



SPATIAL DATA SCIENCE: LO QUE SE PUEDE DECIR (Y LO QUE NO) DESDE LOS DATOS DE COVID-19 EN CHILE



3 opciones para seguir esta charla

1. Expectador

- Suba el audio, relájese y disfrute.

2. Programador Básico

- Link de códigos y mapas (estático)
 - <https://rpubs.com/estebanlp/SpatialDataCience-COVID19>

3. Programador Intermedio

- Link de Rstudio Cloud (interactivo – correr Código)
 - <https://rstudio.cloud/project/1184270>

4. Programador Avanzado

- Link del repositorio para que lo clone y haga un Pull Request o cree un Issue.
 - <https://github.com/estebanlp/Datos-COVID19>

Agenda

- Motivación - Naturaleza del problema
- Datos existentes
- Errores comunes en mostrar información (¿qué no se puede decir?)
 - Graficar mapas en niveles
 - MAUP y Falacia ecológica
 - Sesgo de muestreo vs población
- Cual es la población en riesgo?
 - Población vs población en riesgo
 - Cuarentenas rotativas cambian patrones de movilidad y la población en riesgo
- Datos ideales
 - Que cosas se podrían decir con Spatial data science?
- Desafíos para el futuro

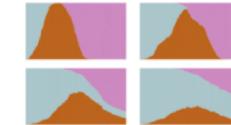
First Law of Geography

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

Waldo Tobler

Friction of Distance

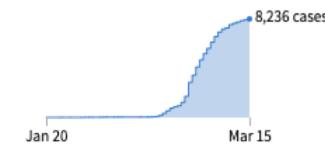
- Activity:
 - Piense en la ruta que generalmente toma para ir de la casa al trabajo (o lugar de estudio, etc.)
- Historias de COVID y Distancia:
 - Coronavirus: How the measures we take matter.
 - Distanciamiento Social
 - The Korean Clusters
 - Interacción Espacial



Health

Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve”

By Harry Stevens March 14, 2020

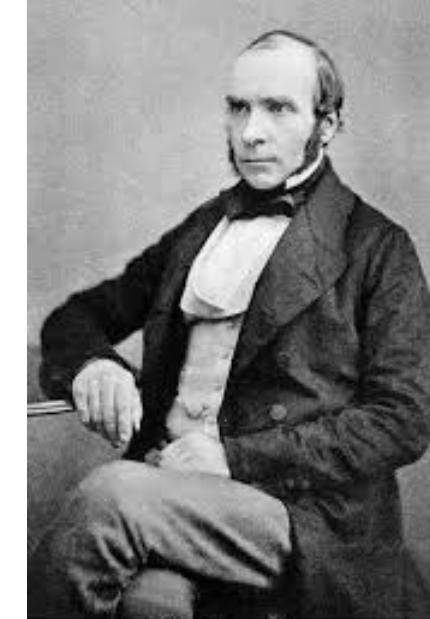


The Korean clusters

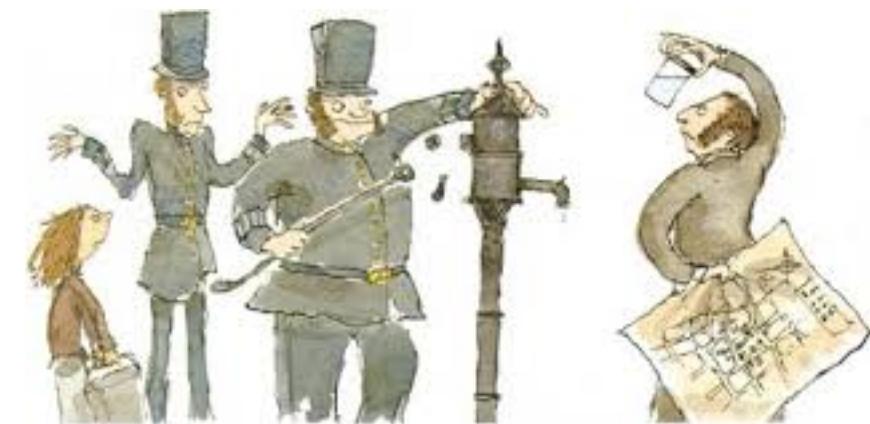
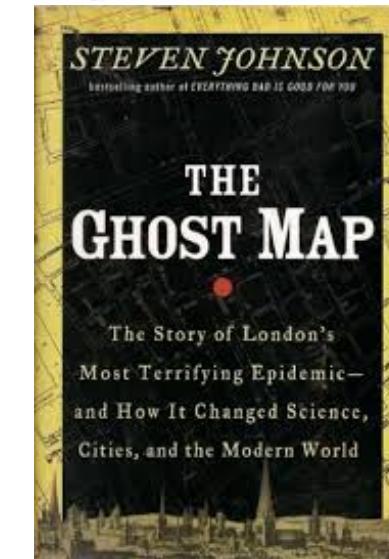
How coronavirus cases exploded in South Korean churches and hospitals

UPDATED MARCH 3, 2020

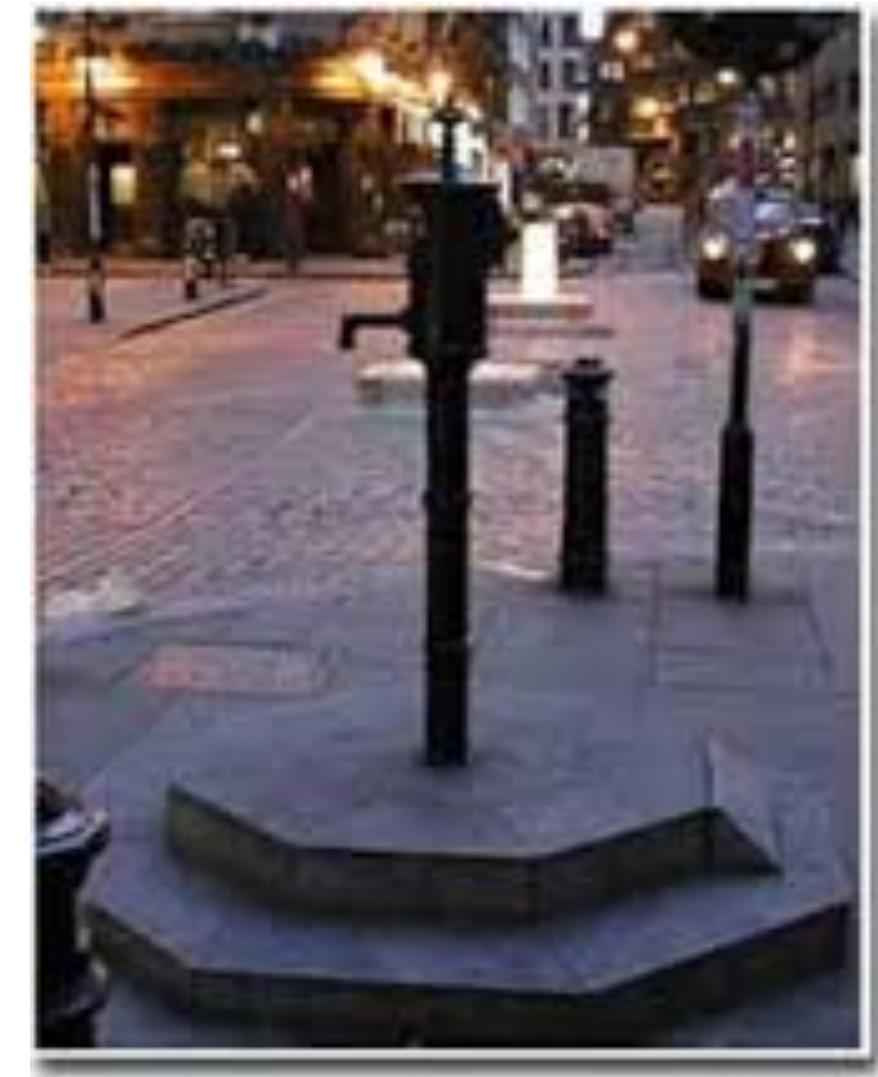
Who is John Snow?



John Snow

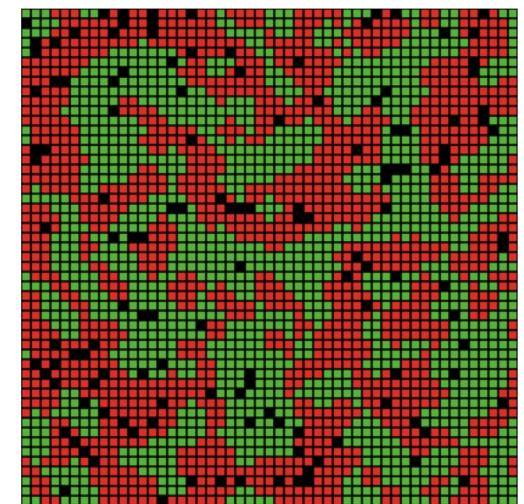
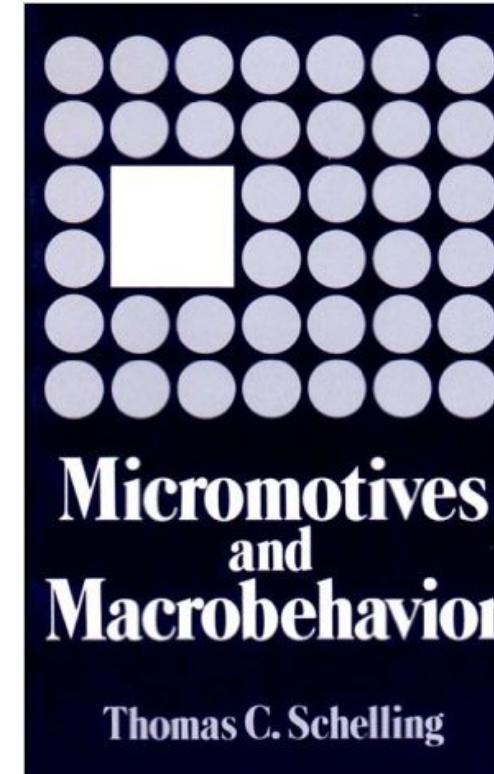


Who is John Snow?



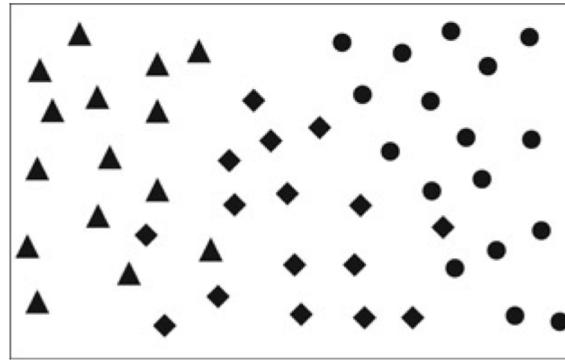
The Schelling Model

- A 1978 model of spatial segregation
 - Book: Micro-motives and Macro-behaviors
- Micro-motive: People move if they are unhappy
- Macro-behavior: Spatial segregation
- Have you seen any other patterns?

