Map Stitching Using Linear Models

Prasun De
BS1826

December 18, 2020

Contents

1 Introduction

We have a region of the area on the earth that can be considered flat ignoring the curvature , but it is big so we cannot fit it in one screen . We want to represent the landmarks on the big map in one single plot and will fit a linear model for this purpose .

2 Data Description

I have described in the other pdf how tha data can collect the data , and generate the map using our code .

3 Mathematical Model

Suppose we have I screenshots and J locations . Let x_{ijk} denote the the x coordinate of the k th click , at the j th landmark in the i th screen shot . We can similarly define y_{ijk} . We shall fit the following model :

$$x_{ijk} = \alpha_i + \beta_j + \epsilon_{ijk}$$
$$y_{ijk} = \alpha'_i + \beta'_i + \epsilon'_{ijk}$$

where $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$, $\epsilon'_{ijk} \sim \mathcal{N}(0, \sigma^2)$ where the errors ϵ'_{ijk} are iid and ϵ_{ijk} are iid .

Note that by the nature of our problem , we will not have all (i, j) pairs in our data . We can see why fitting this model is very intuitive :

Suppose we have some global origin . Let a_i denote the x-coordinate of the origin of the i th screenshot wrt this global origin , and let b_j denote the x-coordinate of the j th landmark wrt the global origin . Then , we can say that the position of the j th landmark wrt the i th screenshot 's origin is $-a_i + b_j$. By adjusting the sign of a_i , we can see that this clearly has the form of the two way anova model . Similar reasoning will hold for the y coordinate . We do not expect any type of interaction . This helps us to guess that we the two-way anova model is a good choice for the data .

4 Rank of the design matrix

The data looks almost like a general scenario of two way anova model , but there is one catch : Not all (i, j) pairs will be there in out data . So this may affect the rank of our design matrix . We explain this as follows :

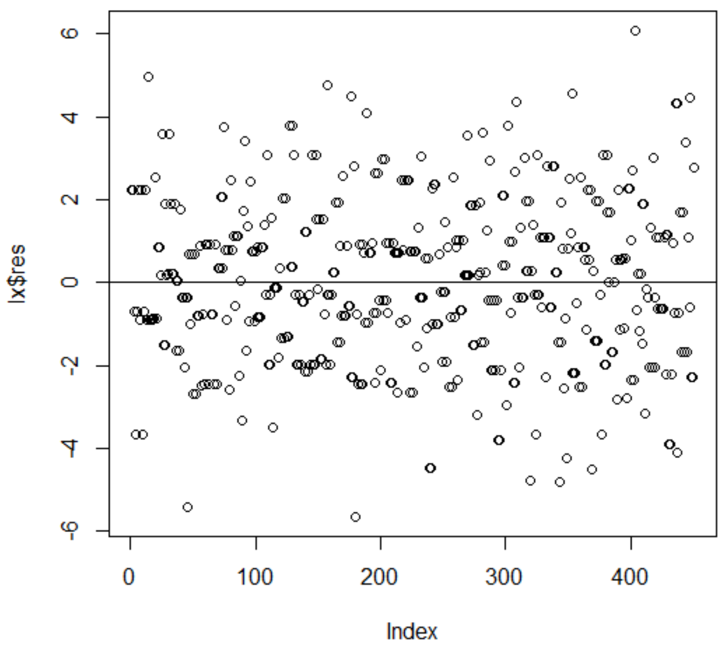
Let G denote a graph with landmarks as vertices . Two vertices are connected iff there is one common screenshot that will contain the landmarks corresponding to the two vertices . Consider a connected component of this graph , and pick a landmark from that connected component C . Suppose it is the q th landmark on the p th screenshot . Say we are first looking at the x coordinate . If we fix α_p , then it will fix β_q as their difference is estimable under the gauss markov setup . Also , this will fix all neighbouring α_i's and β_j's . As it is connected , we can say that fixing one parameter fixes all the other parameters (for x or y coordinates) in that particular component . This tells us that the rank of our design matrix will be $I + J - n$ where n is the number of connected components in G .

If the screen shots are not connected , then it is not possible to draw the map as we require the connectedness of the entire graph G that we defined earlier . So , when we are making the map , we take the screen shots such that the landmarks are connected , and then the rank of the design matrix is simply $I + J - n$. So all (i, j) pairs may not be there , but what matters is the connectivity of the graph . If it is ensured , we can plot the map , and the rank of the design matrix will be $I + J - 1$.

