The Care and Feeding of Spatial Data

Angela Li
November 9, 2019
DC R Conference

About me

- Research Specialist / R Spatial Advocate at Center for Spatial Data Science...

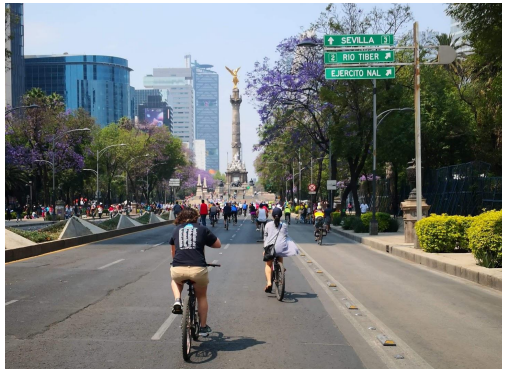
THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

About me

- Research Specialist / R Spatial Advocate at Center for Spatial Data Science...
- R-Ladies Chicago co-organizer…

About me

- Research Specialist / R Spatial Advocate at Center for Spatial Data Science...
- R-Ladies Chicago co-organizer…
- Carpentries instructor...



About me

- Research Specialist / R Spatial Advocate at Center for Spatial Data Science...
- R-Ladies Chicago co-organizer…
- Carpentries instructor...
- and a few other things

Previously...

bit.ly/dc-spatial

ANGELA LI, CENTER FOR SPATIAL DATA SCIENCE

Data Science? Make It Spatial

Back for more!

bit.ly/dc-spatial-2019

How do I get started with spatial data analysis in R?

1. Where do I start?
2. What tools should I use?
3. What can I do?

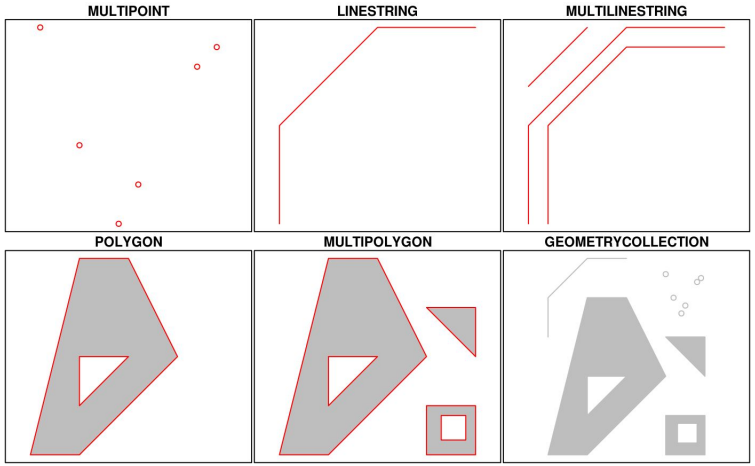
1. Where do I start?

If you
don't love
me at my

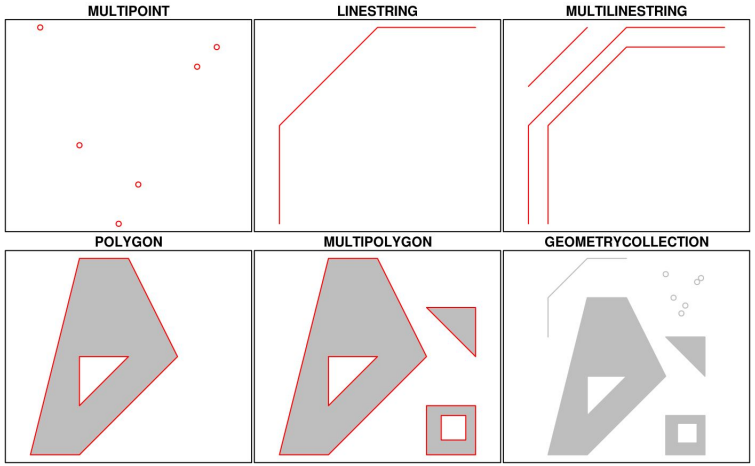

```
{ "type": "FeatureCollection", "features": [{"type":"Feature","geometry":{"type":'
[[[[-73.94377595453821,40.80593168515713],[-73.94369707970962,40.806041543498054],
[-73.94324070094328,40.805850574310156],[-73.94332133188101,40.80573739970471],
[-73.94350012655029,40.80581380922321],[-73.9435020149869,40.80581592361876],
[-73.94350430774979,40.805817793276795],[-73.9435069503635,40.805819372245224],
[-73.94350988241784,40.80582062537496],[-73.94351303164855,40.80582152111309],
[-73.9435163245971,40.8058220387122],[-73.94351968305696,40.805822166427475],
[-73.94377595453821,40.80593168515713]]]]},"properties":{"propname":"Marcus Garvey
Park","us_congres":"13","zipcode":"10027","acres":"0.15435469","location":"124 St
West","gispropnum":"M058","retired":"False","subcategor":"Plgd Within Park","commu
11","precinct":"25","retireddat":null,"omppropid":"M058-07","nys_assemb":"68","sit
Memorial Park","nys_senate":"30","councildis":"9","borough":"M","descriptio":"Marc
Marcus Garvey Memorial Park test"}},{"type":"Feature","geometry":{"type":"MultiPol
[[[[-73.94195823514032,40.80522881326439],[-73.94199537721298,40.80517821943407],
[-73.94199542815677,40.805178236569134],[-73.94228751011488,40.8053031229313],
[-73.94230489011392,40.80531055354029],[-73.94247679131614,40.80538405169119],
[-73.94243293580911,40.80544577338553],[-73.94242886243119,40.805451506626774],
[-73.94240918016563,40.80547920603749],[-73.94239805875212,40.8054948583502],
[-73.9421894362686,40.805406179079235],[-73.94217285586159,40.80539913066549],
[-73.94191398944002,40.805289085664285],[-73.94192490048417,40.805274222128524],
[-73.94195174856038,40.8052376492927],[-73.94195823514032,40.80522881326439]]]]},'
{"propname":"Marcus Garvey Park","us_congres":"13","zipcode":"10027","acres":"0.16
Madison & 124 St","gispropnum":"M058","retired":"False","subcategor":"Plgd Within
Park","communityb":"111","department":"M-11","precinct":"25","retireddat":null,"om
06","nys_assemb":"68","sitename":"Marcus Garvey Memorial
Park","nys_senate":"30","councildis":"9","borough":"M","descriptio":"Marcus Garvey
Memorial Park"}},{"type":"Feature","geometry":{"type":"MultiPolygon","coordinates'
[[[[-73.94297793363395,40.80314765719178],[-73.9430097746168,40.803124291483456],
[-73.94304521893065,40.803104143228225],[-73.94308370696312,40.803087529122124],
[-73.94311033615568,40.80307650034912],[-73.94313898407543,40.80306893408353],
[-73.9431688743432,40.803065036155395],[-73.94319920222738,40.80306490972454],
[-73.94322914767477,40.8030685597899],[-73.94325790257166,40.80307588689895],
[-73.94328468851249,40.80308669436097],[-73.94330878523981,40.80310068916088],
[-73.94332953892928,40.80311749366944],[-73.94334639180876,40.803136653764305],
[-73.94335888688718,40.80315765233478],[-73.9433666880889,40.80317992280391],
[-73.94336958261235,40.803202863534764],[-73.94336749513617,40.80322585494746],
[-73.94335969334733,40.80327529390631],[-73.94334466236197,40.80332376037477],
[-73.9433225899233,40.80337064390407],[-73.94329375263683,40.803415356600574],
```


