GEOGRAPHIC DATA SCIENCE

THE CASE FOR SPACE IN DATA SCIENCE

LEVI JOHN WOLF

levi.john.wolf@bristol.ac.uk
@levijohnwolf
ljwolf.org

WHY ARE GEOGRAPHERS?

GEOGRAPHY IS EVERYWHERE

HOW CAN GEOGRAPHY HELP?

HOW TO LEARN MORE

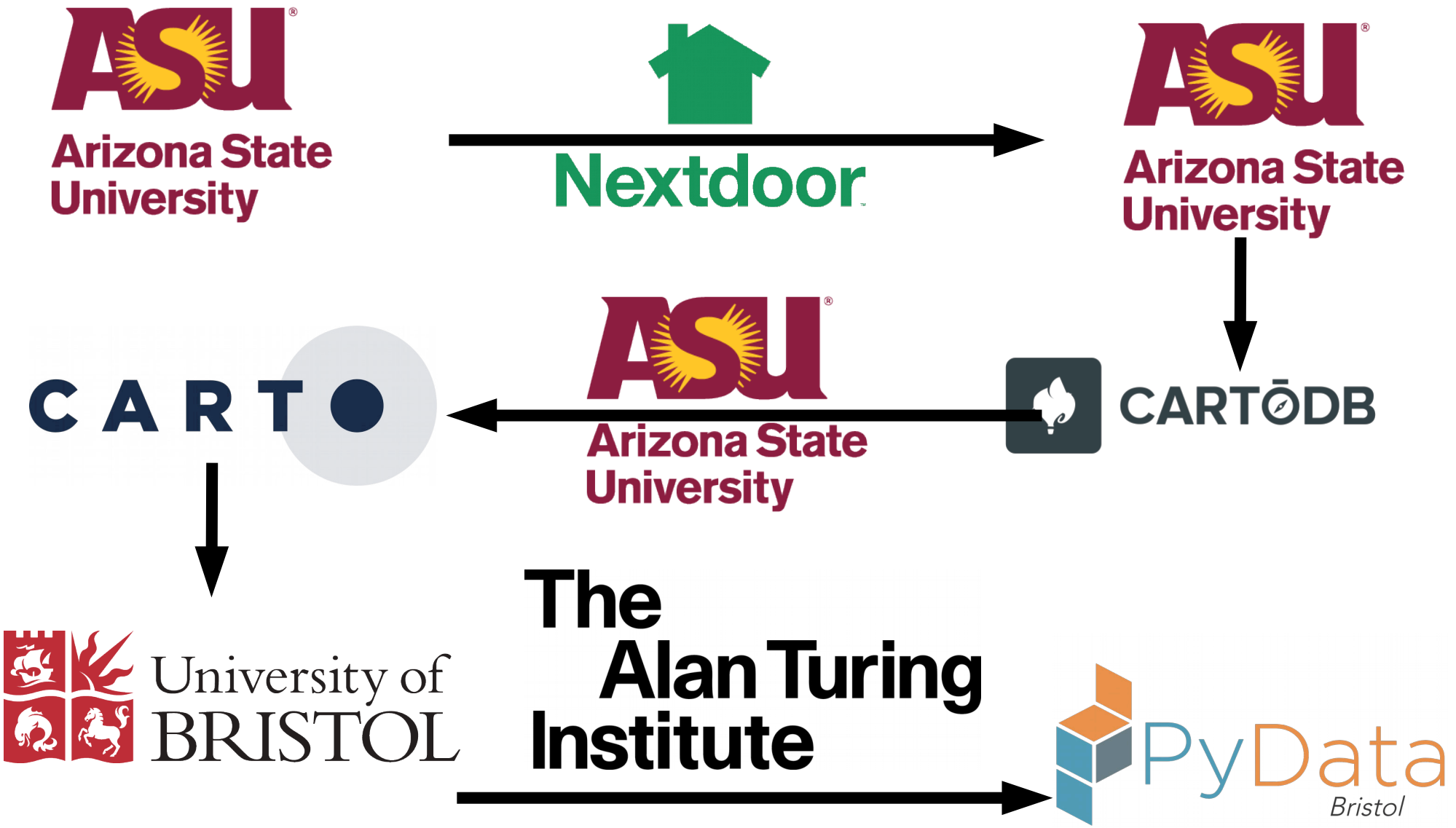
WHY ARE GEOGRAPHERS?

can’t say I know myself

GEOGRAPHY IS EVERYWHERE

HOW CAN GEOGRAPHY HELP?

HOW TO LEARN MORE


Arizona State University
Nextdoor
Arizona State University
Arizona State University
CARTODB
CARTO
Arizona State University
University of BRISTOL
The Alan Turing Institute
PyData
Bristol

GEOGRAPHY IS WHAT GEOGRAPHERS DO

Apocryphal, attributed to Wilson (1970),
also earlier by Johston & Sidaway (2015)

GEOGRAPHY IS WHAT GEOGRAPHERS DO

Apocryphal, attributed to Wilson (1970),
also earlier by Johston & Sidaway (2015)

Arizona State University

University of BRISTOL

GEOGRAPHY

the systematic study of relationships between people and their environments.

Arizona State University

University of BRISTOL

GEOGRAPHY

the systematic study of relationships between people and their environments.

commuting, segregation, climate change, urban planning, campaigns, elections, and voting, housing markets, energy forecasting, operational research, conservation, community dynamics, disease spread, violent conflict, school efficacy, ecology, economic development & cluster policy, digital economy, program evaluation, migration, risk, resilience & social vulnerability, ice sheets, rugby

CARTO

Nextdoor

The Alan Turing Institute

GEOGRAPHERS

use the relationships between people and their environments to solve problems.

CARTO

Nextdoor

GEOGRAPHERS

use the relationships between people and their environments to solve problems.

What drives a healthy digital community, and how can we build them?
How can we drive engagement with our social media platform within and across neighborhoods?
How can we expose relevant commercial opportunities (classified, local advertisements) to users?

CARTO

Nextdoor

The Alan Turing Institute

GEOGRAPHERS

use the relationships between people and their environments to solve problems.

What kind of twitter users engage which hashtags, where, & why?
How do trends spread? Is it just social network structure, or are there other dampeners/drivers?
Can we deliver packages to customers in a more efficient manner?

CARTO

Nextdoor

The Alan Turing Institute

GEOGRAPHERS

use the relationships between people and their environments to solve problems.

How can we extract meaningful knowledge from terabytes of imagery/locational data?
What do you really gain from adding geographical information to machine learning methods?
Can we do a better job modeling how crowds interact with the built environment?
Are there fundamental/consistent structures in cities that emerge across social/cultural contexts?

WHY ARE GEOGRAPHERS?

Geographers use relationships between people & environment to help

GEOGRAPHY IS EVERYWHERE

I apologize for the above pun

HOW CAN GEOGRAPHY HELP?

HOW TO LEARN MORE

NOT EVERYONE SHOULD BE A GEOGRAPHER

EVERYONE SHOULD USE GEOGRAPHY

especially if you've got data!

NYT: ONE NATION TRACKED

https://nyti.ms/2uwMIUZ

Every minute of every day, everywhere on the planet dozens of companies are logging the movements of tens of millions of people.

Without much effort we spotted visitors to the estates of Johnny Depp, Tiger Woods and Arnold Schwarzenegger, connecting the devices' owners to the residences.

NYT: ONE NATION TRACKED

https://nyti.ms/2uwMIUZ

UBIQUITY

Everything has a geographical location & time of occurrence

Every minute of every day, everywhere on the planet dozens of companies are logging the movements of tens of millions of people.

