size = I(0.3),
                         darken = 0.3) +
  stat_density2d(data = crimeHouston,
                 aes(x = lon, y = lat, fill = ..level..),
                 alpha = 0.4,
                 h = 0.025,
                 n = 400,
                 geom = "polygon") +
  scale_fill_gradient(low = "black",
                      high= "red")
>density_ggmap +
  facet_wrap(~offense,
             labeller = labeller(offense = c("aggravated assault"="LESÃO CORPORAL",
                                             "rape" = "ESTUPRO",
                                             "robbery" = "ROUBO",
                                             "burglary" = "ARROMBAMENTO",
                                             "murder" = "ASSASSINATO",
                                             "auto theft" = "FURTO DE VEÍCULO")))
```

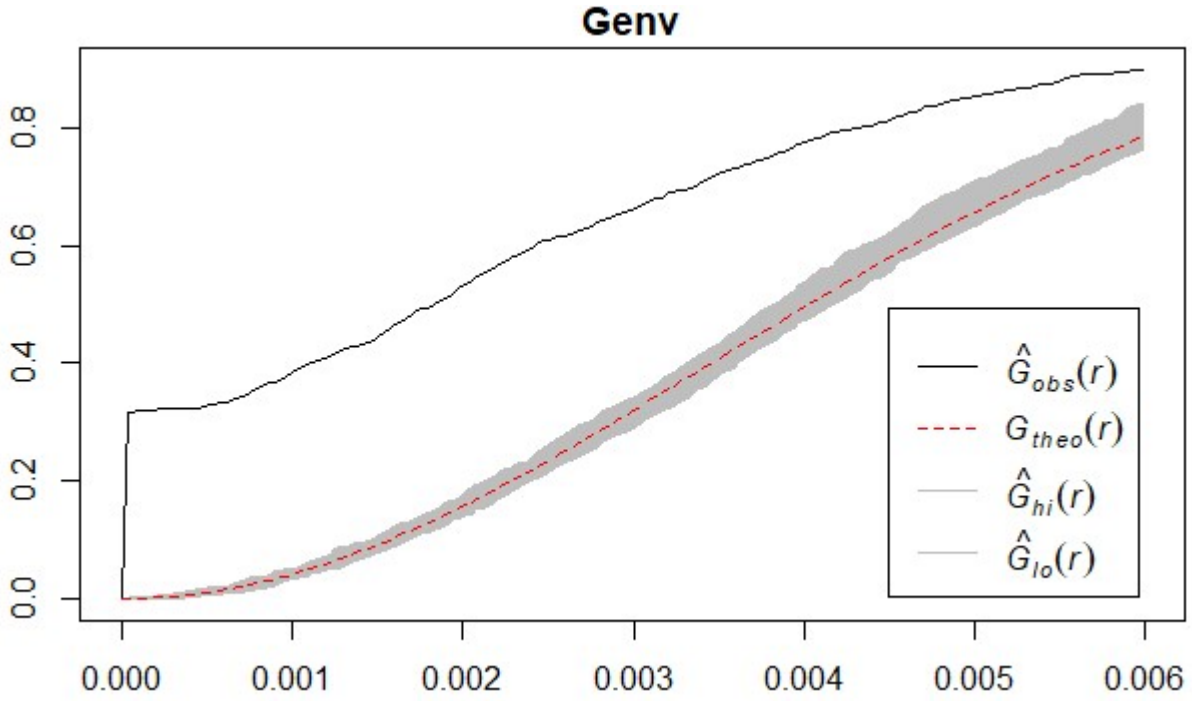