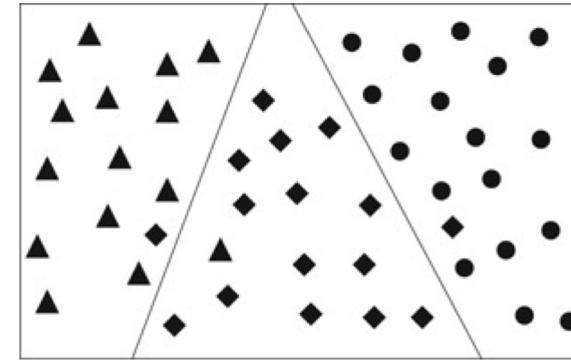


Agent Based
Modelling in
Netlogo

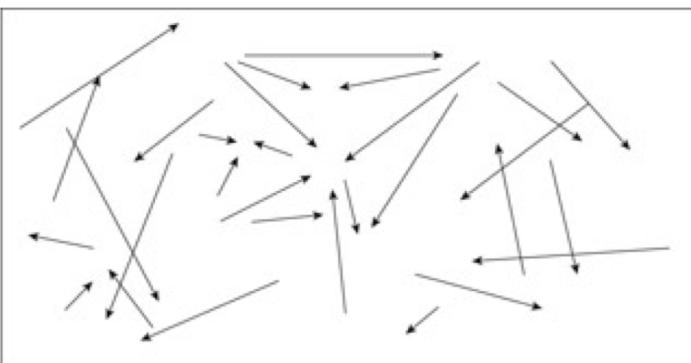
Representations of Spatial patterns



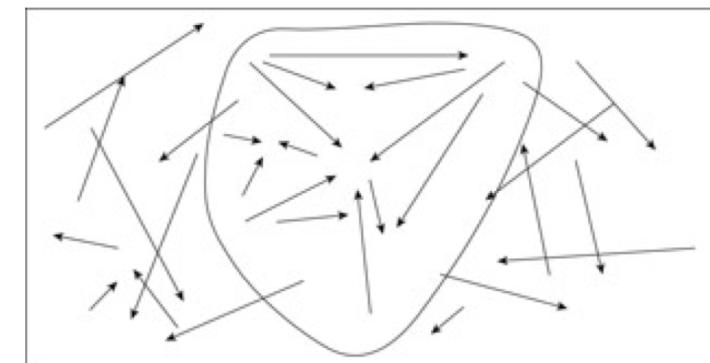
Spatial pattern of economic entities



Delineation of formal regions



Patterns of spatial interaction

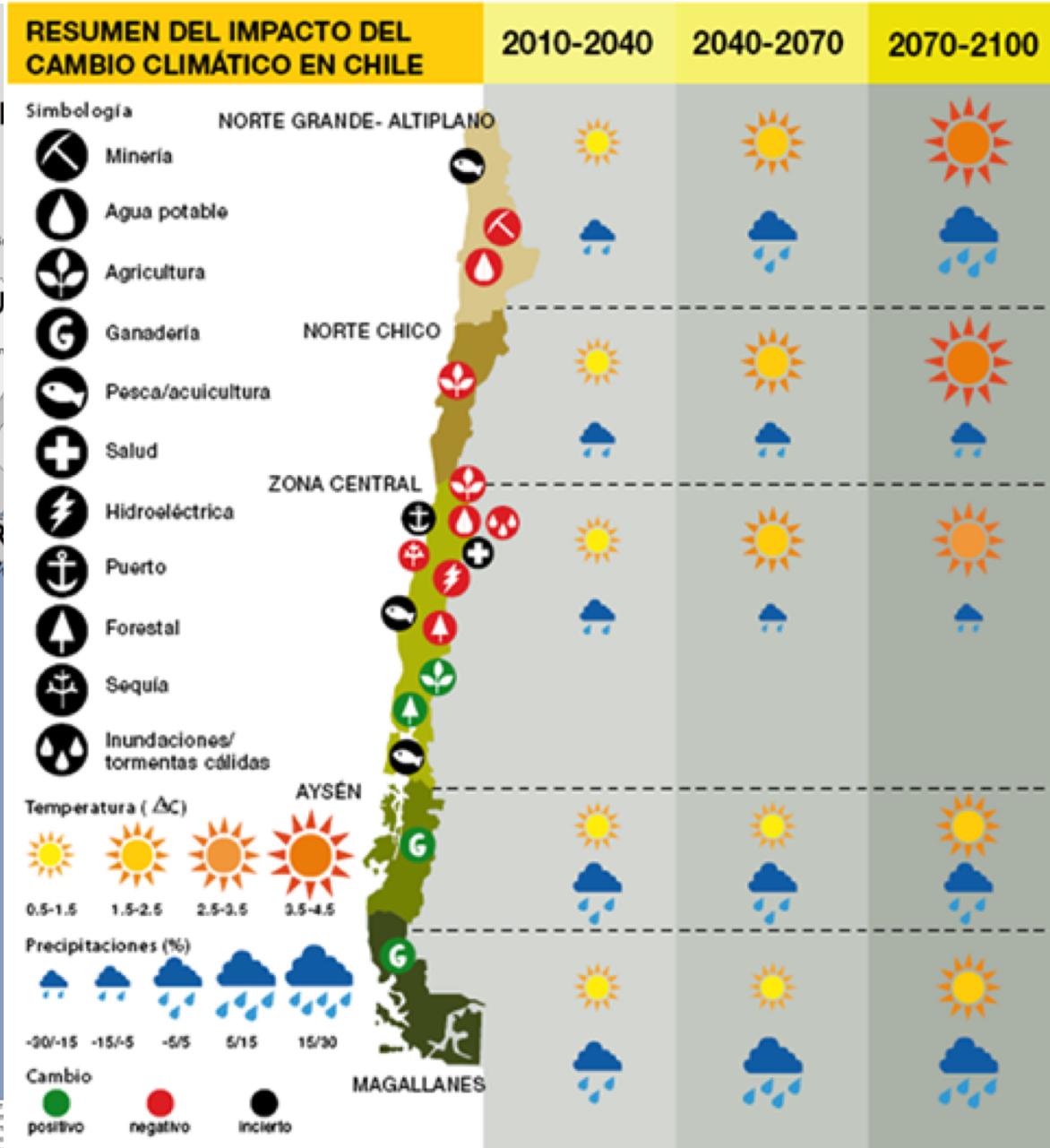


Delineation of functional region

Los agentes económicos (personas), deciden dónde localizarse y actúan en función de su felicidad individual en el espacio dando lugar a patrones espaciales

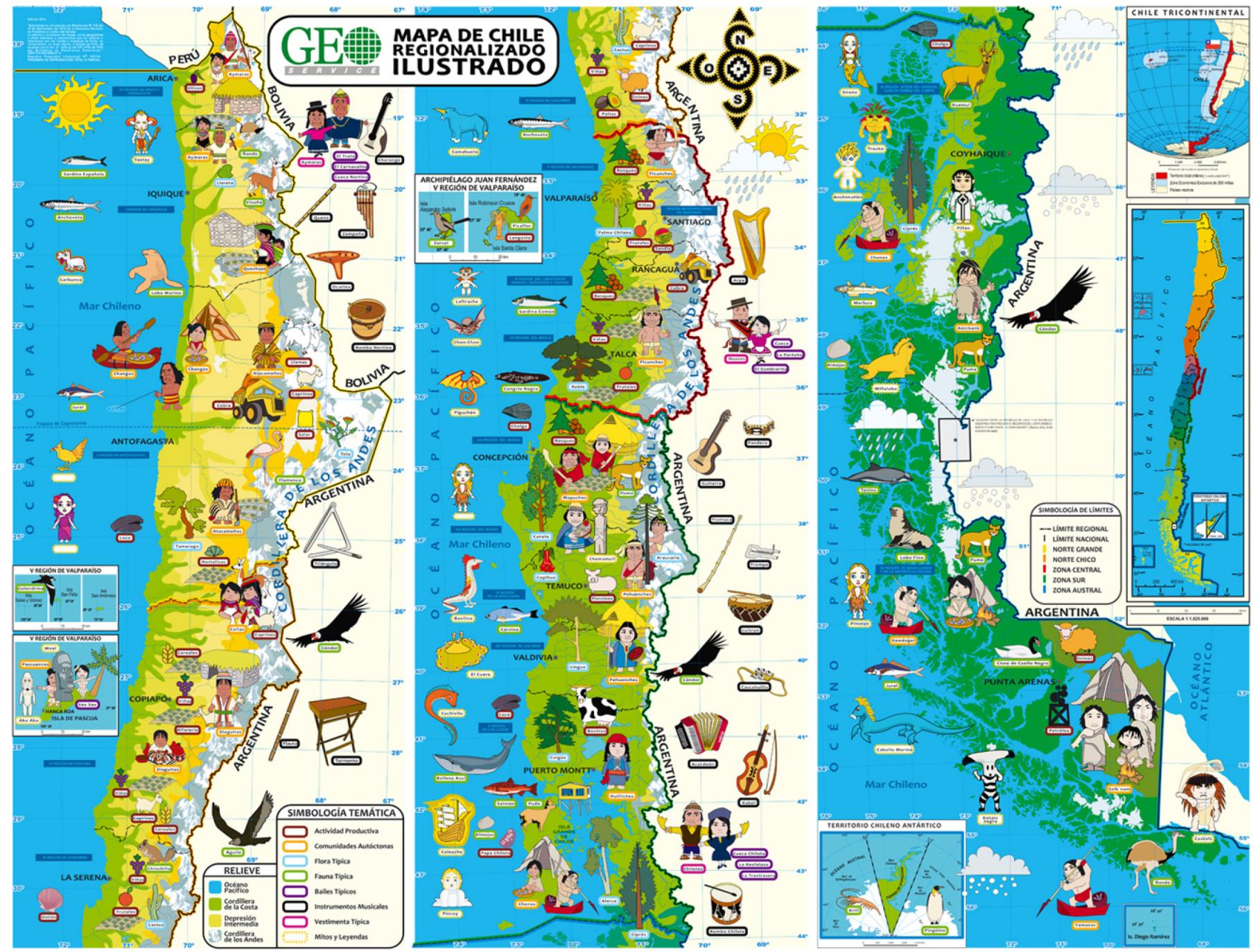


Existe por defecto una disonancia entre las delineaciones administrativas, y cómo los fenómenos (sociales y naturales) ocurren en el espacio.



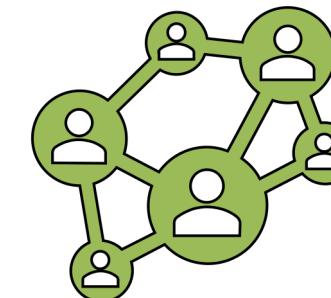


Algunas variables tomarán identidad regional, pero otras curzarán esas fronteras



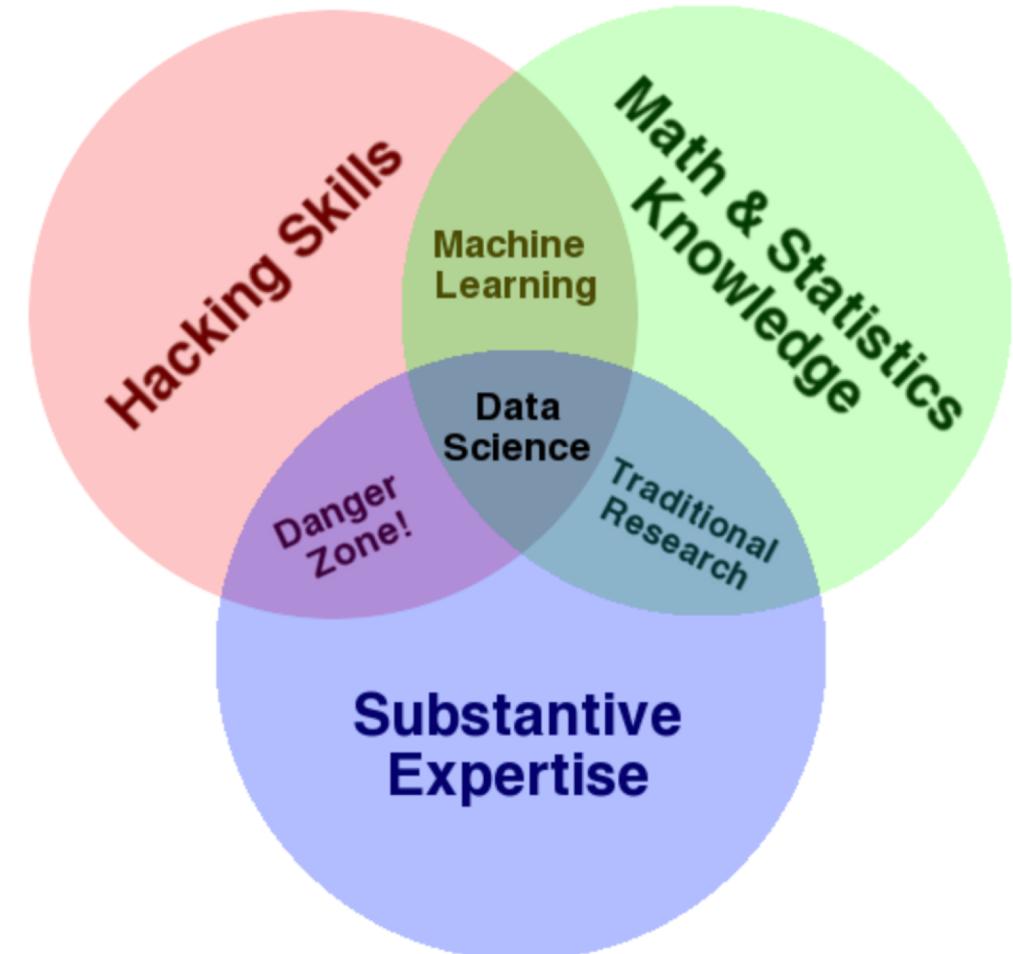
Why Spatial Data Science?

- Substantive
 - From Adam Smith invisible hand to social networks
 - Individual vs. socio-spatial interaction
 - Peer effects, contagion, imitation, trends
 - Spatial externalities
 - Costs/Benefits from activities that impact in a different location
 - Spatial spillovers:
 - Good: neighbor playing the piano
 - Bad: neighbor with COVID-19 and not quarantining
 - Spatial multipliers
 - Investment in a park → housing prices
 - Spatial Mismatch/Disparities
 - Spatial Context
- Practical
 - Data: geo-located observations
 - Private sources
 - Public sources
 - Self reported vs. Web scraped
 - Spatial mismatch between data and social processes
 - Labor markets vs. Cities and Counties → Labor Market areas
 - Epidemic spread vs. Data collection in hospitals
 - Neighborhood Effects
 - Spatial interpolation
 - Change of support problem
 - Data at different spatial levels that don't overlap
 - Ej: Parents' income aggregated at the school level used to predict municipalities' income level

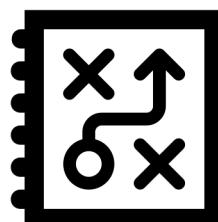
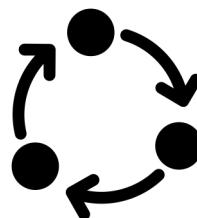
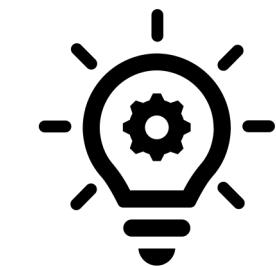


What is “Spatial” Data Science?

- More than just mapping
 - Added value in the explanation
 - Combination of methods, theory, data manipulation
 - Knowledge discovery
 - “from data, to information, to knowledge, to wisdom”
 - Correlation is not causation, and this applies also to spatial analysis.



<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>



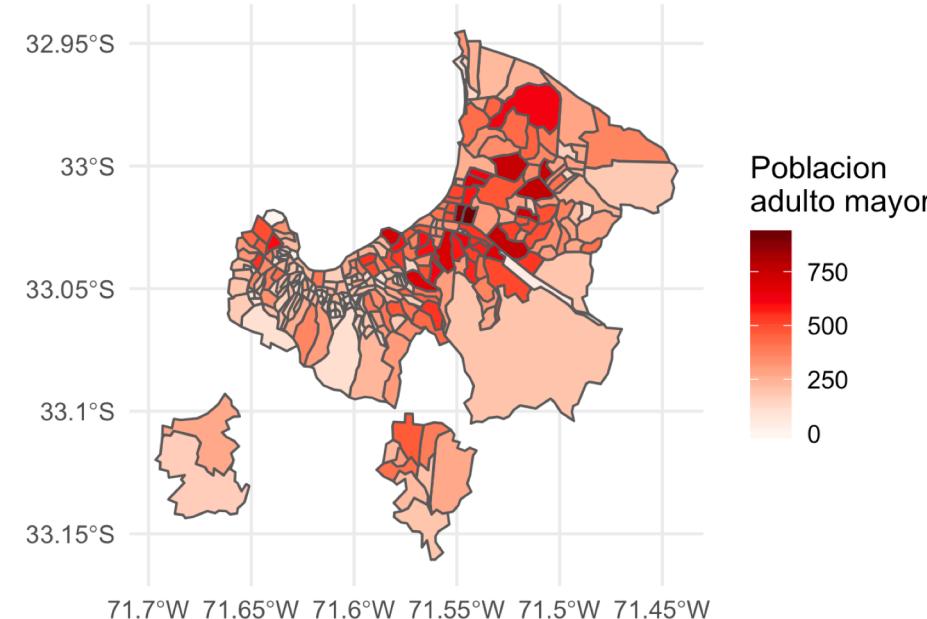
Questions of Spatial Analytics

- Where do things happen
 - Patterns, clusters, hot spots, disparities,..
- Why do things happen
 - Location decisions
- How, things that happen, affect other things (spillovers) and how context affect what happens (interaction)
- Where should things be happening/be located
 - Optimization

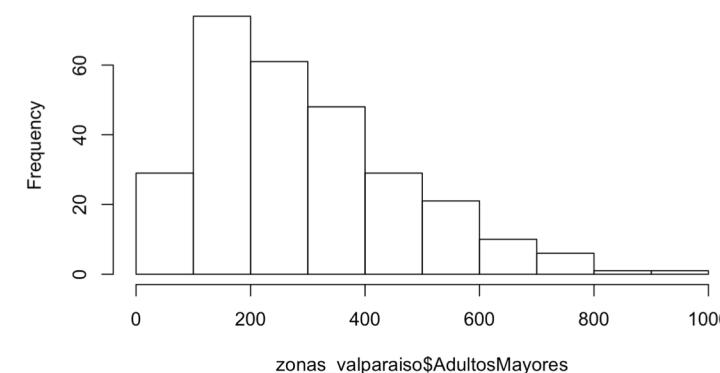
¿Cuál de estos dos mapas es real?