CONTEXT

Location enhances information that is already available

Without much effort we spotted visitors to the estates of Johnny Depp, Tiger Woods and Arnold Schwarzenegger, connecting the devices' owners to the residences.

Everything has a geographical location & time of occurrence

Everything has a geographical location & time of occurrence

Every Monday without fail, a final-year undergrad emails me at 17:55 PM, right as I'm packing up to leave my office, to ask if they can meet to talk about their dissertation.

Everything has a geographical location & time of occurrence

Event E happens

Everything has a geographical location & time of occurrence

occurrence

Event E happens

at 17:55

every Monday

before I leave my office

Everything has a geographical location & time of occurrence

occurrence
periodicity

Event E happens

at 17:55
every Monday
before I leave my office

Everything has a geographical location & time of occurrence

occurrence
periodicity
causality

Event E happens

at 17:55
every Monday
before I leave my office

Everything has a geographical location & time of occurrence

occurrence
periodicity
(G-)causality

Event E happens

at 17:55
every Monday
before I leave my office

Everything has a geographical location & time of occurrence

Event E happens
at 17:55
every Monday
before I leave my office

occurrence
periodicity
(G-)causality

TIME: MORE THAN CLOCK POSITION

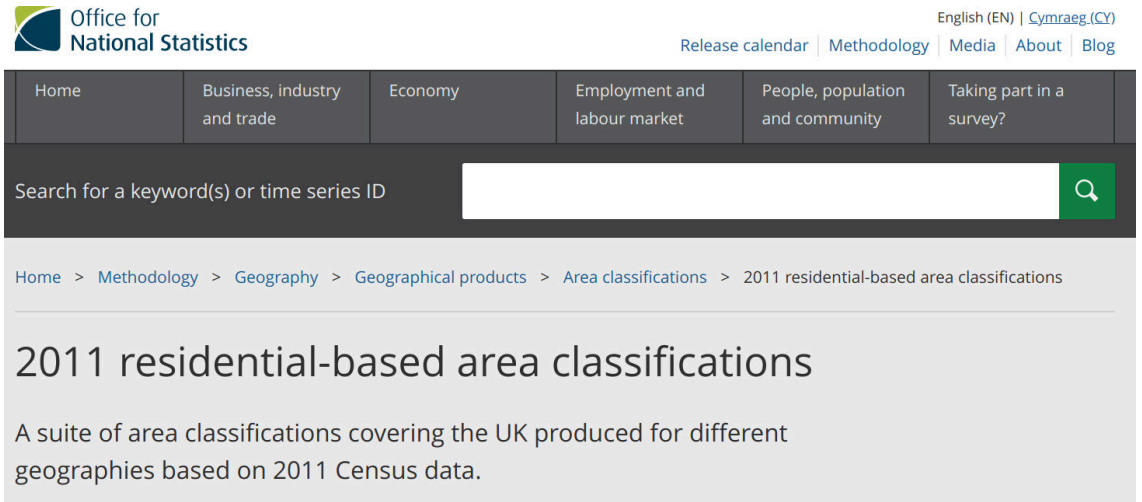
Everything has a geographical location & time of occurrence

Event E happens
at 51°27'N 2°35'W
in Bristol
near me

location
containment
proximity

SPACE: MORE THAN EARTH POSITION

EVEN IF YOU DON'T COLLECT IT, YOU HAVE IT



Massive amounts of government-collected contextual data is OK for commercial use

New, high-quality imagery is increasingly cheap

Enrich from Data Observatory

Leverage the CARTO Data Observatory to put your data into context

CARTO's Data Observatory enables you to discover hidden patterns and gain comprehensive insights about unforeseen opportunities through data enrichment and analysis.

The Data Observatory provides demographic, financial, real estate, transportation, and various population segment data, directly through CARTO Builder (or SQL). See the Data Observatory catalog for a list of all available measures and regions of the world where there is coverage.

Enrich your Data

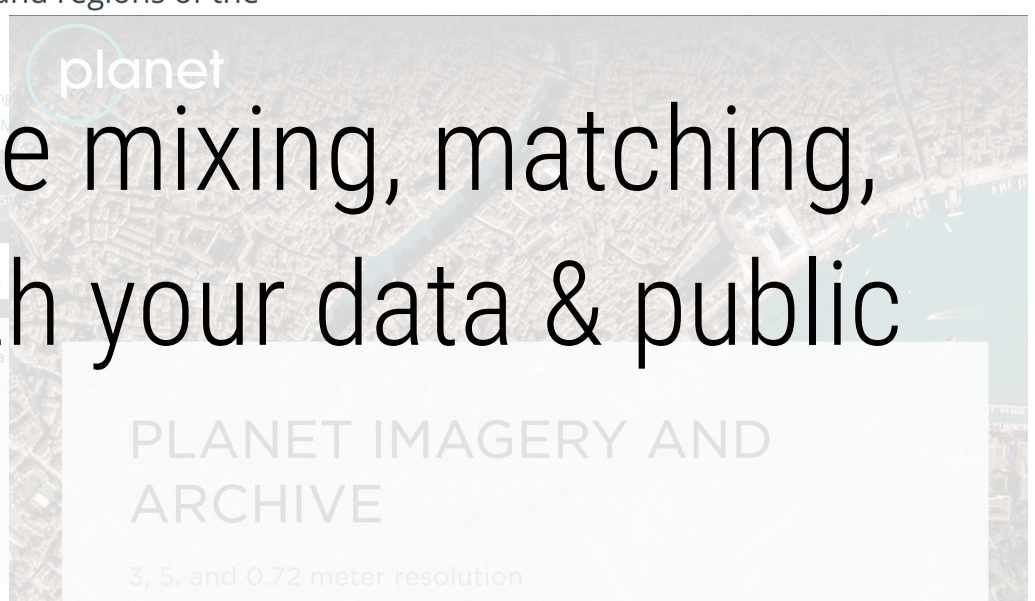
Add Multiple Measurements

Style By Value

Limits

External Resources

Download resources



Companies can do the mixing, matching, & merging for you with your data & public data automatically.

WHY ARE GEOGRAPHERS?

Geographers use relationships between people & environment to help

GEOGRAPHY IS EVERYWHERE

Geography, like time, is ubiquitous & helps us contextualize knowledge

HOW CAN GEOGRAPHY HELP?

I’m not sure, ask your neighbor.

HOW TO LEARN MORE

GEOGRAPHIC DATA SCIENCE

COMMODITY:

Do standard analysis on your standard data
& chuck the results into Tableau/Alteryx to map

GEOGRAPHIC DATA SCIENCE

COMMODITY:

Do standard analysis on your standard data
& chuck the results into Tableau/Alteryx to map

ENRICHED:

Do standard analysis on data after you have
augmented it using *spatial feature engineering*

EXTENDED:

Use methods that explicitly learn from, analyze, or summarize the geographical structure in your data

GEOGRAPHICDATA.SCIENCE/BOOK

COMMODITY:

Do standard analysis on your standard data
& chuck the results into Tableau/Alteryx to map

ENRICHED:

Do standard analysis on data after you have
augmented it using *spatial feature engineering*

EXTENDED:

Use methods that explicitly learn from, analyze, or summarize the geographical structure in your data

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

Say you're modelling the price-per-head of an Airbnb

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

Say you're modelling the price-per-head of an Airbnb

How many pubs are within a 200m walk of this Airbnb?

How far is this Airbnb from the closest noisy major road?

How pretty is the area this Airbnb is in?

What's the burglary rate in the ward the Airbnb is in?

