# # Conceptos básicos de los SIG

{#gis-basics}

## Resumen

Los aspectos espaciales de sus datos pueden proporcionar mucha información sobre la situación del brote, y responder a preguntas como:

* ¿Dónde están los focos actuales de la enfermedad?
* ¿Cómo han cambiado los puntos conflictivos con el tiempo?
* ¿Cómo es el acceso a las instalaciones sanitarias? ¿Se necesitan mejoras?

El enfoque actual de esta página de los SIG (GIS, Sistemas de información geográfico) es abordar las necesidades de la epidemiología aplicada en su respuesta a los brotes. Exploraremos los métodos básicos de visualización de datos espaciales utilizando los paquetes **tmap** y **ggplot2**. También recorreremos algunos de los métodos básicos de gestión y consulta de datos espaciales con el paquete **sf**. Por último, abordaremos brevemente conceptos de estadística espacial como las relaciones espaciales, la autocorrelación espacial y la regresión espacial utilizando el paquete **spdep**.

## Términos clave

A continuación presentamos algunos términos clave. Para una introducción completa a los SIG y al análisis espacial, le sugerimos que revise uno de los tutoriales o cursos más largos que aparecen en la sección de Referencias.

**Sistema de Información Geográfica (SIG)** - Un SIG es un marco o entorno para recopilar, gestionar, analizar y visualizar datos espaciales.

### Software GIS

Algunos de los programas de SIG más conocidos permiten la interacción "apuntar y clicar" para el desarrollo de mapas y el análisis espacial. Estas herramientas tienen ventajas como no tener que aprender código y la facilidad de seleccionar y colocar manualmente los iconos y características en un mapa. He aquí dos de los más populares:

**ArcGIS** - Un software comercial de SIG desarrollado por la empresa ESRI, que es muy popular pero bastante caro

**QGIS** - Un software SIG gratuito de código abierto que puede hacer casi todo lo que ArcGIS puede hacer. Puedes [descargar QGIS aquí](https://qgis.org/en/site/forusers/download.html)

El uso de R como SIG puede parecer más intimidante al principio porque, en lugar de "apuntar y clicar", tiene una "interfaz de línea de comandos" (hay que codificar para adquirir el resultado deseado). Sin embargo, esto es una gran ventaja si necesita producir mapas repetidamente o crear un análisis que sea reproducible.

### Datos espaciales

Las dos formas principales de datos espaciales utilizadas en los SIG son los datos vectoriales y los ráster:

**Datos vectoriales** - El formato más común de datos espaciales utilizado en los SIG, los datos vectoriales se componen de características geométricas de vértices y trayectorias. Los datos espaciales vectoriales pueden dividirse a su vez en tres tipos ampliamente utilizados:

* Puntos - Un punto consiste en un par de coordenadas (x,y) que representa una ubicación específica en un sistema de coordenadas. Los puntos son la forma más básica de datos espaciales, y pueden utilizarse para denotar un caso (por ejemplo, el domicilio de un paciente) o una ubicación (por ejemplo, un hospital) en un mapa.
* Líneas - Una línea está compuesta por dos puntos conectados. Las líneas tienen una longitud y pueden utilizarse para indicar cosas como carreteras o ríos.
* Polígonos - Un polígono está compuesto por al menos tres segmentos de línea conectados por puntos. Las características de los polígonos tienen una longitud (es decir, el perímetro del área) así como una medida de área. Los polígonos pueden utilizarse para señalar una zona (por ejemplo, un pueblo) o una estructura (por ejemplo, la superficie real de un hospital).

**Datos ráster** - Un formato alternativo para los datos espaciales, los datos ráster son una matriz de celdas (por ejemplo, píxeles) en la que cada celda contiene información como la altura, la temperatura, la pendiente, la cubierta forestal, etc. Suelen ser fotografías aéreas, imágenes de satélite, etc. Los rásteres también pueden utilizarse como "mapas base" debajo de los datos vectoriales.

### Visualización de datos espaciales

Para representar visualmente los datos espaciales en un mapa, el software SIG requiere que se proporcione suficiente información sobre dónde deben estar las diferentes características, en relación unas con otras. Si se utilizan datos vectoriales, como ocurre en la mayoría de los casos, esta información suele almacenarse en un archivo shapefile:

**Shapefiles** - Un shapefile es un formato de datos común para almacenar datos espaciales "vectoriales" consistentes en líneas, puntos o polígonos. Un shapefile es en realidad una colección de al menos tres archivos - .shp, .shx y .dbf. Todos estos archivos subcomponentes deben estar presentes en un determinado directorio (carpeta) para que el shapefile sea legible. Estos archivos asociados pueden comprimirse en una carpeta ZIP para enviarlos por correo electrónico o descargarlos de un sitio web.

El shapefile contendrá información sobre las características en sí mismas, así como su ubicación en la superficie de la Tierra. Esto es importante porque, aunque la Tierra es un globo terráqueo, los mapas suelen ser bidimensionales; las decisiones sobre cómo "aplanar" los datos espaciales pueden tener un gran impacto en el aspecto y la interpretación del mapa resultante.

**Sistemas de referencia de coordenadas (SRC)** - Un SRC es un sistema basado en coordenadas que se utiliza para localizar accidentes geográficos en la superficie de la Tierra. Tiene unos cuantos componentes clave:

* Sistema de coordenadas - Hay muchos sistemas de coordenadas diferentes, así que asegúrate de saber de qué sistema son tus coordenadas. Los grados de latitud/longitud son comunes, pero también podría ver coordenadas [UTM](https://www.maptools.com/tutorials/utm/quick_guide).
* Unidades - Saber cuáles son las unidades de su sistema de coordenadas (por ejemplo, grados decimales, metros)
* Datum - Una versión particular modelada de la Tierra. Estos han sido revisados a lo largo de los años, así que asegúrate de que tus capas de mapa utilizan el mismo datum.
* Proyección - Referencia a la ecuación matemática que se utilizó para proyectar la tierra realmente redonda sobre una superficie plana (mapa).

