

lab_09

November 29, 2015

1 Points

Points are spatial entities that can be understood in two fundamentally different ways. On the one hand, points can be seen as fixed objects in space, which is to say their location is taken as given (*exogenous*). In this case, analysis of points is very similar to that of other types of spatial data such as polygons and lines. On the other hand, points can be seen as the occurrence of an event that could theoretically take place anywhere but only manifests in certain locations. This is the approach we will adopt in the rest of the notebook.

When points are seen as events that could take place in several locations but only happen in a few of them, a collection of such events is called a *point pattern*. In this case, the location of points is one of the key aspects of interest for analysis. A good example of a point pattern is crime events in a city: they could technically happen in many locations but we usually find crimes are committed only in a handful of them. Point patterns can be *marked*, if more attributes are provided with the location, or *unmarked*, if only the coordinates of where the event occurred are provided. Continuing the crime example, an unmarked pattern would result if only the location where crimes were committed was used for analysis, while we would be speaking of a marked point pattern if other attributes, such as the type of crime, the extent of the damage, etc. was provided with the location.

Point pattern analysis is thus concerned with the description, statistical characterization, and modeling of point patterns, focusing specially on the generating process that gives rise and explains the observed data. *What's the nature of the distribution of points? Is there any structure we can statistically discern in the way locations are arranged over space? Why do events occur in those places and not in others?* These are all questions that point pattern analysis is concerned with.

This notebook aims to be a gentle introduction to working with point patterns in Python. As such, it covers how to read, process and transform point data, as well as several common ways to visualize point patterns.

```
In [1]: %matplotlib inline

import numpy as np
import pandas as pd
import geopandas as gpd
import pysal as ps
import seaborn as sns
import matplotlib.pyplot as plt
from shapely.geometry import Point
```

1.1 Data

We are going to dip our toes in the lake of point data by looking at a sample of geo-referenced tweets in the city of Liverpool. The dataset we will be playing with contains the location of over 130,000 messages posted on Twitter from January to the end of October of 2012. A detailed description of the variables included is provided in the “Datasets” section of the course website, as well as instructions to download it.

Once you have downloaded it and extracted the compressed .zip file, let us first set the paths to the shapefile. In addition, we will also be using the LSOA Census and geography dataset we already know, so let us add the path in advance to make things easier later on:

```
In [2]: # This will be different on your computer and will depend on where
        # you have downloaded the files

        # Twitter dataset
        tw_path = '../../../../../data/tweets/tweets_liverpool/tweets_liverpool.shp'
        # LSOAs polygons
        lsoas_path = '../../../../../data/Liverpool/shapefiles/Liverpool_lsoa11.shp'
        # Total population counts from Census Geodata Pack
        pop_path = '../../../../../data/Liverpool/tables/CT0010_lsoa11.csv'
```

Since the data are stored in a shapefile, loading it is in the same way as usual:

```
In [3]: %%time
        # Read the file
        tw = gpd.read_file(tw_path)
        # Create a brief summary of the columns in the table
        tw.info()
```

```
<class 'geopandas.geodataframe.GeoDataFrame'>
Int64Index: 131209 entries, 0 to 131208
Data columns (total 12 columns):
DAY          131209 non-null int64
DOW          131209 non-null int64
HOUR         131209 non-null int64
LAT          131209 non-null float64
LON          131209 non-null float64
LSOA11CD     131209 non-null object
MINUTE       131209 non-null int64
MONTH        131209 non-null int64
X            131209 non-null float64
Y            131209 non-null float64
YEAR         131209 non-null int64
geometry     131209 non-null object
dtypes: float64(4), int64(6), object(2)
memory usage: 13.0+ MB
CPU times: user 22.9 s, sys: 507 ms, total: 23.4 s
Wall time: 23.6 s
```

Note how we have also added the command `%%time` at the top of the cell. Once the cell has run, this provides an accurate measurement of the time it took the computer to run the code. We are adding this because, as you can see in the description of the columns, this is a fairly large table, with 131,209 rows.

Depending on the running time of the cell above, it is recommended you do not use the full dataset but instead you shorten it and consider only a random sample of tweets (which retains the same properties). If it took your computer longer than 20 seconds to read the file (as indicated at the end of the cell output, `total`), you are strongly encouraged to subset your data by taking a random sample. This will make the rest of the analysis run much more smoothly on your computer and will result in a better experience. See below for details on how to do this.