then you don't
deserve
me at my

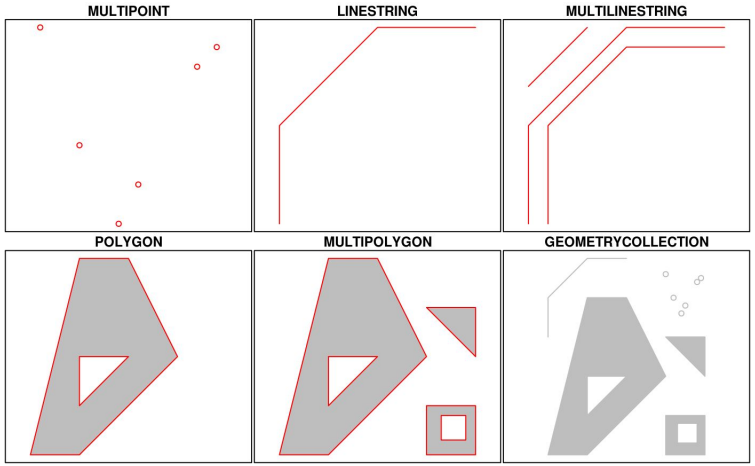
Vector data
Raster data



Points, lines, polygons
Grids, pixels, cells



Points, lines, polygons
Grids, pixels, cells



Not really: addresses, zip codes, counties

```
> (ers <- read_excel("data/Emergency-Departments.xlsx"))
```

```
> (ers <- read_excel("data/Emergency-Departments.xlsx"))
# A tibble: 65 x 3
   Hospital                                          Location        Region
   <chr>                                             <chr>            <dbl>
 1 Mercyhealth Javon Bea Hospital - Rockton          Rockford, IL         1
 2 Mercyhealth Javon Bea Hospital - Riverside        Rockford, IL         1
 3 OSF St. Anthony Medical Center                    Rockford, IL         1
 4 Swedish American Health System                    Rockford, IL         1
 5 Advocate BroMenn Medical Center                   Normal, IL           2
 6 Genesis Medical Center - Silvis                   Silvis, IL           2
 7 Unity Point Health - Methodist Medical Center     Peoria, IL           2
 8 OSF St. Francis Medical Center                    Peoria, IL           2
 9 OSF St. Joseph Medical Center                     Bloomington, IL      2
10 Unity Point Health – Trinity Medical Center       Rock Island, IL      2
# … with 55 more rows
```

```
> (ers <- read_excel("data/Emergency-Departments.xlsx"))
# A tibble: 65 x 3
   Hospital                                          Location        Region
   <chr>                                             <chr>            <dbl>
 1 Mercyhealth Javon Bea Hospital - Rockton          Rockford, IL         1
 2 Mercyhealth Javon Bea Hospital - Riverside        Rockford, IL         1
 3 OSF St. Anthony Medical Center                    Rockford, IL         1
 4 Swedish American Health System                    Rockford, IL         1
 5 Advocate BroMenn Medical Center                   Normal, IL           2
 6 Genesis Medical Center - Silvis                   Silvis, IL           2
 7 Unity Point Health - Methodist Medical Center     Peoria, IL           2
 8 OSF St. Francis Medical Center                    Peoria, IL           2
 9 OSF St. Joseph Medical Center                     Bloomington, IL      2
10 Unity Point Health – Trinity Medical Center       Rock Island, IL      2
# … with 55 more rows
```

Geocode it!

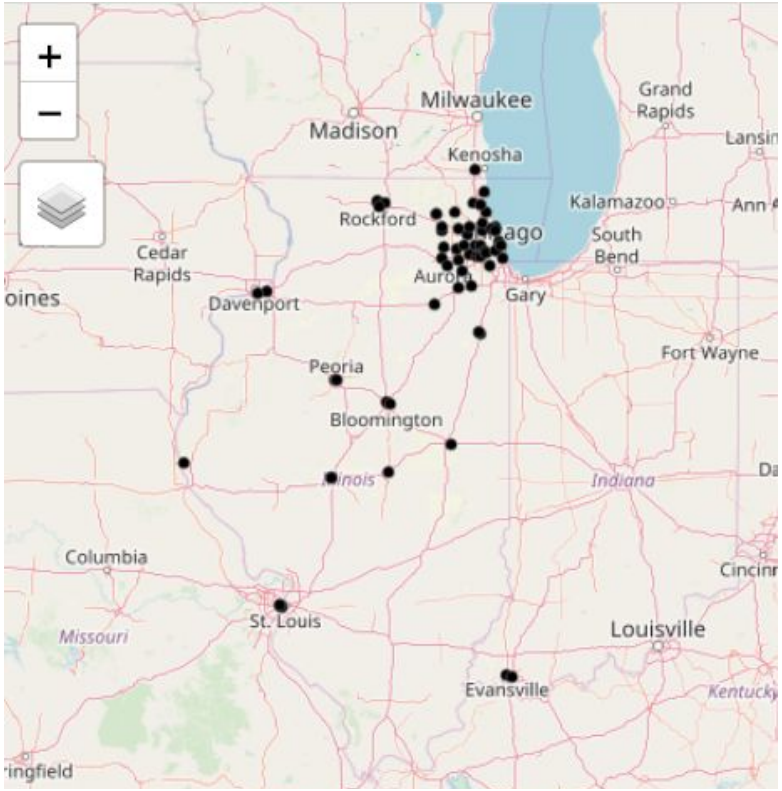
Geocode it!

```
> tmaptools::geocode_OSM("Georgetown University")
$query
[1] "Georgetown University"

$coords
        x         y
-77.07458  38.90894

$bbox
     xmin      ymin      xmax      ymax
-77.07926  38.90513 -77.06897  38.91268
```

```
> (ers_geocoded <- read_csv("data-output/geocoded/ers_to_geocode_1570813925_geocoded.csv"))
```

```
> (ers_geocoded <- read_csv("data-output/geocoded/ers_to_geocode_1570813925_geocoded.csv"))
# A tibble: 65 x 7
      ID ADDRESS                      CITY     STATE Longitude Latitude `Match Score`
   <dbl> <chr>                        <chr>    <chr>     <dbl>    <dbl>         <dbl>
 1     1 Mercyhealth Javon Bea Hos… Rockford IL         -89.1     42.3         100
 2     2 Mercyhealth Javon Bea Hos… Rockford IL         -89.1     42.3          88.2
 3     3 OSF St. Anthony Medical C… Rockford IL         -89.0     42.3          96.2
 4     4 Swedish American Health S… Rockford IL         -89.1     42.2          79.1
 5     5 Advocate BroMenn Medical … Normal   IL         -89.0     40.5          95.7
 6     6 Genesis Medical Center - … Silvis   IL         -90.4     41.5         100
 7     7 Unity Point Health - Meth… Peoria   IL         -89.6     40.7          85.9
 8     8 OSF St. Francis Medical C… Peoria   IL         -89.6     40.7          96.2
 9     9 OSF St. Joseph Medical Ce… Bloomin… IL         -89.0     40.5          96.1
10    10 Unity Point Health – Trin… Rock Is… IL         -90.5     41.5          84.4
# … with 55 more rows
```

```
> (ers_geocoded <- read_csv("data-output/geocoded/ers_to_geocode_1570813925_geocoded.csv"))
# A tibble: 65 x 7
      ID ADDRESS                      CITY     STATE Longitude Latitude Match Score`
   <dbl> <chr>                        <chr>    <chr>     <dbl>    <dbl>        <dbl>
 1     1 Mercyhealth Javon Bea Hos… Rockford IL        -89.1     42.3         100
 2     2 Mercyhealth Javon Bea Hos… Rockford IL        -89.1     42.3          88.2
 3     3 OSF St. Anthony Medical C… Rockford IL        -89.0     42.3          96.2
 4     4 Swedish American Health S… Rockford IL        -89.1     42.2          79.1
 5     5 Advocate BroMenn Medical … Normal   IL        -89.0     40.5          95.7
 6     6 Genesis Medical Center - … Silvis   IL        -90.4     41.5         100
 7     7 Unity Point Health - Meth… Peoria   IL        -89.6     40.7          85.9
 8     8 OSF St. Francis Medical C… Peoria   IL        -89.6     40.7          96.2
 9     9 OSF St. Joseph Medical Ce… Bloomin… IL        -89.0     40.5          96.1
10    10 Unity Point Health – Trin… Rock Is… IL        -90.5     41.5          84.4
# … with 55 more rows
```