Tener en cuenta la referencia geográfica de los datos es crucial para tomar decisiones en situaciones de movilidad (pandemias)

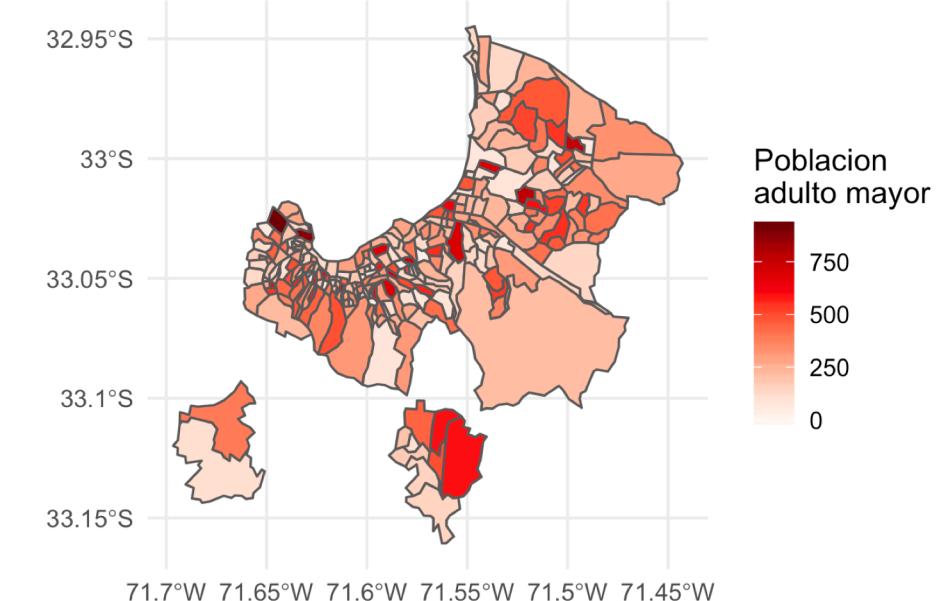
Poblacion de 65 años y más
Valparaíso y Viña del Mar



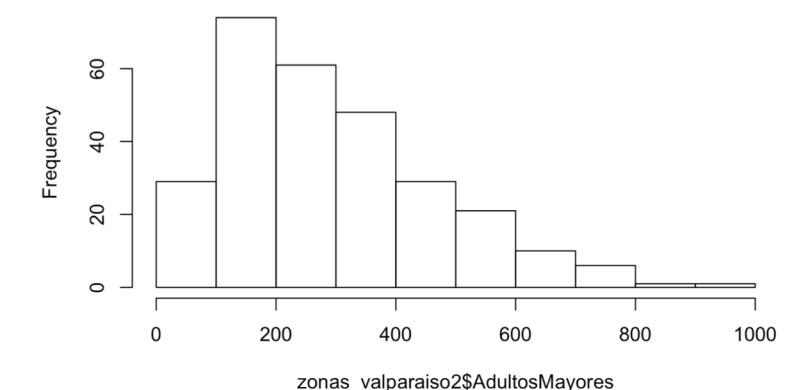
Histograma Adultos Mayores Viña-Valpo



Poblacion de 65 años y más
Valparaíso y Viña del Mar

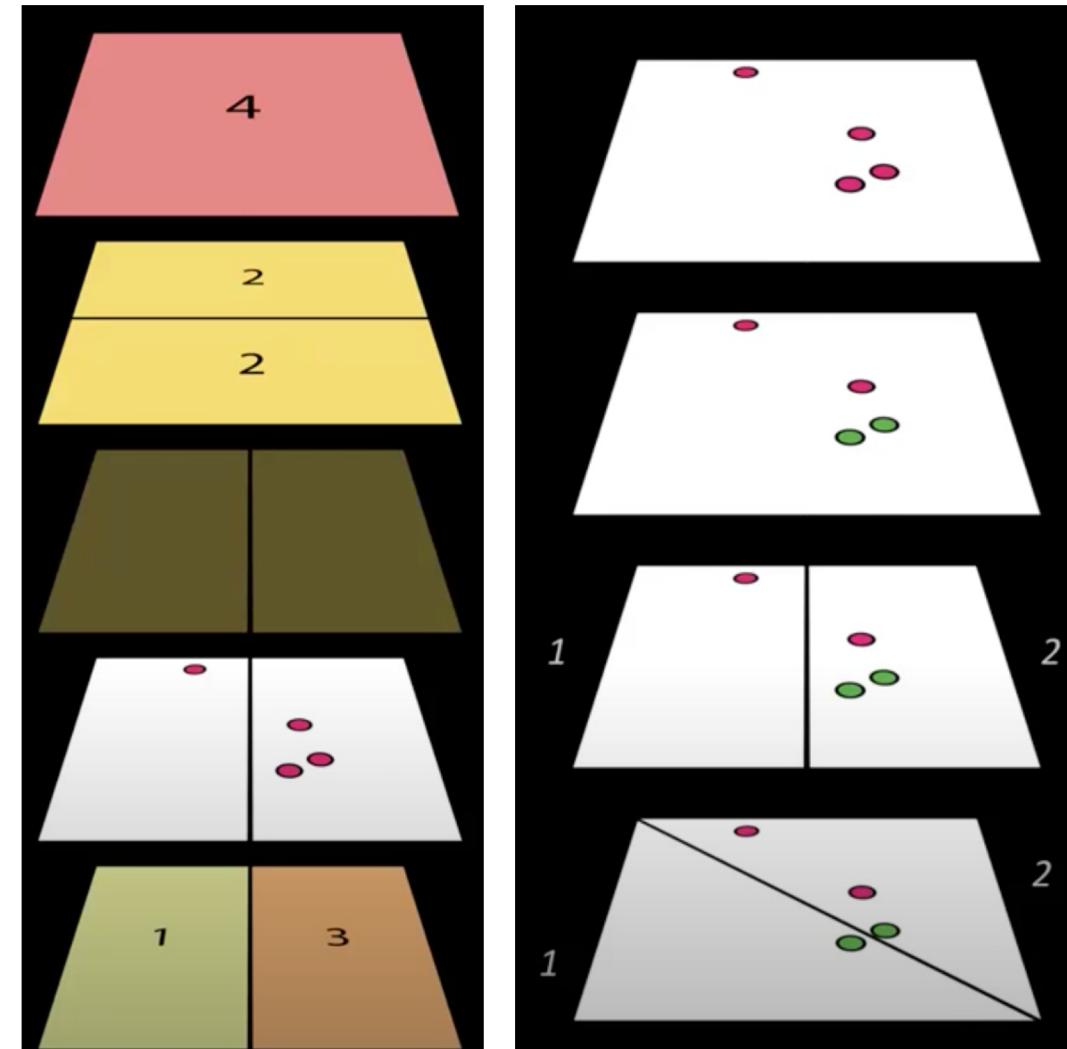


Histograma Adultos Mayores Viña-Valpo



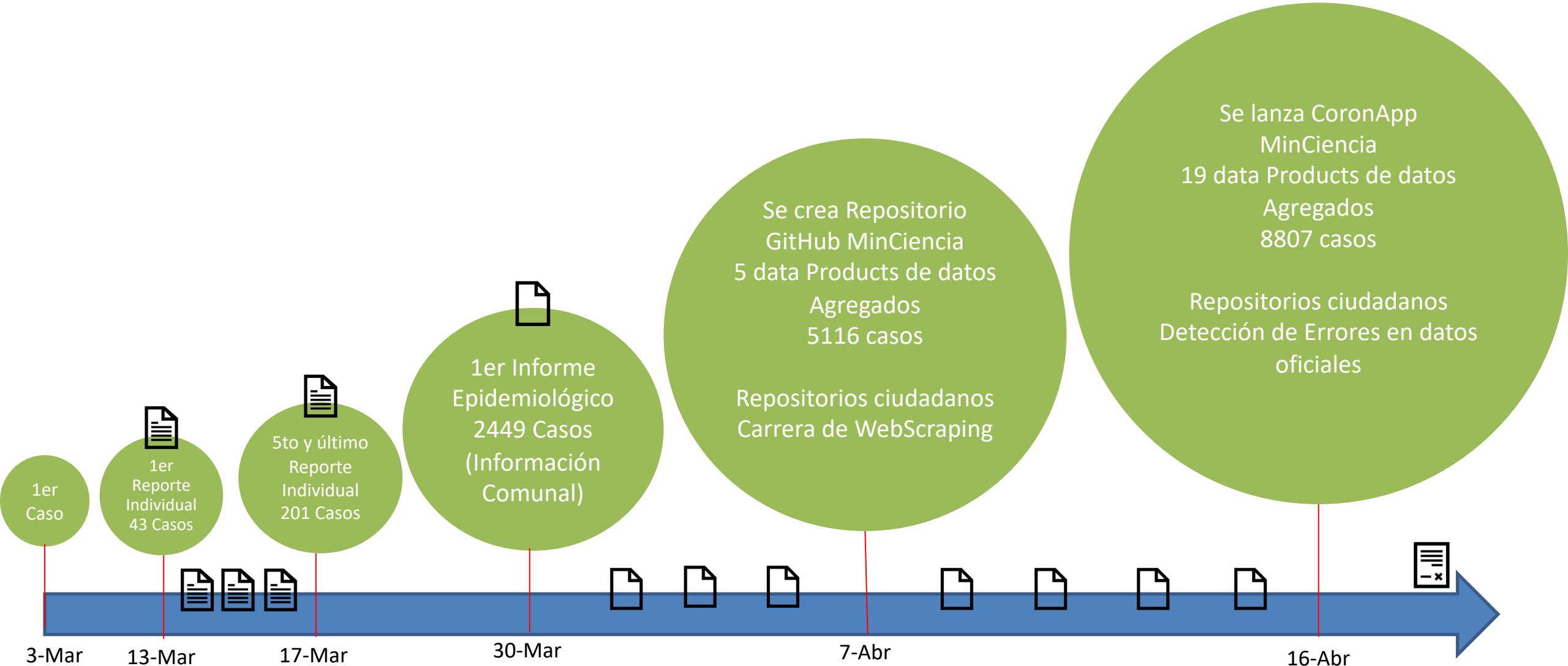
Important questions to ask

- Data source
 - Sampling vs. population?
 - any (spatial) selection bias?
- Spatial units
 - Discrete vs. continuous
 - Is the problem or research question (or variable of interest) measured in the same scale I want to make inference?
 - Are my results changing with the scale of aggregation? (MAUP)
 - Locations are given vs. random
 - Crime data – ‘random’ point locations
 - Purchase data – ‘given’ at store locations
- Ecological fallacy
 - Drawing conclusions about individual behavior from aggregate data analysis
 - Multi-level modelling



DATOS EXISTENTES

Evolución de los datos del COVID19 en Chile





Search or jump to...

/

Pull requests Issues Marketplace Explore

MinCiencia / Datos-COVID19

Watch ▾ 17

Star 92

Fork 72

Code

Issues 14

Pull requests 0

Actions

Projects 0

Wiki

Security 0

Insights

En formato estándar

345 commits

9 branches

0 packages

1 release

6 contributors

MIT

Branch: master ▾

New pull request

Create new file

Upload files

Find file

Clone or download ▾



fzmolina Update TotalesNacionales.csv ...