1.1.1 Random sample of tweets

Once we have loaded the data, taking a random sample is a relative simple operation. Let us first perform the computations and then delve into the steps, one by one.

```
In [4]: # Set the "seed" so every run produces the generates the same random numbers
np.random.seed(1234)
# Create a sequence of length equal to the number of rows in the table
ri = np.arange(len(tw))
# Randomly reorganize (shuffle) the values
np.random.shuffle(ri)
# Reindex the table by using only the first 10,000 numbers
# of the (now randomly arranged) sequence
tw = tw.iloc[ri[:10000], :]
# Display summary of the new table
tw.info()
```

```
<class 'geopandas.geodataframe.GeoDataFrame'>
Int64Index: 10000 entries, 50049 to 86084
Data columns (total 12 columns):
DAY          10000 non-null int64
DOW          10000 non-null int64
HOUR         10000 non-null int64
LAT          10000 non-null float64
LON          10000 non-null float64
LSOA11CD     10000 non-null object
MINUTE       10000 non-null int64
MONTH        10000 non-null int64
X            10000 non-null float64
Y            10000 non-null float64
YEAR         10000 non-null int64
geometry     10000 non-null object
dtypes: float64(4), int64(6), object(2)
memory usage: 1015.6+ KB
```

Let us walk through the strategy taken to randomly sample the table:

- First we create a separate sequence of numbers starting from zero (Python always starts counting on zero, not one) as long as the number of rows in the table we want to subset. At this point, this list starts on 0, then 1, 2, 3, 4, 5, ..., $N-1$ (with N the length of the table, that is 131,209).
- Then, in line 4, the list is randomly rearranged. After this, the length is still the same -131,209- but the order has changed from the original sequence to a completely random one.
- At this point, we can subset the table, which we do in line 7. This command is composed of two elements: one (`ri[:10000]`) in which we keep only the first 10,000 elements of the randomly ordered list (if you wanted to subset the table to have a different number of observations, change that in here); the second (`tw.iloc`) is a standard subsetting query as we have been doing so far.

The trick here is that by querying the table on the subset of 10,000 numbers obtained from a random draw of the entire set, we are only keeping the rows indexed on those numbers. This attains two things: one, it returns only 10,000 observations instead of the total 131,209; two, the subset that it does keep is entirely random, as the index used for it has been randomly “shuffled”.

1.2 Visualization of a Point Pattern

We will spend the rest of this notebook learning different ways to visualize a point pattern. In particular, we will consider to main strategies: one relies on aggregating the points into polygons, while the second one is based on creating continuous surfaces using kernel density estimation.

1.2.1 Points meet polygons

Having learned about visualization of lattice (polygon) data, the most straightforward way to visualize point patterns is to “turn” them into polygons and apply techniques like choropleth mapping to visualize their

spatial distribution. To do that, we will overlay a polygon layer on top of the point pattern, *join* the points to the polygons by assigning to each point the polygon where they fall into, and create a choropleth of the counts by polygon. This approach is very intuitive but of course raises the following question: *what polygons do we use to aggregate the points?* Ideally, we want a boundary delineation that matches as closely as possible the point generating process and partitions the space into areas with a similar internal intensity of points. However, that is usually not the case, no less because one of the main reasons we typically want to visualize the point pattern is to learn about such generating process, so we would typically not know a priori whether a set of polygons match it. If we cannot count on the ideal set of polygons to begin with, we can adopt two more realistic approaches: using a set of pre-existing irregular areas or create a artificial set of regular polygons.

Irregular lattices To exemplify this approach, we will use the areas of the LSOAs that we have been working with throughout the course. So, before anything else, let us load them up into an object we will call `lsoas`:

```
In [5]: lsoas = gpd.read_file(lsoas_path).set_index('LSOA11CD')
```