Milwaukee
Grand Rapids
Madison
Kenosha
Kalamazoo
Rockford
South Bend
Cedar Rapids
Aurora
Gary
Davenport
Fort Wayne
Peoria
Bloomington
Indiana
Columbia
St. Louis
Missouri
Louisville
Evansville

Spatial data formats are confusing

The most common spatial data formats

- **.shp** (shapefile… actually 4 files?)
- **.geojson** (lots of web-based mapping)
- **.gpkg** (new and good, but not common yet)
- **.csv** (collaboration with non-geo ppl)
- **.tiff** (when I deal with raster data)

And many more read by GDAL (Geographic Data Abstraction Library)

GDAL

Search docs

- Download
- Programs
- Raster drivers
- Vector drivers
- User
- API
- Tutorials
- Development
- Community
- How to contribute?
- FAQ
- License

GDAL documentation

Edit on GitHub

Next

GDAL

GDAL is a translator library for raster and vector geospatial data formats that is released under an X/MIT style Open Source License by the Open Source Geospatial Foundation. As a library, it presents a single raster abstract data model and single vector abstract data model to the calling application for all supported formats. It also comes with a variety of useful command line utilities for data translation and processing. The NEWS page describes the October 2019 GDAL/OGR 3.0.2 release.

OSGeo Project

See Software using GDAL

This documentation is also available as a PDF file

To use these file types in R

```
library(sf)
st_read("ers.shp")
```

To use these file types in R

```
library(sf)
st_read("ers.shp")

read_sf("ers.shp")
# if you prefer tibbles
```

Creates a tidy sf dataframe

```
Simple feature collection with 65 features and 5 fields
geometry type:  POINT
dimension:      XY
bbox:           xmin: -91.39832 ymin: 37.96421 xmax: -87.50401 ymax: 42.56531
epsg (SRID):    NA
proj4string:    NA
# A tibble: 65 x 6
      ID ADDRESS                        CITY     STATE `Match Score`              geometry
   <dbl> <chr>                          <chr>    <chr>         <dbl>               <POINT>
 1     1 Mercyhealth Javon Bea Hospita… Rockford IL            100     (-89.09792 42.29837)
 2     2 Mercyhealth Javon Bea Hospita… Rockford IL             88.2   (-89.09792 42.29837)
 3     3 OSF St. Anthony Medical Center Rockford IL             96.2   (-89.00814 42.26944)
 4     4 Swedish American Health System Rockford IL             79.1   (-89.08095 42.24607)
 5     5 Advocate BroMenn Medical Cent… Normal   IL             95.7   (-88.99096 40.49739)
 6     6 Genesis Medical Center - Silv… Silvis   IL            100     (-90.41868 41.49378)
 7     7 Unity Point Health - Methodis… Peoria   IL             85.9   (-89.59565 40.70003)
```

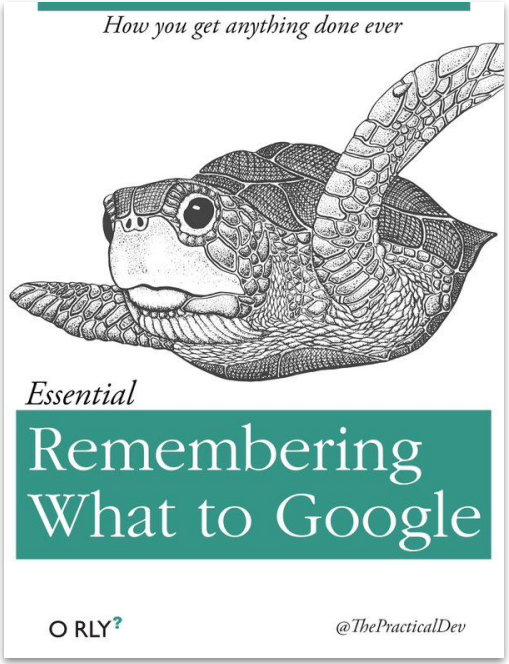
Creates a tidy sf dataframe

```
Simple feature collection with 65 features and 5 fields
geometry type:  POINT
dimension:      XY
bbox:           xmin: -91.39832 ymin: 37.96421 xmax: -87.50401 ymax: 42.56531
epsg (SRID):    NA
proj4string:    NA
# A tibble: 65 x 6
      ID ADDRESS                          CITY     STATE `Match Score`  geometry
   <dbl> <chr>                            <chr>    <chr>         <dbl>  <POINT>
 1     1 Mercyhealth Javon Bea Hospita… Rockford IL            100     (-89.09792 42.29837)
 2     2 Mercyhealth Javon Bea Hospita… Rockford IL             88.2   (-89.09792 42.29837)
 3     3 OSF St. Anthony Medical Center Rockford IL             96.2   (-89.00814 42.26944)
 4     4 Swedish American Health System Rockford IL             79.1   (-89.08095 42.24607)
 5     5 Advocate BroMenn Medical Cent… Normal   IL             95.7   (-88.99096 40.49739)
 6     6 Genesis Medical Center - Silv… Silvis   IL            100     (-90.41868 41.49378)
 7     7 Unity Point Health - Methodis… Peoria   IL             85.9   (-89.59565 40.70003)
```

How to Google Search for spatial data

"park district **boundaries**"
"food vendor **shapefile**"
"administrative district **polygons**"
"road network **geometries**"
"abandoned car **GIS file**"

"DC boundary shapefile geospatial .shp .geojson"



DC boundary shapefile geospatial .shp .geojson .kml

All | Images | News | Shopping | Videos | More | Settings | Tools

10 results (0.67 seconds)

Washington DC Boundary | Open Data DC
https://opendata.dc.gov › datasets
Jun 25, 2019 - The dataset is a polygon representing the District of Columbia **boundary**, created as part of the **DC** Geographic Information System (**DC GIS**) for ...
Missing: ~~.geojson~~

Boundaries - Open Data DC - Code for DC
data.codefordc.org › dataset › tags=Boundaries
DC Maps - **GIS** Datasets Source https://github.com/benbalter/**dc**-maps/ Original Source http://opendata.**dc**.gov/ Shipstead-Luce Act **Boundary GeoJSON** ...

