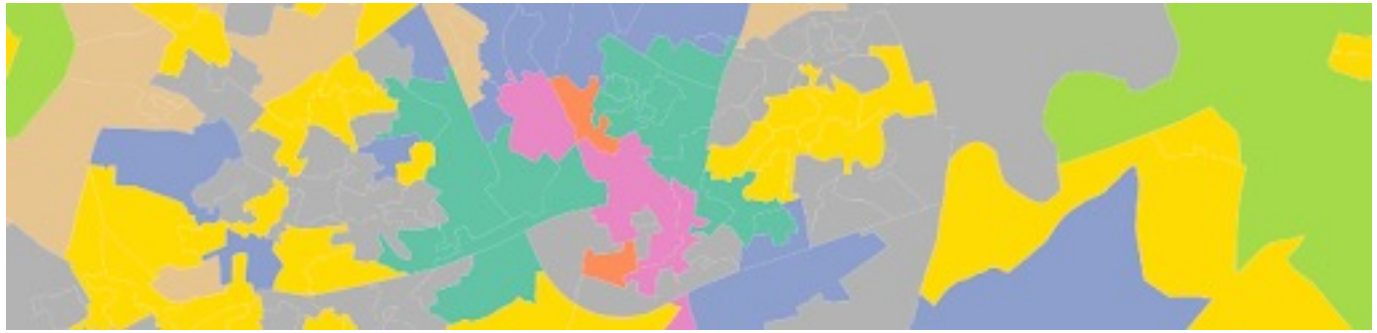
Principles of Spatial Analysis

WEEK 04: SPATIAL AUTOCORRELATION



UCL

This week

- Questions arising from last week's tutorial
- Spatial dependence and distance decay
- Measuring spatial autocorrelation: Global Moran's I, Getis-Ord Gi*, Local Moran's I
- Defining neighbours
- Bonus section on organising your data as 'tidy data'

CASA0005

- Areas of significant overlap – but these are important basics.
- From this week there will be more differentiation; some of the same topics will still be covered more in-depth or with with different applications.
- Complete different topics covered: geostatisics using kriging, spatial interaction models, geodemographic classification, and network analysis.

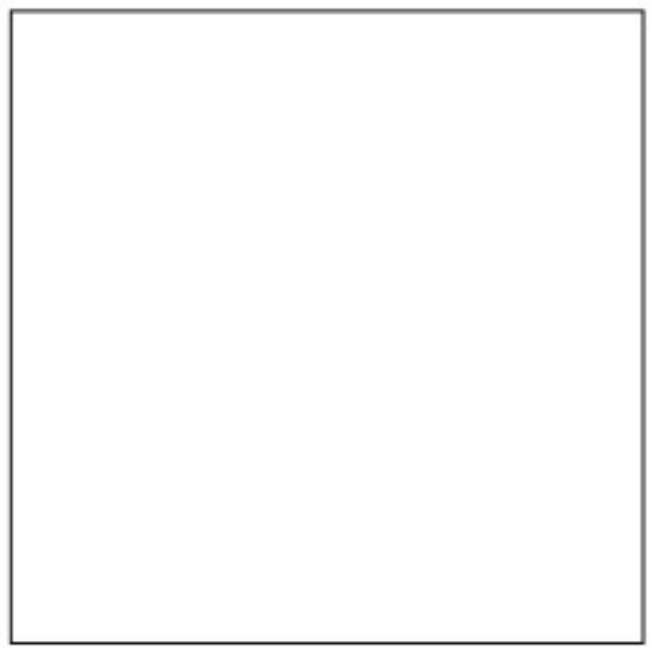
`here` library

- `here` library requires a project, RStudio has a setting to open the most recently opened project automatically (e.g. CASA)
- update on the tutorial material: advice to create a new project and use the `here` library or to use absolute paths or relative paths
- absolute path: `/Users/Name/Directory/W03/Analysis/`
- relative path: `data/london.shp`
- relative path: `../W04/Analysis`

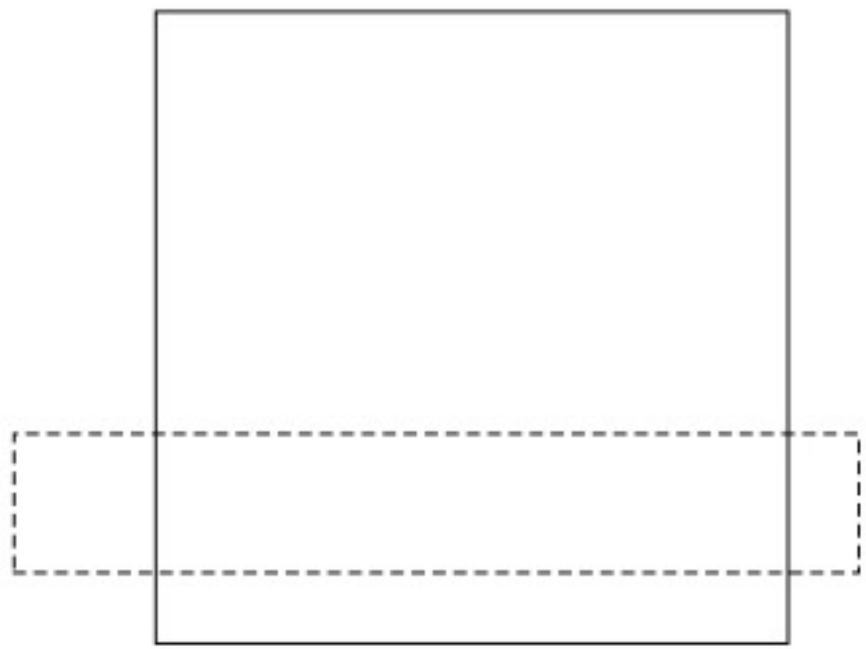
Different types of intersect?