Latest commit 2a8fdf5 9 hours ago

.github/workflows

updated pipeline to run new code

5 days ago

input

updated ReporteDiario

10 hours ago

output

Update TotalesNacionales.csv

9 hours ago

src

producto23 added

2 days ago

.gitignore

all conflicts fixed

4 days ago

LICENSE

added license

14 days ago

README.md

Update README.md

10 hours ago

requirements.txt

added xlrd to requirements

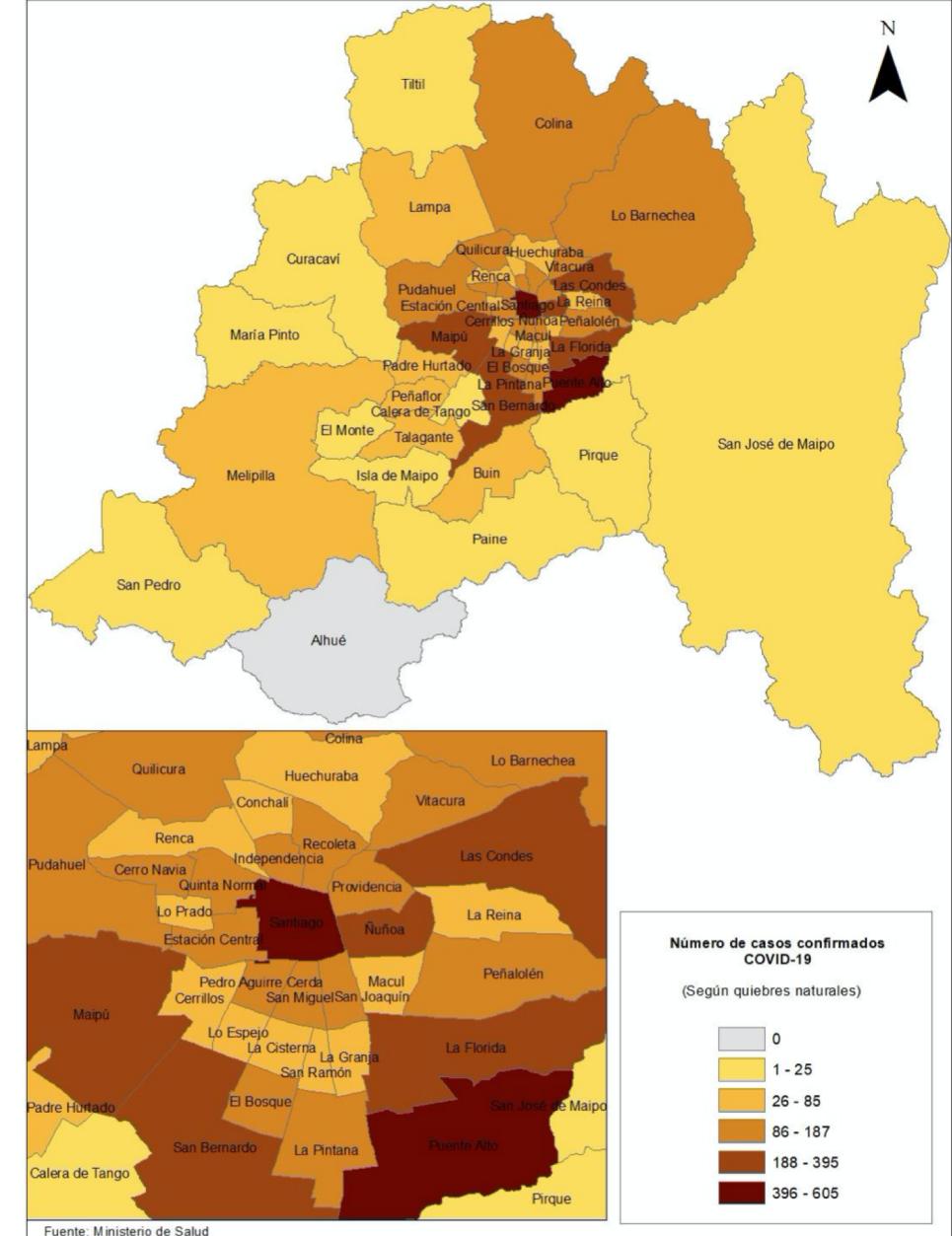
11 days ago



INFORME EPIDEMIOLÓGICO
ENFERMEDAD POR SARS-CoV-2
(COVID-19)
CHILE 20-04-2020

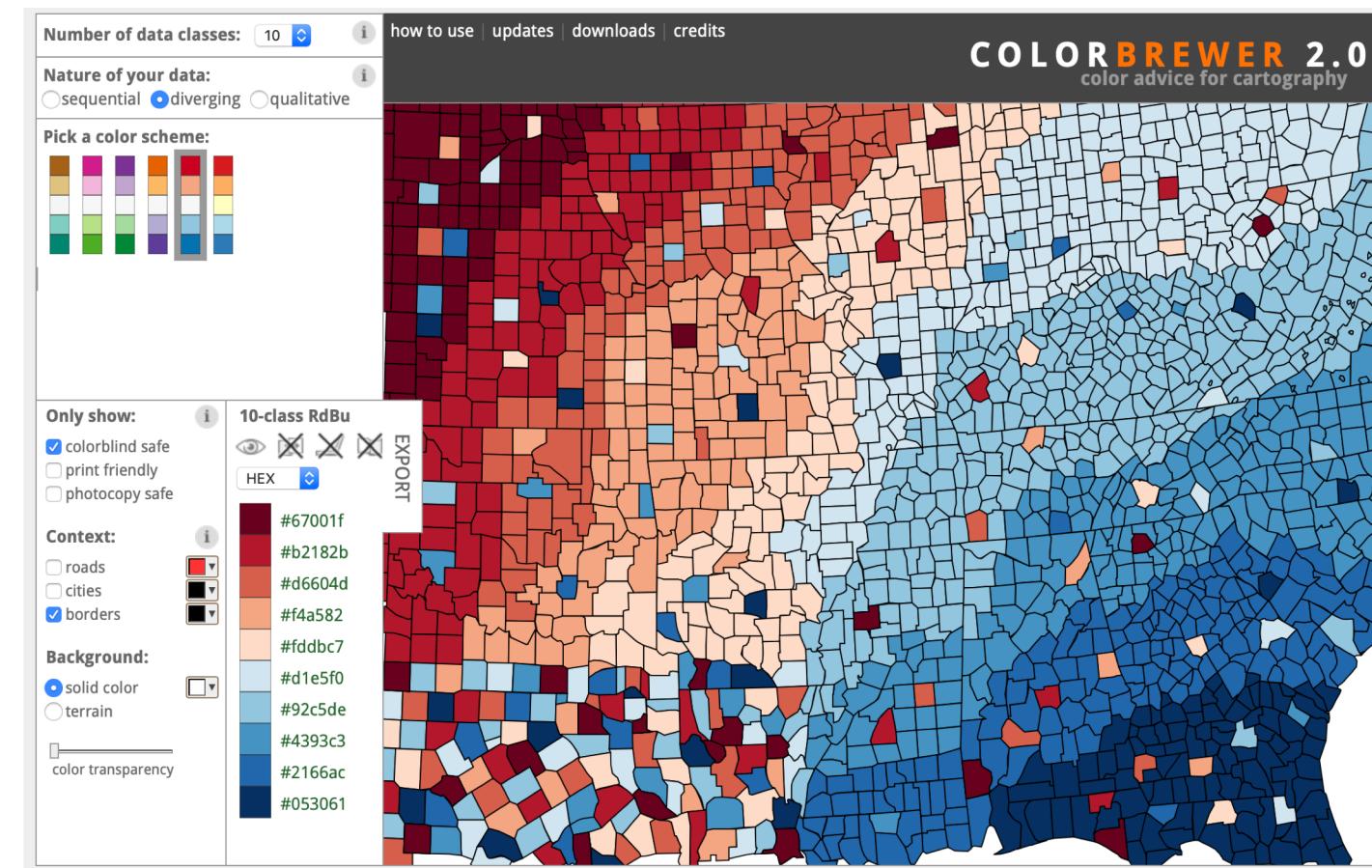
Departamento de Epidemiología

Número de casos COVID-19 según comuna de residencia
Región Metropolitana, 19 de abril de 2020



ESDA: Describe spatial distributions

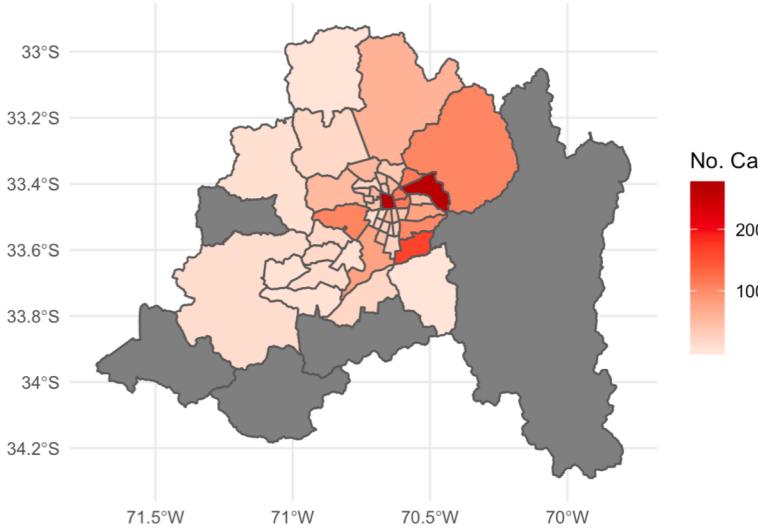
- Map classifications
 - Defined Intervals
 - Manual
 - Equal interval
 - Natural Breaks (Jenks)
 - Statistical based Intervals
 - Quantile (equal share)
 - Standard Deviation
 - Extreme values
 - Keep in mind when mapping
 - Color of choice (colors have meanings!)
 - Projection of choice (distance vs. shape-based representations)
 - Definition of categories (non-ordered, ordinal, divergent, etc.)



La elección de los quiebres importan!

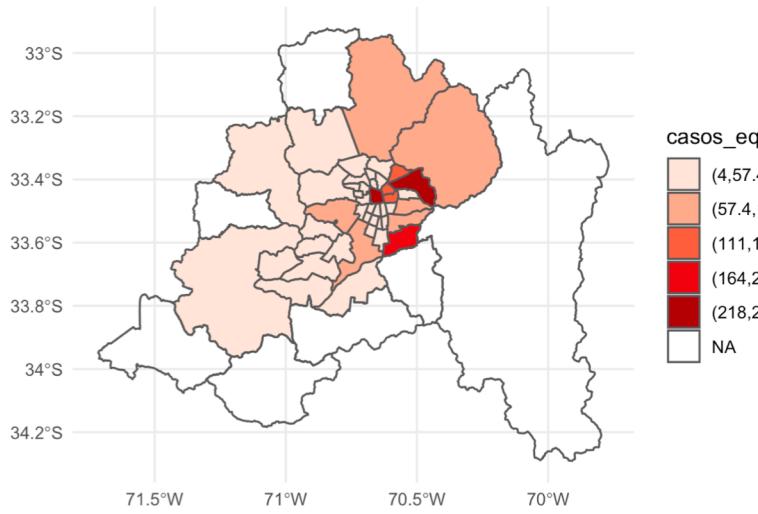
Casos Confirmados

Región Metropolitana - 2020-04-08



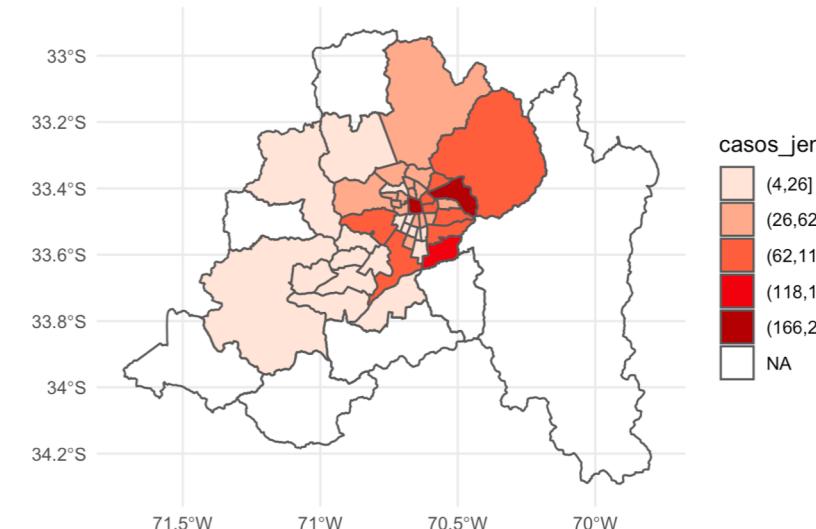
Casos Confirmados

Región Metropolitana - 2020-04-08



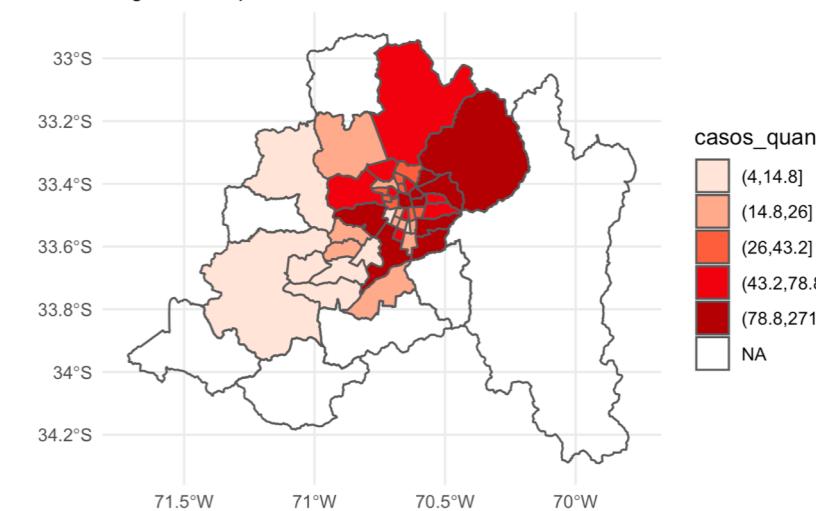
Casos Confirmados

Región Metropolitana - 2020-04-08



Casos Confirmados

Región Metropolitana - 2020-04-08

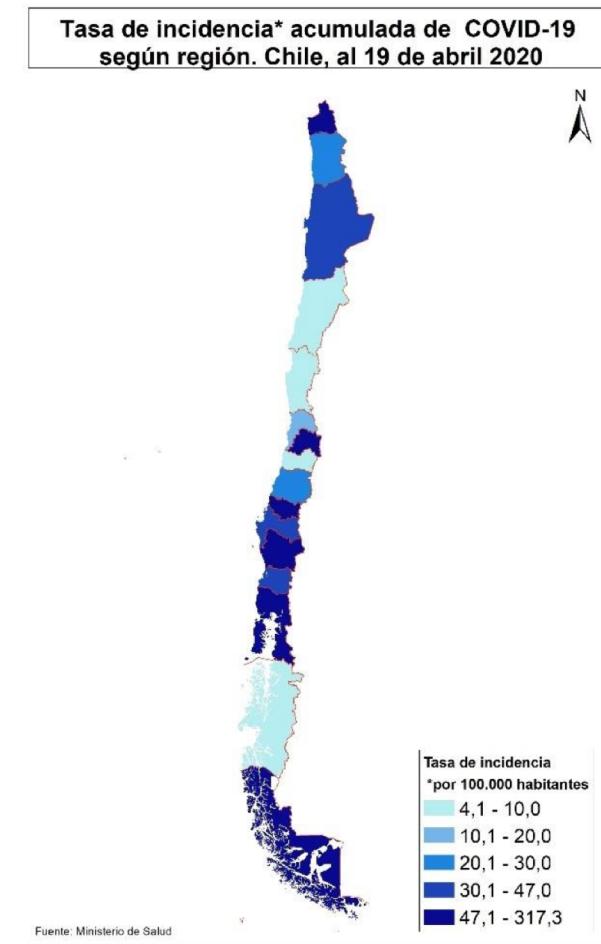
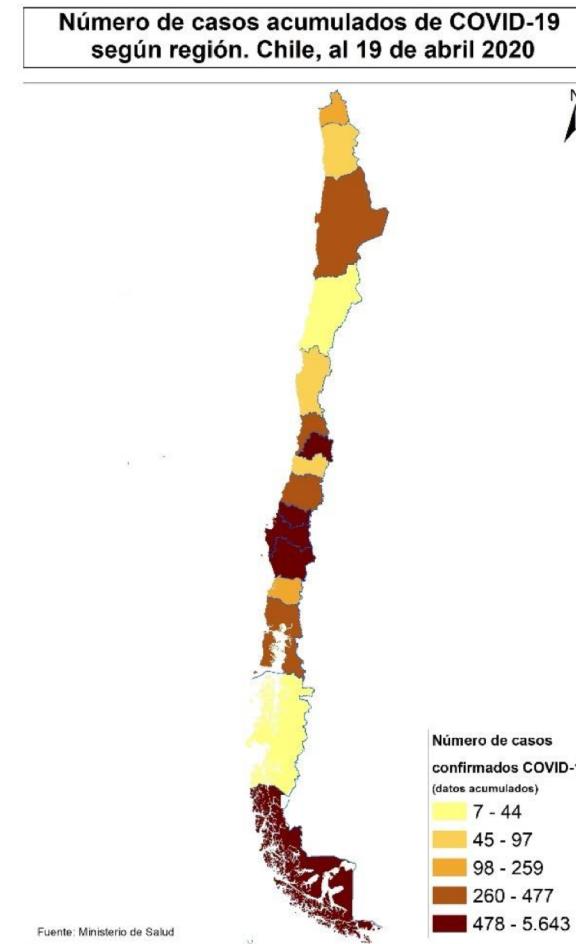


Además de la referencia geográfica de los datos es crucial entender la distribución subyacente y el propósito del análisis para mapear los datos, buscar patrones, y tomar decisiones

Exploratory Spatial Data Analysis (ESDA)

Describe spatial distributions – Mapping rates

- Riesgo y tasas
 - Riesgo: probabilidad de que un evento ocurra
 - El riesgo real no se observa
 - Solo se observa eventos
- Estimación del Riesgo
 - Tasa cruda (raw rate)
 - O_i : No. de Eventos
 - P_i : Población en riesgo
 - r_i : O_i/P_i
 - Foco → Heterogeneidad Espacial
 - El riesgo es uniforme a través del espacio?
 - Dónde hay riesgos elevados?
 - Qué causa esas altas tasas de riesgo?