Recuerde que puede resumir los datos espaciales sin utilizar las herramientas cartográficas que se muestran a continuación. A veces basta con una simple tabla por zonas geográficas (por ejemplo, distrito, país, etc.).

## Introducción a los SIG

Hay un par de elementos clave que deberá tener y en los que deberá pensar para hacer un mapa. Entre ellos están:

* unos datos: puede estar en un formato de datos espaciales (como los shapefiles, como se ha indicado anteriormente) o puede no estar en un formato espacial (por ejemplo, sólo como un csv).
* Si su conjunto de datos no está en formato espacial, también necesitará unos datos **de referencia**. Los datos de referencia consisten en la representación espacial de los datos y los **atributos** relacionados, que incluirían el material que contiene la información de ubicación y dirección de características específicas.
  + Si está trabajando con límites geográficos predefinidos (por ejemplo, regiones administrativas), los archivos shape de referencia suelen estar disponibles para su descarga de forma gratuita desde una agencia gubernamental u organización de intercambio de datos. En caso de duda, un buen punto de partida es buscar en Google "[regiones] shapefile"
  + Si dispone de información de direcciones, pero no de latitud y longitud, puede que tenga que utilizar un **motor de geocodificación** para obtener los datos de referencia espacial de sus registros.
* Una idea sobre **cómo quiere presentar** la información de sus conjuntos de datos a su público objetivo. Hay muchos tipos diferentes de mapas, y es importante pensar qué tipo de mapa se ajusta mejor a sus necesidades.

### Tipos de mapas para visualizar sus datos

**Mapa coroplético**: tipo de mapa temático en el que se utilizan colores, sombreados o patrones para representar regiones geográficas en relación con su valor de un atributo. Por ejemplo, un valor mayor podría indicarse con un color más oscuro que un valor menor. Este tipo de mapa es particularmente útil cuando se visualiza una variable y cómo cambia a través de regiones o áreas geopolíticas definidas.

**Mapa térmico de densidad de casos**: un tipo de mapa temático en el que se utilizan colores para representar la intensidad de un valor, pero que no utiliza regiones definidas ni límites geopolíticos para agrupar los datos. Este tipo de mapa se suele utilizar para mostrar "puntos calientes" o zonas con una alta densidad o concentración de puntos.

**Mapa de densidad de puntos**: un tipo de mapa temático que utiliza puntos para representar los valores de los atributos en sus datos. Este tipo de mapa es el más adecuado para visualizar la dispersión de los datos y buscar clusters visualmente.

Mapa de **símbolos proporcionales (mapa de símbolos graduados)**: es un mapa temático similar a un mapa coroplético, pero en lugar de utilizar el color para indicar el valor de un atributo, utiliza un símbolo (normalmente un círculo) en relación con el valor. Por ejemplo, un valor mayor podría indicarse con un símbolo mayor que un valor menor. Este tipo de mapa se utiliza mejor cuando se quiere visualizar el tamaño o la cantidad de los datos en las distintas regiones geográficas.

También puede combinar varios tipos de visualizaciones diferentes para mostrar patrones geográficos complejos. Por ejemplo, los casos (puntos) del mapa siguiente están coloreados según su centro de salud más cercano (véase la leyenda). Los círculos rojos grandes muestran las zonas de captación de los centros de salud de un determinado radio, y los puntos rojos brillantes los que estaban fuera de cualquier zona de captación:

Nota: El enfoque principal de esta página del SIG se basa en el contexto de la respuesta a los brotes de campo. Por lo tanto, el contenido de la página cubrirá las manipulaciones, visualizaciones y análisis básicos de datos espaciales.

## Preparación

### Cargar paquetes

Este trozo de código muestra la carga de los paquetes necesarios para los análisis. En este manual destacamos p\_load() de **pacman**, que instala el paquete si es necesario y lo carga para su uso. También puede cargar los paquetes instalados con library() de **.** Consulta la página sobre [los fundamentos de R](#r-basics) para obtener más información sobre los paquetes de R.

Puedes ver un resumen de todos los paquetes de R que se ocupan de los datos espaciales en la ["Vista de tareas espaciales" de CRAN](https://cran.r-project.org/web/views/Spatial.html).

### Datos del caso de muestra

Para fines de demostración, trabajaremos con una muestra aleatoria de 1000 casos del dataframe del listado de la epidemia de ébola simulada (computacionalmente, trabajar con menos casos es más fácil de mostrar en este manual). Si quiere seguir el proceso, [clica para descargar linelist "limpia"](https://github.com/epirhandbook/Epi_R_handbook/raw/master/data/case_linelists/linelist_cleaned.rds) (como archivo .rds).

Dado que estamos tomando una muestra aleatoria de los casos, sus resultados pueden ser ligeramente diferentes de los que se muestran aquí cuando ejecutes el código por tu cuenta.

Importar datos con la función import() del paquete **rio** (maneja muchos tipos de archivos como .xlsx, .csv, .rds - ver la página de [importación y exportación](#import-and-export) para más detalles).

A continuación, seleccionamos una muestra aleatoria de 1000 filas utilizando sample() de la **base** R.

Ahora queremos convertir esta lista de líneas, que es de tipo dataframe, en un objeto de tipo "sf" (spatial features). Dado que linelist tiene dos columnas "lon" y "lat" que representan la longitud y latitud de la residencia de cada caso, esto será fácil.