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

MAP MATCHING

Using spatial relationships between two datasets to transfer information

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

Say you're modelling the price-per-head of an Airbnb

How many other Airbnbs are within a 200m walk?

Are there any Airbnbs within a 200m walk that are kid-friendly?

What's the average number of people accommodated by nearby Airbnbs?

Is this Airbnb co-located with other similar Airbnbs that compete?

GEOGRAPHICDATA.SCIENCE/BOOK

ENRICHED:

Do standard analysis on data after you have augmented it using *spatial feature engineering*

synthesizing information using spatial relationships within or across data.

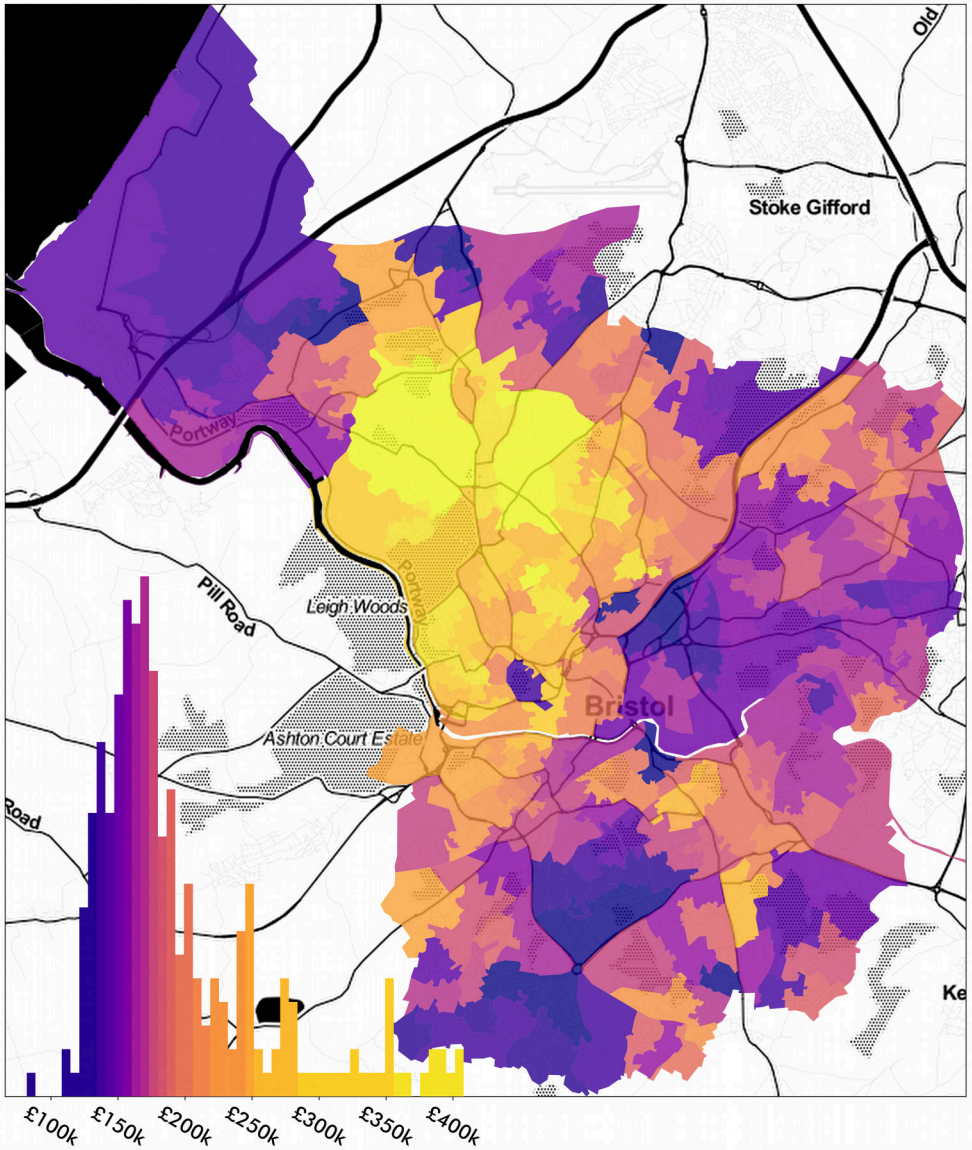
MAP SYNTHESIS

Using spatial relationships between observations in a dataset to synthesize new features

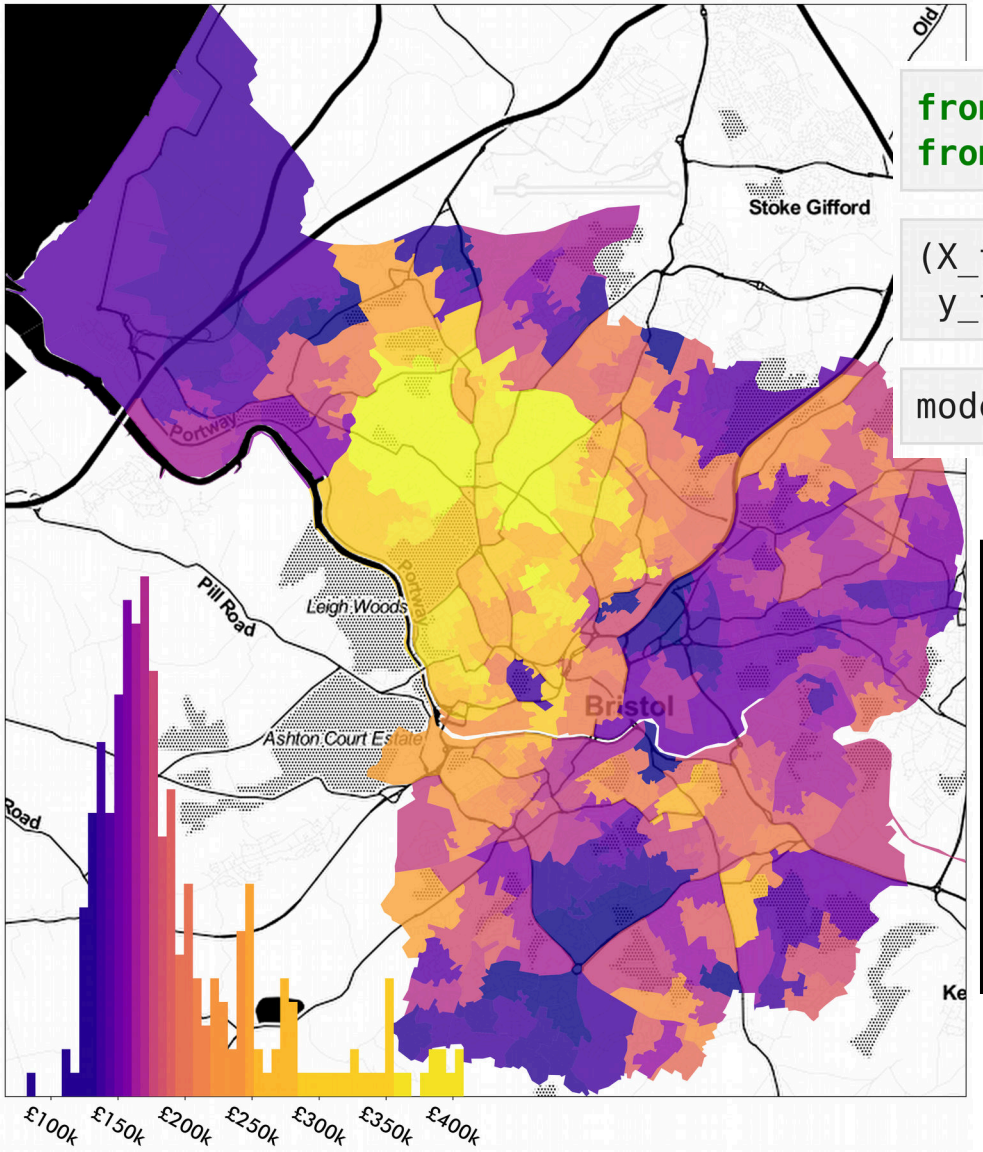
GEOGRAPHICDATA.SCIENCE/BOOK

EXTENDED:

Use methods that explicitly learn from, analyze, or summarize the geographical structure in your data



USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE



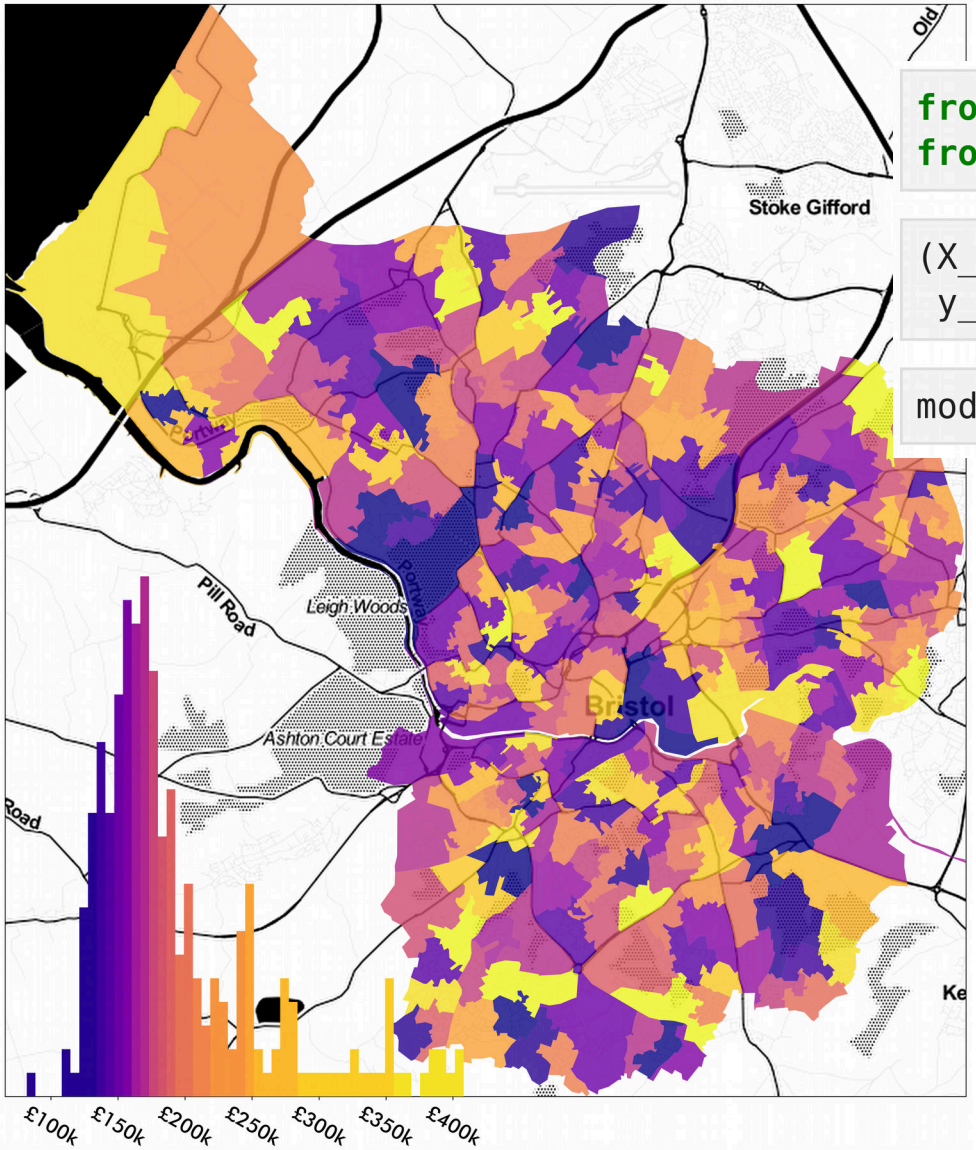

```python
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```python
(X_train, X_test,
 y_train, y_test) = train_test_split(X,y, test_size=.2)
```

```python
model = LinearRegression().fit(X_train, y_train)
```

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,P} \\ x_{2,1} & x_{2,2} & \dots & x_{2,P} \\ \vdots & & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,P} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE



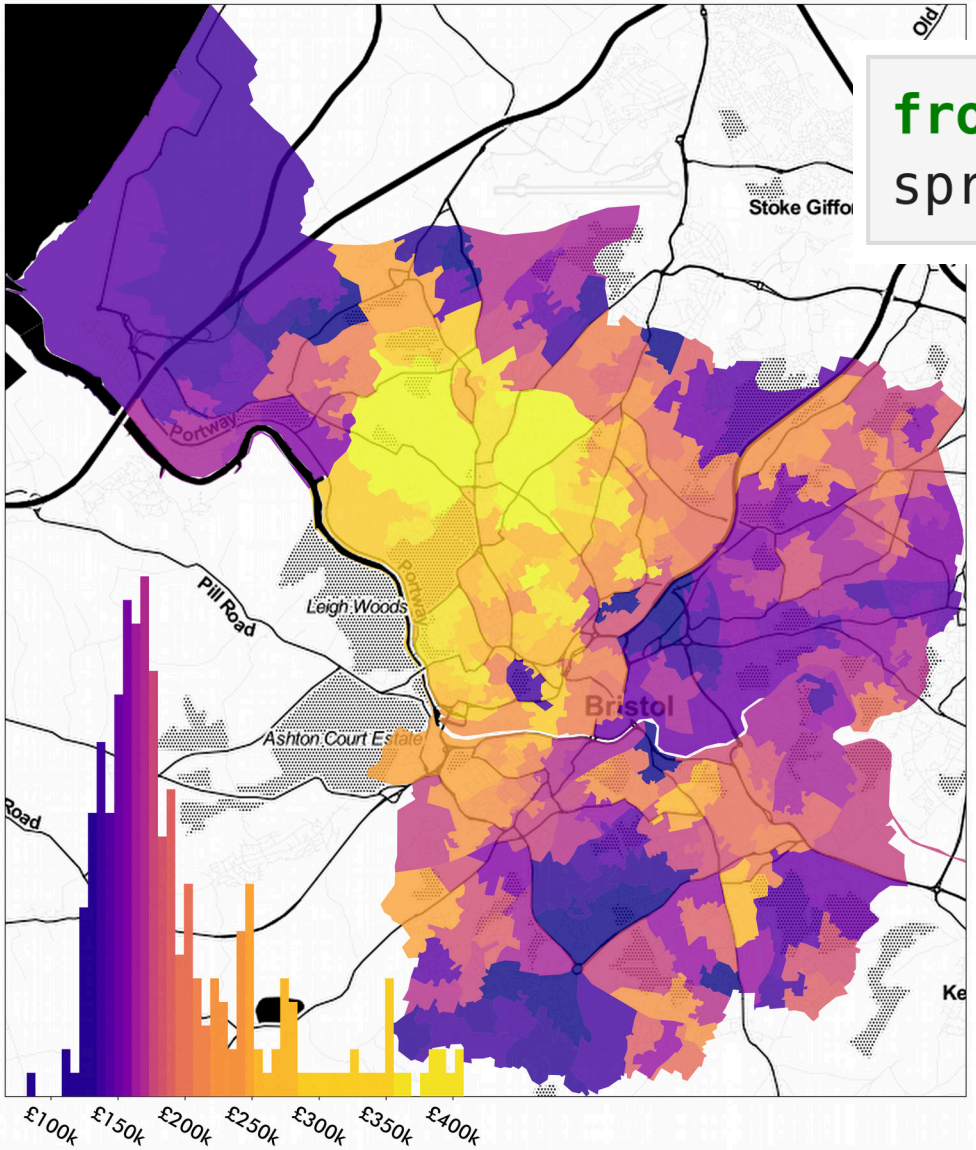

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```
(X_train, X_test,
 y_train, y_test) = train_test_split(X,y, test_size=.2)
```