- Difference between `st_intersects() and st_intersection()`
- Stackoverflow is a great resource for these type of questions
- In tutorial there is no difference between the two because we using point data in combination with polygon data.

```r
library(sf)
library(dplyr)

# create square
s <- rbind(c(1, 1), c(10, 1), c(10, 10), c(1, 10), c(1, 1)) %>%
  list %>%
  st_polygon %>%
  st_sfc

plot(s)
```



```r
# create rectangle
r <- rbind(c(-1, 2), c(11, 2), c(11, 4), c(-1, 4), c(-1, 2)) %>%
  list %>%
  st_polygon %>%
  st_sfc

plot(r, add= TRUE, lty = 2)
```



```r
# intersect points and square with st_intersection
i <- st_intersection(s, r)

plot(i, add = TRUE, lty = 2, col = "red")
```

```
i
#> Geometry set for 1 feature
#> geometry type:  POLYGON
#> dimension:      XY
#> bbox:           xmin: 1 ymin: 2 xmax: 10 ymax: 4
#> epsg (SRID):    NA
#> proj4string:    NA
#> POLYGON ((10 4, 10 2, 1 2, 1 4, 10 4))
```

st_intersection()

```
r[which(unlist(st_intersects(s, r)) == 1)]
#> Geometry set for 1 feature
#> geometry type:  POLYGON
#> dimension:      XY
#> bbox:           xmin: -1 ymin: 2 xmax: 11 ymax: 4
#> epsg (SRID):    NA
#> proj4string:    NA
#> POLYGON ((-1 2, 11 2, 11 4, -1 4, -1 2))
```

st_intersects()

`sparse=TRUE`

- Usage of `sparse=TRUE` in `st_intersects()`
- Check documentation for an explanation of the parameters or try to figure out what happens if you change the parameter values.

sparse=TRUE

If sparse=FALSE, st_predicate (with predicate e.g. "intersects") returns a dense logical matrix with element i,j TRUE when predicate(x[i], y[j]) (e.g., when geometry of feature i and j intersect); if sparse=TRUE, an object of class sgbp with a sparse list representation of the same matrix, with list element i an integer vector with all indices j for which predicate(x[i],y[j]) is TRUE (and hence a zero-length integer vector if none of them is TRUE). From the dense matrix, one can find out if one or more elements intersect by apply(mat, 1, any), and from the sparse list by lengths(lst) > 0, see examples below.

How to deal with this?

- Check the documentation of your function
- Google, Stack Overflow / Stack Exchange, GitHub
- Sanity check #1: plot the results
- Sanity check #2: inspect the values in the data frame