SHP - Open Data DC - Code for DC
data.codefordc.org › dataset › res_format=SHP
SHP · **GeoJSON** ... District of Columbia Administrative **Boundaries Shapefile** (142 k) These ArcGIS **shapefiles** ... paper maps and later added to the **DC** Geographic Information System (**DC GIS**)... XML · dBase · prj · sbn · sbx · **SHP** · shx · **KML** ...

2. What tools should I use?

My toolkit

My toolkit

There are *many* R packages out there for spatial data

CRAN Task View: Analysis of Spatial Data

Maintainer: Roger Bivand
Contact: Roger.Bivand at nhh.no
Version: 2019-06-08
URL: https://CRAN.R-project.org/view=Spatial

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial data, where observations can be identified with geographical locations, and where additional information about these locations may be retrieved if the location is recorded with care.

Base R functions are complemented by contributed packages, some of which are on CRAN, and others are still in development. One location is Github . Some key packages including sf and stars are grouped under r-spatial , others including raster under rspatial . Maintenance of the sp is continuing here: sp.

Another set of locations for the development and maintenance of packages on R-Forge , which lists "Spatial Data and Statistics" projects in its project tree . Information on R-spatial packages was until 2016 posted on the R-Forge rspatial project website , including a visualisation gallery.

The contributed packages address two broad areas: moving spatial data into and out of R, and analysing spatial data in R.

The R-SIG-Geo mailing-list is a good place to begin for obtaining help and discussing questions about both accessing data, and analysing it. The mailing list is a good place to search for information about relevant courses. Further information about courses may be found under the "Events" tab of this blog .

There are a number of contributed tutorials and introductions; a recent one is Introduction to visualising spatial data in R by Robin Lovelace and James Cheshire.

The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please fork the task view repository and provide a pull request in ctv format for the ctv/Spatial.ctv file.

Classes for spatial data and metadata

Because many of the packages importing and using spatial data have had to include objects of storing data and functions for visualising it, an initiative is in progress to construct shared classes and plotting functions for spatial data.

Complementary initiatives are ongoing to support better handling of geographic metadata in R.

Spatial data - general

- The sp package provides classes and methods for dealing with spatial data and is discussed in a note in R News .
- sf is a newer package now on CRAN, and is being actively developed here: sf, providing Simple Features for R, in compliance with the OGC Simple Feature standard. The development of the package is being supported by the R Consortium . It provides simple features access for vector data, and as such is a modern implementation and standardization of parts of sp. It is documented in an R Journal article.

CRAN Task View: Analysis of Spatial Data

Maintainer: Roger Bivand
Contact: Roger.Bivand at nhh.no
Version: 2019-06-08
URL: https://CRAN.R-project.org/view=Spatial

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial data, where observations can be identified with geographical locations, and where additional information about these locations may be retrieved if the location is recorded with care.

Base R functions are complemented by contributed packages, some of which are on CRAN, and others are still in development. One location is Github . Some key packages including sf and stars are grouped under r-spatial , others including raster under rspatial . Maintenance of the sp is continuing here: sp.

Another set of locations for the development and maintenance of packages on R-Forge , which lists "Spatial Data and Statistics" projects in its project tree . Information on R-spatial packages was until 2016 posted on the R-Forge rspatial project website , including a visualisation gallery.

https://cran.r-project.org/web/views/Spatial.html

The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please fork the task view repository and provide a pull request in ctv format for the ctv/Spatial.ctv file.

Classes for spatial data and metadata

Because many of the packages importing and using spatial data have had to include objects of storing data and functions for visualising it, an initiative is in progress to construct shared classes and plotting functions for spatial data.

Complementary initiatives are ongoing to support better handling of geographic metadata in R.

Spatial data - general

- The sp package provides classes and methods for dealing with spatial data and is discussed in a note in R News .
- sf is a newer package now on CRAN, and is being actively developed here: sf, providing Simple Features for R, in compliance with the OGC Simple Feature standard. The development of the package is being supported by the R Consortium . It provides simple features access for vector data, and as such is a modern implementation and standardization of parts of sp. It is documented in an R Journal article.

CRAN Task View: Analysis of Spatial Data

Maintainer: Roger Bivand
Contact: Roger.Bivand at nhh.no
Version: 2019-06-08
URL: https://CRAN.R-project.org/view=Spatial

Base R includes many functions that can be used for reading, visualising, and analysing spatial data. The focus in this view is on "geographical" spatial data, where observations can be identified with geographical locations, and where additional information about these locations may be retrieved if the location is recorded with care.

Base R functions are complemented by contributed packages, some of which are on CRAN, and others are still in development. One location is Github . Some key packages including sf and stars are grouped under r-spatial , others including raster under rspatial . Maintenance of the sp is continuing here: sp.

Another set of locations for the development and maintenance of packages on R-Forge , which lists "Spatial Data and Statistics" projects in its project tree . Information on R-spatial packages was until 2016 posted on the R-Forge rspatial project website , including a visualisation gallery.

The contributed packages address two broad areas: moving spatial data into and out of R, and analysing spatial data in R.

The R-SIG-Geo mailing-list is a good place to begin for obtaining help and discussing questions about both accessing data, and analysing it. The mailing list is a good place to search for information about relevant courses. Further information about courses may be found under the "Events" tab of this blog .

There are a number of contributed tutorials and introductions; a recent one is Introduction to visualising spatial data in R by Robin Lovelace and James Cheshire.

The packages in this view can be roughly structured into the following topics. If you think that some package is missing from the list, please fork the task view repository and provide a pull request in ctv format for the ctv/Spatial.ctv file.