```
model = LinearRegression().fit(X_train, y_train)
```

$$\begin{bmatrix} y_{23} \\ y_{2596} \\ \dots \\ y_{834} \end{bmatrix} = \begin{bmatrix} x_{23,1} & x_{23,2} & \dots & x_{23,P} \\ x_{2596,1} & x_{2596,2} & \dots & x_{2596,P} \\ \vdots & & \ddots & \vdots \\ x_{834,1} & x_{834,2} & \dots & x_{834,P} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_{23} \\ \epsilon_{2596} \\ \vdots \\ \epsilon_{834} \end{bmatrix}$$

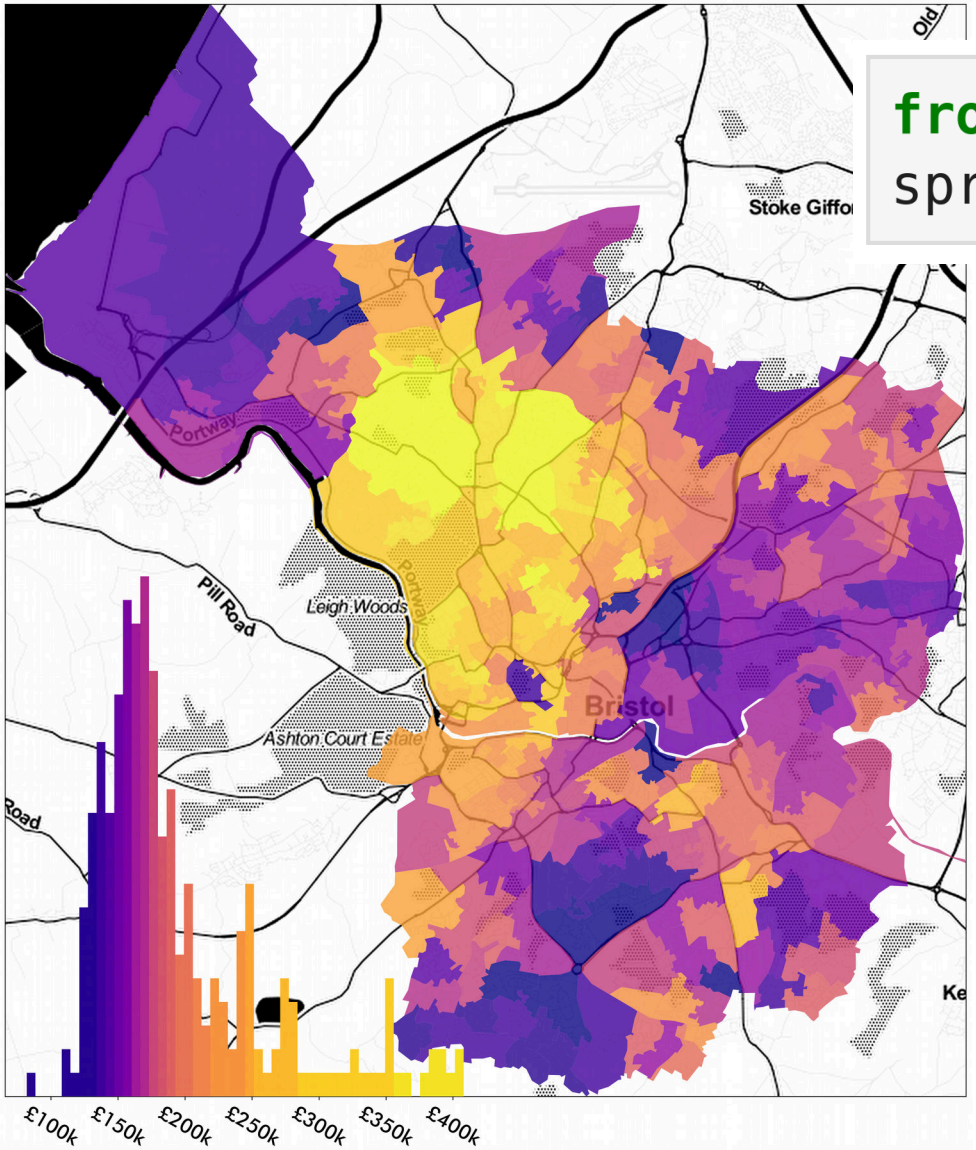
USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE