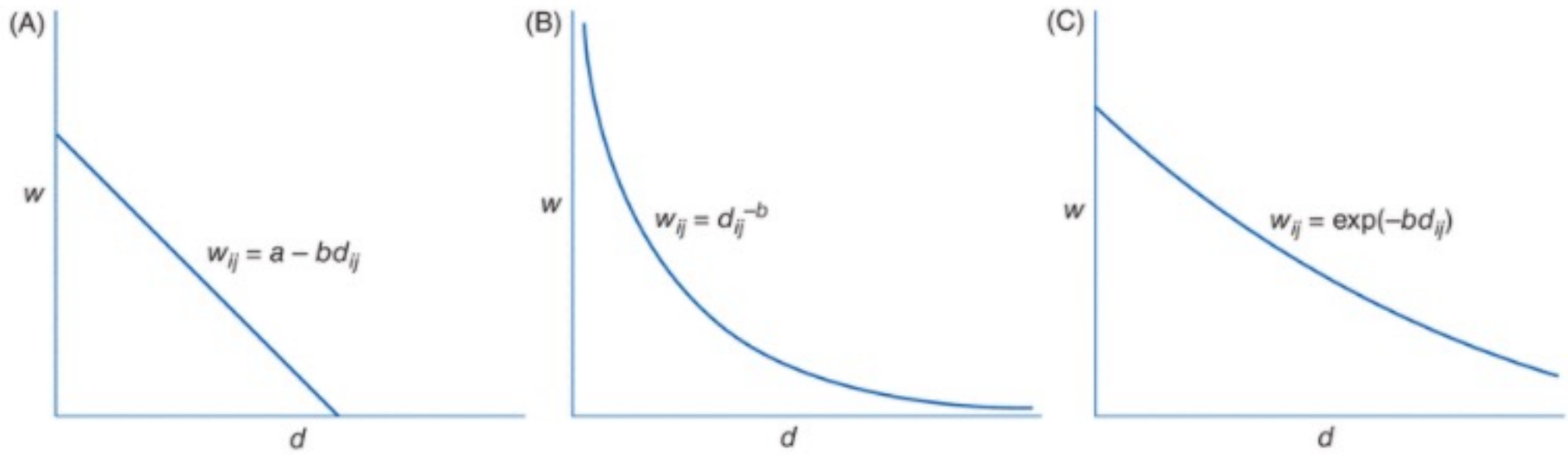
Back to Tobler

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

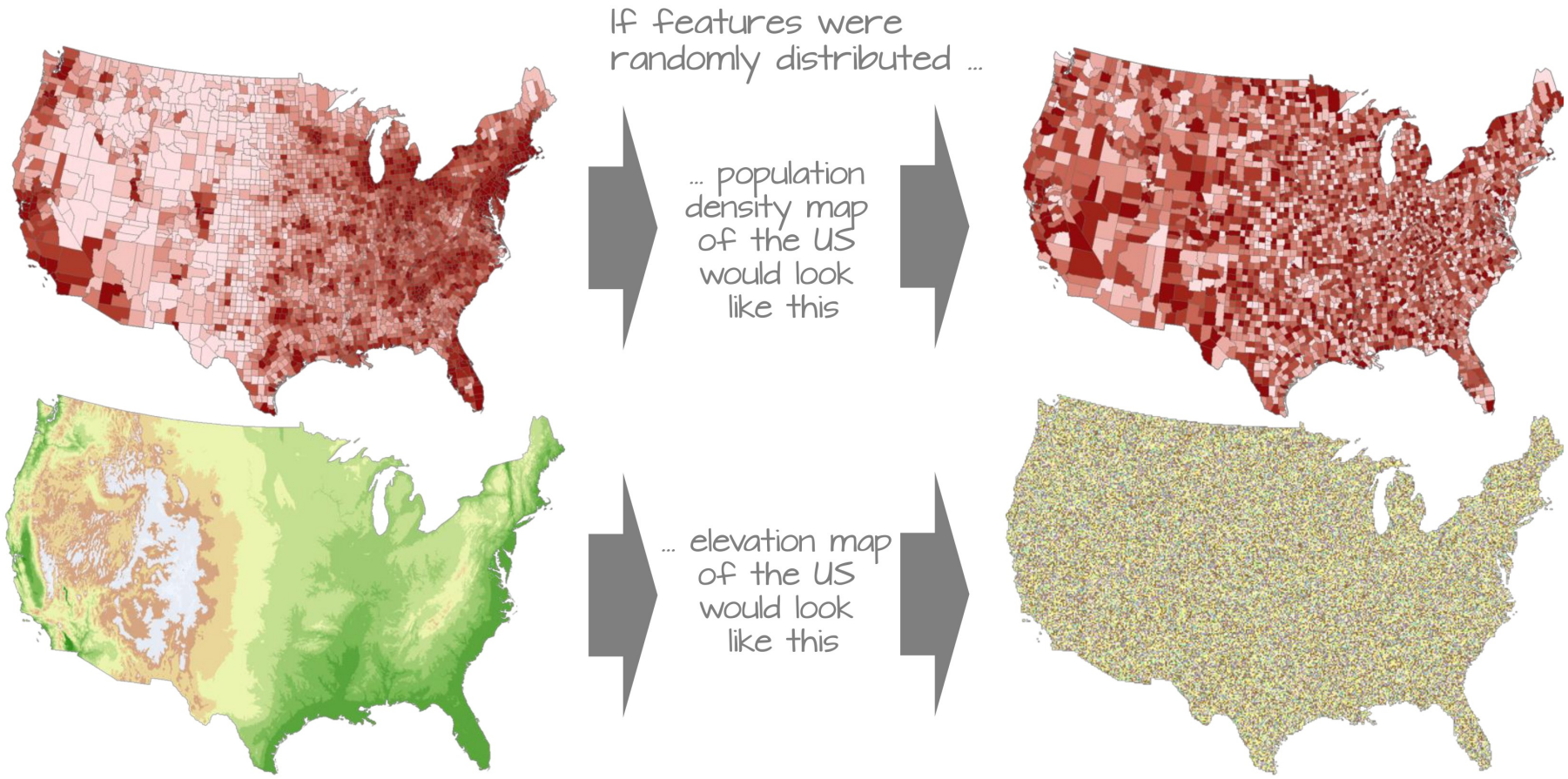
Spatial dependence I

- Spatial dependence is the idea that the observed value of a variable in one location is dependent (to some degree) on the observed value of the same variable in a nearby location.
- Often understood as distance decay – the idea which is used in many geographical applications (e.g. spatial interpolation, spatial interaction).

Spatial dependence II



Spatial autocorrelation



Gimond, M. 2021. Intro to GIS and Spatial Analysis. [online]
https://mgimond.github.io/Spatial/introGIS.html

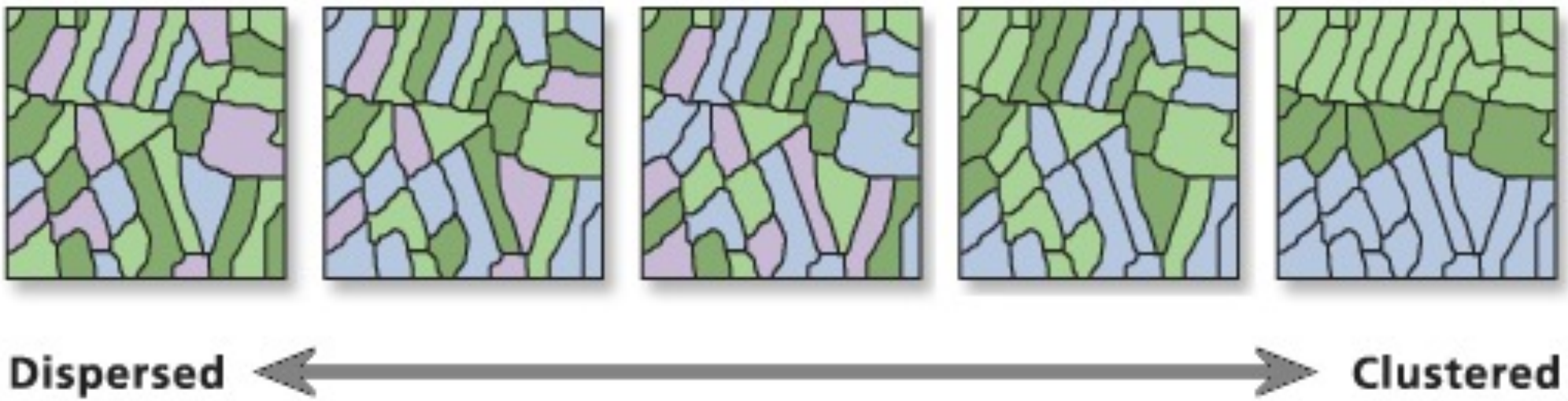
Spatial autocorrelation II

- Spatial dependence exists for both incident (event) and attribute data.
- Questions: Is our data clustered, randomly distributed or dispersed? Can we find hotspots of high values vs. low values? Can we find areas where high values exist directly next to low values?
- In most cases, the distribution of attribute values will seldom show evidence of Complete Spatial Randomness (CSR).
- Complete Spatial Randomness means that a pattern is completely made up by chance. Important concept for understanding ‘significance’ in a spatial context.

Spatial autocorrelation III

- We can test our data for spatial autocorrelation, with some form of statistical measure of the similarity of attributes of our data.
- We want to distinguish between areas of positively autocorrelated patterns (in which high values are surrounded by high values, and low values by low values, i.e. clusters); random patterns (in which neighbouring values are independent of each other, i.e. CSR); and dispersed patterns (in which high values tend to be surrounded by low values and vice versa).

Spatial autocorrelation IV



Measuring spatial autocorrelation

Two ways:

1) Global: What is the overall spatial dependence across the entire data set area? Studying at a global level will tell you how clustered, dispersed or random the data is distributed over the entire area studied.
2) Local: What is the difference between each unit of analysis (e.g. areal unit) and its neighbours? Studying at the local level, you can find areas of greater contrast by seeing if places are quantifiably more like or dislike their neighbours than the average other place.

Global Moran's I

- The most commonly used indicator of global spatial autocorrelation - it was initially suggested by Moran (1948), and popularised through the work on spatial autocorrelation by Cliff and Ord (1973).
- Works through identifying neighbours for each target feature (e.g. polygon) and summarising their values by computing their means (spatially lagged variable value).
- We can then plot the target feature's value against the spatially lagged mean value, repeat this for all features in your data set, and fit an OLS regression.
- The resulting β estimate is your Moran's I statistic.