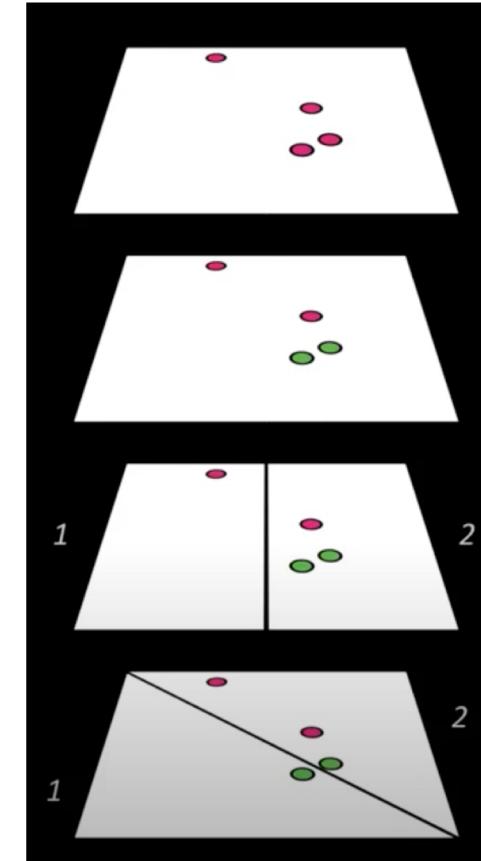
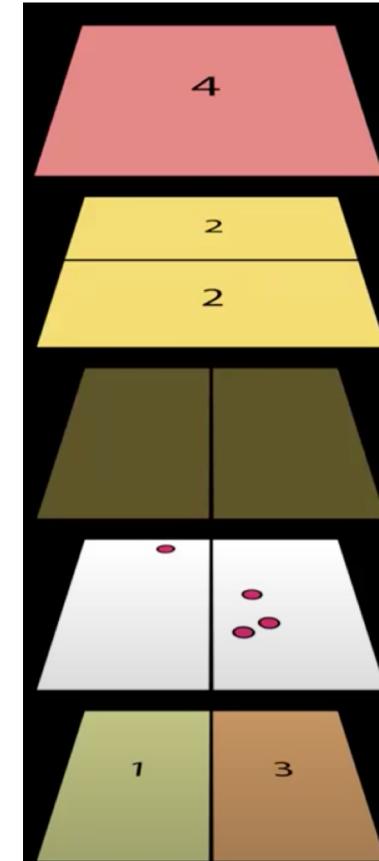


ESDA: Describe spatial distributions –Mapping rates

- Riesgo promedio (AR)
 - $AR = \sum(O_i) / \sum(P_i)$

- Eventos esperados
 - $E_i = AR * P_i$

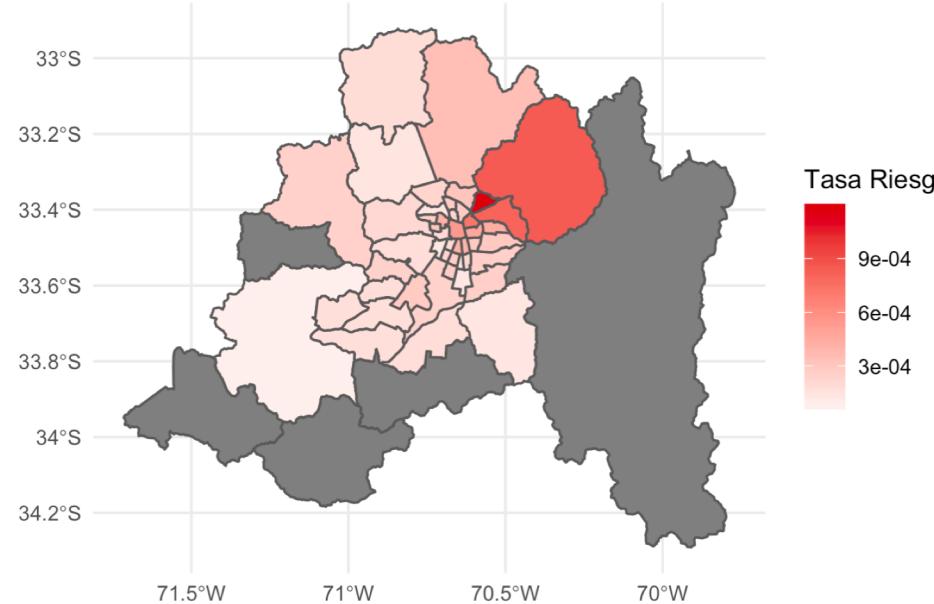
- Riesgo Relativo
 - $R_i = O_i / E_i$
- Exceso de Riesgo
 - Idea básica: riesgo mayor que algún estandar
 - Estandar: algun grupo de referencia
 - e.g: Población de la comuna, Población móvil, etc.
 - $R_i > 1 \rightarrow$ Exceso de Riesgo (más eventos que el promedio)
 - $R_i < 1 \rightarrow$ Bajo Riesgo (menos eventos que el promedio)





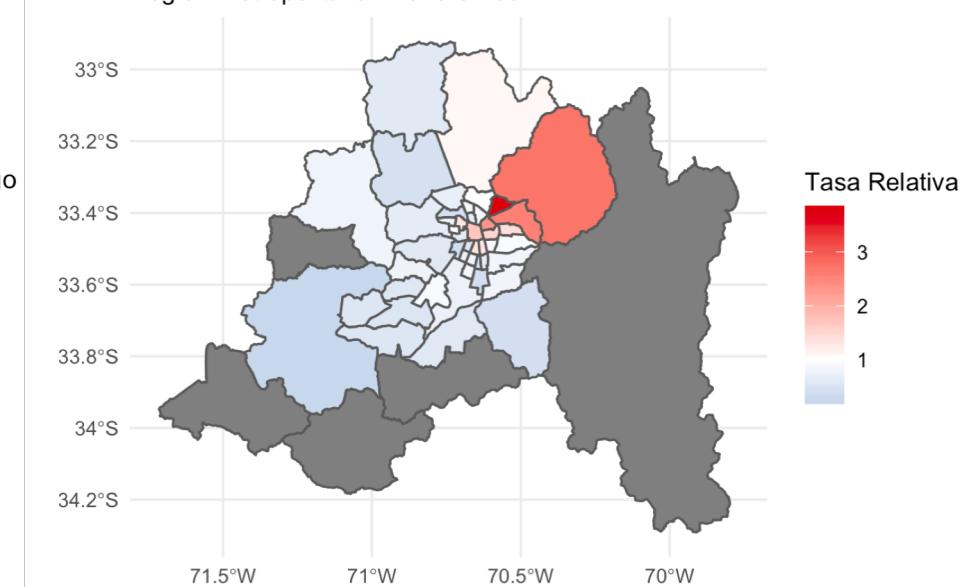
Tasa Cruda de Riesgo

Región Metropolitana - 2020-04-08

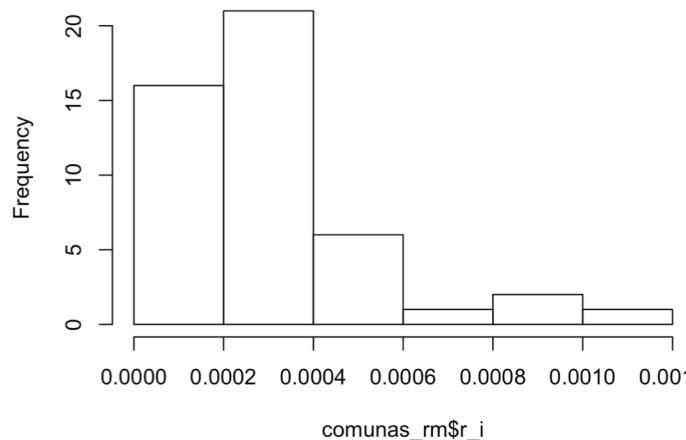


Tasa Relativa de Riesgo

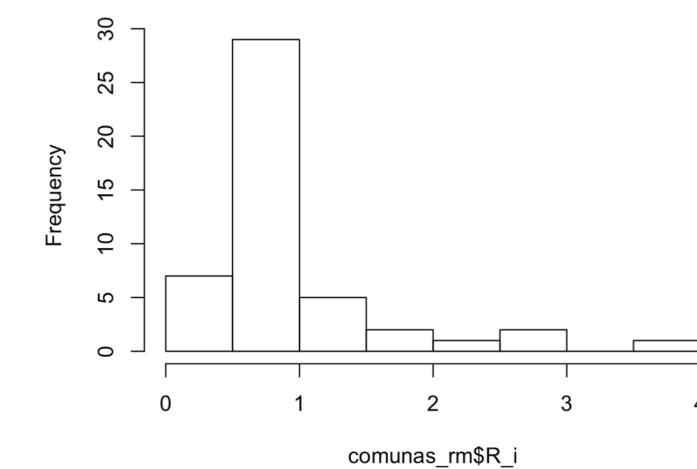
Región Metropolitana - 2020-04-08



Tasa Cruda de Riesgo



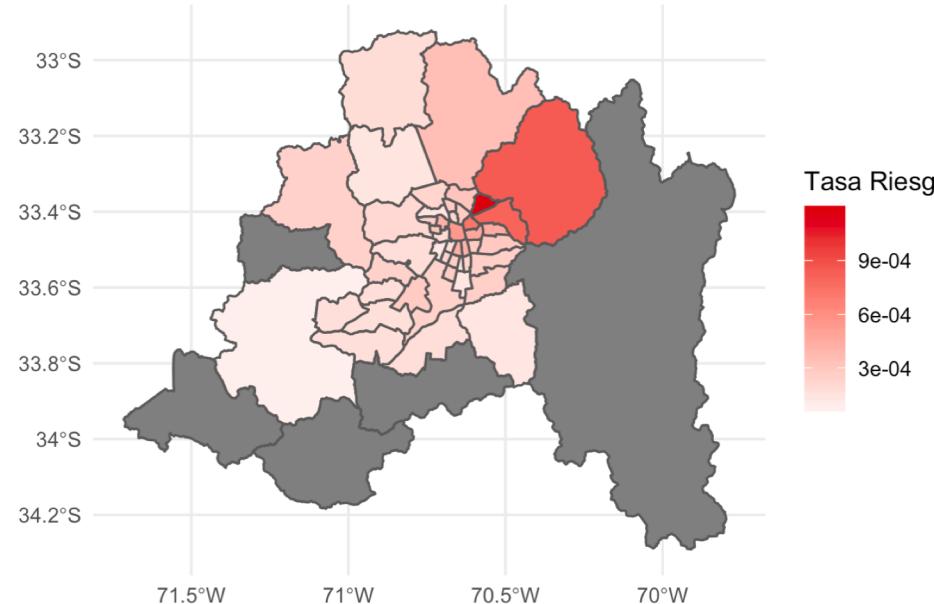
Tasa Relativa de Riesgo





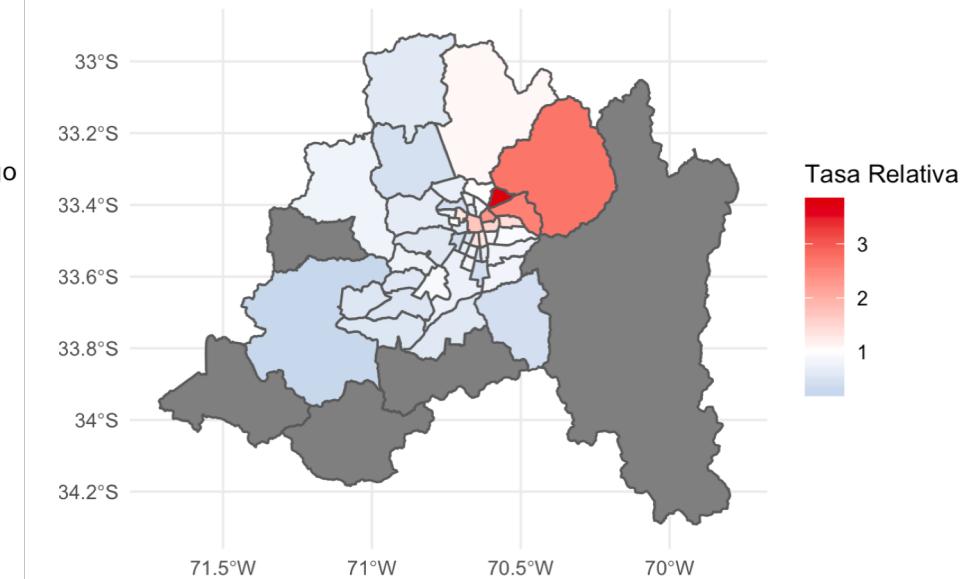
Tasa Cruda de Riesgo

Región Metropolitana - 2020-04-08

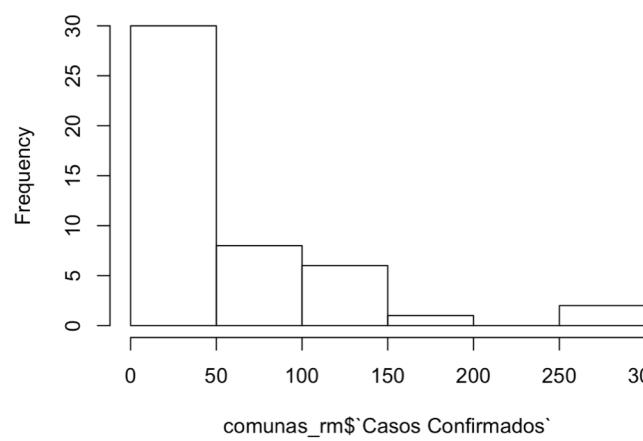


Tasa Relativa de Riesgo

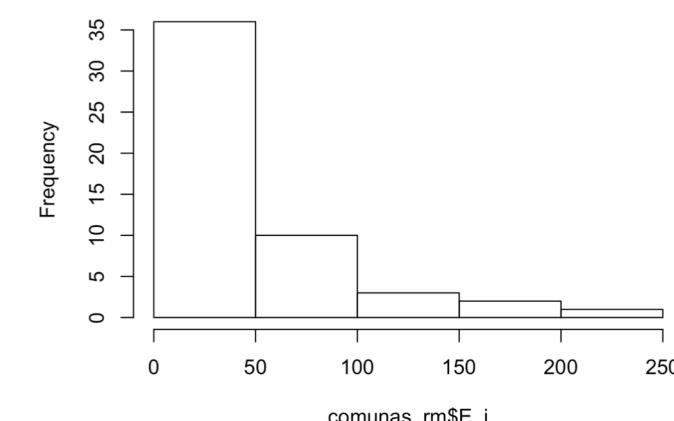
Región Metropolitana - 2020-04-08



Casos Confirmados (Reportados)