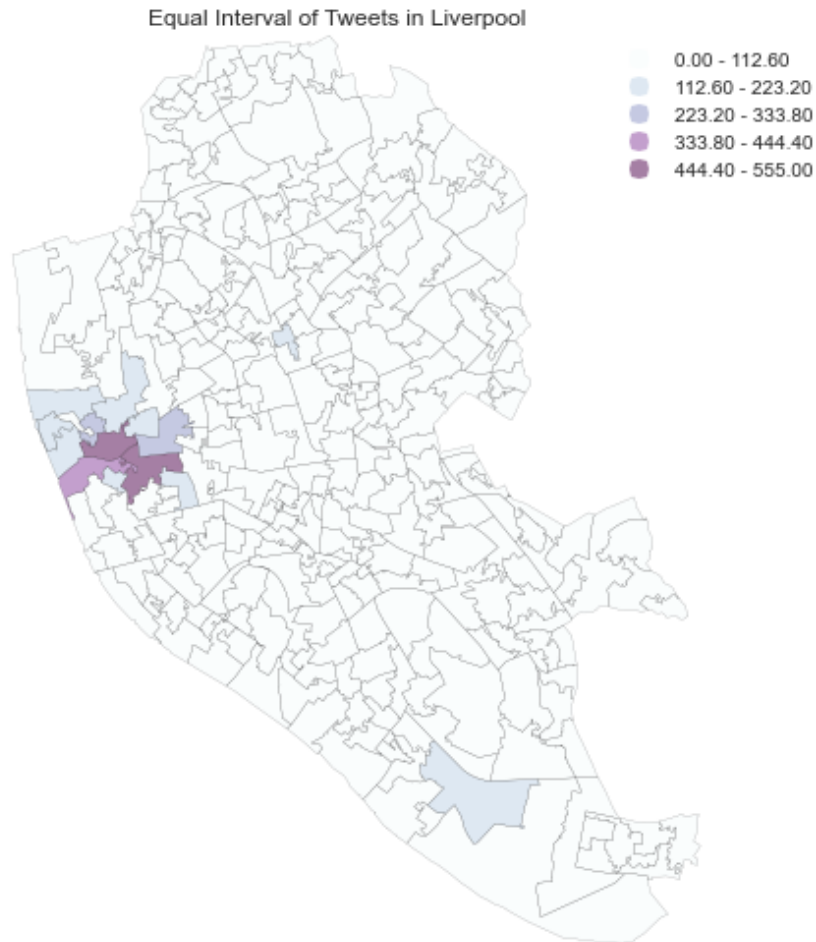
The next step we require is to assign to each tweet the LSOA where it was posted from. This can be done through a standard GIS operation called point-in-polygon. For the sake of keeping the focus on the visualization of points, the tweet dataset already includes the LSOA identifier where each tweet falls into in the column `LSOA11CD`. However, if you were exploring a brand new dataset and had to join it by yourself, you could do this in QGIS using the point-in-polygon tool available on the Vector menu (**Vector** → **Data Management Tools** → **Join Attributes by Location**). Alternatively, you could also perform this operation using `geopandas` and its “spatial join” extension. Although the latter is a bit more involved and advanced, it is also more efficient and fast.

Once we have the ID of the polygon where each tweet falls into, creating the count of tweets by polygon is one line of code away. Again, we rely on the `groupby` operator which takes all the tweets in the table and “groups” them “by” their LSOA code. Once grouped, we apply the method `size`, which counts how many elements each group has and returns a column indexed on the LSOA code with all the counts as its values. To make the mapping easier, we also assign the counts to a newly created column in the `lsoas` table.

```
In [6]: # Create counts
tw_lsoa = tw.groupby('LSOA11CD').size()
# Assign counts into a column in the LSOAS table
lsoas['tweet_count'] = tw_lsoa
```

At this point, we are ready to map the counts. Technically speaking, this is a choropleth just as we have seen many times before (see Lab 4 if you need a refresher):

```
In [7]: # Set up figure and axis
f, ax = plt.subplots(1, figsize=(9, 9))
# Plot the equal interval choropleth and add a legend
lsoas.plot(column='tweet_count', scheme='equal_interval', legend=True, \
           axes=ax, colormap='BuPu', linewidth=0.1)
# Remove the axes
ax.set_axis_off()
# Set the title
ax.set_title("Equal Interval of Tweets in Liverpool")
# Keep axes proportionate
plt.axis('equal')
# Draw map
plt.show()
```



[Optional exercise]

Create a similar choropleth as above but use a quantile or Fisher-Jenks classification instead of equal interval. What are the main differences? Why do you think this is the case? How does it relate to the distribution of counts by polygons?

The map above clearly shows a concentration of tweets in the city centre of Liverpool. However, it is important to remember that the map is showing *raw* counts of tweets. At this point it is useful to remember what we discussed in Labs 3 and 4 about mapping raw counts. In the case to tweets, as with many other phenomena that affect to only a portion of the population, it is crucial to keep in mind the underlying population. Although tweets could theoretically take place anywhere on the map, they really can only appear in areas where there are people who can post the messages. If population is not distributed equally (and most often it is not) and we ignore its spatial pattern, the map of raw counts will most likely simply display the overall pattern of the underlying population. In this example, if all we map is raw counts of

tweets, we are showing a biased picture towards areas with high levels of population because, everything else equal, the more people the more potential for tweets to appear.