Utilizamos el paquete **sf** (spatial features) y su función st\_as\_sf() para crear el nuevo objeto que llamamos linelist\_sf. Este nuevo objeto tiene esencialmente el mismo aspecto que linelist, pero las columnas lon y lat han sido designadas como columnas de coordenadas, y se ha asignado un sistema de referencia de coordenadas (CRS) para cuando se muestren los puntos. El 4326 identifica nuestras coordenadas como basadas en el [Sistema Geodésico Mundial 1984 (WGS84)](https://gisgeography.com/wgs84-world-geodetic-system/) - que es el estándar para las coordenadas GPS.

Este es el aspecto del dataframe original del listado. En esta demostración, sólo utilizaremos la columna date\_onset y geometry (que se construyó a partir de los campos de longitud y latitud anteriores y es la última columna del dataframe).

### Archivos shapefiles de los límites de la administración

**Sierra Leona: Archivos shapefiles de los límites administrativos**

De antemano, hemos descargado todos los límites administrativos de Sierra Leona del [sitio web de](https://data.humdata.org/dataset/sierra-leone-all-ad-min-level-boundaries) Humanitarian Data Exchange (HDX) [aquí](https://data.humdata.org/dataset/sierra-leone-all-ad-min-level-boundaries). Como alternativa, puede descargar estos y todos los demás datos de ejemplo para este manual a través de nuestro paquete R, como se explica en la página [descargando el manual y los datos](#download-handbook-and-data).

Ahora vamos a hacer lo siguiente para guardar el shapefile del nivel 3 de administración en R:

1. Importar el shapefile
2. Limpiar los nombres de las columnas
3. Filtrar las filas para mantener sólo las áreas de interés

Para importar un shapefile utilizamos la función read\_sf() de **sf**. Se le proporciona la ruta del archivo a través de here(). - En nuestro caso el archivo se encuentra dentro de nuestro proyecto R en las subcarpetas "data", "gis" y "shp", con nombre de archivo "sle\_adm3.shp" (para más información ver las páginas sobre [Importación y exportación](#import-and-export) y [proyectos R](#r-projects)). Tendrá que proporcionar su propia ruta de archivo.

A continuación utilizamos clean\_names() del paquete **janitor** para estandarizar los nombres de las columnas del shapefile. También utilizamos filter() para mantener sólo las filas con admin2name de "Western Area Urban" o "Western Area Rural".

A continuación puede ver el aspecto del shapefile después de la importación y la limpieza. Desplácese a la derecha para ver cómo hay columnas con el nivel de administración 0 (país), el nivel de administración 1, el nivel de administración 2 y, finalmente, el nivel de administración 3. Cada nivel tiene un nombre de carácter y un identificador único "pcode". El pcode se expande con cada nivel de administración creciente, por ejemplo, SL (Sierra Leona) -> SL04 (Occidental) -> SL0410 (Zona Occidental Rural) -> SL040101 (Koya Rural).

### Datos de población

**Sierra Leona: Población por ADM3**

Estos datos pueden descargarse de nuevo de HDX (enlace [aquí](https://data.humdata.org/dataset/sierra-leone-population)) o a través de nuestro paquete R **epirhandbook** como se explica [en esta página](#download-handbook-and-data). Utilizamos import() para cargar el archivo .csv. También pasamos el archivo importado a clean\_names() para estandarizar la sintaxis de los nombres de las columnas.

Este es el aspecto del archivo de población. Desplácese a la derecha para ver cómo cada jurisdicción tiene columnas con la población masculina, la población femenina, la población total y el desglose de la población en columnas por grupos de edad.

### Instalaciones sanitarias

**Sierra Leona: Datos de los centros de salud de OpenStreetMap**

Una vez más, hemos descargado las ubicaciones de los centros de salud desde el HDX [aquí](https://data.humdata.org/dataset/hotosm_sierra_leone_health_facilities) o mediante las instrucciones de la página [descarga de manuales y datos](#download-handbook-and-data).

Importamos el shapefile de puntos de las instalaciones con read\_sf(), limpiamos de nuevo los nombres de las columnas y filtramos para mantener sólo los puntos etiquetados como "hospital", "clínica" o "médicos".

Aquí está el dataframe resultante - desplácese a la derecha para ver el nombre de la instalación y las coordenadas geométricas.

## Trazado de coordenadas

La forma más sencilla de trazar las coordenadas X-Y (longitud/latitud, puntos), en este caso de los casos, es dibujarlas como puntos directamente desde el objeto linelist\_sf que creamos en la sección de preparación.

El paquete **tmap** ofrece capacidades de mapeo simples tanto para el modo estático (modo "plot") como para el interactivo (modo "view") con sólo unas pocas líneas de código. La sintaxis de tmap es similar a la de **ggplot2**, de forma que los comandos se añaden unos a otros con +. Lea más detalles en esta [viñeta](https://cran.r-project.org/web/packages/tmap/vignettes/tmap-getstarted.html).

1. Establezca el modo **tmap**. En este caso utilizaremos el modo "plot", que produce salidas estáticas.

A continuación, los puntos se trazan solos. tm\_shape() se proporciona con los objetos linelist\_sf. A continuación, añadimos puntos mediante tm\_dots(), especificando el tamaño y el color. Como linelist\_sf es un objeto sf, ya hemos designado las dos columnas que contienen las coordenadas lat/long y el sistema de referencia de coordenadas (CRS):

Por sí solos, los puntos no nos dicen mucho. Así que también deberíamos trazar los límites administrativos:

De nuevo utilizamos tm\_shape() (ver [documentación](https://www.rdocumentation.org/packages/tmap/versions/3.3/topics/tm_shape)) pero en lugar de proporcionar el shapefile de los puntos del caso, proporcionamos el shapefile de los límites administrativos (polígonos).

Con el argumento bbox = (bbox significa "bounding box") podemos especificar los límites de las coordenadas. Primero mostramos la visualización del mapa sin bbox, y luego con él.