What I mainly use

Spatial package
sf
tmap
spdep
sp

What I mainly use

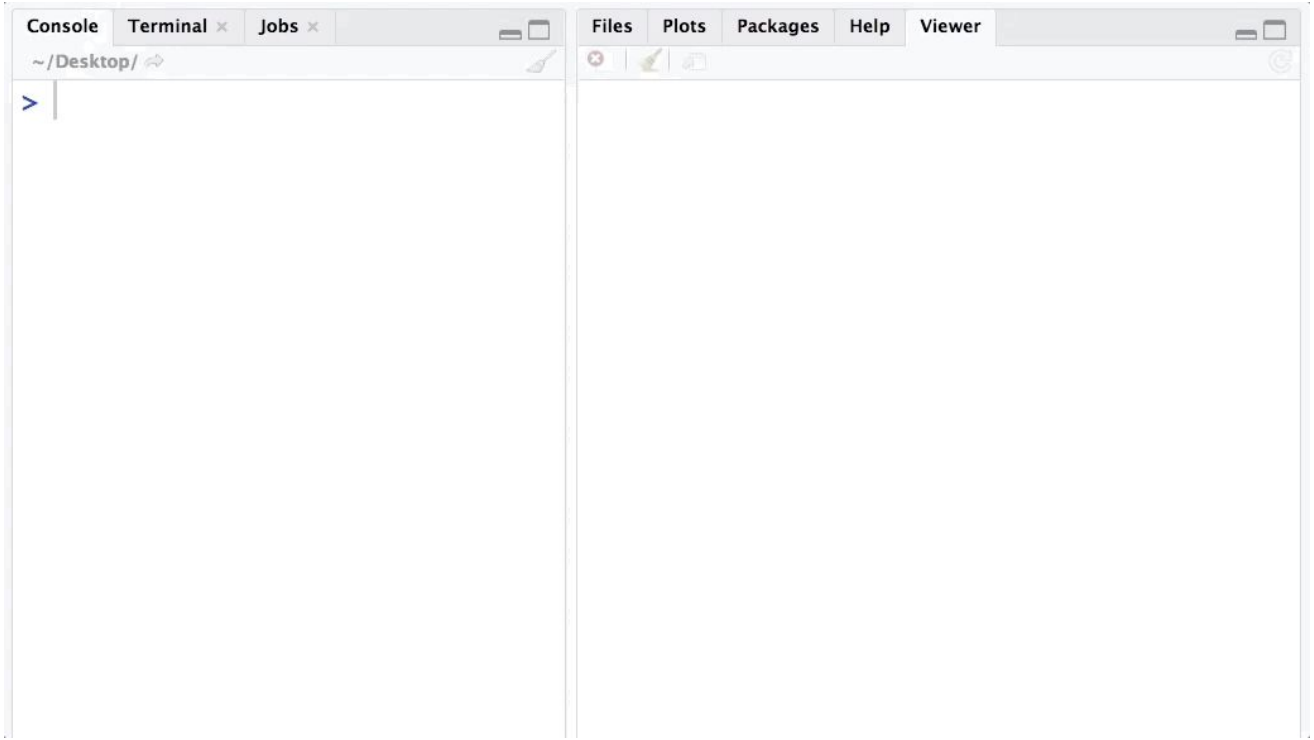
Spatial package	Similar R package
sf	readr, tidyr, dplyr
tmap	ggplot2
spdep	glmnet
sp	data.table

What I mainly use

Spatial package	Similar R package	Why analogous?
sf	readr, tidyr, dplyr	Tidy data ingestion, manipulation, cleaning
tmap	ggplot2	Custom layered visualization
spdep	glmnet	Modeling, not used all the time
sp	data.table	Different way of handling data, may be better for your use case

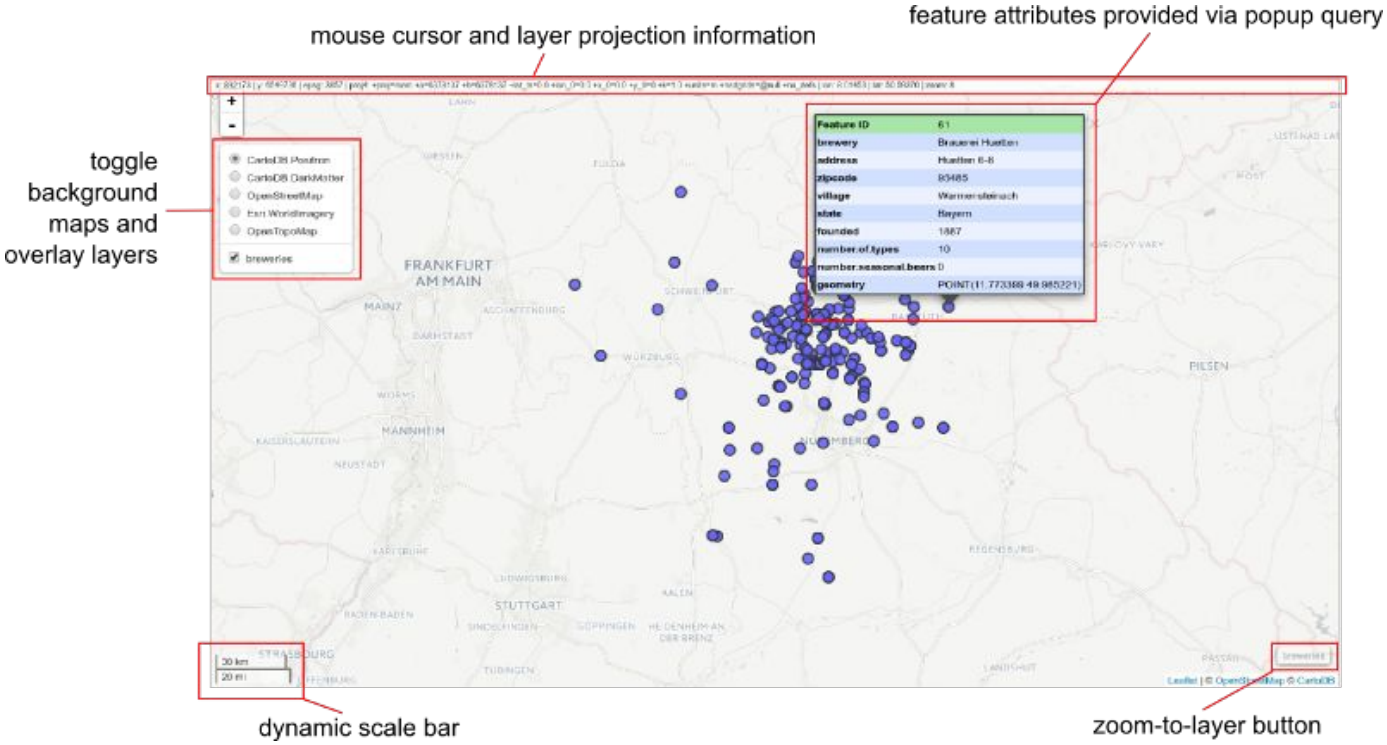
3. What can I do?