```
from pysal.model import spreg
spreg.GM_Lag(X_train, y_train, w=w)
```

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,P} \\ x_{2,1} & x_{2,2} & \dots & x_{2,P} \\ \vdots & & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,P} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE

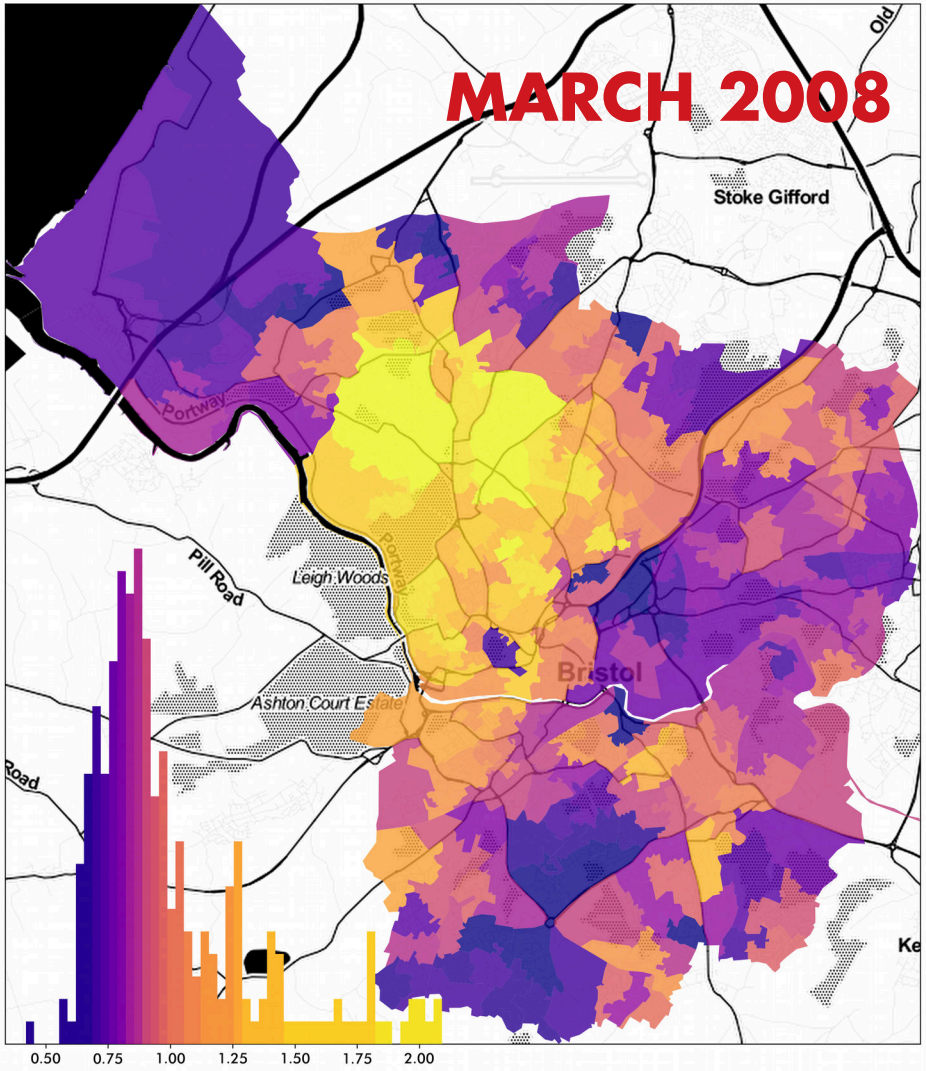


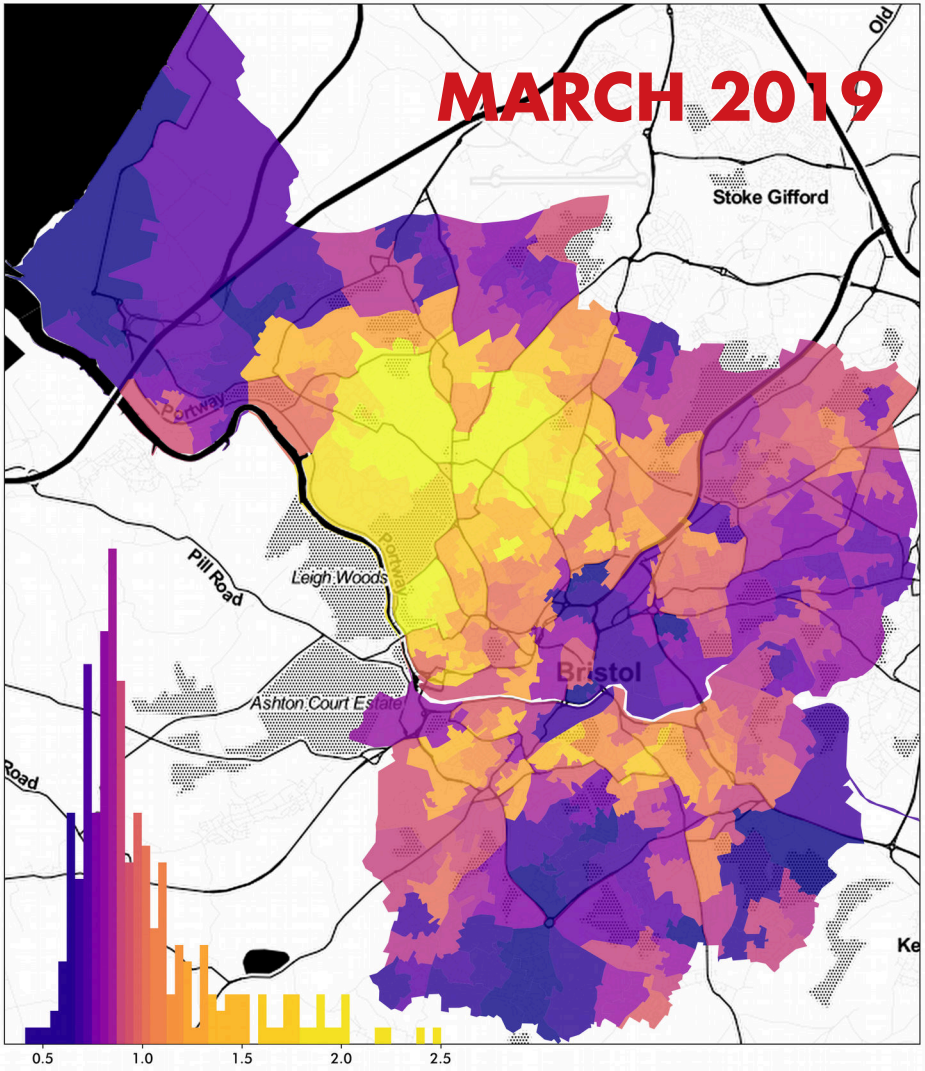