Y ahora tanto los puntos como los polígonos juntos:

Para leer una buena comparación de las opciones de mapeo en R, consulta esta [entrada del blog](https://rstudio-pubs-static.s3.amazonaws.com/324400_69a673183ba449e9af4011b1eeb456b9.html).

## Uniones espaciales

Es posible que esté familiarizado con la unión de datos de unos datos a otro. En la página [unión de datos](#joining-data) de este manual se tratan varios métodos. Una unión espacial tiene un propósito similar, pero aprovecha las relaciones espaciales. En lugar de confiar en los valores comunes de las columnas para hacer coincidir correctamente las observaciones, puede utilizar sus relaciones espaciales, como que una característica esté dentro de otra, o que sea la vecina más cercana de otra, o que esté dentro de un búfer de un determinado radio de otra, etc.

El paquete **sf** ofrece varios métodos para las uniones espaciales. Consulta más documentación sobre el método st\_join y los tipos de uniones espaciales en esta [referencia](https://r-spatial.github.io/sf/reference/geos_binary_pred.html).

### Puntos en el polígono

**Asignación espacial de unidades administrativas a los casos**

Aquí se plantea un interesante dilema: la lista de casos no contiene ninguna información sobre las unidades administrativas de los casos. Aunque lo ideal es recoger dicha información durante la fase inicial de recogida de datos, también podemos asignar unidades administrativas a los casos individuales basándonos en sus relaciones espaciales (es decir, el punto se cruza con un polígono).

A continuación, intersecaremos espacialmente las ubicaciones de nuestros casos (puntos) con los límites del ADM3 (polígonos):

1. Comience con linelist (puntos)
2. Unión espacial a los límites, estableciendo el tipo de unión en "st\_intersects"
3. Utiliza select() para mantener sólo algunas de las nuevas columnas de los límites administrativos

Todas las columnas de sle\_adms se han añadido a linelist. Cada caso tiene ahora columnas que detallan los niveles administrativos a los que pertenece. En este ejemplo, sólo queremos mantener dos de las nuevas columnas (nivel administrativo 3), así que seleccionamos() los nombres de las columnas antiguas y sólo las dos adicionales de interés:

A continuación, sólo a efectos de visualización, puede ver los diez primeros casos y que sus jurisdicciones de nivel de administración 3 (ADM3) que se han adjuntado, basado en donde el punto espacialmente se cruzó con las formas de polígono.

Ahora podemos describir nuestros casos por unidad administrativa, algo que no podíamos hacer antes de la unión espacial.

También podemos crear un gráfico de barras de los recuentos de casos por unidad administrativa.

En este ejemplo, comenzamos el ggplot() con el linelist\_adm, de modo que podemos aplicar funciones de factor como fct\_infreq() que ordena las barras por frecuencia (véase la página sobre [Factores](#factors) para obtener consejos).

### Vecino más cercano

**Encontrar el centro de salud/zona de captación más cercano**

Podría ser útil saber dónde se encuentran los centros de salud en relación con los focos de la enfermedad.

Podemos utilizar el método st\_nearest\_feature join de la función st\_join() (paquete **sf**) para visualizar el centro sanitario más cercano a los casos individuales.

1. Comenzamos con el shapefile linelist linelist\_sf
2. Unimos espacialmente con sle\_hf, que es la ubicación de los centros de salud y las clínicas (puntos)

Podemos ver a continuación (primeras 50 filas) que cada caso tiene ahora datos sobre la clínica/hospital más cercano

Podemos ver que "Den Clinic" es el centro de salud más cercano para aproximadamente el 30% de los casos.

Para visualizar los resultados, podemos utilizar **tmap** - esta vez en modo interactivo para facilitar la visualización

### Búferes

También podemos explorar cuántos casos se encuentran a menos de 2,5 km (~30 minutos) de distancia a pie del centro de salud más cercano.

Nota: Para un cálculo más preciso de la distancia, es mejor reproyectar su objeto sf al respectivo sistema de proyección cartográfica local, como UTM (Tierra proyectada sobre una superficie plana). En este ejemplo, para simplificar, nos ceñiremos al sistema de coordenadas geográficas del Sistema Geodésico Mundial (WGS84) (la Tierra representada en una superficie esférica / redonda, por lo que las unidades están en grados decimales). Utilizaremos una conversión general de: 1 grado decimal = ~111km.

Vea más información sobre proyecciones cartográficas y sistemas de coordenadas en este [artículo de esri](https://www.esri.com/arcgis-blog/products/arcgis-pro/mapping/gcs_vs_pcs/). Este [blog](http://www.geo.hunter.cuny.edu/~jochen/gtech201/lectures/lec6concepts/map coordinate systems/how to choose a projection.htm) habla de los diferentes tipos de proyecciones cartográficas y de cómo se puede elegir una proyección adecuada en función del área de interés y del contexto de su mapa/análisis.

**En primer lugar**, se crea un buffer circular con un radio de ~2,5km alrededor de cada centro de salud. Esto se hace con la función st\_buffer() de **tmap**. Como la unidad del mapa está en grados decimales lat/long, así es como se interpreta "0,02". Si su sistema de coordenadas del mapa está en metros, el número debe proporcionarse en metros.

A continuación, trazamos las zonas de amortiguación propiamente dichas, con el :

En segundo lugar, intersectamos estos buffers con los casos (puntos) utilizando *st\_join()* y el tipo de unión st\_intersects\*. Es decir, los datos de los buffers se unen a los puntos con los que se cruzan.

Ahora podemos contar los resultados: nrow(linelist\_sf\_hf\_2k[is.na(linelist\_sf\_hf\_2k$osm\_id.y),]) de 1000 casos no se cruzan con ningún buffer (falta ese valor), y por tanto viven a más de 30 minutos a pie del centro de salud más cercano.