Global Moran's I

- In essence, it is a cross-product statistic between a variable and its spatial lag, with the variable expressed in deviations from its mean.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

where z_i is the deviation of an attribute for feature i from its mean $(x_i - \bar{X})$, $w_{i,j}$ is the spatial weight between feature i and j, n is the total number of features, and S_0 is the aggregate of all spatial weights.

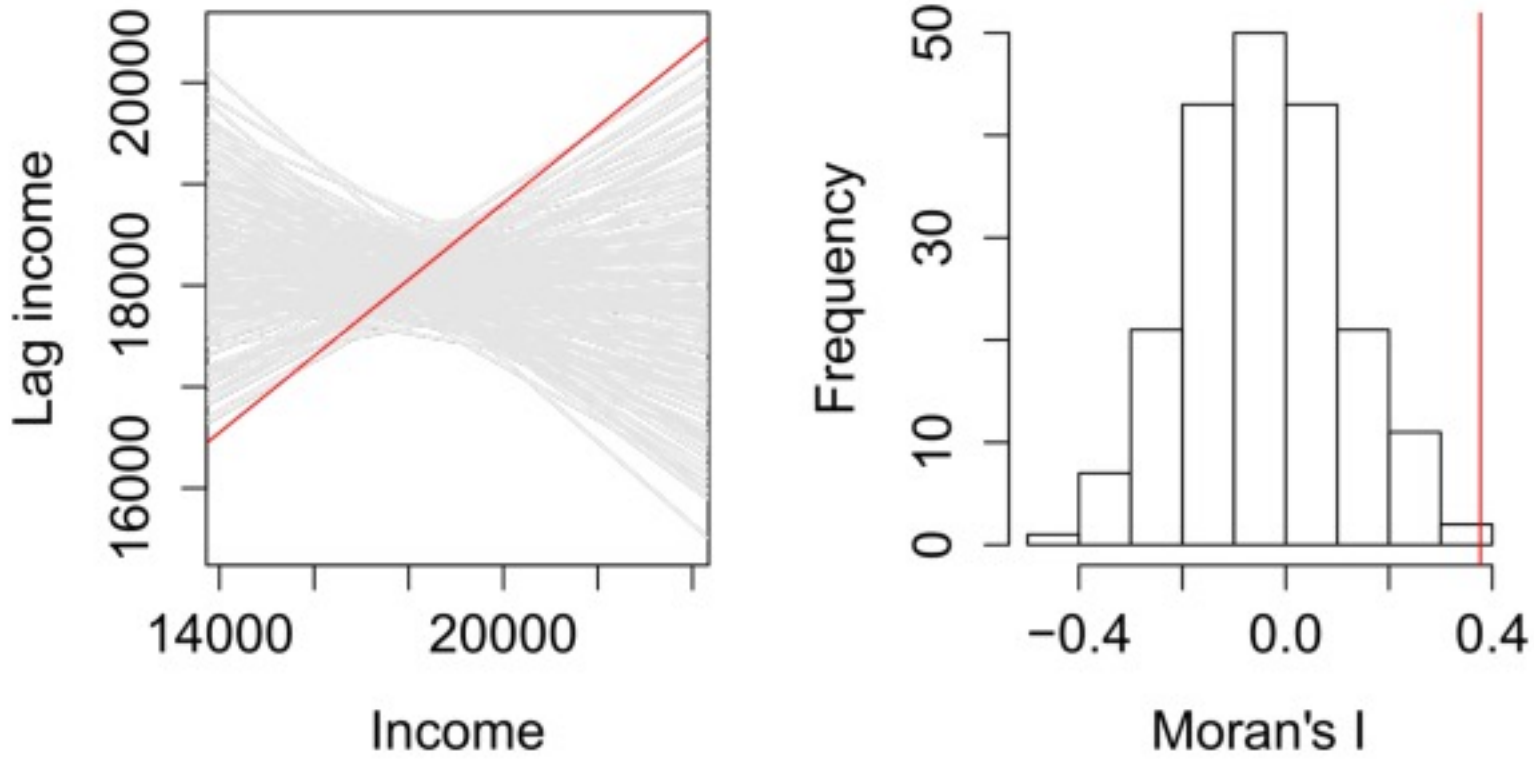
Global Moran's I

- Output is the global Moran's I statistic, which is the correlation coefficient for the relationship between a variable (attribute) and its surrounding values.
- The statistic evaluates whether the pattern is clustered, dispersed or random.
- How to generate the significance of the relationship?

Global Moran's I

- To understand whether our relationship is significant, we can use either an analytical approach or a computational approach. The latter is the preferred option as it does not require making any assumption about the shape and layout of our data set – for this we can use a Monte Carlo test.
- This approach randomly and repeatedly assigns values to polygons in the data set.
- The output is a sampling distribution of Moran's I values under the (null) hypothesis that attribute values are randomly distributed across the study area.