```
from pysal.model import spreg
spreg.GM_Lag(X_train, y_train, w=w)
```

Graph that encodes information about adjacency, proximity, etc. between obs.

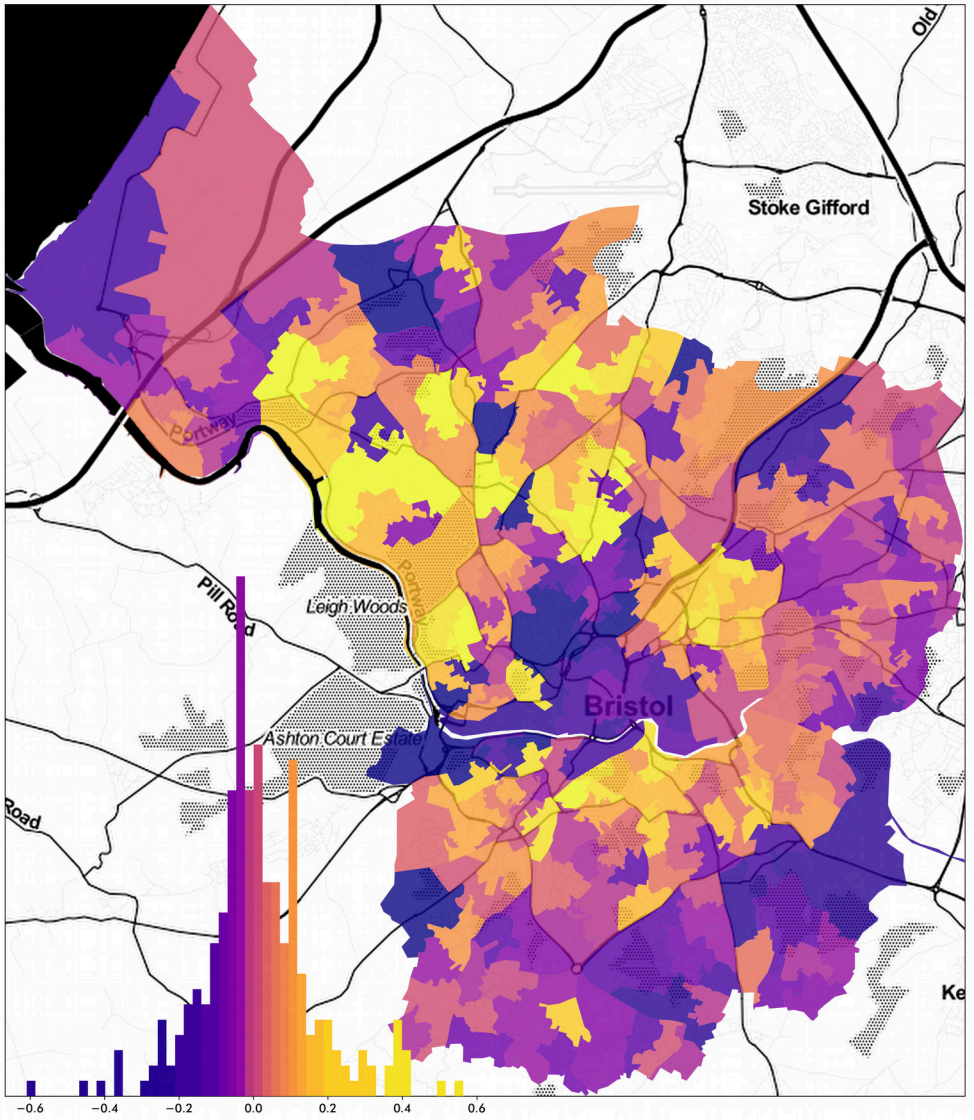
$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,P} \\ x_{2,1} & x_{2,2} & \dots & x_{2,P} \\ \vdots & & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,P} \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE



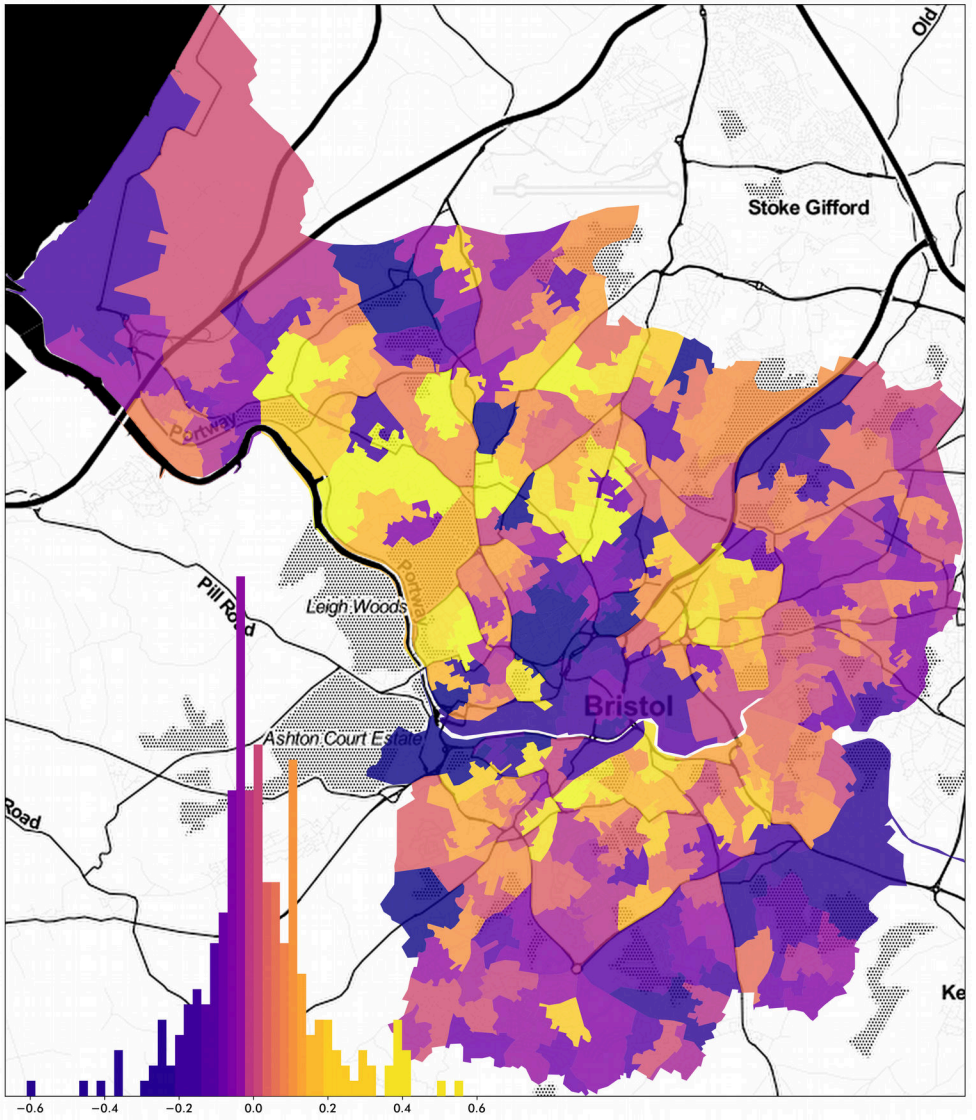


USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE

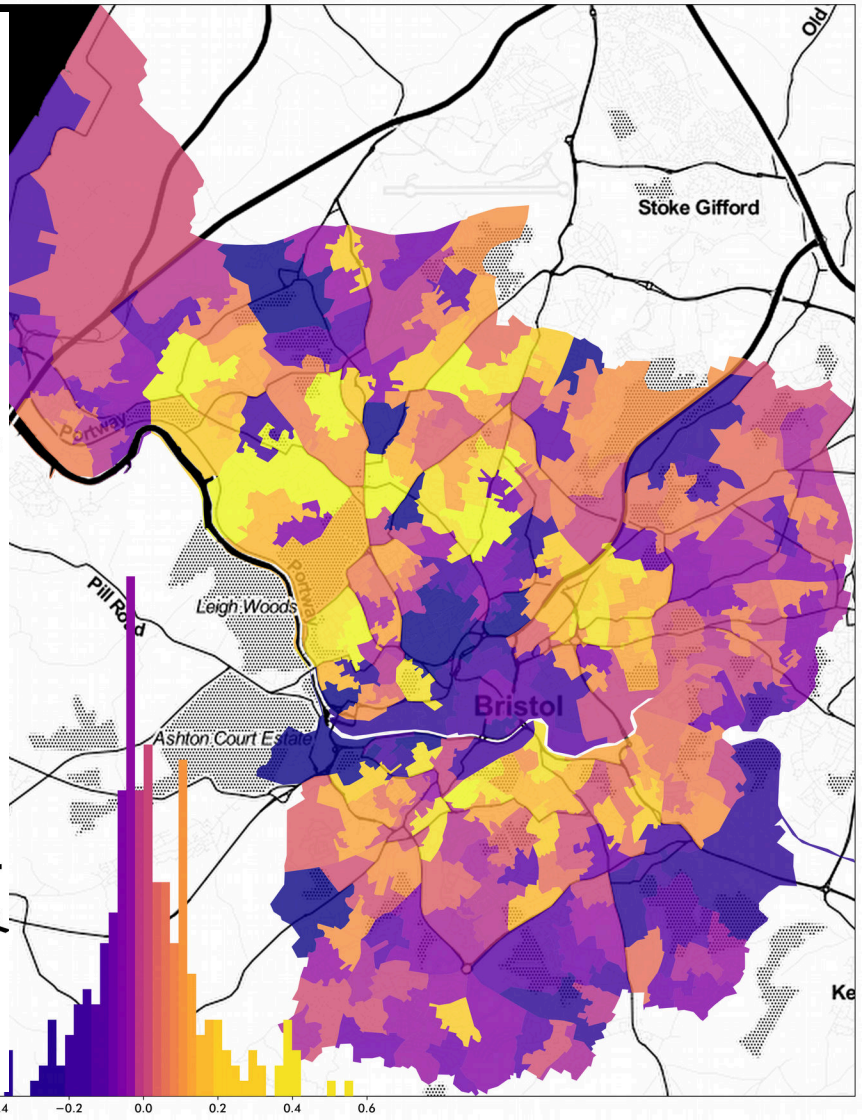


USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE

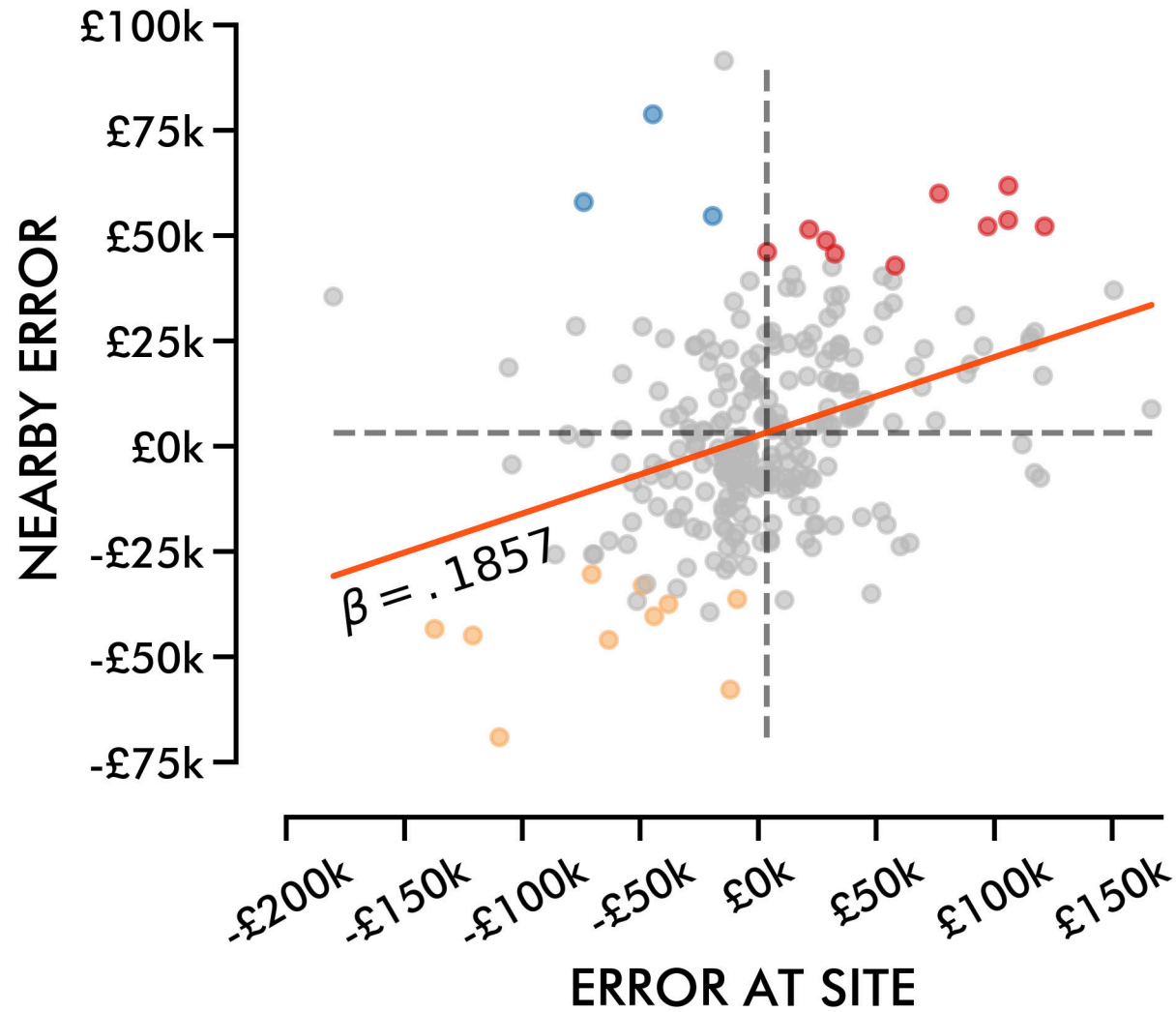
IS ERROR CLUSTERED?

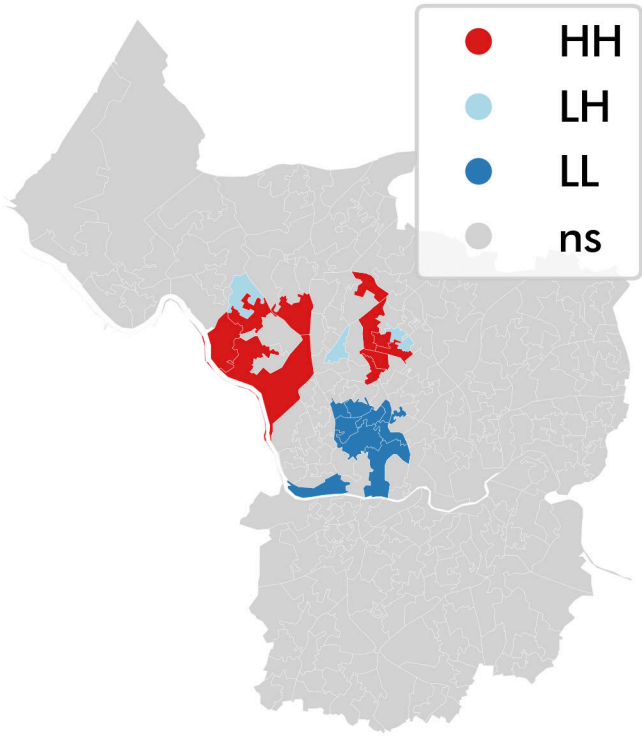


USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE



USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE





