IV SEMINÁRIO INTERNACIONAL DE ESTATÍSTICA COM R
R & PYTHON E AS TENDÊNCIAS DE COLABORAÇÃO
NITERÓI, 21 A 23 DE MAIO DE 2019

QSPATIAL: DESENVOLVIMENTO DE UM PACOTE DE ESTATÍSTICA ESPACIAL PARA O R

Ricardo Junqueira de Souza[1], Jony Arrais Pinto Junior[2]

Introdução

A Estatística Espacial é um conjunto de métodos de análise nos quais a localização espacial dos dados é utilizada de forma explicita. Os dados espaciais podem ser classificados de três formas: de área, pontuais ou de superfície contínua, cada qual dotado de características específicas e abordados por metodologias de análise distintas.

Os dados de padrão de pontos são definidos como um conjunto de localizações distribuídas em uma região definida cuja ocorrência é o resultado de um mecanismo estocástico (DIGGLE, 2014). Este tipo de dado tem sua localização espacial exata conhecida por meio de coordenadas (latitude e longitude, por exemplo), sendo utilizado na análise da distribuição espacial de eventos de interesse.

Nem sempre é possível se obter a localização exata dos eventos, devido aos requerimentos de confidencialidade e privacidade de informações. Deste modo, a divulgação de dados espaciais tem se tornado cada vez mais comum sob a forma de contagens agregadas em unidades de área (OYANA, 2016). Este tipo de dado é denominado de área, nele as contagens podem ser divididas em sub-regiões da área de interesse de acordo com a dimensão do problema. Os exemplos mais comuns deste tipo de dado são contagens de crimes ou de casos de uma doença em subdivisões de uma região.

Quando pesquisadores analisam dados como os citados anteriormente, naturalmente surge o questionamento sobre a existência ou não de um padrão espacial no fenômeno estudado. Para responder tal indagação, mapas e análises para quantificar a existência de dependência espacial são as ferramentas mais utilizadas neste contexto.

Análises para estes dois tipos de dados que possam corroborar a hipótese de existência de padrão espacial podem ser realizadas por meio do software R (R Core Team, 2018). Para dados pontuais existe o pacote *spatstat*, que produz análises estatísticas completas, com visualização, análise exploratória e modelagem (BADDELEY; RUBAK;

[1] Universidade Federal Fluminense (UFF), rjunqueira@id.uff.br

[2] Universidade Federal Fluminense (UFF), jarrais@id.uff.br

TURNER, 2016). Já para dados de área está disponível o pacote *spdep*, que permite sua análise por meio de testes globais e locais e modelagem (BIVAND; PEBESMA; GÓMEZ, 2013).

Este trabalho propõe a criação do *qspatial*: um pacote que integre as capacidades de análise do *spdep* e do *spatstat* com as ferramentas de visualização do *ggplot2* e que simplifique as escolhas envolvidas de modo a tornar estas análises mais acessíveis (WICKHAM, 2016).

Para demonstrar as funcionalidades do pacote proposto, será considerado um conjunto de dados sobre roubos na cidade de São Paulo. Os dados foram tratados e analisados como padrão de pontos e então agregados nos distritos da cidade e analisados sob a ótica de dados de área.

Objetivos

O objetivo deste trabalho é o desenvolvimento de um novo pacote para o R que facilite a visualização e análise exploratória de dados espaciais por indivíduos que não possuam tanto conhecimento com os diversos pacotes existentes para o tratamento destes tipos de dados no R. Bem como, o pacote tenta produzir os gráficos com melhores layouts que os pacotes já existentes, uma vez que utiliza as ferramentas do *ggplot2*.

O pacote trará funções específicas para tratar tanto de dados de padrão de pontos quanto dados de área, simplificando as escolhas envolvidas e gerando todas as visualizações com o *ggplot2*. Estas funções foram pensadas de forma que necessitem apenas do *shapefile* e de um vetor com as contagens, no caso de dados de área, ou de dois vetores com as coordenadas para dados pontuais.

De uma maneira geral, as funções têm como padrão a utilização das metodologias mais comuns na literatura de dados espaciais, mas também permite que usuários mais avançados modifiquem as análises e os mapas de acordo com sua vontade (BADDELEY, A; RUBAK., E; TURNER, R., 2015; BIVAND R.; PEBESMA, E.; GÓMEZ, V., 2013; DIGGLE, 2016). O enfoque do pacote é visualização do espaço e do fenômeno de interesse, bem como, indícios de existência ou não de um padrão espacial.

Material e Método

O primeiro passo deste trabalho foi a realização de uma revisão da literatura sobre estatística espacial, concentrando-se em dois tipos de dados: padrão de pontos e de área.

Durante esta revisão foram elaborados materiais contendo exemplos e a metodologia para realizar uma análise exploratória dentro dos dois contextos, apresentando uma visualização do espaço e do fenômeno e apresentando medidas que ajudam a indicar a existência ou não de um padrão espacial no fenômeno estudado.

Durante esta etapa percebeu-se a necessidade de um número de passos adicionais para criar as visualizações dos mapas utilizando o pacote *ggplot2*. Enquanto dados de padrão de pontos eram suportados pela função *geom_point*, dados de área necessitavam da integração das variáveis de interesse no *shapefile* e uma transformação em *dataframe* para permitir o uso da variável como *fill* do mapa.

Em relação a análise espacial de ambos os tipos de dado notou-se que existem uma série de escolhas a serem feitas nas etapas preliminares. Para se analisar dados de área deve-se primeiro criar uma estrutura de vizinhança e uma matriz de pesos, sendo que existem diversas opções acessíveis por funções do pacote *spdep*. As escolhas feitas nesta etapa têm impacto direto no resultado das análises (BIVAND, PEBESMA, GÓMEZ, 2013).