Podemos visualizar los resultados de forma que los casos que no se cruzan con ningún búfer aparecen en rojo.

### Otras uniones espaciales

Los valores alternativos para el argumento join incluyen (de la [documentación](https://r-spatial.github.io/sf/reference/st_join.html))

* st\_contiene\_propiamente
* st\_contiene
* st\_covered\_by
* st\_covers
* st\_crosses
* st\_disjoint
* st\_equals\_exact
* st\_equals
* st\_está\_dentro\_de\_la\_distancia
* st\_nearest\_feature
* st\_overlaps
* st\_touches
* st\_within

## Mapas coropléticos

Los mapas coropléticos pueden ser útiles para visualizar los datos por áreas predefinidas, normalmente unidades administrativas o áreas de salud. En la respuesta a los brotes, esto puede ayudar a dirigir la asignación de recursos a zonas específicas con altas tasas de incidencia, por ejemplo.

Ahora que tenemos los nombres de las unidades administrativas asignados a todos los casos (véase la sección sobre uniones espaciales, más arriba), podemos empezar a mapear los recuentos de casos por zonas (mapas coropléticos).

Como también tenemos datos de población por ADM3, podemos añadir esta información a la tabla case\_adm3 creada anteriormente.

Comenzamos con el dataframe creado en el paso anterior case\_adm3, que es una tabla resumen de cada unidad administrativa y su número de casos.

1. Los datos de la población sle\_adm3\_pop se unen utilizando un left\_join() de **dplyr sobre la** base de valores comunes a través de la columna admin3pcod en el dataframe case\_adm3, y la columna adm\_pcode en el dataframe sle\_adm3\_pop. Véase la página sobre la [unión de datos](#joining-data)).
2. select() se aplica al nuevo dataframe, para mantener sólo las columnas útiles - total es la población total
3. Los casos por cada 10.000 habitantes se calculan como una nueva columna con mutate()

Unir esta tabla con el shapefile de polígonos ADM3 para la cartografía

Trazado de los resultados

También podemos mapear las tasas de incidencia

## Mapeo con ggplot2

Si ya está familiarizado con el uso de **ggplot2**, puede utilizar ese paquete para crear mapas estáticos de sus datos. La función geom\_sf() dibujará diferentes objetos en función de las características (puntos, líneas o polígonos) que haya en sus datos. Por ejemplo, puede utilizar geom\_sf() en un ggplot() utilizando datos sf con geometría de polígonos para crear un mapa coroplético.

Para ilustrar cómo funciona esto, podemos empezar con el archivo shape de polígonos ADM3 que hemos utilizado antes. Recordemos que se trata de regiones de nivel administrativo 3 en Sierra Leona:

Podemos utilizar la función left\_join() de **dplyr** para añadir los datos que queremos mapear al objeto shapefile. En este caso, vamos a utilizar el dataframe case\_adm3 que creamos anteriormente para resumir los recuentos de casos por región administrativa; sin embargo, podemos utilizar este mismo enfoque para mapear cualquier dato almacenado en un dataframe.

Para hacer un gráfico de columnas de los recuentos de casos por región, utilizando **ggplot2,** podríamos llamar a geom\_col() de la siguiente manera:

Si queremos utilizar **ggplot2** para hacer un mapa de coropletas de los recuentos de casos, podemos utilizar una sintaxis similar para llamar a la función geom\_sf():

A continuación, podemos personalizar la apariencia de nuestro mapa utilizando una gramática que sea consistente en **ggplot2**, por ejemplo:

Para los usuarios de R que se sientan cómodos trabajando con **ggplot2**, geom\_sf() ofrece una implementación simple y directa que es adecuada para las visualizaciones básicas de mapas. Para saber más, lea la [viñeta de geom\_sf()](https://ggplot2.tidyverse.org/reference/ggsf.html) o el [libro de ggplot2](https://ggplot2-book.org/maps.html).

## Mapas base

### OpenStreetMap

A continuación describimos cómo conseguir un mapa base para un mapa **ggplot2** utilizando las características de OpenStreetMap. Métodos alternativos incluyen el uso de **ggmap** que requiere el registro gratuito con Google ([detalles](https://www.earthdatascience.org/courses/earth-analytics/lidar-raster-data-r/ggmap-basemap/)).

[**OpenStreetMap**](https://en.wikipedia.org/wiki/OpenStreetMap) es un proyecto de colaboración para crear un mapa editable y gratuito del mundo. Los datos de geolocalización subyacentes (por ejemplo, ubicaciones de ciudades, carreteras, características naturales, aeropuertos, escuelas, hospitales, caminos, etc.) se consideran el resultado principal del proyecto.

Primero cargamos el paquete **OpenStreetMap**, del que obtendremos nuestro mapa base.

A continuación, creamos el objeto mapa, que definimos mediante la función openmap() del paquete **OpenStreetMap** ([documentación](https://www.rdocumentation.org/packages/OpenStreetMap/versions/0.3.4/topics/openmap)). Proporcionamos lo siguiente:

* upperLeft y lowerRight Dos pares de coordenadas que especifican los límites del mosaico del mapa base
  + En este caso hemos puesto los máximos y mínimos de las filas del listado, para que el mapa responda dinámicamente a los datos
* zoom = (si es nulo se determina automáticamente)
* tipo = qué tipo de mapa base - aquí hemos enumerado varias posibilidades y el código utiliza actualmente la primera ([1]) "osm"
* mergeTiles = elegimos TRUE para que los basetiles se fusionen en uno solo

Si trazamos este mapa base ahora mismo, usando autoplot.OpenStreetMap() del paquete **OpenStreetMap**, verás que las unidades en los ejes no son coordenadas de latitud/longitud. Se está utilizando un sistema de coordenadas diferente. Para mostrar correctamente las residencias del caso (que se almacenan en lat/long), esto debe ser cambiado.