Global Moran's I



Global Moran's I

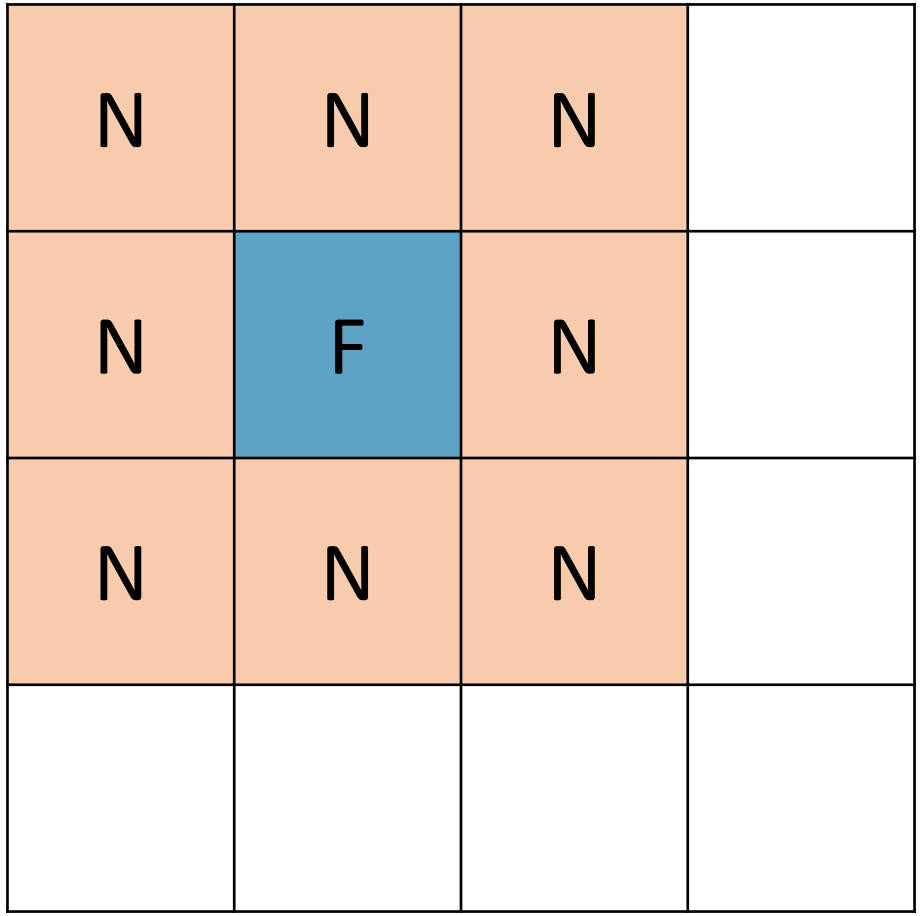
- In the scenario on the previous slide: the Moran's I value (in red) lies outside the main distribution of values in our null hypothesis – therefore we can assume that this is not a value that would occur if our data set was randomly distributed.
- A pseudo p-value is generated from the simulation results.
- For instance: if out of 199 simulations, just one simulation result is more extreme than our observed statistic , *p* is equal to (1 + 1) / (199 + 1) = 0.01. This is interpreted as "there is a 1% probability that we would be wrong in rejecting the null hypothesis."
- Be aware, that the pseudo *p*-value is only a summary of the results from the reference distribution and should not be interpreted as an analytical *p*-value (assumption of normality and normal distribution).

Getis-Ord Gi*

- Where do “high” and “low” values cluster in space.
- More commonly known as hot-spot analysis – finding statistically significant clusters of high and low attribute values: “what is the probability that a spatial distribution of values is not random?”

The Getis-Ord Gi* asks if each feature's neighborhood is significantly different from the study area by comparing each feature’s neighborhood mean value against the global mean value of all features combined.

Neighbourhood



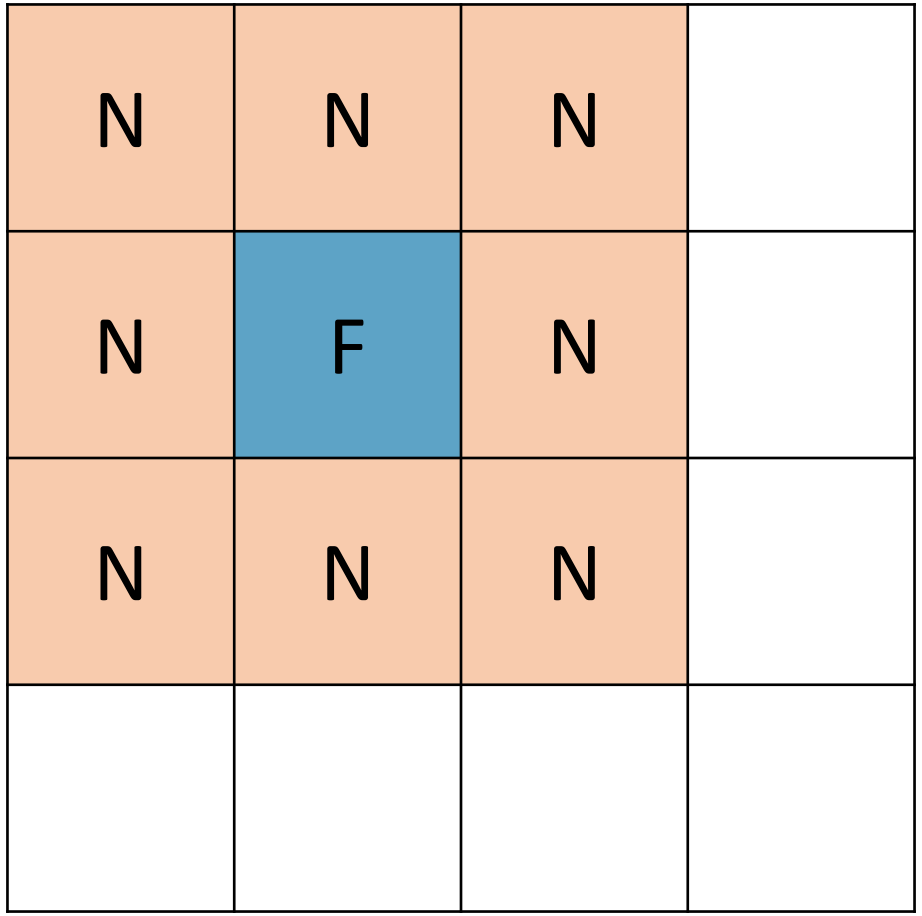
Local Moran's I

- A localised measure of autocorrelation.
- More commonly known as cluster and outlier analysis.

The Local Moran's I asks if each feature is significantly different from all the other features and if the neighborhood significantly different from all the other neighbourhoods. Four types: high-high, low-low, but also outliers: high-low, low-high.

E.g. if a feature's value is significantly higher than all other features and the neighborhood's values are significantly higher than all the other neighborhood's values: high-high cluster.

Neighbourhood



Again:

- The Getis-Ord Gi* asks if each feature's neighborhood is significantly different from the study area by comparing each feature’s neighborhood mean value against the global mean value of all features combined.
- The Local Moran’s I asks if each feature is significantly different from all the other features and is the neighborhood significantly different from all the other neighbourhoods.

Keep in mind

- Spatial autocorrelation is all about the values of some feature and whether there is spatial patterning (clustering) for these values – and whether they show some form of spatial patterning.
- For point or event data this means that you will need some form of aggregation (e.g. counts of the event within a grid or within an administrative boundary) to run these measures.