As funções do *qspatial* foram criadas com isto em mente, para simplificar as escolhas envolvidas as opções utilizadas como *default* das funções foram as mais comumente vistas na literatura. Como o foco atual do pacote está em dados pontuais e de área, foram criadas duas funções principais para fazer uma visualização e análise recebendo apenas o *shapefile* da região de interesse e as coordenadas/variável de interesse.

Para dados pontuais foi criada a função *qmpattern*, que a partir dos dados recebidos ela gera um mapa de padrão de pontos, um mapa de densidade e dois gráficos com análises de completa aleatoriedade espacial. Existem argumentos na função para modificar o raio utilizado no mapa de densidade, o tamanho dos pontos e a paleta de cor utilizada.

Para dados de área foi criada a função *lmoranmap*, ela recebe apenas o *shapefile* e a variável de interesse para gerar um mapa coroplético, um mapa com os resultados do teste de moran local, um com as regiões cujo resultado do teste tiveram p-valor significante e um com as categorias que cada região pertenceria na função *moran.plot* do pacote *spdep*. Para a escala de cor foi utilizada a função *scale_fill_viridis_c* do *ggplot2*, que gera escalas distinguíveis até mesmo em preto e branco (WICKHAM, 2017).

O *qspatial* foi desenvolvido no *R Studio* com utilização do pacote *devtools*, também foi criado um repositório no *github* para controle de versões e disponibilização para a comunidade (WICKHAM, 2015). Para demonstrar seu funcionamento foi utilizado um exemplo de dados de roubos na cidade de São Paulo (SSP, 2016).

Os dados possuíam 17.669 observações, com as informações de latitude e longitude de cada evento. Para analisar sob o ponto de vista de padrão de pontos foi utilizada a

função *qmpattern*, que gerou a visualização e realizou os testes de completa aleatoriedade espacial. O passo seguinte foi agregar os casos de acordo com os distritos da cidade e então analisa-los através da função *lmoranmap*. Utilizando o *spdep* seriam necessárias as funções: *poly2nb* para criar as vizinhanças, *nb2listw* para criar a matriz de pesos e *localmoran* para a análise espacial. Os mapas seriam criados com a função *spplot* e teriam o layout básico de gráficos do R, que é visualmente inferior ao do *ggplot2*.

Resultados e Discussão

O *qspatial* propõe a simplificação do acesso a visualizações e análises espaciais para dados pontuais e de áreas, integrando as análises feitas pelos pacotes *spatstat* e *spdep* com as ferramentas gráficas do *ggplot2*. De forma geral, outros pacotes de estatística espacial necessitam de uma série de passos preliminares que podem ser pouco amigáveis para usuários sem intimidade com o R ou com a estatística espacial.

As funções criadas geram todos os resultados automaticamente, realizando os preparativos para uso do *ggplot2* internamente e retornando os mapas prontos. Para usuários mais avançados também foram adicionados argumentos nas funções que permitem maior manipulação dos mapas criados. Para estes usuários também foi criada a função *plotgg*, que a partir de um *shapefile* e uma variável de interesse retorna a sintaxe pronta para criação do mapa com o *ggplot2*.

A facilidade de uso do *qspatial* foi demonstrada com dados de crimes ocorridos no estado de São Paulo que foram tratados sob a ótica de padrão de pontos e de dados de área. As funções utilizadas precisaram apenas de um *shapefile* e uma variável de interesse/coordenadas para gerar uma análise estatística completa. Na documentação destas funções foram adicionadas instruções quanto a forma de interpretação e também na maneira que a alteração dos argumentos influenciam nos resultados.

Conclusão

O pacote buscou simplificar o processo de visualização e análise de dados espaciais. A aplicação da versão atual do pacote a um conjunto de dados que poderia ser analisado tanto sob a forma de padrão de pontos quanto agregado para dados de área se demonstrou promissora. Ele está disponível no endereço: https://github.com/qspatialR/qspatial e pode ser instalado no R através do comando *devtools::install_github("qspatialR/qspatial")*.

As funções criadas para o pacote têm como *default* as opções avaliadas como as de melhor custo benefício, elas garantem resultados de fácil interpretação e com baixo custo

computacional. Para usuários mais íntimos do R o pacote também traz mais opções de customização das funções, sendo possível, por exemplo, escolher quais funções serão utilizadas para a análise das propriedades de segunda ordem dos dados pontuais.

Este trabalho foi feito como parte de um projeto de iniciação a pesquisa cujo objetivo é criar um pacote do R para realizar análises exploratórias dos diferentes tipos de dados espaciais. O pacote ainda está em desenvolvimento, serão consideradas novas funcionalidades e a expansão para visualização e análise de dados de superfície contínua.

Referências

BADDELEY, A.; RUBAK, E.; TURNER, R. Spatial Point Patterns Methodology and Applications with R. Londres: Chapman and Hall/CRC Press, 2015.

BIVAND, R. S.; PEBESMA, E.; RUBIO-GÓMEZ, V. Applied Spatial Data Analysis with R. 2. Ed. Springer, 2013.

DIGGLE, P. J. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. CRC Press, 3a. ed, 2016.

OYANA, T. J.; MARGAI, F. M. Spatial Analysis: Statistics, Visualization, and Computational Methods. CRC Press, 2016.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria., 2018.

SSP. Dados estatísticos do estado de São Paulo. Disponível em: https://www.ssp.sp.gov.br/Estatistica/Mapas.aspx. Acesso em 25 de fev. 2019

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

WICKHAM, H. R packages: Organize, test, document and share your code. O'Reilly Media, 2015.

WICKHAM, H.; HESTER, J.; CHANG, W. devtools: Tools to make developing R packages easier. R package version 2.0.1, 2018. https://cran.r-project.org/package=devtools