To obtain a more accurate picture, what we would like to see is a map of the *intensity* of tweets, not of raw counts. To do this, ideally we want to divide the number of tweets per polygon by the total number of potential population who could tweet at any given moment. This of course is not always available, so we have to resort to proxy variables. For the sake of this example, we will use the residential population. Let us first load it up:

```
In [8]: # Load table with population counts (and other variables too)
pop = pd.read_csv(pop_path, index_col=0)
# Total Population is 'CT00100001'
pop = pop['CT00100001']
pop.head()
```

```
Out[8]: GeographyCode
E01006512    1880
E01006513    2941
E01006514    2108
E01006515    1208
E01006518    1696
Name: CT00100001, dtype: int64
```

Now we can insert it as a new column in the `lsoas` table:

```
In [9]: lsoas['Total_Pop'] = pop
```

At this point, we can easily calculate the ratio of tweets per resident of each of the areas:

```
In [10]: lsoas['tweet_int'] = lsoas['tweet_count'] / lsoas['Total_Pop']
lsoas.head()
```

```
Out[10]:
```

	geometry	tweet_count	\
LSOA11CD			
E01006512	POLYGON ((336103.358 389628.58, 336103.416 389...	171	
E01006513	POLYGON ((335173.781 389691.538, 335169.798 38...	494	
E01006514	POLYGON ((335495.676 389697.267, 335495.444 38...	101	
E01006515	POLYGON ((334953.001 389029, 334951 389035, 33...	47	
E01006518	POLYGON ((335354.015 388601.947, 335354 388602...	15	

	Total_Pop	tweet_int
LSOA11CD		
E01006512	1880	0.090957
E01006513	2941	0.167970
E01006514	2108	0.047913
E01006515	1208	0.038907
E01006518	1696	0.008844

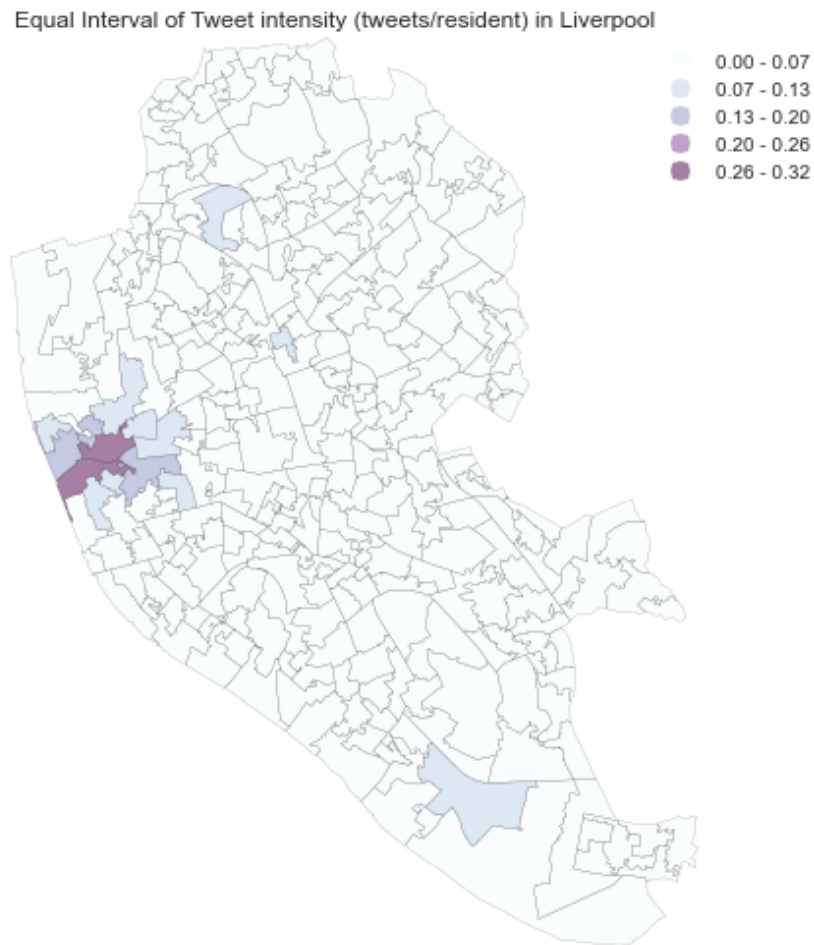
With the intensity at hand, creating the new choropleth takes exactly the same as above:

```
In [11]: # Set up figure and axis
f, ax = plt.subplots(1, figsize=(9, 9))
# Plot the equal interval choropleth and add a legend
lsoas.plot(column='tweet_int', scheme='equal_interval', legend=True, \
           axes=ax, colormap='BuPu', linewidth=0.1)
# Remove the axes
ax.set_axis_off()
```

```

# Set the title
ax.set_title("Equal Interval of Tweet intensity (tweets/resident) in Liverpool")
# Keep axes proportionate
plt.axis('equal')
# Draw map
plt.show()

```



In this case, the pattern in the raw counts is so strong that the adjustment by population does not have a huge effect, but in other contexts mapping rates can yield very different results than mapping simple counts.