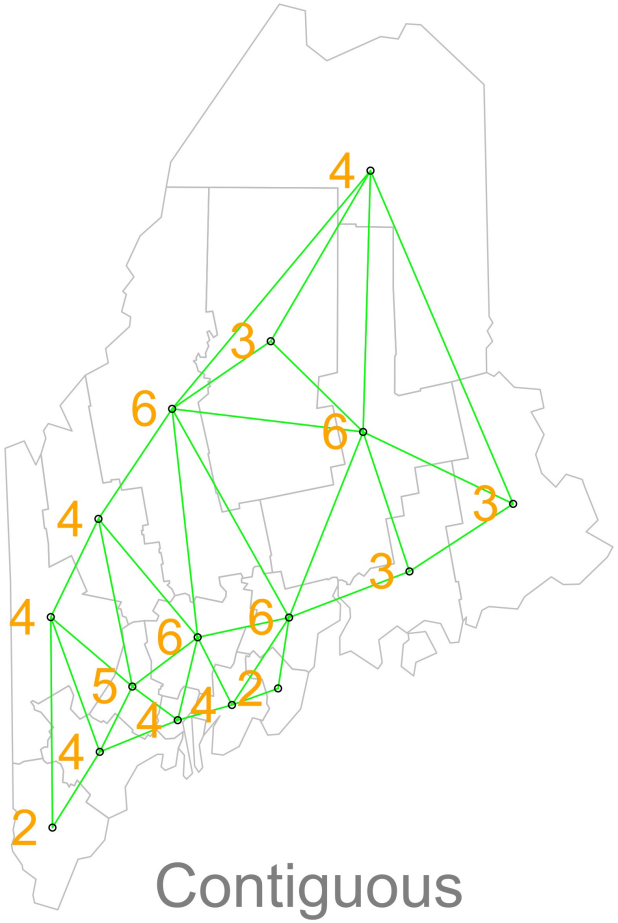
Defining neighbourhoods I

- To calculate any measure of spatial autocorrelation, we need to understand how our spatial units relate to each other as neighbours, i.e. how do we conceive their spatial relationship with one another.
- There are two approaches to defining neighbours, either through contiguity or through distance (proximity).
- This relationship is then used to create an Adjacency Network – or spatial weights network – within the calculation.

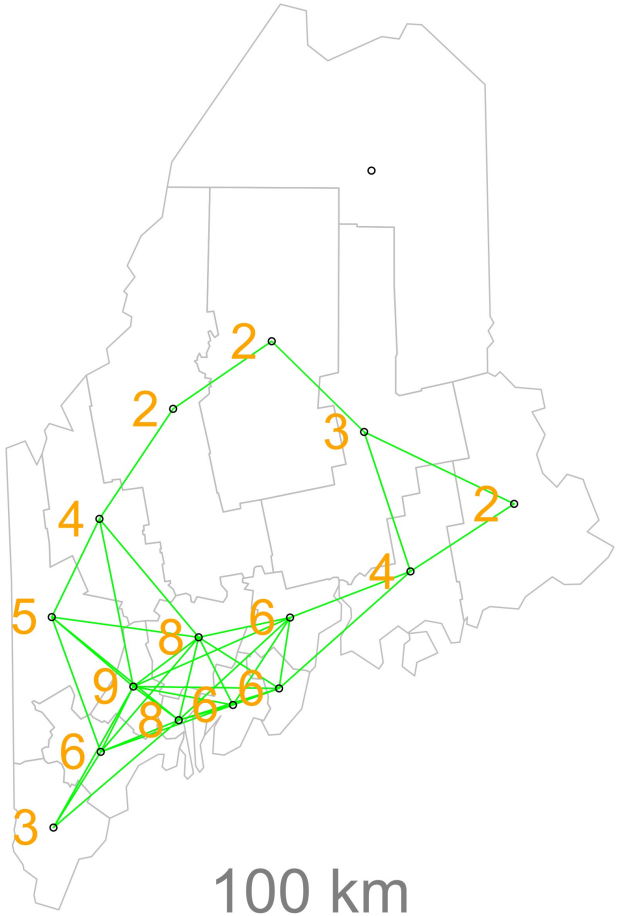
Defining neighbourhoods II

- Contiguity: units that are spatially next to one another, and connect to one another via the same side (edge), vertex or both (careful with simplification of geometries!).
- Distance (proximity): units that are within a specific distance of one another, which itself can be defined in numerous ways: Fixed Distance; Inverse Distance, Travel Time, K Nearest Neighbours.

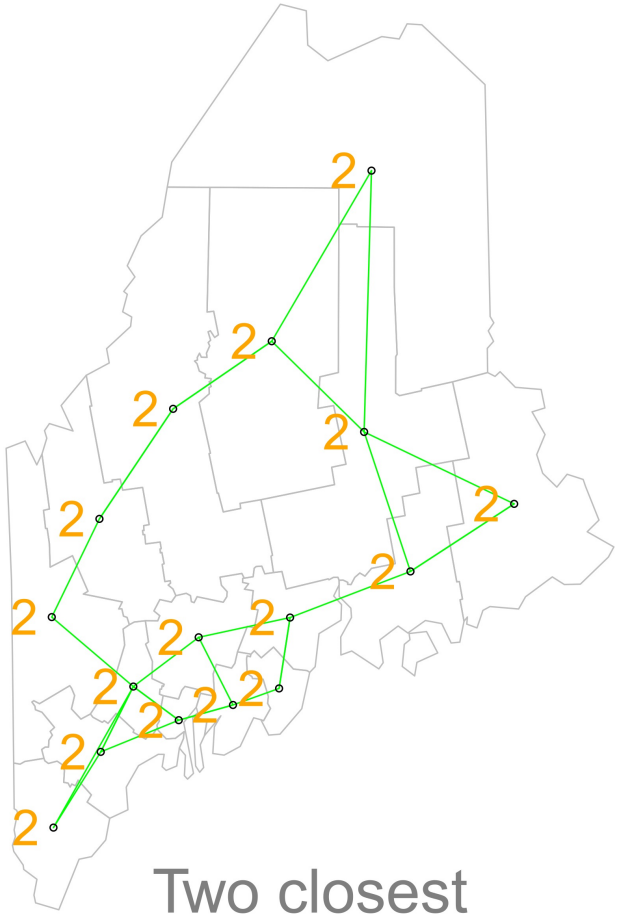
Defining neighbourhoods III



Contiguous



100 km



Two closest

Defining neighbourhoods IV

Conclusion

- Measuring spatial autocorrelation is important for understanding spatial relationships.
- The specification of the neighbourhood can impact the results of these tests; the formulation of the neighbourhood should be grounded in a particular theory or rationale that sets expectations for the spatial form of the process under investigation.
- Next week we will further explore how to 'deal' with / utilise spatial autocorrelation by exploring spatial models.