One nice thing about tmap

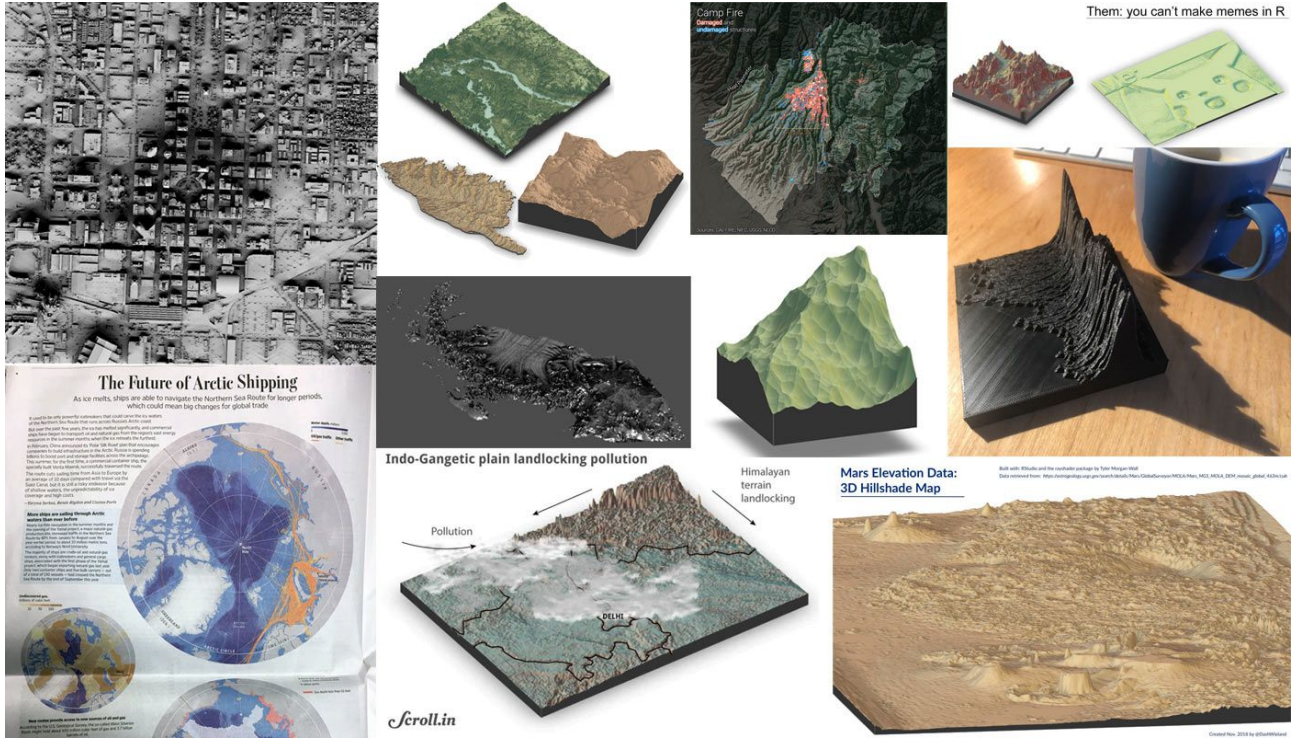


Alternate between interactive exploration and static mapping!

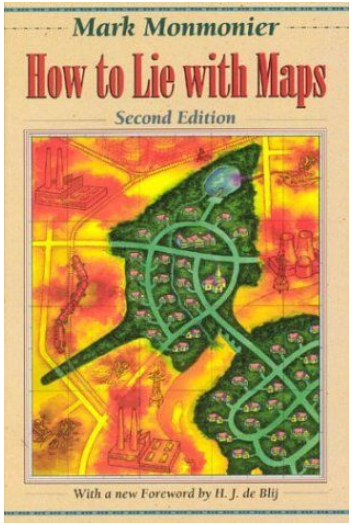
Check out mapview for a GIS-like experience



And rayshader for 3-D visualization and graphics



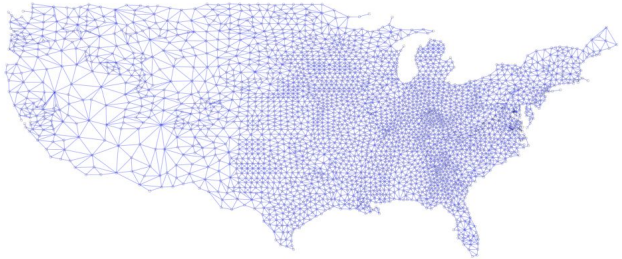
Think beyond “dots on a map”



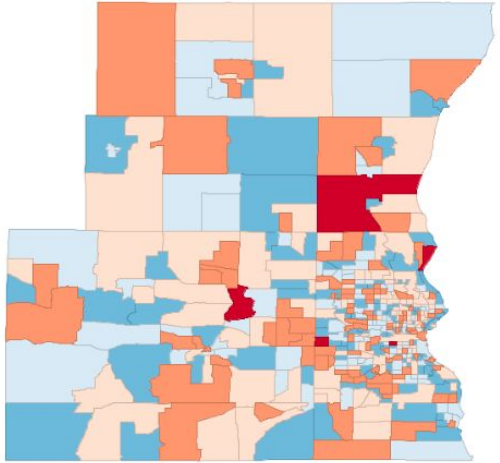
Why does **where** matter?

Tobler's First Law of Geography

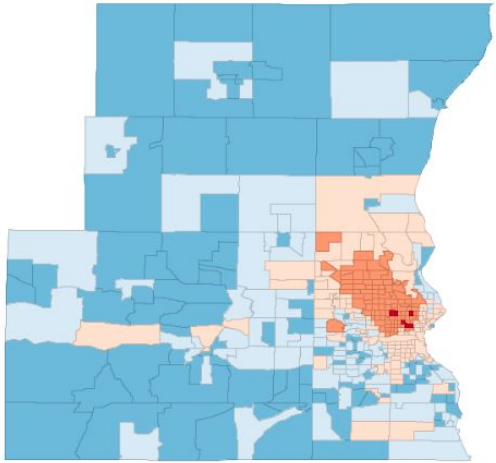
Everything is related to everything else, but near things are more related than distant things.



Spatial pattern detection

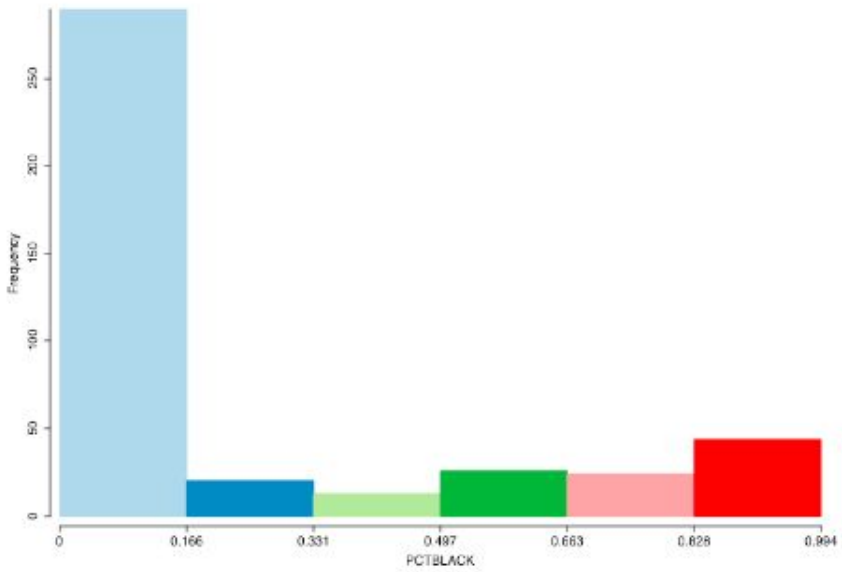


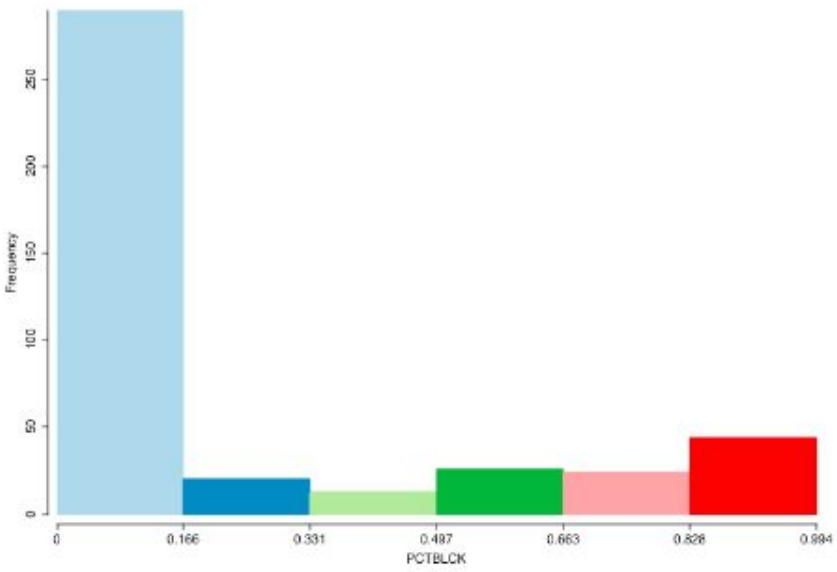
Hinge=1.5: PCTBLACK
- Lower outlier (0)
- < 25% (104)
- 25% - 50% (104)
- 50% - 75% (105)
- > 75% (99)
- Upper outlier (5)