[Optional exercise]

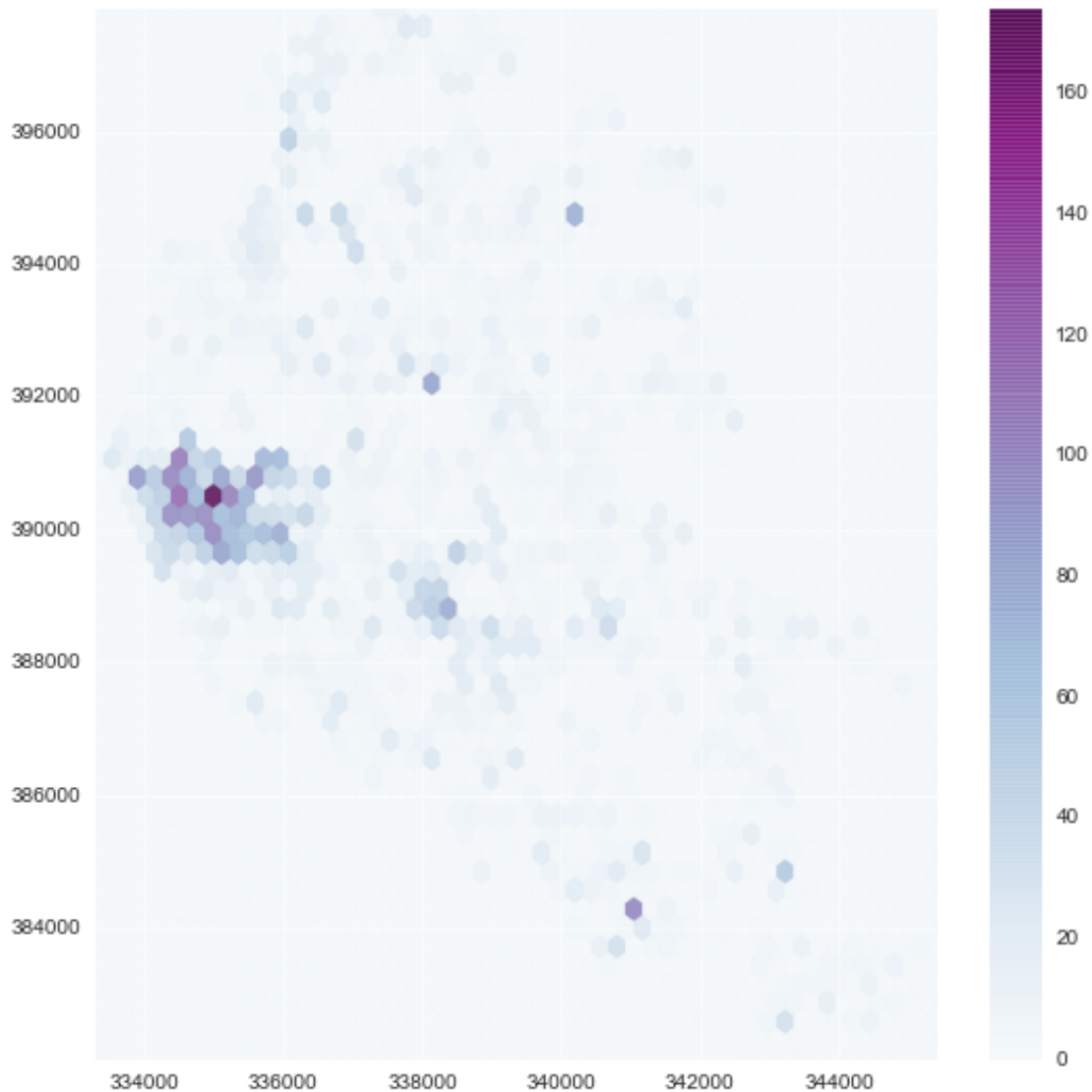
Create a similar choropleth as above but use a quantile or Fisher-Jenks classification instead of equal interval. What are the main differences? Why do you think this is the case? How does it relate to the distribution of counts by polygons?

Regular lattices: hex-binning Sometimes we either do not have any polygon layer to use or the ones we have are not particularly well suited to aggregate points into them. In these cases, a sensible alternative is to create an artificial topology of polygons that we can use to aggregate points. There are several ways to do this but the most common one is to create a grid of hexagons. This provides a regular topology (every polygon is of the same size and shape) that, unlike circles, cleanly exhausts all the space without overlaps and has more edges than squares, which alleviates edge problems.