Considerando efeitos de segunda ordem, como uma possível dependência espacial com pontos vizinhos, uma maneira de tirar conclusões estatísticas sobre o agrupamento ou a regularidade das ocorrências é analisando a função G (distribuição de probabilidade acumulada em relação às distâncias).



A curva em azul representa uma função G teórica de pontos totalmente aleatórios. A distância mediana do vizinho mais próximo (d, G(d) = 0.5) nesta curva seria de aproximadamente 0.004. Como a função G real é representada por uma curva acima da curva teórica, pode-se intuir que existe um padrão de agrupamento (distância menores são mais prováveis de acontecer que distâncias maiores).

Mesmo com possível variações da curva teórica estimada, criando assim um intervalo de confiança (em cinza, abaixo), observa-se a mesma conclusão anterior.



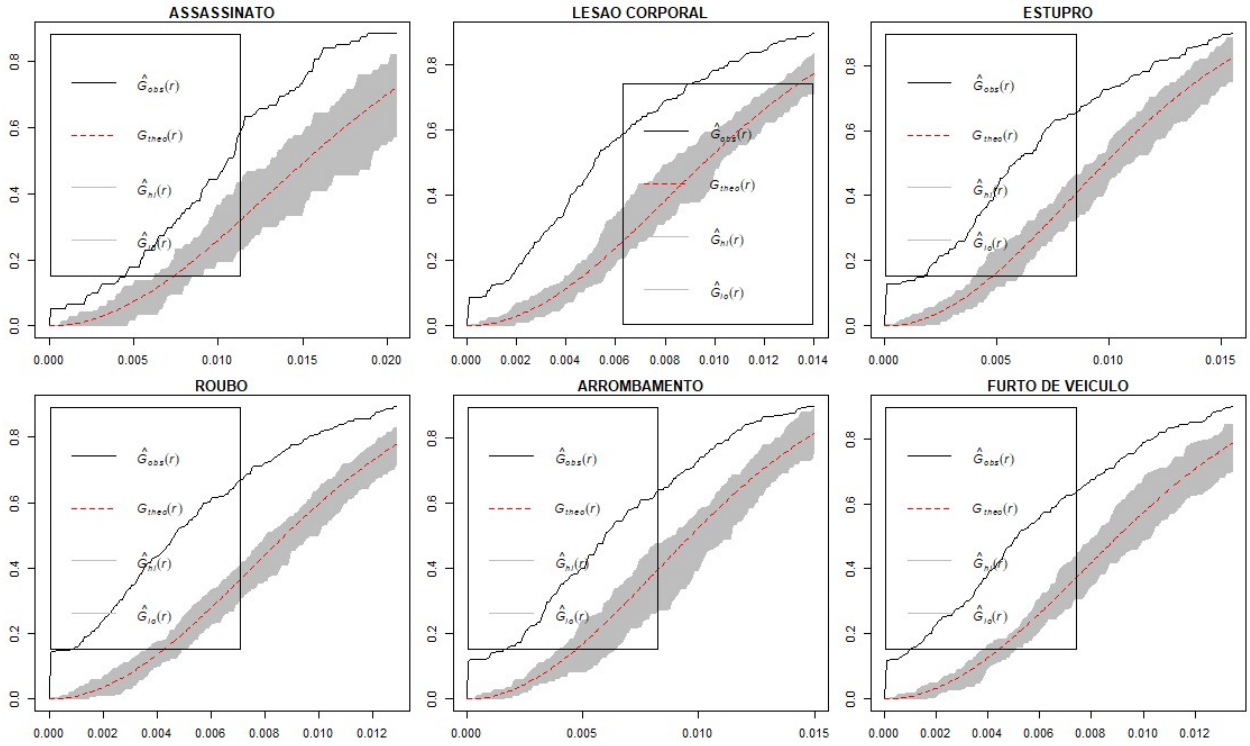
Para realizar a conclusão formal, pode-se executar o teste de aleatoriedade completa de Clark-Evans. Ele considera a aleatoriedade espacial completa como a hipótese nula (H0) e conclui (com p-value próximo a 0) o agrupamento nos dados (“Clustered R<1):