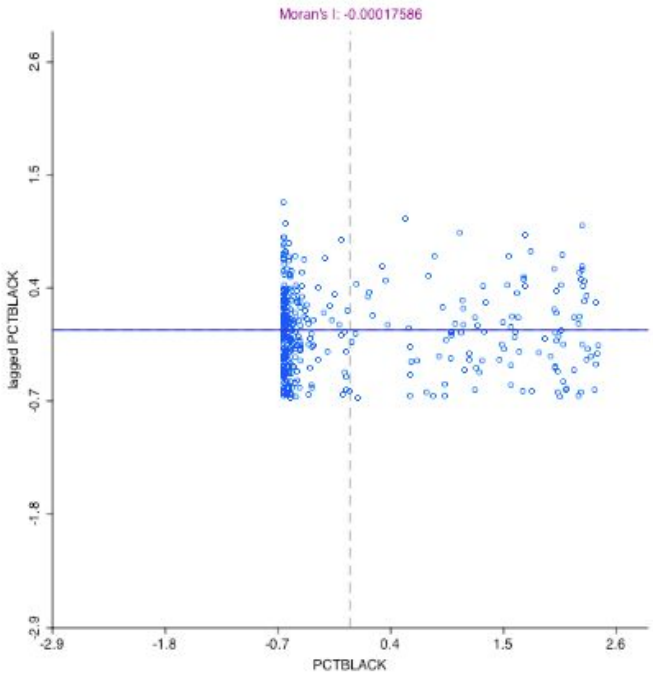
Hinge=1.5: PCTBLCK
- Lower outlier (0)
- < 25% (104)
- 25% - 50% (104)
- 50% - 75% (105)
- > 75% (99)
- Upper outlier (5)