Así, queremos convertir el mapa a latitud/longitud con la función openproj() del paquete **OpenStreetMap**. Proporcionamos el mapa base y también el Sistema de Referencia de Coordenadas (CRS) que queremos. Lo hacemos proporcionando la cadena de caracteres "proj.4" para la proyección WGS 1984, pero también se puede proporcionar el CRS de otras maneras. (vea [esta página](https://www.earthdatascience.org/courses/earth-analytics/spatial-data-r/understand-epsg-wkt-and-other-crs-definition-file-types/) para entender mejor qué es una cadena proj.4)

Ahora cuando creamos el gráfico vemos que a lo largo de los ejes están las coordenadas de latitud y longitud. El sistema de coordenadas ha sido convertido. Ahora nuestros casos se trazarán correctamente si se superponen.

Consulta los tutoriales [aquí](http://data-analytics.net/cep/Schedule_files/geospatial.html) y [aquí](https://www.rdocumentation.org/packages/OpenStreetMap/versions/0.3.4/topics/autoplot.OpenStreetMap) para obtener más información.

## Mapas térmicos de densidad contorneada

A continuación describimos cómo conseguir un mapa de calor de densidad contorneada de casos, sobre un mapa base, comenzando con un listado (una fila por caso).

1. Crear un mosaico de mapa base a partir de OpenStreetMap, como se ha descrito anteriormente
2. Trazar los casos del listado utilizando las columnas de latitud y longitud
3. Convierta los puntos en un mapa de calor de densidad con stat\_density\_2d() de **ggplot2**,

Cuando tenemos un mapa base con coordenadas de latitud y longitud, podemos trazar nuestros casos encima utilizando las coordenadas de latitud y longitud de su residencia.

Partiendo de la función autoplot.OpenStreetMap() para crear el mapa base, las funciones de **ggplot2 se añaden** fácilmente encima, como se muestra con geom\_point() a continuación:

El mapa anterior puede ser difícil de interpretar, especialmente con los puntos superpuestos. Por lo tanto, puede trazar un mapa de densidad en 2d utilizando la función **ggplot2** stat\_density\_2d(). Se siguen utilizando las coordenadas lat/lon del listado, pero se realiza una estimación de la densidad del núcleo en 2D y los resultados se muestran con líneas de contorno - como un mapa topográfico. Lea la [documentación](https://ggplot2.tidyverse.org/reference/geom_density_2d.html) completa [aquí](https://ggplot2.tidyverse.org/reference/geom_density_2d.html).

### Mapa de calor de las series temporales

El mapa de calor de la densidad anterior muestra los casos acumulados. Podemos examinar el brote a lo largo del tiempo y del espacio haciendo un facetado del mapa de calor basado en el mes de inicio de los síntomas, como se deriva del listado.

Comenzamos en linelist, creando una nueva columna con el Año y el Mes de inicio. La función format() de la **base** R cambia la forma en que se muestra una fecha. En este caso queremos "AAAA-MM".

Ahora, simplemente introducimos el facetado a través de **ggplot2** en el mapa de calor de la densidad. Se aplica facet\_wrap(), utilizando la nueva columna como filas. Fijamos el número de columnas de facetas en 3 para mayor claridad.

## Estadísticas espaciales

La mayor parte de nuestra discusión hasta ahora se ha centrado en la visualización de datos espaciales. En algunos casos, también puede interesarle utilizar la estadística espacial para cuantificar las relaciones espaciales de los atributos de sus datos. En esta sección se ofrece una breve visión general de algunos conceptos clave de la estadística espacial y se sugieren algunos recursos que resultarán útiles si desea realizar análisis espaciales más exhaustivos.

### Relaciones espaciales

Antes de poder calcular cualquier estadística espacial, tenemos que especificar las relaciones entre las características de nuestros datos. Hay muchas formas de conceptualizar las relaciones espaciales, pero un modelo sencillo y comúnmente aplicable es el de la adyacencia, es decir, que esperamos una relación geográfica entre las zonas que comparten una frontera o son "vecinas" unas de otras.

Podemos cuantificar las relaciones de adyacencia entre los polígonos de las regiones administrativas en los datos sle\_adm3 que hemos estado utilizando con el paquete **spdep**. Especificaremos la contigüidad de reinas, lo que significa que las regiones serán vecinas si comparten al menos un punto a lo largo de sus fronteras. La alternativa sería la contigüidad de la torre, que requiere que las regiones compartan un borde - en nuestro caso, con polígonos irregulares, la distinción es trivial, pero en algunos casos la elección entre reina y torre puede ser influyente.

La matriz impresa arriba muestra las relaciones entre las 9 regiones de nuestros datos sle\_adm3. Una puntuación de 0 indica que dos regiones no son vecinas, mientras que cualquier valor distinto de 0 indica una relación de vecindad. Los valores de la matriz se han escalado para que cada región tenga un peso total de 1 en la fila.

Una mejor manera de visualizar estas relaciones de vecindad es trazarlas:

Hemos utilizado un enfoque de adyacencia para identificar los polígonos vecinos; los vecinos que identificamos también se denominan a veces **vecinos por contigüidad**. Pero ésta es sólo una forma de elegir qué regiones se espera que tengan una relación geográfica. Los enfoques alternativos más comunes para identificar las relaciones geográficas generan vecinos basados en la **distancia**; brevemente, estos son:

* **K-vecinos más cercanos** - Basándose en la distancia entre los centroides (el centro ponderado geográficamente de cada región poligonal), selecciona las n regiones más cercanas como vecinas. También se puede especificar un umbral de proximidad de distancia máxima. En **spdep, se** puede utilizar knearneigh() (ver [documentación](https://r-spatial.github.io/spdep/reference/knearneigh.html)).
* **Vecinos de umbral de distancia** - Selecciona todos los vecinos dentro de un umbral de distancia. En **spdep**, estas relaciones de vecindad pueden ser identificadas usando dnearneigh() (ver [documentación](https://www.rdocumentation.org/packages/spdep/versions/1.1-7/topics/dnearneigh)).