Casos Esperados

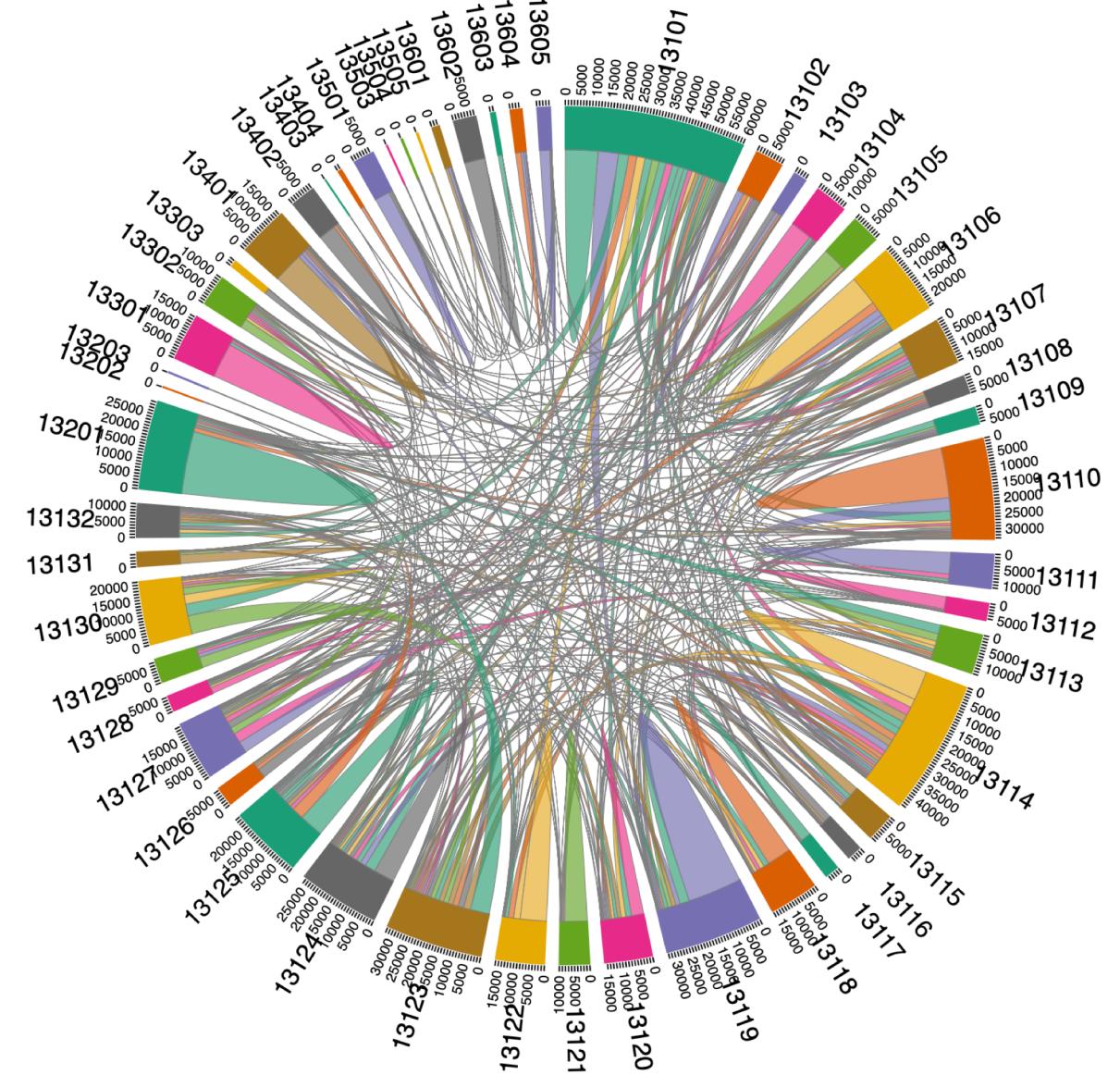


Qué pasa con el análisis si consideramos ...

MOVILIDAD DE LOS TRABAJADORES

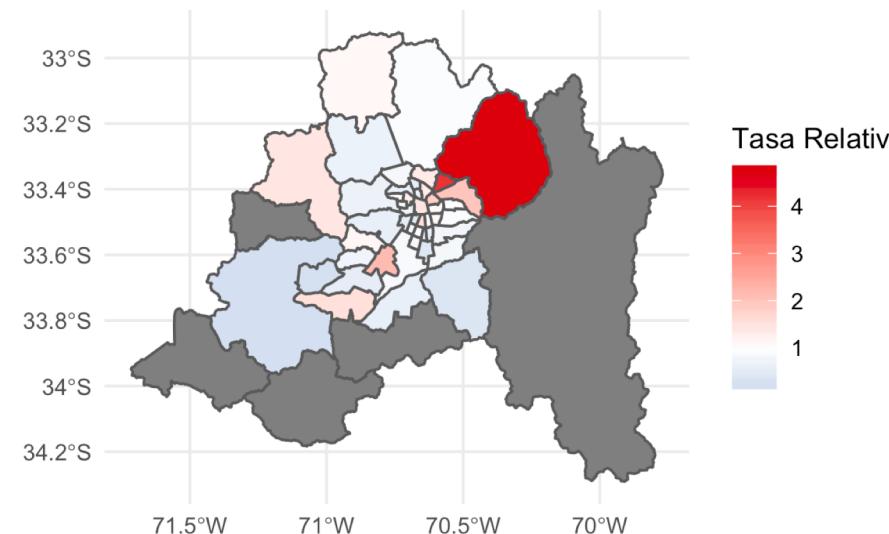
Patrones de Conmutación

- Commutación
 - Proceso de Viaje Casa-Trabajo
 - Distancia, tiempo, ruta, modo de viaje, imprevistos, etc.
 - Muchas personas (especialmente en grandes ciudades) viven en una comuna, pero trabajan en otra comuna.

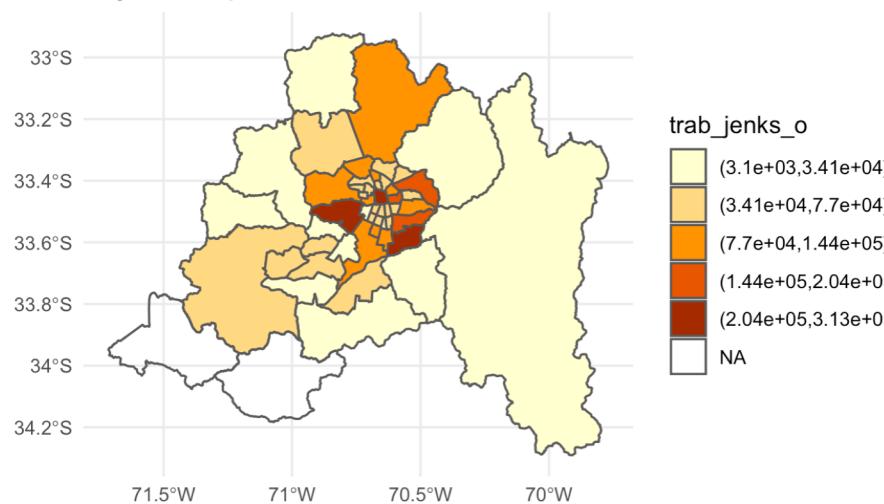




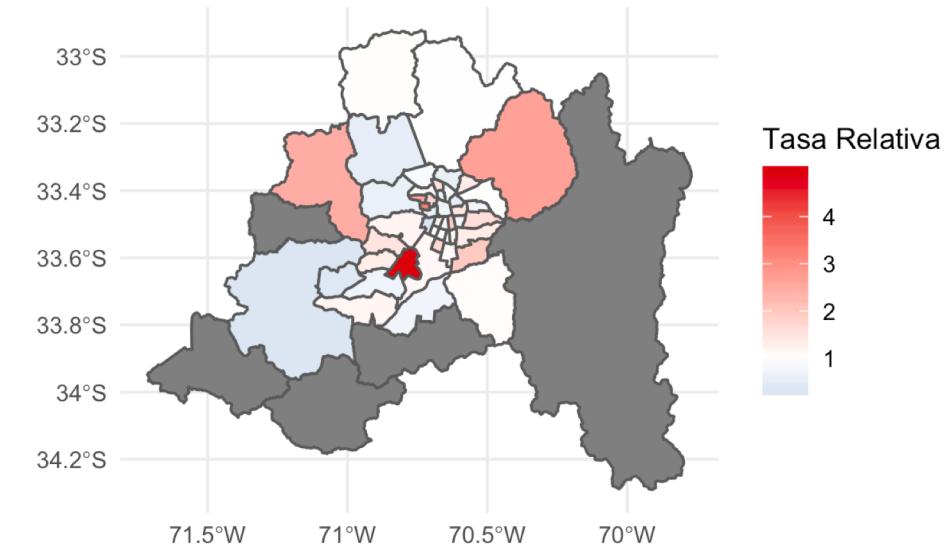
Tasa Relativa de Riesgo - Trabajadores Origen
Región Metropolitana - 2020-04-08



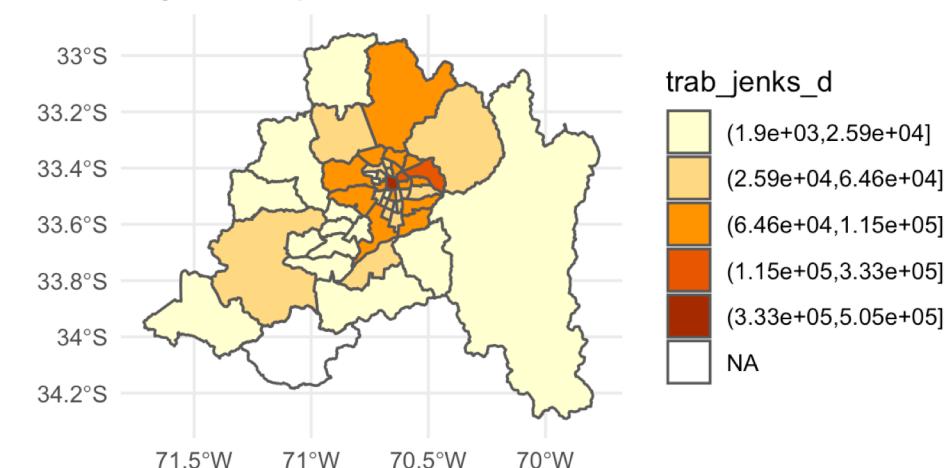
Trabajadores por Origen
Región Metropolitana - 2020-04-08



Tasa Relativa de Riesgo - Trabajadores Destino
Región Metropolitana - 2020-04-08



Trabajadores por Destino
Región Metropolitana - 2020-04-08

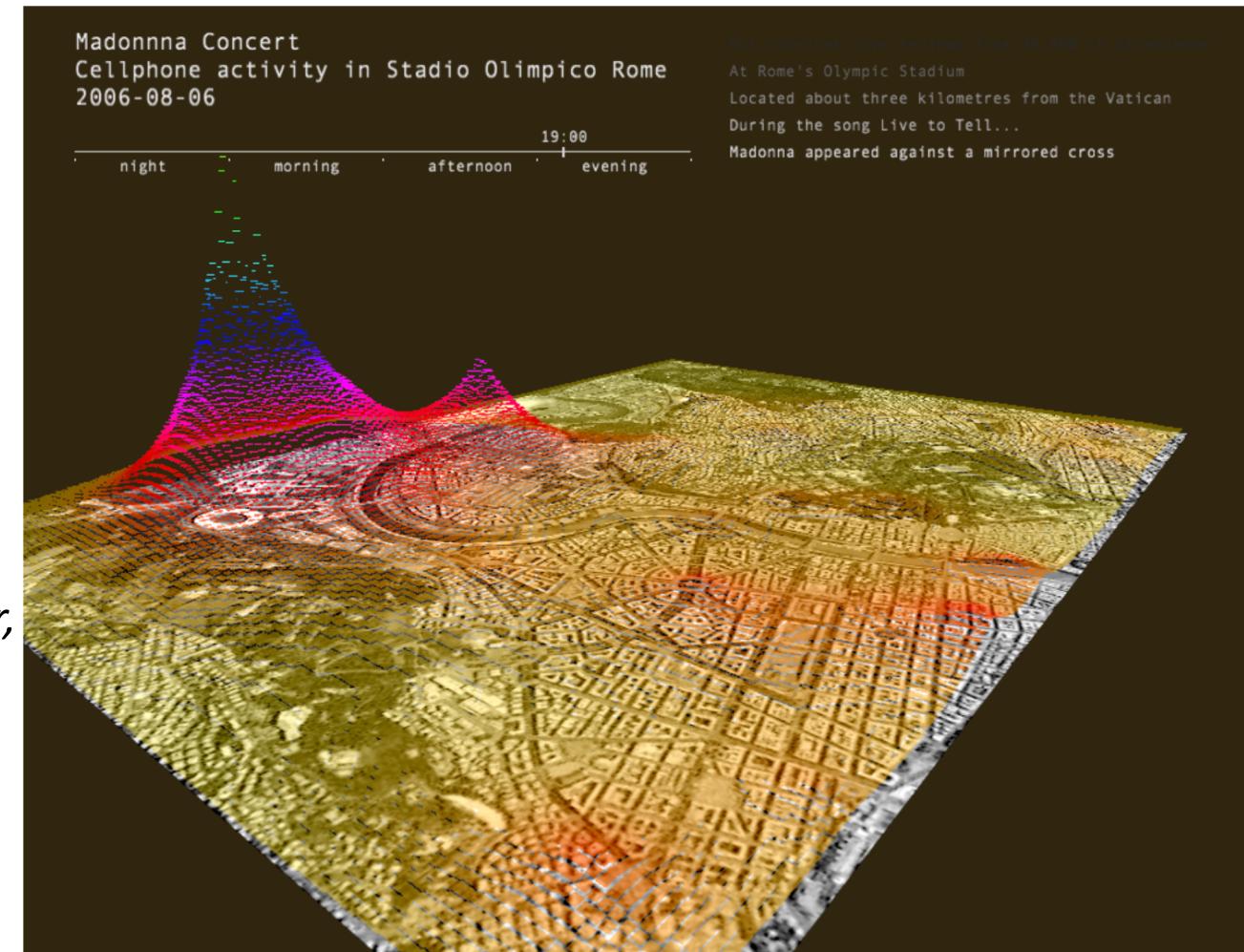


¿Qué más se podría hacer con mejores datos?