BONUS
BONUS
BONUS
BONUS

Tidy data I

- Wickham 2014
- 80 percent of your time goes to data cleaning and preparation (‘data wrangling’).
- Tidy data refers to the structure and organisation of your data set.
- The idea boils down to three principles.

Tidy data II

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Each variable must have its own column

Tidy data III

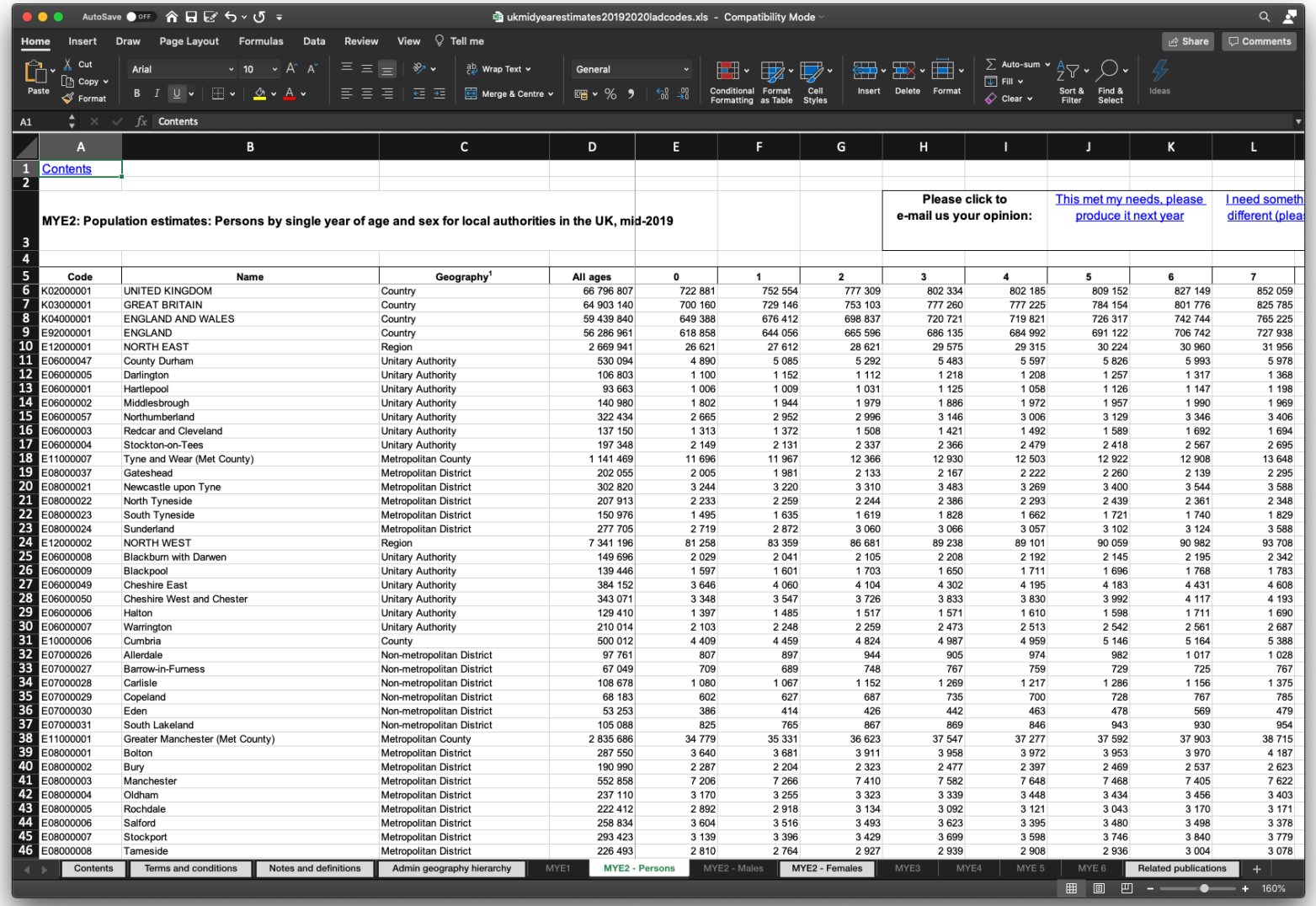
Each observation must have its own row

Tidy data IV

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Each value must have its own cell

Tidy data V



ukmidyearestimates20192020ladcodes.xls - Compatibility Mode

Contents

MYE2: Population estimates: Persons by single year of age and sex for local authorities in the UK, mid-2019

Please click to e-mail us your opinion: | This met my needs, please produce it next year | I need someth different (plea