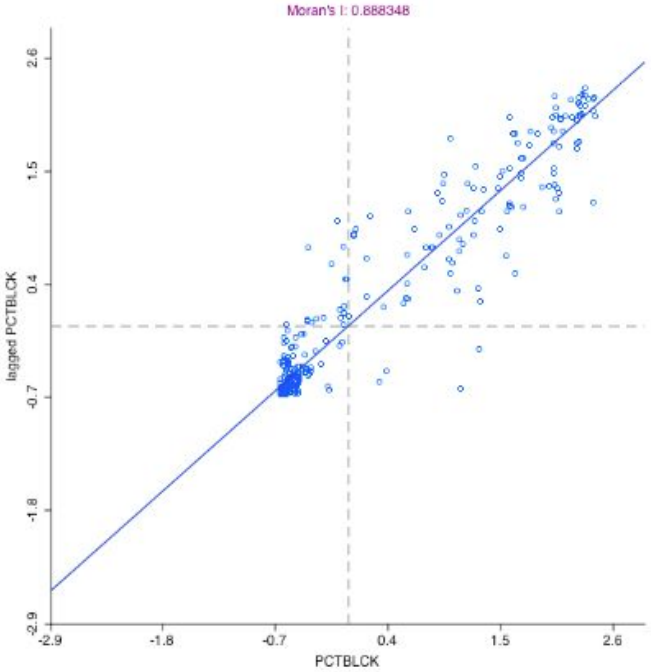
Spatial pattern detection



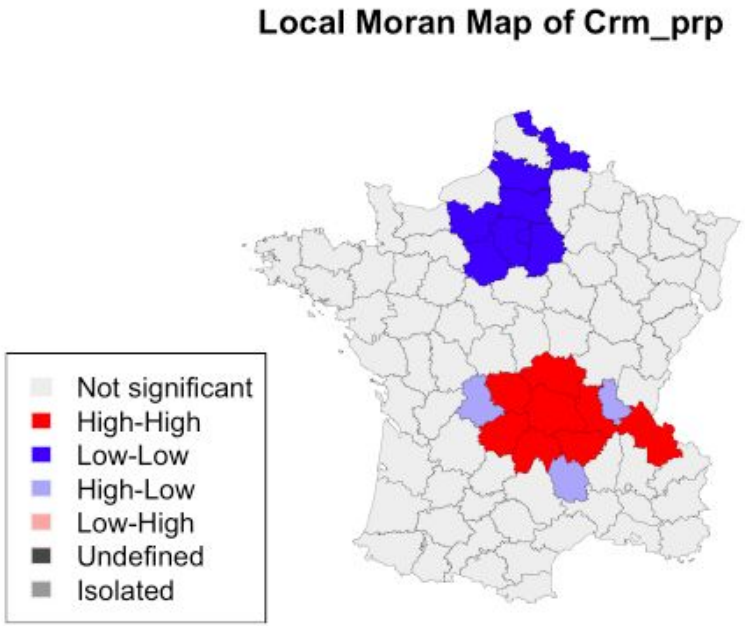


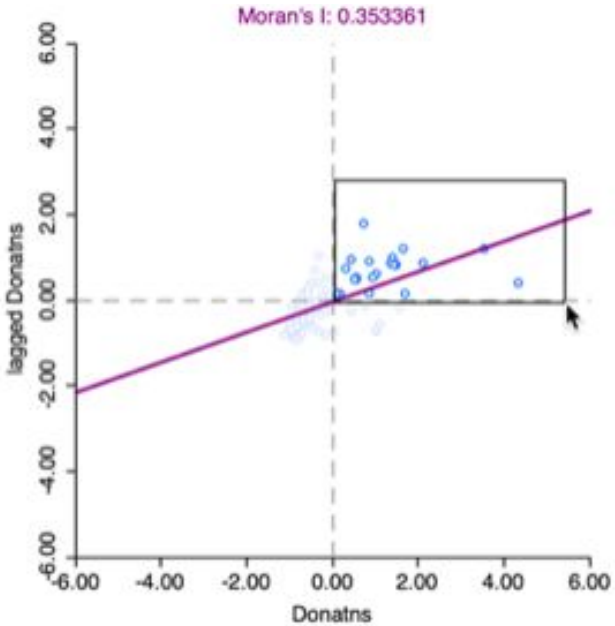
Spatial pattern detection



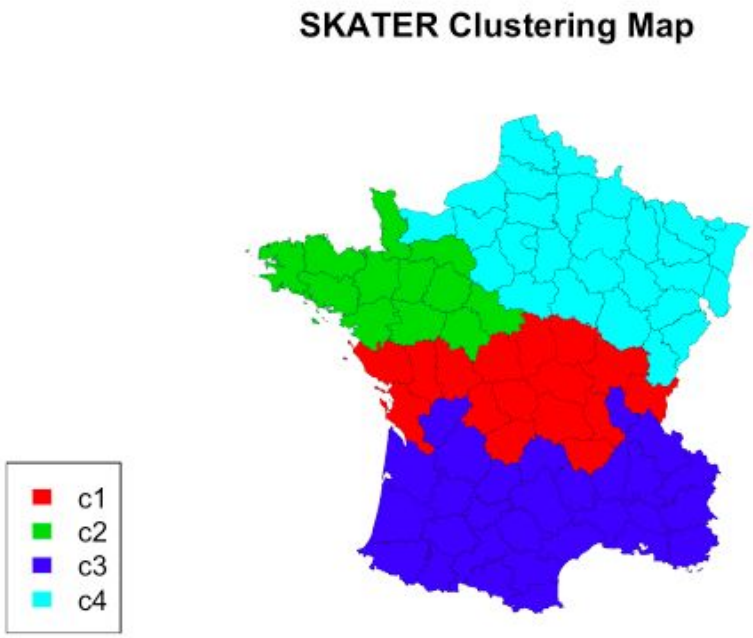


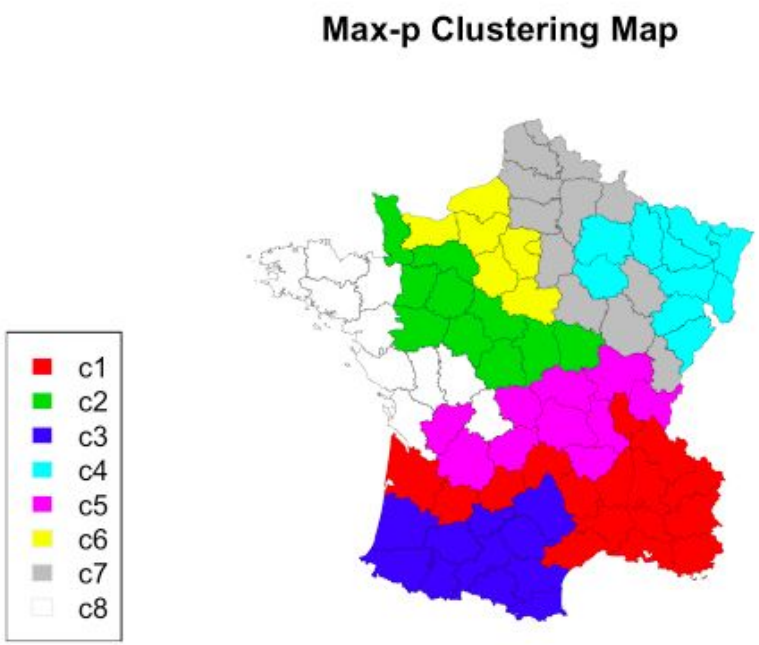
Local spatial autocorrelation





Spatially constrained clustering





Pre-alpha release of rgeoda

https://rgeoda.github.io/rgeoda-book/
https://github.com/lixun910/rgeoda

Stay in touch!

#rspatial on Twitter
Modern Geospatial Analysis Tutorial at rstudio::conf
Geospatial Track at useR! 2020 (submit tutorials!)

ali6@uchicago.edu
angela-li.github.io
@CivicAngela on Twitter

I may be back in DC in May, stay tuned for more…

Appendix

Choose your own (geocoding) adventure

- Batch geocoding: can't beat university/industry geocoders
  - Talk to your GIS librarian today!
  - Or pay $$$ for an ArcGIS license and use their geocoding service
- Free R package alternatives (rate limits on use):
  - `tmaptools::geocode_OSM()`
  - `ggmap::geocode()`
  - `opencage::opencage_forward()`

My top 5 sources for spatial data

1. US Census, through `library(tidycensus)`
2. Open data portals (“DC Open Data”)
3. OpenStreetMap
4. CSVs with latitude and longitude I then transform
5. Data I inherited with spatial features

Also feel free to try out...

README.md

geodaData

The goal of geodaData is to store sample spatial datasets. These datasets are intended to be used to teach basic spatial analysis concepts. They are used in the weekly R Spatial Workshop at the Center for Spatial Data Science at UChicago, and are based off of the GeoDa workbook and data site developed by Luc Anselin and team. Datasets are stored in the sf

github.com/spatialanalysis/geodaData

```
remotes::install_github("spatialanalysis/geodaData")
```

Usage

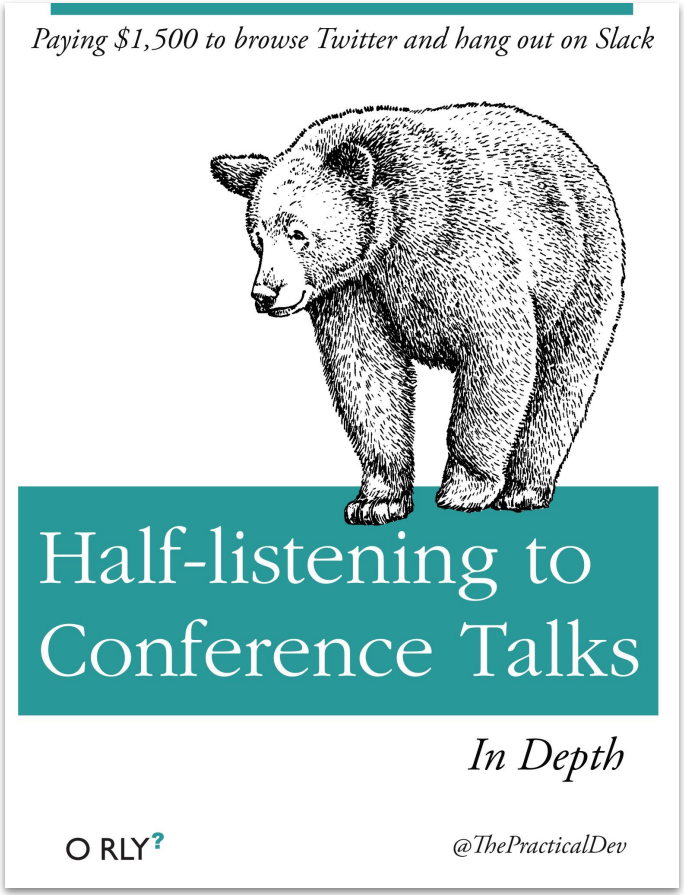
To use geodaData in a workshop, first load `sf`, then load the package:

```
library(sf) # can use without sf, but datasets will print weirdly in console
library(geodaData)
```

Spatial error message whispering

```
* installing *source* package ‘classInt’ ...
** package ‘classInt’ successfully unpacked and MD5 sums checked
** using staged installation
** libs
xcrun: error: invalid active developer path (/Library/Developer/CommandLineTools),
missing xcrun at: /Library/Developer/CommandLineTools/usr/bin/xcrun
ERROR: compilation failed for package ‘classInt’
* removing ‘/Library/Frameworks/R.framework/Versions/3.6/Resources/library/classInt’
Warning in install.packages :
  installation of package ‘classInt’ had non-zero exit status
```

“It’s not you, it’s the spatial library dependency”

Easter egg!

(thanks for listening)