USE METHODS THAT EXPLICITLY LEARN FROM GEOGRAPHIC STRUCTURE

WHY ARE GEOGRAPHERS?

Geographers use relationships between people & environment to help

GEOGRAPHY IS EVERYWHERE

Geography, like time, is ubiquitous & contextualizes knowledge

HOW CAN GEOGRAPHY HELP?

Spatial feature engineering & explicitly-geographic learning helps analysis

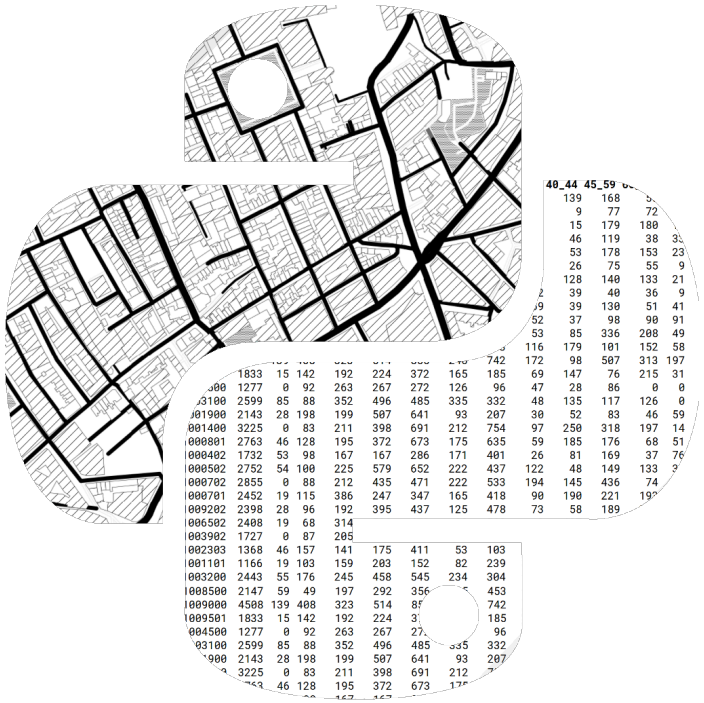
HOW TO LEARN MORE

AUTOMATING GIS PROESSES

Henrikki Tenkanen & Vuokko Heikinheimo

Solving Geographic Information Processing problems
with high-performance Python. CRC Press, 202X.

automating-gis-processes.github.io



GEOGRAPHIC DATA SCIENCE

Sergio Rey

Daniel Arribas-Bel

Levi John Wolf

Methods & models that do (un)supervised learning with Python using PySAL & Scikit-Learn. CRC Press, 2022.

geographicdata.science/book