Code	Name	Geography[1]	All ages	0	1	2	3	4	5	6	7
K02000001	UNITED KINGDOM	Country	66 796 807	722 881	752 554	777 309	802 334	802 185	809 152	827 149	852 059
K03000001	GREAT BRITAIN	Country	64 903 140	700 160	729 146	753 103	777 260	777 225	784 154	801 776	825 785
K04000001	ENGLAND AND WALES	Country	59 439 840	649 388	676 412	698 837	720 721	719 821	726 317	742 744	765 225
E92000001	ENGLAND	Country	56 286 961	618 858	644 056	665 596	686 135	684 992	691 122	706 742	727 938
E12000001	NORTH EAST	Region	2 669 941	26 621	27 612	28 621	29 575	29 315	30 224	30 960	31 956
E06000047	County Durham	Unitary Authority	530 094	4 890	5 085	5 292	5 483	5 597	5 826	5 993	5 978
E06000005	Darlington	Unitary Authority	106 803	1 100	1 152	1 112	1 218	1 208	1 257	1 317	1 368
E06000001	Hartlepool	Unitary Authority	93 663	1 006	1 009	1 031	1 125	1 058	1 126	1 147	1 198
E06000002	Middlesbrough	Unitary Authority	140 980	1 802	1 944	1 979	1 886	1 972	1 957	1 990	1 969
E06000057	Northumberland	Unitary Authority	322 434	2 665	2 952	2 996	3 146	3 006	3 129	3 346	3 406
E06000003	Redcar and Cleveland	Unitary Authority	137 150	1 313	1 372	1 508	1 421	1 492	1 589	1 692	1 694
E06000004	Stockton-on-Tees	Unitary Authority	197 348	2 149	2 131	2 337	2 366	2 479	2 418	2 567	2 695
E11000007	Tyne and Wear (Met County)	Metropolitan County	1 141 469	11 696	11 967	12 366	12 930	12 503	12 922	12 908	13 648
E08000037	Gateshead	Metropolitan District	202 055	2 005	1 981	2 133	2 167	2 222	2 260	2 139	2 295
E08000021	Newcastle upon Tyne	Metropolitan District	302 820	3 244	3 220	3 310	3 483	3 269	3 400	3 544	3 588
E08000022	North Tyneside	Metropolitan District	207 913	2 233	2 259	2 244	2 386	2 293	2 439	2 361	2 348
E08000023	South Tyneside	Metropolitan District	150 976	1 495	1 635	1 619	1 828	1 662	1 721	1 740	1 829
E08000024	Sunderland	Metropolitan District	277 705	2 719	2 872	3 060	3 066	3 057	3 102	3 124	3 588
E12000002	NORTH WEST	Region	7 341 196	81 258	83 359	86 681	89 238	89 101	90 059	90 982	93 708
E06000008	Blackburn with Darwen	Unitary Authority	149 696	2 029	2 041	2 105	2 208	2 192	2 145	2 195	2 342
E06000009	Blackpool	Unitary Authority	139 446	1 597	1 601	1 703	1 650	1 711	1 696	1 768	1 783
E06000049	Cheshire East	Unitary Authority	384 152	3 646	4 060	4 104	4 302	4 195	4 183	4 431	4 608
E06000050	Cheshire West and Chester	Unitary Authority	343 071	3 348	3 547	3 726	3 833	3 830	3 992	4 117	4 193
E06000006	Halton	Unitary Authority	129 410	1 397	1 485	1 517	1 571	1 610	1 598	1 711	1 690
E06000007	Warrington	Unitary Authority	210 014	2 103	2 248	2 259	2 473	2 513	2 542	2 561	2 687
E10000006	Cumbria	County	500 012	4 409	4 459	4 824	4 987	4 959	5 146	5 164	5 388
E07000026	Allerdale	Non-metropolitan District	97 761	807	897	944	905	974	982	1 017	1 028
E07000027	Barrow-in-Furness	Non-metropolitan District	67 049	709	689	748	767	759	729	725	767
E07000028	Carlisle	Non-metropolitan District	108 678	1 080	1 067	1 152	1 269	1 217	1 286	1 156	1 375
E07000029	Copeland	Non-metropolitan District	68 183	602	627	687	735	700	728	767	785
E07000030	Eden	Non-metropolitan District	53 253	386	414	426	442	463	478	569	479
E07000031	South Lakeland	Non-metropolitan District	105 088	825	765	867	869	846	943	930	954
E11000001	Greater Manchester (Met County)	Metropolitan County	2 835 686	34 779	35 331	36 623	37 547	37 277	37 592	37 903	38 715
E08000001	Bolton	Metropolitan District	287 550	3 640	3 681	3 911	3 958	3 972	3 953	3 970	4 187
E08000002	Bury	Metropolitan District	190 990	2 287	2 204	2 323	2 477	2 397	2 469	2 537	2 623
E08000003	Manchester	Metropolitan District	552 858	7 206	7 266	7 410	7 582	7 648	7 468	7 405	7 622
E08000004	Oldham	Metropolitan District	237 110	3 170	3 255	3 323	3 339	3 448	3 434	3 456	3 403
E08000005	Rochdale	Metropolitan District	222 412	2 892	2 918	3 134	3 092	3 121	3 043	3 170	3 171
E08000006	Salford	Metropolitan District	258 834	3 604	3 516	3 493	3 623	3 395	3 480	3 498	3 378
E08000007	Stockport	Metropolitan District	293 423	3 139	3 396	3 429	3 699	3 598	3 746	3 840	3 779
E08000008	Tameside	Metropolitan District	226 493	2 810	2 764	2 927	2 939	2 908	2 936	3 004	3 078

Contents | Terms and conditions | Notes and definitions | Admin geography hierarchy | MYE1 | MYE2 - Persons | MYE2 - Males | MYE2 - Females | MYE3 | MYE4 | MYE 5 | MYE 6 | Related publications

Common errors

- Column headers are values rather than variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple observational units are stored in the same column.
- A single observation is stored in multiple tables.

Tidy ?

country	year	type	count
Afghanistan	2019	cases	745
Afghanistan	2019	population	19 987 071
Afghanistan	2020	cases	2 666
Afghanistan	2020	population	20 595 360
Brazil	2019	cases	3,7737
Brazil	2019	population	172 006 362
Brazil	2020	cases	80 488
Brazil	2020	population	174 504 898
China	2019	cases	212 258
China	2019	population	1 272 915 272
China	2020	cases	213 766
China	2020	population	1 280 428 583

Tidy ?

country	year	rate
Afghanistan	2019	745 / 19,987,071
Afghanistan	2020	2,666 / 20,595,360
Brazil	2019	3,7737 / 172,006,362
Brazil	2020	80,488 / 174,504,898
China	2019	212,258 / 1,272,915,272
China	2020	213,766 / 1,280,428,583

Tidy ?

Cases

country	2019	2020
Afghanistan	745	2 666
Brazil	3,7737	80 488
China	212 258	213 766

Population

country	2019	2020
Afghanistan	19 987 071	20 595 360
Brazil	172 006 362	174 504 898
China	1 272 915 272	1 280 428 583

Tidy ?

country	year	cases	population
Afghanistan	2019	745	19 987 071
Afghanistan	2020	2 666	20 595 360
Brazil	2019	3,7737	172 006 362
Brazil	2020	80 488	174 504 898
China	2019	212 258	1 272 915 272
China	2020	213 766	1 280 428 583

Questions

Justin van Dijk

j.t.vandijk@ucl.ac.uk