DATOS IDEALES

Datos Ideales

- Atributos
 - Desagregados a nivel individual
 - Desagregados a nivel espacial
 - Ubicación del punto, o manzana
 - Vinculables con variables relevantes
 - Sexo, Edad, Fecha de Síntomas, Fecha de Examen, Lugar de Examen, *sector laboral, condiciones de habitabilidad, riesgo financiero, situación demográfica del hogar, etc.*
 - Trazabilidad
 - Aseguramiento de la veracidad del dato
 - Fechas confiables de publicación
 - Confidencialidad
 - Aseguramiento de la confidencialidad de los datos de los pacientes y familiares



<http://senseable.mit.edu/realtimerome/>



Acerca de los datos en esta página

1. Los datos que publica diariamente el INS se obtienen tras recopilar la información de pruebas realizadas por los laboratorios autorizados en el país. Estos datos se consignan en el aplicativo SisMuestras, donde se analizan y se descargan para compartir con la ciudadanía.
2. Todos los días, la tabla de datos abiertos será publicada antes de las 6pm. Los cambios que ocurren posteriormente se verán reflejados en la tabla del día siguiente. Si llega a existir la necesidad de un ajuste extraordinario, se hará a las 9pm de ese día. Se anunciará que se trata de un ajuste extraordinario. Con esto se evita que ajustes posteriores, no se vean reflejados en la información que publica Minsalud con corte único a las 4pm.
3. Las tablas de datos individuales son un reporte preliminar de vigilancia epidemiológica. Esta información es publicada en tiempo real y extraída del total de casos probables reportados. A medida que las entidades territoriales amplían el trabajo en campo y la información reportada a las instituciones de salud, la tabla es actualizada.
4. Antes del 15 de abril, el sistema SisMuestras no existía. Por ello, tras el inicio de operación de este aplicativo, los laboratorios cargaron datos de muestras que no habían reportado y por ello las discrepancias registradas. Desde el 15 de abril se utiliza como única fuente el reporte diario de pruebas procesadas en SisMuestras.
5. Desde el 22 de marzo, hay otros laboratorios diferentes al INS y al LDSP de Bogotá realizando pruebas diagnósticas del SARS-CoV2. Todos estos laboratorios tienen la responsabilidad de cargar periódicamente la información al aplicativo SisMuestras y deben hacer las correcciones necesarias según sus resultados.
6. Las correcciones fruto de errores de digitación o notificación se corregirán posteriormente. Cada noche se publicará una fe de erratas que hará explícito el error corregido y la fecha de modificación.
Por esto, es importante tener presente que los datos seguirán cambiando de manera periódica, pues por nuestro Sistema de Vigilancia irá mejorando la calidad de los mismos. Esto sucede en todos los países y es un proceso natural de la recolección de datos. Los datos definitivos serán cerrados 12 meses después del fin del registro.
7. Los datos de Casos Diarios deben siempre analizarse con la fecha registrada en la columna FIS (Fecha de Inicio de Síntomas). La fecha de diagnóstico no refleja el inicio del caso. Es importante tener en cuenta que esta fecha de diagnóstico se ve afectada por (a) el tiempo que toma una muestra en llegar al laboratorio; (b) el tiempo de procesamiento de la muestra en el laboratorio; (c) el tiempo entre el procesamiento de la muestra y la expedición del resultado. Las fechas de diagnóstico, además, están sufriendo variaciones toda vez que los laboratorios están haciendo revisión y actualización de sus datos en el nuevo sistema SisMuestras.
8. La gráfica de acumulado diario de pruebas realizadas se basa en la información cargada por todos los laboratorios autorizados para realizar el diagnóstico de SARS-CoV2 en el país. En el momento, los laboratorios de Bogotá y Antioquia se encuentran completando las fechas de realización de pruebas, por lo que la gráfica variará en la medida en la que se completen los datos.
9. El ID de la tabla corresponde al número de identificación de caso en el orden en el que el INS recibe la alerta por parte de una entidad territorial. La fecha de confirmación del caso no tiene un orden cronológico en relación a este ID, pues esta fecha responde al momento en el que el laboratorio específico da el reporte de la muestra.
10. Es importante saber que hay 4 fechas en la tabla:
 1. Fecha de reporte a Sivigila -Fecha Notificación
 2. Fecha de diagnóstico (la da el laboratorio)
 - Fecha Diagnóstico
 3. Fecha de inicio de síntomas - FIS
 4. Fecha de reporte en Web-Fecha Web

*la fecha de diagnóstico hasta hace 14 días coincidía con el registro en web. Ahora ha variado por la presencia de los nuevos laboratorios que cargan en diferentes momentos sus resultados en SisMuestras.

Por eso se crea la variable fecha Web separada de la fecha de diagnóstico.
11. A continuación publicamos el diccionario de variables para que se pueda comprender el significado de cada una y asegurar el uso correcto de las mismas.

Caso 1: Colombia

Datos abiertos individuales sin ubicación desagregada



Caso 2: China

Datos abiertos individuales, con ubicación desagregada e historia de contactos

www.nature.com/scientificdata/

SCIENTIFIC DATA

OPEN DATA DESCRIPTOR **Epidemiological data from the COVID-19 outbreak, real-time case information**

beoutbreakprepared / nCoV2019

Code Issues 26 Pull requests 2 Actions Projects 0 Wiki Security 0 Insights

Location for summaries and analysis of data related to n-CoV 2019, first reported in Wuhan, China

232 commits 8 branches 0 packages 0 releases 9 contributors MIT

Branch: master New pull request Create new file Upload files Find file Clone or download

tbrewer-healthmap Merge pull request #57 from beoutbreakprepared/dev ... Latest commit 32c3411 7 hours ago

.github/workflows code requirements 6 days ago

bulletins_sitreps Further updates to time series and start of some US States 19 days ago

co-morbidities Link Typo Fix 14 days ago

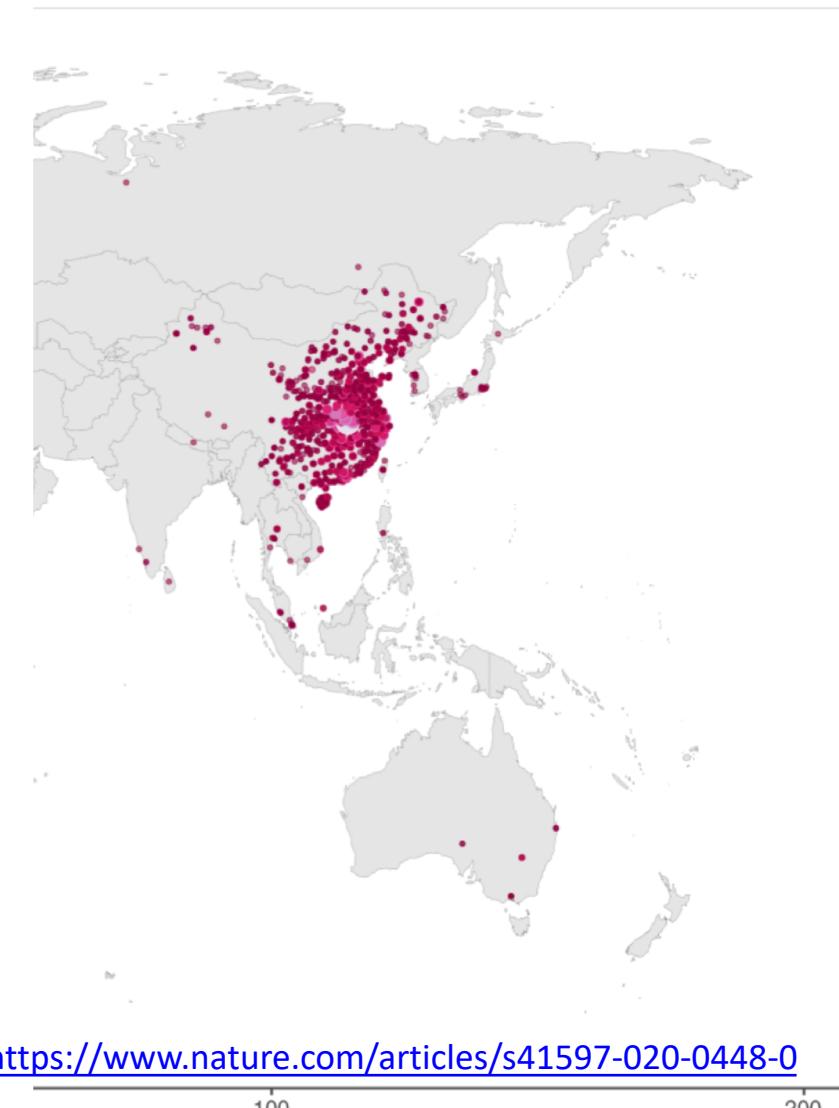
code line count check 8 hours ago

covid19 Update README.md 2 months ago

dataset_archive data update 17 hours ago

demographics Added GBD demographics 2 months ago

error_reports data update 17 hours ago

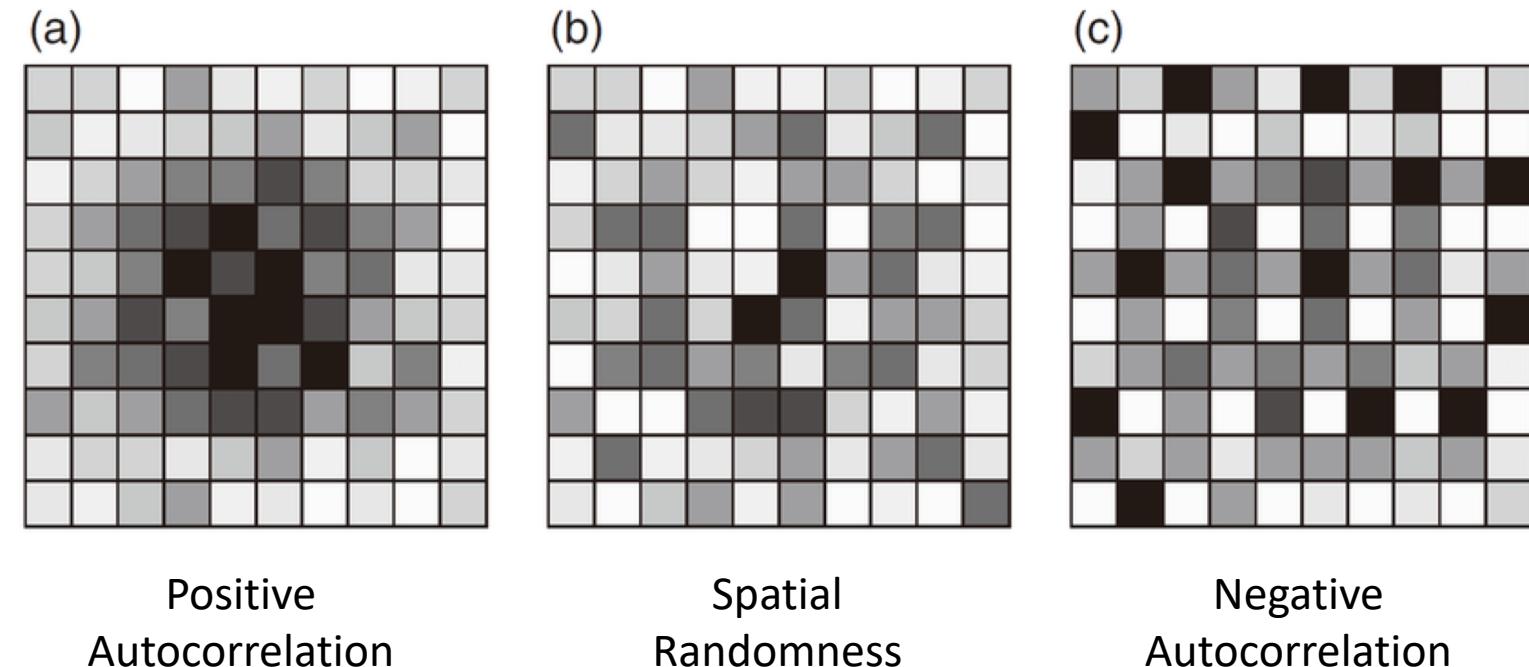


Que cosas se podrían decir con Spatial Data Science?

- ESDA - Spatial Dependence
- Spatial point pattern analysis
 - John Snow – patrones de puntos aleatorios?
- Spatial inference (regression)
 - EJ: Explicación de cómo otras variables contribuyen al contagio (hacinamiento, temperatura, comportamiento, etc)
- Spatial simulation (interpolation)
- Space-time dynamics

Spatial association – Spatial Dependence

- Spatial randomness (null)
- Spatial autocorrelation
 - Clustering pattern vs. Finding Clusters
 - Positive
 - Negative
- Common issues
 - Scale of Aggregation - MAUP
- Statistics
- Spatial weights



Source:

https://www.researchgate.net/publication/327345078_Modelling_Irregular_Spatial_Patterns_using_Graph_Convolutional_Neural_Networks

Spatial 'auto'-correlation?

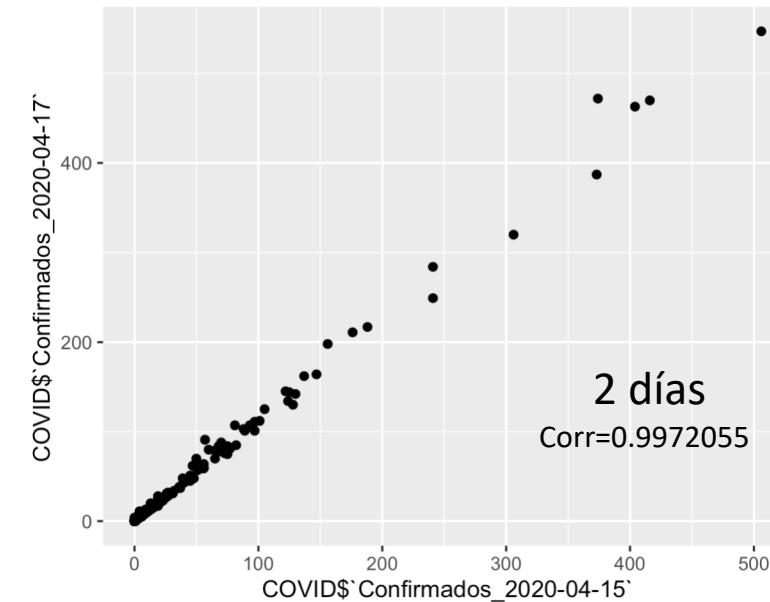
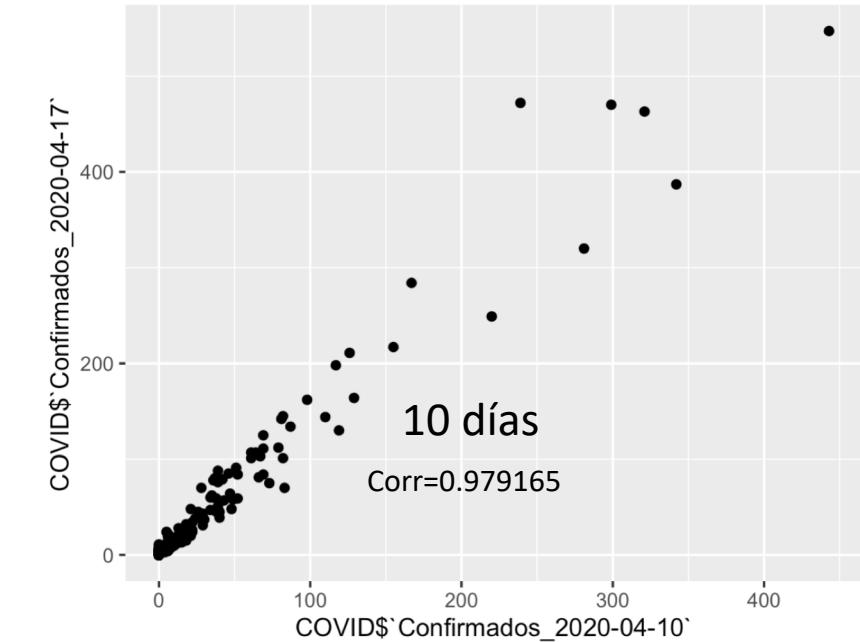
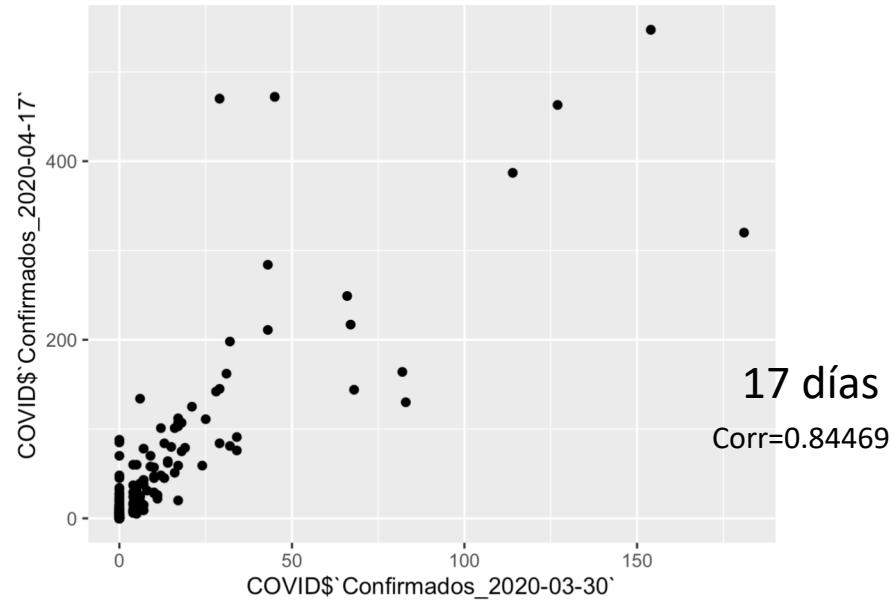
Attribute similarity