Python has a simplified way to create this hexagon layer *and* aggregate points into it in one shot thanks to the method `hexbin`, which is available in every axis object (e.g. `ax`). Let us first see how you could create a map of the hexagon layer alone:

```
In [12]: # Setup figure and axis
f, ax = plt.subplots(1, figsize=(9, 9))
# Add hexagon layer that displays count of points in each polygon
hb = ax.hexbin(tw.X, tw.Y, gridsize=50, alpha=0.8, cmap='BuPu')
# Add a colorbar (optional)
plt.colorbar(hb)
```

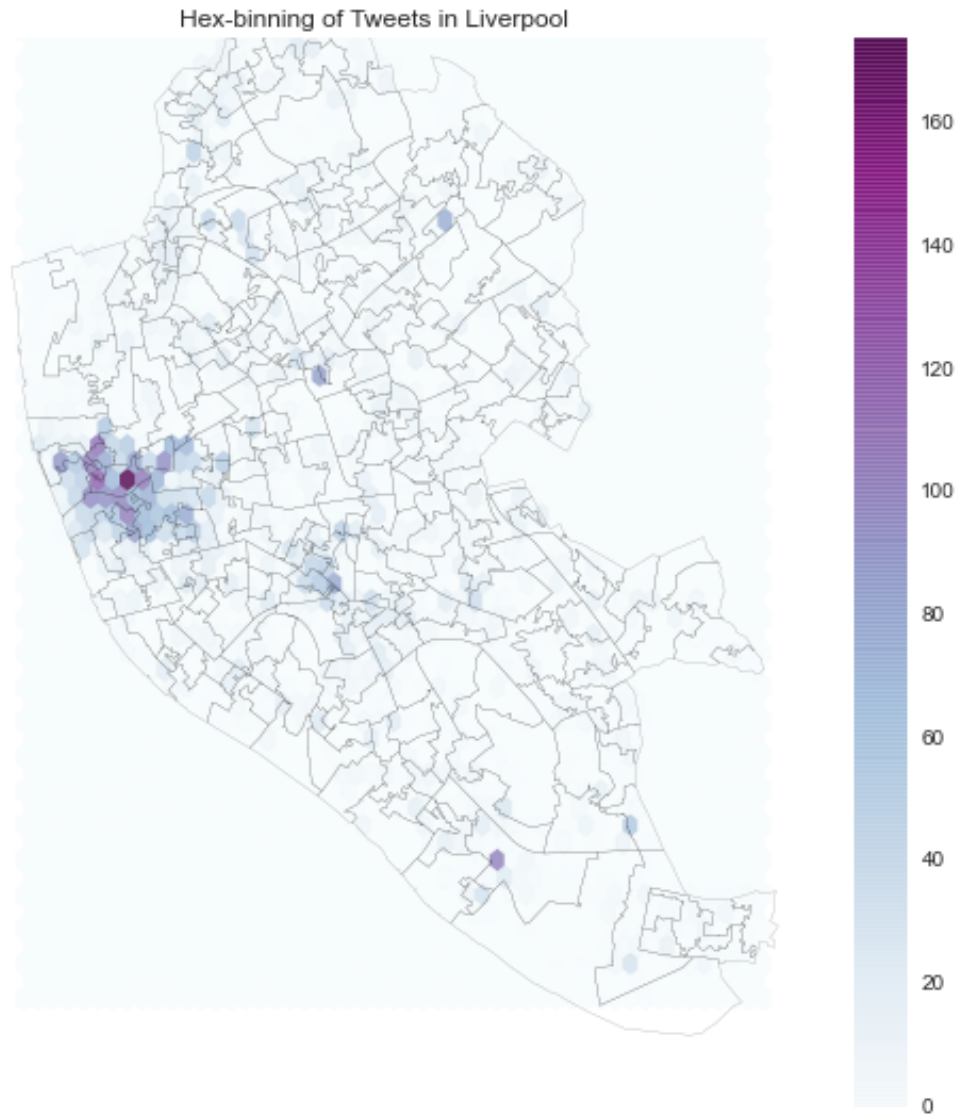
```
Out[12]: <matplotlib.colorbar.Colorbar instance at 0x116f52a70>
```



See how all it takes is to set up the figure and call `hexbin` directly using the set of coordinate columns (`tw.X` and `tw.Y`). Additional arguments we include is the number of hexagons by axis (`gridsize`, 50 for a 50 by 50 layer), the transparency we want (80%), and the colormap of our choice (BuPu in our case). Additionally, we include a colorbar to get a sense of what colors imply. Note that we need to pass the name of the object that includes the `hexbin` (`hb` in our case), but keep in mind this is optional, you do not need to always create one.