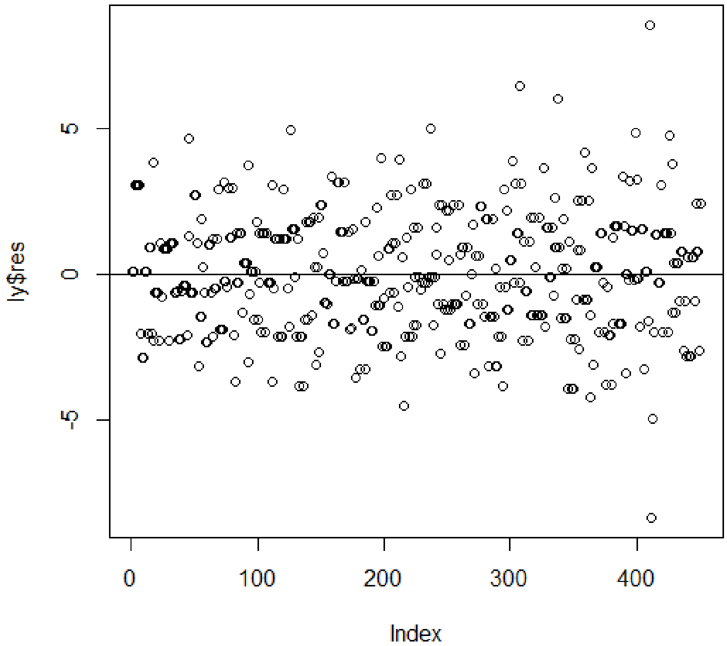
5 Diagnosis

The residual plots look good , there are no obvious outliers (for x residuals) and there is randomness in the plots . We can use the R inbuilt function 'influence.measures' to detect possible influential points . To detect outliers that may occur due to user error , we can try to use the idea of inter quartile ranges . If $Q1$ is the first quartile of out data , $Q3$ is the third quartile , then we can define interquartile range as $IQR = Q3 - Q1$. We can say that a point is an outlier if it lies outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$. From the fitted data, we did not observe any outliers according to the latter method . But , there were some influential points observed . A lot of these inluential points were in multiple screeenshots , which explains why they are influential . Also some points that were only in one single screenshots were seen to be influential . Also the qq plot of the residuals looked good which justifies our assumption for normality of errors .

Plot of the residuals :



It looks random , with no such obvious outliers



There is one obvious outlier here , which we see lies after the 400th entry somewhere . So we looked at the residuals from entry 400 to 451 and found the culprit . There had been an inadvertent error made by me in the data entry , and we found that out by observing the residual plot .

```
    x.coordinate y.coordinate          Landmark Screenshot.No
411     79.76863     398.1369 isi football court            11
412     74.70483     381.2576 isi football court            11
>
```

We observe that there is a lot of difference in the two y cordinates . We have actually considered multiple clicks for a given landmark in a given screenshot . But there had been an error in clicking on ISI football court . Thus we found out the error .

We found that the following instances of data might be 'influential' :

```
> tmp = influence.measures(lx)
> dat[which(apply(tmp$is.inf,1,sum) > 0),]
    x.coordinate y.coordinate               Landmark Screenshot.No
14      110.8745     411.1102             isi subway             1
45      751.5660     788.0494         maa kali temple            1
108    1160.0458     627.6958          laxmi bhandar             2
112    1160.0458     631.0717          laxmi bhandar             2
143     518.6312     308.6764 new kalpana bhujia store           3
157     251.9378     399.8248  car parking area of isi           3
158     246.8740     399.8248  car parking area of isi           3
159     246.8740     403.2007  car parking area of isi           3
161     246.8740     398.1369  car parking area of isi           3
180     532.1347     487.5973            isi nursery             4
317     682.3607     156.7625         intex asha care            8
318     682.3607     153.3866         intex asha care            8
320     677.2969     155.0745         intex asha care            8
321     682.3607     153.3866         intex asha care            8
323    1016.5715     207.4005 anshika training pvt ltd           8
325    1016.5715     204.0246 anshika training pvt ltd           8
327    1016.5715     209.0884 anshika training pvt ltd           8
343     940.6145     524.7319     glamour hair cutting           9
369     918.6714     212.4643        the french street           9
370     354.9017     247.9109            ifl gold loan           9
383     365.0293     349.1868       isi transport unit          10
387     365.0293     345.8110       isi transport unit          10
404     754.9419     177.0177                 amrapali          11
```

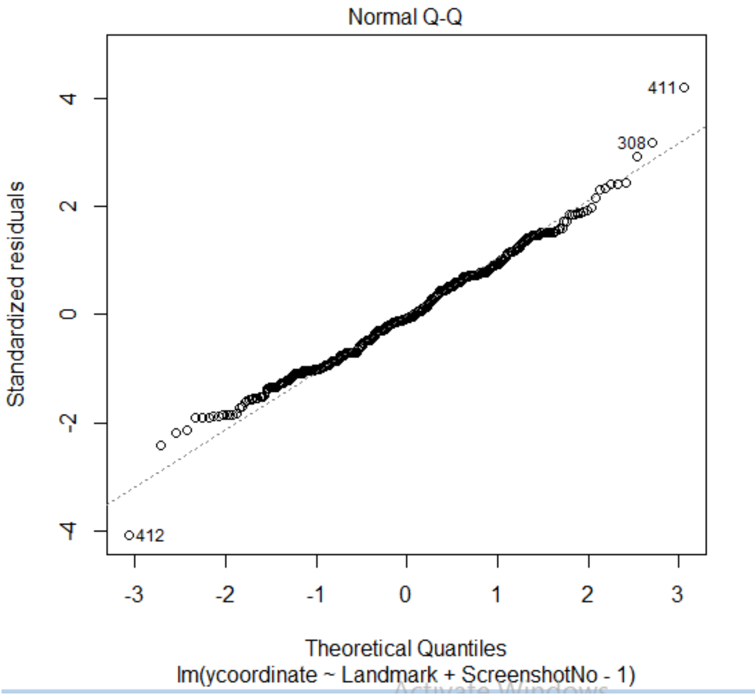
Mainly these points are either in a lot of common screenshots or very less number of screenshots . In a dataset of so many points , we can have many influential points .