```
> clarkevans.test(Houstonppp, alternative = "less")

     Clark-Evans test
     No edge correction
     Z-test

data:  Houstonppp
R = 0.63085, p-value < 2.2e-16
alternative hypothesis: clustered (R < 1)
```

Tal conclusão considera todos os pontos observados, ou seja, todas as ocorrências de crimes independentemente de seu tipo. Portanto, vale a pena dividir o conjunto de dados por tipo de delito e realizar a mesma análise separadamente.

```
#### Dividindo o conjunto de dados por tipo de crime

#ASSASSINATOS
>murderHouston <- subset(crimeHouston,offense=="murder")
>murderppp = ppp( x = murderHouston$lon,
                  y = murderHouston$lat,
                  window = HoustonOWin)
>Genv_murder <- envelope(murderppp, fun = Gest, nsim = 20)

#LESAO CORPORAL
>assaultHouston <- subset(crimeHouston,offense=="aggravated assault")
>assaultppp = ppp(x = assaultHouston$lon,
                  y = assaultHouston$lat,
                  window = HoustonOWin)
>Genv_assault <- envelope(assaultppp, fun = Gest, nsim = 20)

#ESTUPRO
>rapeHouston <- subset(crimeHouston,offense=="rape")
>rapeppp <- ppp(x = rapeHouston$lon,
                  y = rapeHouston$lat,
                  window = HoustonOWin)
>Genv_rape <- envelope(rapeppp, fun = Gest, nsim = 20)

#ROUBO
>robberyHouston <- subset(crimeHouston,offense=="robbery")
>robberyppp <- ppp(x = robberyHouston$lon,
              y = robberyHouston$lat,
              window = HoustonOWin)
>Genv_robbery <- envelope(robberyppp, fun = Gest, nsim = 20)

#ARROMBAMENTO
>burglaryHouston <- subset(crimeHouston,offense=="burglary")
>burglaryppp <- ppp(x = burglaryHouston$lon,
                  y = burglaryHouston$lat,
                  window = HoustonOWin)
>Genv_burglary <- envelope(burglaryppp, fun = Gest, nsim = 20)

#FURTO DE VEICULO
>autotheftHouston <- subset(crimeHouston,offense=="auto theft")
>autotheftppp <- ppp(x = autotheftHouston$lon,
                  y = autotheftHouston$lat,
                  window = HoustonOWin)
>Genv_autotheft <- envelope(autotheftppp, fun = Gest, nsim = 20)

#Comparando as curvas G
>par(mfrow=c(2,3))
```

```
>plot(Genv_murder, main="ASSASSINATO")
>plot(Genv_assault, main="LESAO CORPORAL")
>plot(Genv_rape, main="ESTUPRO")
>plot(Genv_robbery, main="ROUBO")
>plot(Genv_burglary, main="ARROMBAMENTO")
>plot(Genv_autotheft, main="FURTO DE VEICULO")
```



Para cada um dos tipos de crime observados, a curva G se apresenta acima da curva de aleatoriedade espacial completa e seu intervalo de confiança. Logo, intui-se que há tendência de agrupamento em cada um dos seis subgrupos. Tal conclusão é corroborada pelos testes formais de Clark-Evans.

```
> clarkevans.test(murderppp, alternative = "less")

      Clark-Evans test
      No edge correction
      Z-test

data:  murderppp
R = 0.78624, p-value = 1.778e-07
alternative hypothesis: clustered (R < 1)

> clarkevans.test(assaultppp, alternative = "less")

      Clark-Evans test
      No edge correction
```

```
      Z-test

data:  assaultppp
R = 0.71394, p-value < 2.2e-16
alternative hypothesis: clustered (R < 1)

> clarkevans.test(rapeppp, alternative = "less")

      Clark-Evans test
      No edge correction
      Z-test

data:  rapeppp
R = 0.81568, p-value = 6.287e-12
alternative hypothesis: clustered (R < 1)

> clarkevans.test(robberyppp, alternative = "less")

      Clark-Evans test
      No edge correction
      Z-test

data:  robberyppp
R = 0.66876, p-value < 2.2e-16
alternative hypothesis: clustered (R < 1)

> clarkevans.test(burglaryppp, alternative = "less")

      Clark-Evans test
      No edge correction
      Z-test

data:  burglaryppp
R = 0.76614, p-value < 2.2e-16
alternative hypothesis: clustered (R < 1)

> clarkevans.test(autotheftppp, alternative = "less")

      Clark-Evans test
      No edge correction
      Z-test

data:  autotheftppp
R = 0.74985, p-value < 2.2e-16
alternative hypothesis: clustered (R < 1)
```

Com p-value próximo a zero nos seis diferentes testes, conclui-se que o comportamento das ocorrências dos crimes ocorre de forma similar para todos os tipos classificados na base de dados de Houston.