### Autocorrelación espacial

La tan citada primera ley de la geografía de Tobler afirma que "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las lejanas". En epidemiología, esto suele significar que el riesgo de un determinado resultado sanitario en una región determinada es más similar al de sus regiones vecinas que al de las lejanas. Este concepto se ha formalizado como **autocorrelación espacial**: la propiedad estadística de que las características geográficas con valores similares se agrupan en el espacio. Las medidas estadísticas de autocorrelación espacial pueden utilizarse para cuantificar el alcance de la agrupación espacial en sus datos, localizar dónde se produce la agrupación e identificar patrones compartidos de autocorrelación espacial entre distintas variables en sus datos. Esta sección ofrece una visión general de algunas medidas comunes de autocorrelación espacial y cómo calcularlas en R.

I de **Moran** - Se trata de una estadística de resumen global de la correlación entre el valor de una variable en una región y los valores de la misma variable en las regiones vecinas. La estadística I de Moran suele oscilar entre -1 y 1. Un valor de 0 indica que no hay ningún patrón de correlación espacial, mientras que los valores más cercanos a 1 o -1 indican una mayor autocorrelación espacial (valores similares cercanos) o dispersión espacial (valores disímiles cercanos), respectivamente.

Como ejemplo, calcularemos un estadístico I de Moran para cuantificar la autocorrelación espacial en los casos de Ébola que hemos mapeado antes (recordemos que se trata de un subconjunto de casos del dataframe del listado de la epidemia simulada). El paquete **spdep** tiene una función, moran.test, que puede hacer este cálculo por nosotros:

La salida de la función moran.test() nos muestra un estadístico I de Moran de round(moran\_i$estimate[1],2). Esto indica la presencia de autocorrelación espacial en nuestros datos, en concreto, que es probable que las regiones con un número similar de casos de ébola estén próximas entre sí. El valor p proporcionado por moran.test() se genera mediante la comparación con la expectativa bajo la hipótesis nula de ausencia de autocorrelación espacial, y puede utilizarse si se necesita informar de los resultados de una prueba de hipótesis formal.

I de **Moran local** - Podemos descomponer el estadístico I de Moran (global) calculado anteriormente para identificar la autocorrelación espacial localizada; es decir, para identificar grupos específicos en nuestros datos. Este estadístico, que a veces se denomina **indicador local de asociación espacial (LISA)**, resume el grado de autocorrelación espacial alrededor de cada región individual. Puede ser útil para encontrar puntos "calientes" y "fríos" en el mapa.

Para mostrar un ejemplo, podemos calcular y mapear la I de Moran local para los recuentos de casos de Ébola utilizados anteriormente, con la función local\_moran() de **spdep**:

**Getis-Ord Gi\*** - Esta es otra estadística que se utiliza comúnmente para el análisis de puntos calientes; en gran parte, la popularidad de esta estadística se relaciona con su uso en la herramienta de análisis de puntos calientes en ArcGIS. Se basa en la suposición de que, normalmente, la diferencia del valor de una variable entre regiones vecinas debería seguir una distribución normal. Utiliza un enfoque de puntuación z para identificar las regiones que tienen valores significativamente más altos (punto caliente) o significativamente más bajos (punto frío) de una variable específica, en comparación con sus vecinos.

Podemos calcular y asignar la estadística Gi\* utilizando la función localG() de **spdep**:

Como puede ver, el mapa de Getis-Ord Gi\* tiene un aspecto ligeramente diferente del mapa de Moran local que elaboré anteriormente. Esto refleja que el método utilizado para calcular estas dos estadísticas es ligeramente diferente; cuál de ellas debe utilizar depende de su caso de uso específico y de la pregunta de investigación de interés.

**Prueba L de Lee** - Es una prueba estadística de correlación espacial bivariada. Permite comprobar si el patrón espacial de una determinada variable x es similar al patrón espacial de otra variable, y, que se supone que está relacionada espacialmente con x.

Para dar un ejemplo, vamos a probar si el patrón espacial de los casos de ébola de la epidemia simulada está correlacionado con el patrón espacial de la población. Para empezar, necesitamos tener una variable de población en nuestros datos sle\_adm3. Podemos utilizar la variable total del dataframe sle\_adm3\_pop que hemos cargado anteriormente.

Podemos visualizar rápidamente los patrones espaciales de las dos variables una al lado de la otra, para ver si se parecen:

Visualmente, los patrones parecen disímiles. Podemos utilizar la función lee.test() de **spdep** para comprobar estadísticamente si el patrón de autocorrelación espacial de las dos variables está relacionado. El estadístico L será cercano a 0 si no hay correlación entre los patrones, cercano a 1 si hay una fuerte correlación positiva (es decir, los patrones son similares), y cercano a -1 si hay una fuerte correlación negativa (es decir, los patrones son inversos).

La salida anterior muestra que el estadístico L de Lee para nuestras dos variables fue round(lee\_test$estimate[1],2), lo que indica una débil correlación negativa. Esto confirma nuestra evaluación visual de que el patrón de los casos y la población no están relacionados entre sí, y proporciona pruebas de que el patrón espacial de los casos no es estrictamente un resultado de la densidad de población en las zonas de alto riesgo.