We can repeat the process for the y cooordinates :

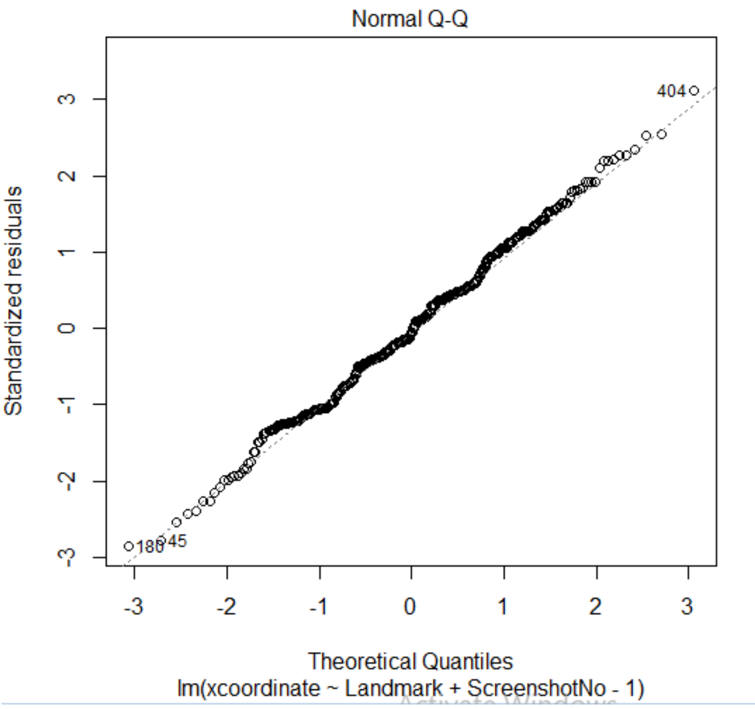
```
> tmp = influence.measures(ly)
> dat[which(apply(tmp$is.inf,1,sum) > 0),]
    x.coordinate y.coordinate                Landmark Screenshot.No
84    1188.74066     490.9732               new hotel             2
85    1190.42860     490.9732               new hotel             2
101    832.58678     740.7873        isi hostel campus            2
108   1160.04580     627.6958           laxmi bhandar             2
109   1163.42167     627.6958           laxmi bhandar             2
110   1158.35787     627.6958           laxmi bhandar             2
126    994.62836     173.6418              isi subway             3
145    522.00710     310.3644 new kalpana bhujia store            3
147    516.94330     310.3644 new kalpana bhujia store            3
156    245.18607     399.8248  car parking area of isi            3
157    251.93781     399.8248  car parking area of isi            3
158    246.87401     399.8248  car parking area of isi            3
237   1065.52155     384.6334               dj gagcix             6
308    392.03625     291.7971      tyre and tube repair           8
320    677.29694     155.0745          intex asha care            8
338   1127.97507     519.6681                    msme             9
367    923.73517     212.4643        the french street            9
368    925.42310     212.4643        the french street            9
369    918.67137     212.4643        the french street            9
374    353.21379     249.5988           ifl gold loan             9
391    444.36218     198.9608          isi green house           10
393    446.05011     198.9608          isi green house           10
399    359.96552     781.2977        isi football court          10
411     79.76863     398.1369        isi football court          11
412     74.70483     381.2576        isi football court          11
413     74.70483     384.6334        isi football court          11
426    381.90865     745.8511                amrapali            12
> |
```

Note that as expected, we have entries corresponding to the ISI FOOTBALL COURT as we had made an error while clicking on the landmark once (out of multiple clicks for each screenshot).

We observe the QQ plots :



The two major deviations in the end correspond to the two outliers that we explained . So , removing those values will make it a good qq plot . We have the following qq plot for the x residuals .



This is a pretty good fit .

Similar analysis must be done by the user rather than relying too much on any fixed procedure .

6 Scope for improvement

We could have also implemented a procedure for doing map plotting where we included different zoom levels . It could be easily done by providing a scale for each image so that the user can click on both ends of the scale to estimate the zoom . But in this situation , there is a potential problem that can come up : error variance may not be constant if we have widely varying levels of zoom where a small amount of human error can upscale into a big error !

Also , it will also make the process of error analysis more cumbersome as there are more ways in which the user might mess up in giving the input namely : incorrectly specifying zoom level for any zoomed image can give us unexpected results .

Also , it wold have been nice we could have worked with screenshots of different rotations .