Once we know the basics, we can insert it into the usual plotting routine we have been using to generate a complete hex-bin map of tweets in Liverpool:

```
In [13]: # Set up figure and axis
         f, ax = plt.subplots(1, figsize=(9, 9))
         # Add a base layer with the LSOA geography
         for poly in lsoas['geometry']:
             gpd.plotting.plot_multipolygon(ax, poly, facecolor='white', alpha=0, linewidth=0.1)
         # Add hexagon layer that displays count of points in each polygon
         hb = ax.hexbin(tw.X, tw.Y, gridsize=50, alpha=0.8, cmap='BuPu')
         # Add a colorbar (optional)
         plt.colorbar(hb)
         # Remove axes
         ax.set_axis_off()
         # Add title of the map
         ax.set_title("Hex-binning of Tweets in Liverpool")
         # Keep map proportionate
         plt.axis('equal')
         # Draw the map
         plt.show()
```



1.2.2 Kernel Density Estimation

NOTE: It is recommended that, for this section, you use the random subset of tweets rather than the entire batch of 131,209.

Using a hexagonal binning can be a quick solution when we do not have a good polygon layer to overlay the points directly and some of its properties, such as the equal size of each polygon, can help alleviate some of the problems with a “bad” irregular topology (one that does not fit the underlying point generating process). However, it does not get around the issue of the modifiable areal unit problem (M.A.U.P., see Lecture 4): at the end of the day, we are still imposing arbitrary boundary lines and aggregating based on them, so the possibility of mismatch with the underlying distribution of the point pattern is very real.

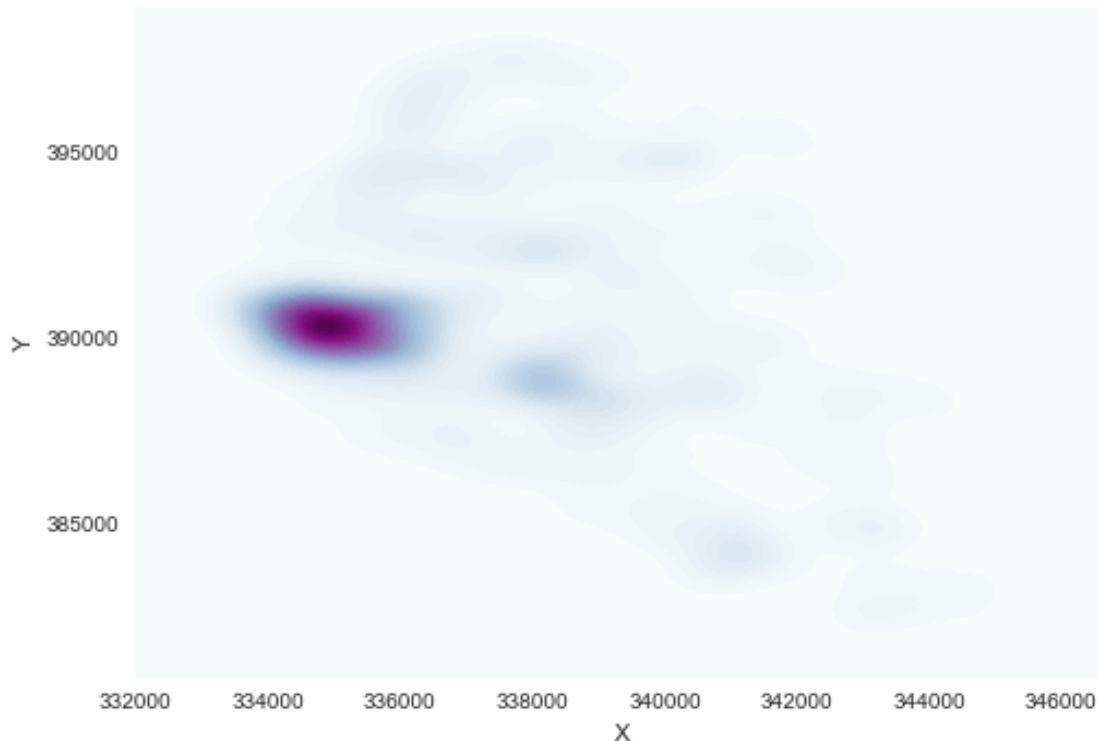
One way to work around this problem is to avoid aggregating altogether. Instead, we can aim at estimating the *continuous* observed probability distribution. The most commonly used method to do this is the so called *kernel density estimate* (KDE). The idea behind KDEs is to count the number of points in a *continuous* way. Instead of using discrete counting, where you include a point in the count if it is inside a certain boundary

and ignore it otherwise, KDEs use functions (kernels) that include points but give different weights to each one depending of how far of the location where we are counting the point is.