- Time: correlation of a variable with a realization of the same variable at a different time
- Space: Correlation of a variable with itself, considering cases at different locations
 - Main idea: is there any correlation between values of a variable considering their locational similarity?
- Similarity measures
 - Cross-product:
 - y_i, y_j . (linear) larger values are closer → large product
- Dissimilarity measures:
 - Squared differences: $(y_i - y_j)^2$
 - Absolute differences: $|y_i - y_j|$
 - the smaller the more similar



Time correlation

*¿Qué tanto el
pasado puede
decir sobre el
presente?*

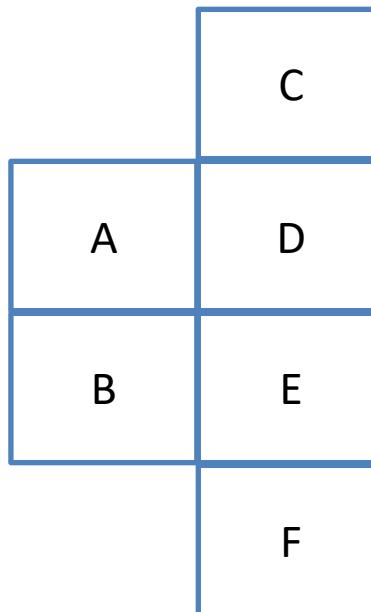


Spatial ‘auto’-correlation?

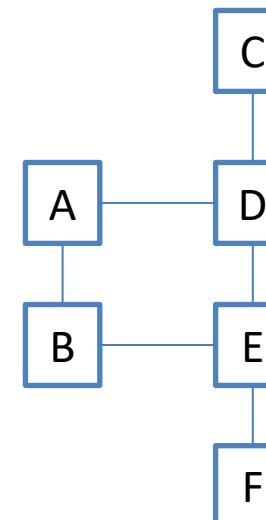
Statistic - spatial weights

- Capturing interaction
 - n spatial units, can be interacting with all n observations
 - This is an n^2 problem, but we only have n observations
 - Classic solution
 - Let’s impose a structure of how interaction works → spatial structure
 - Exclude some interactions
 - Capture in a single parameter → spatial autocorrelation coefficient
- Spatial weights
 - Let W be an n by n matrix indicating the level of association between n spatial observations w_{ij}
 - Let $f(y_i, y_j)$ be an attribute similarity measure between i and j
 - Then:
 - Statistic: $\sum_{ij} f(y_i, y_j) * w_{ij}$

Spatial structure

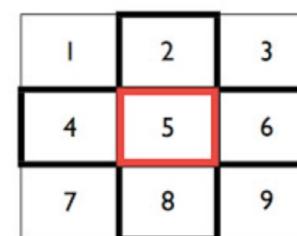


Imposed Spatial linking structure

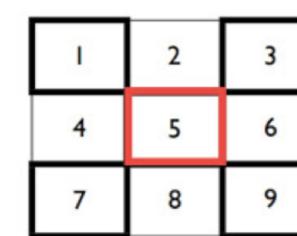


Spatial Weight Matrix

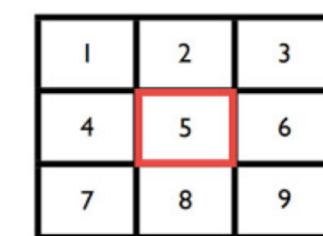
$$W = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$



rook contiguity - edges only
2, 4, 6, 8 are neighbors of 5



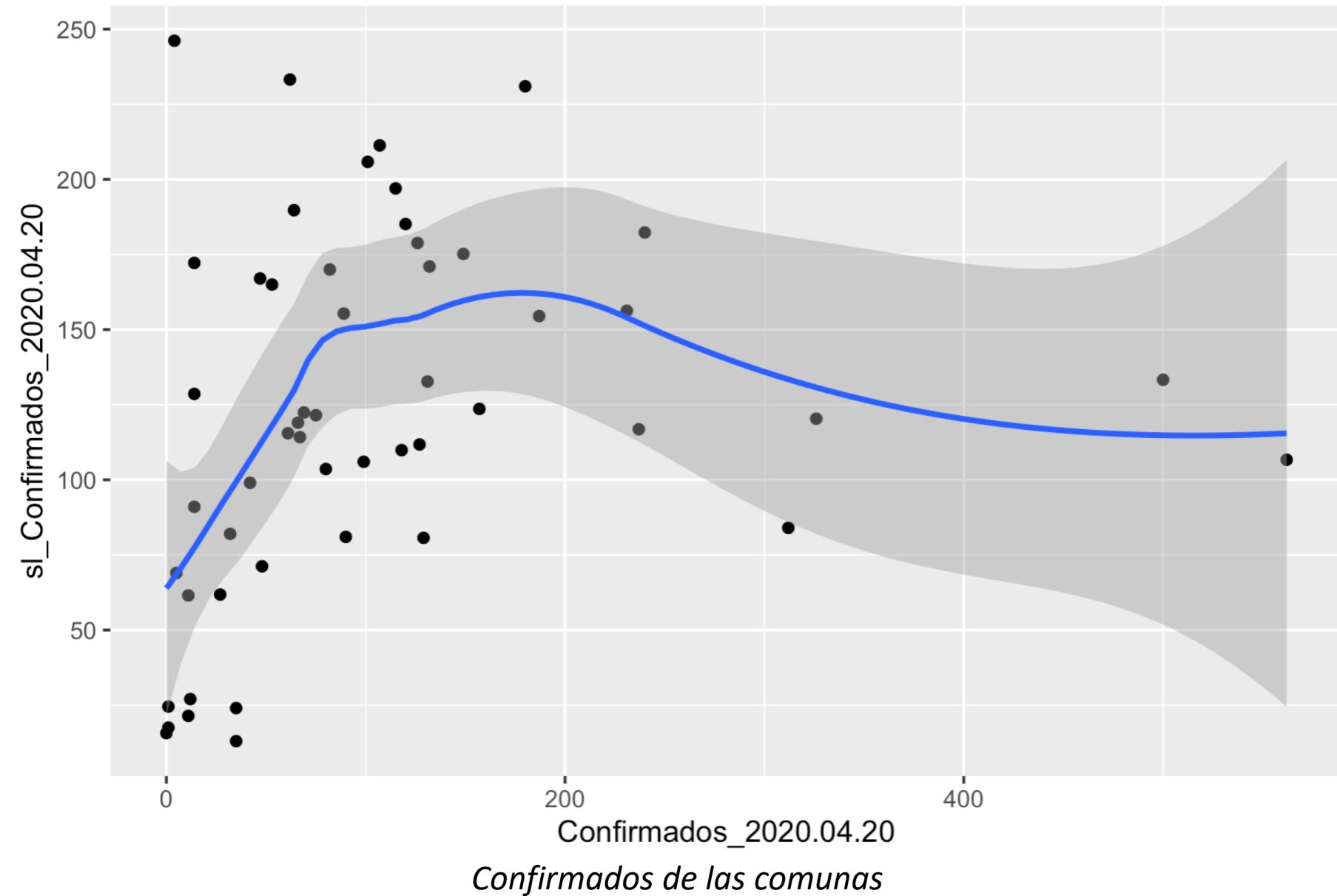
bishop contiguity - corners only
1, 3, 7, 9 are neighbors of 5



queen contiguity - edges and corners
5 has eight neighbors

Rezago Espacial de los casos de Coronavirus (RM)

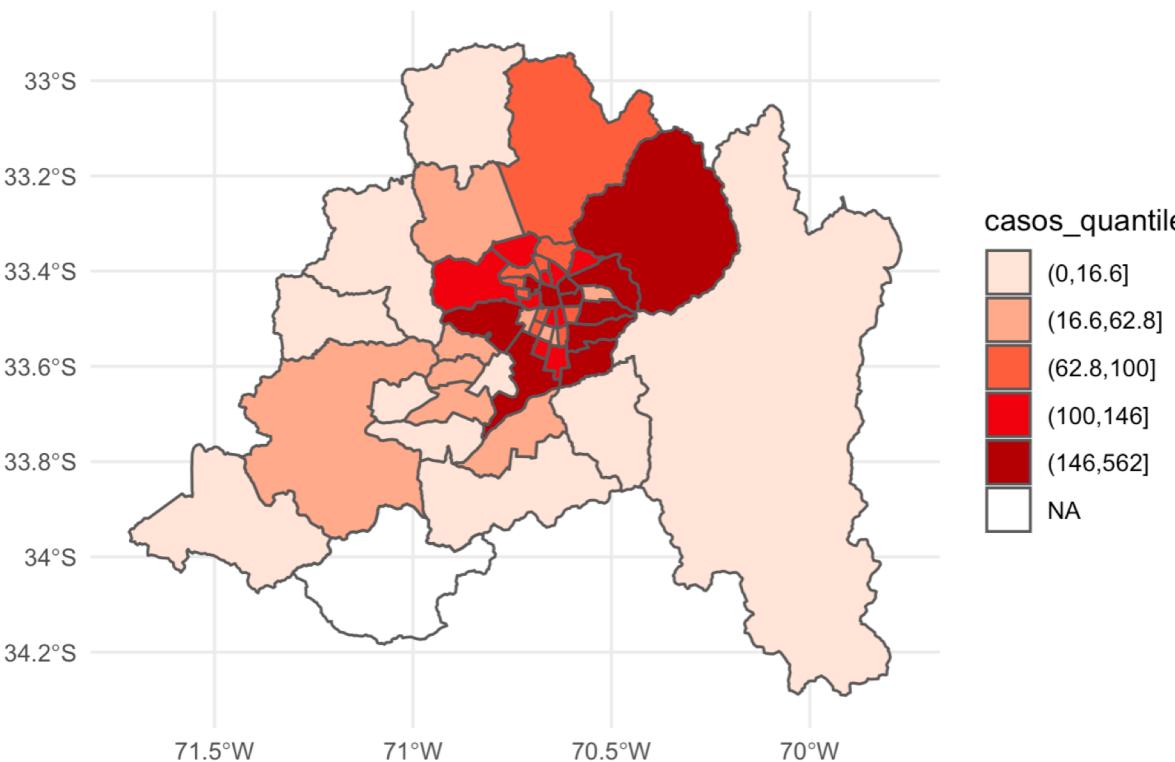
*Confirmados
Promedio de
las comunas
vecinas*



Casos por vecindad espacial a nivel comunal

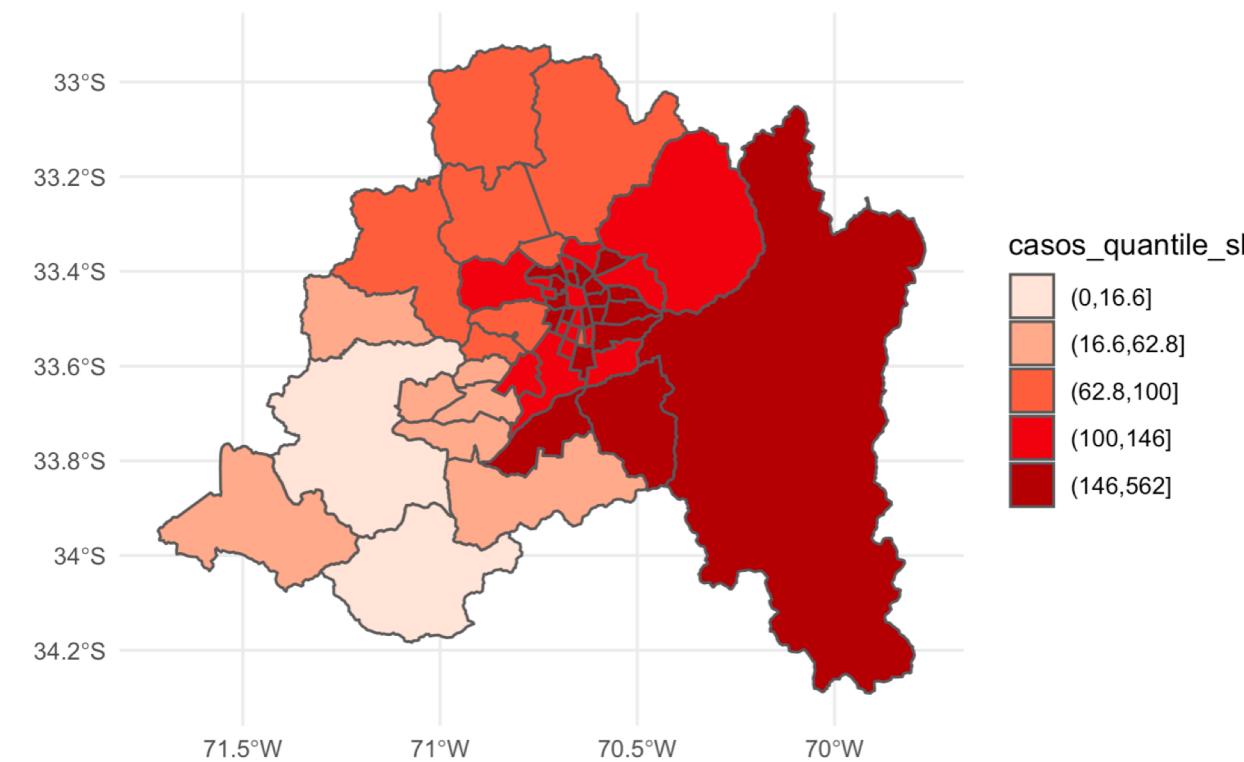
Casos Confirmados

Región Metropolitana - 2020.04.20



Promedio Casos Confirmados Comunas Vecinas

Región Metropolitana - 2020.04.20





SPATIAL DATA SCIENCE: LO QUE SE PUEDE DECIR (Y LO QUE NO) DESDE LOS DATOS DE COVID-19 EN CHILE