El estadístico L de Lee puede ser útil para hacer este tipo de inferencias sobre la relación entre variables distribuidas espacialmente; sin embargo, para describir la naturaleza de la relación entre dos variables con más detalle, o ajustar por confusión, se necesitarán técnicas de regresión espacial. Éstas se describen brevemente en la siguiente sección.

### Regresión espacial

Es posible que desee hacer inferencias estadísticas sobre las relaciones entre las variables de sus datos espaciales. En estos casos, es útil considerar las técnicas de regresión espacial, es decir, los enfoques de regresión que consideran explícitamente la organización espacial de las unidades en sus datos. Algunas de las razones por las que puede necesitar considerar modelos de regresión espacial, en lugar de modelos de regresión estándar como los GLM, incluyen:

* Los modelos de regresión estándar suponen que los residuos son independientes entre sí. En presencia de una fuerte autocorrelación espacial, es probable que los residuos de un modelo de regresión estándar también estén autocorrelacionados espacialmente, violando así este supuesto. Esto puede dar lugar a problemas de interpretación de los resultados del modelo, en cuyo caso sería preferible un modelo espacial.
* Los modelos de regresión también suelen suponer que el efecto de una variable x es constante en todas las observaciones. En el caso de la heterogeneidad espacial, los efectos que deseamos estimar pueden variar a lo largo del espacio, y podemos estar interesados en cuantificar esas diferencias. En este caso, los modelos de regresión espacial ofrecen más flexibilidad para estimar e interpretar los efectos.

Los detalles de los enfoques de regresión espacial están fuera del alcance de este manual. En su lugar, esta sección ofrece una visión general de los modelos de regresión espacial más comunes y sus usos, y remite a referencias que pueden ser útiles si se desea profundizar en este ámbito.

**Modelos de error espacial** - Estos modelos suponen que los términos de error entre unidades espaciales están correlacionados, en cuyo caso los datos violarían los supuestos de un modelo OLS estándar. Los modelos de error espacial también se denominan a veces **modelos autorregresivos simultáneos (SAR)**. Pueden ajustarse utilizando la función errorsarlm() del paquete **spatialreg** (funciones de regresión espacial que solían formar parte de **spdep**).

**Modelos de rezago espacial** - Estos modelos suponen que la variable dependiente de una región i está influida no sólo por el valor de las variables independientes en i, sino también por los valores de esas variables en las regiones vecinas a i. Al igual que los modelos de error espacial, los modelos de rezago espacial también se describen a veces como **modelos autorregresivos simultáneos (SAR)**. Pueden ajustarse utilizando la función lagsarlm() del paquete **spatialreg**.

El paquete **spdep** contiene varias pruebas de diagnóstico útiles para decidir entre los modelos OLS estándar, de retardo espacial y de error espacial. Estas pruebas, denominadas diagnósticos del multiplicador de Lagrange, pueden utilizarse para identificar el tipo de dependencia espacial en sus datos y elegir el modelo más apropiado. La función lm.LMtests() puede utilizarse para calcular todas las pruebas del multiplicador de Lagrange. Anselin (1988) también proporciona una útil herramienta de diagrama de flujo para decidir qué modelo de regresión espacial utilizar basándose en los resultados de las pruebas del multiplicador de Lagrange:

**Modelos jerárquicos bayesianos**: los enfoques bayesianos se utilizan habitualmente para algunas aplicaciones del análisis espacial, sobre todo para la [cartografía de enfermedades](https://pubmed.ncbi.nlm.nih.gov/15690999/). Se prefieren en los casos en los que los datos de los casos están escasamente distribuidos (por ejemplo, en el caso de un resultado raro) o son estadísticamente "ruidosos", ya que pueden utilizarse para generar estimaciones "suavizadas" del riesgo de enfermedad al tener en cuenta el proceso espacial latente subyacente. Esto puede mejorar la calidad de las estimaciones. También permiten que el investigador especifique previamente (mediante la elección de la prioridad) los complejos patrones de correlación espacial que pueden existir en los datos, que pueden dar cuenta de la variación espacialmente dependiente e independiente en las variables independientes y dependientes. En R, los modelos jerárquicos bayesianos pueden ajustarse utilizando el paquete **CARbayes** (véase [la viñeta](https://cran.r-project.org/web/packages/CARBayes/vignettes/CARBayes.pdf)) o R-INLA (véase [el sitio web](https://www.r-inla.org/home) y el [libro de texto](https://becarioprecario.bitbucket.io/inla-gitbook/)). R también puede utilizarse para llamar a software externo que realiza estimaciones bayesianas, como JAGS o WinBUGS.

## Recursos

* Funciones simples de R y [viñeta del](https://cran.r-project.org/web/packages/sf/vignettes/sf1.html) paquete sf
* [Viñeta del](https://cran.r-project.org/web/packages/tmap/vignettes/tmap-getstarted.html) paquete R tmap
* ggmap: [Visualización espacial con ggplot2](https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf)
* [Introducción a la elaboración de mapas con R, visión general de los diferentes paquetes](https://bookdown.org/nicohahn/making_maps_with_r5/docs/introduction.html)
* Datos espaciales en R [(curso EarthLab)](https://www.earthdatascience.org/courses/earth-analytics/spatial-data-r/)
* [Libro de texto](https://link.springer.com/book/10.1007/978-1-4614-7618-4) Applied Spatial Data Analysis in R
* **SpatialEpiApp** - una [aplicación Shiny que se puede descargar como un paquete de R](https://github.com/Paula-Moraga/SpatialEpiApp), lo que le permite proporcionar sus propios datos y llevar a cabo la cartografía, el análisis de conglomerados y las estadísticas espaciales.
* [Taller de](http://www.econ.uiuc.edu/~lab/workshop/Spatial_in_R.html) introducción a la econometría espacial en R