The actual algorithm to estimate a kernel density is not trivial but its application in Python is extremely simplified by the use of Seaborn's `kdeplot` command. Same as above, let us first see how to create the simplest possible KDE and then we will create a full-fledge map.

```
In [14]: sns.kdeplot(tw['X'], tw['Y'], n_levels=50, shade=True, cmap='BuPu')
```

```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x11d0d8d10>
```



Seaborn greatly streamlines the process and boils it down to a single line. The method `sns.kdeplot` (which we can also use to create a KDE of a single variable) takes the X and Y coordinate of the points as the only compulsory attributes. In addition, we specify the number of levels we want the color gradient to have (`n_levels`), whether we want to color the space in between each level (`shade`, yes), and the colormap of choice.

Once we know how the basic logic works, we can insert it into the usual mapping machinery to create a more complete plot. The main difference here is that we now have to tell `sns.kdeplot` where we want the surface to be added (`ax` in this case).

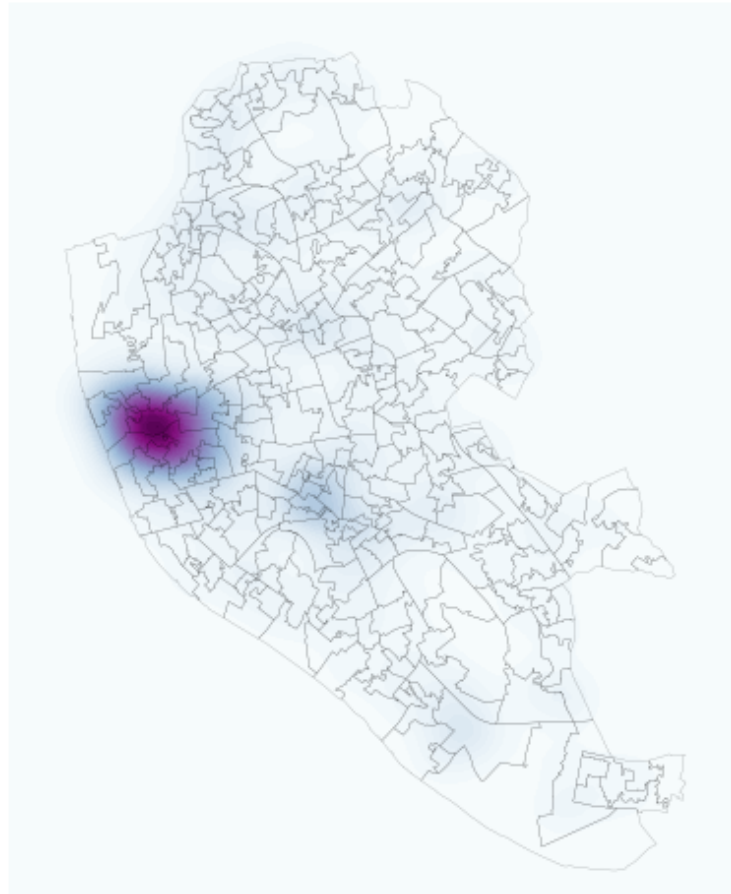
```
In [15]: # Set up figure and axes
f, ax = plt.subplots(1, figsize=(9, 9))
# Add a base layer with the LSOA geography
for poly in lsoas['geometry']:
    gpd.plotting.plot_multipolygon(ax, poly, facecolor='white', alpha=0, linewidth=0.1)
# Generate KDE
sns.kdeplot(tw['X'], tw['Y'], ax=ax, \
            n_levels=50, shade=True, cmap='BuPu')
# Remove axes
```

```

ax.set_axis_off()
# Add title
ax.set_title("Hex-binning of Tweets in Liverpool")
# Keep axes proportionate
plt.axis('equal')
# Draw map
plt.show()

```

Hex-binning of Tweets in Liverpool



1.3 Optional exercise (if time permits)

Reproduce the point analysis above with a different dataset of your choice. This involves:

- Obtain the data.
- Load the data in a notebook.
- If you can find a suitable polygon layer to which aggregate the points:
 - Perform a spatial join using QGIS.

- Aggregate points into the polygon geography by obtaining counts of points per polygon.
- Create a raw count choropleth.
- If you have a potential measure of the underlying population, create the ratios and generate a new choropleth.
- Create a Hex binning map of the points.
- Compute and display a kernel density estimate (KDE) of the distribution of the points.

As a suggestion, you can use the following additional datasets:

- **TO DO:** as Michalis for house prices.
- **TO DO:** find crime data.

Geographic Data Science'15 - Lab 9 by Dani Arribas-Bel is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.