REPORT

LS3204

Instructor – Prof. Anindita Bhadra,
Prof. Anuradha Bhat,
Prof. Radhika Venkatesan

Dog microhabitats and effect of human establishment.

By

Harsh Bardhan Gupta & Sattwik Mohanty

19MS019 19MS20

Introduction

The term habitat summarizes the array of resources, physical and biotic factors that are present in an area, such as to support the survival and reproduction of a particular species. A species habitat can be seen as the physical dimension of its ecological niche. A habitat is the part of an organism's niche and is the physical space it occupies to live, grow and reproduce. A species' ecological niche can be defined as the range of resources and conditions allowing the species to maintain a viable population. But habitat is a larger measure of space, which corresponds to the living space occupied a population of species. And within the habitat, not every place is distributed evenly with organisms of the species. Parts of the population, i.e groups of some individuals seems to occupy a particular space within the habitat of the species. The smaller space within the habitat may differ slightly from the space around it, which suits the individual better. Here comes the concept of microhabitat.

Microhabitat may also mean the habitat of microbial organism, but here we refer to the small space within habitat. Microhabitat may be suggested as a subset of the habitat. A habitat can be subdivided into regions with different environmental conditions. These subdivisions are called microhabitat. For example, in a pond, some organisms are surface dwellers while some others are bottom dwellers. A microhabitat is a small area which differs somehow from the surrounding habitat. Its unique conditions may be home to unique species that may not be found in the larger region. But, also within a species, there may be a specific number of individuals occupying a specific microbial habitat. And this type of microhabitat formation within a population can lead to intraspecific competition in the population and by the competition between groups it can lead to territoriality where the group inhibits the other group to occupy its territory. The primary motivation of the study came from the intuition we had from the everyday life in our campus where we observe more dogs at some places than others. Besides that, a similar study on microhabitats was done by Simonetti et al. in Chile where they showed the pattern of microhabitat formation which is affected by food resource, predation risk and density of the herbs. In our study we make use of the Google maps to locate the organisms of an area and we analyze the formation of microhabitat using cluster analysis.

Clustering:

K-means clustering:

K-means clustering is a method in vector quantization, where we make 'k' number of clusters, from our dataset ,with 'n' number of data points. In this method, each observation or data point belongs to the cluster centroid or mean. The whole data set has k number of clusters and the mean of the cluster represents a cluster and the data point is classified into the cluster based on the distance of the point from the mean. Here we used constrained K-means algorithm with a fixed minimum 'n'.

Elbow method:

Elbow method is used in clustering to get the optimum K-value. In the Elbow method, we are actually varying the number of clusters (K). For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters. Our intuition suggest that if we take more K-value the data representation might get better, but further increase won't improve the representation and can lead to over fitting.

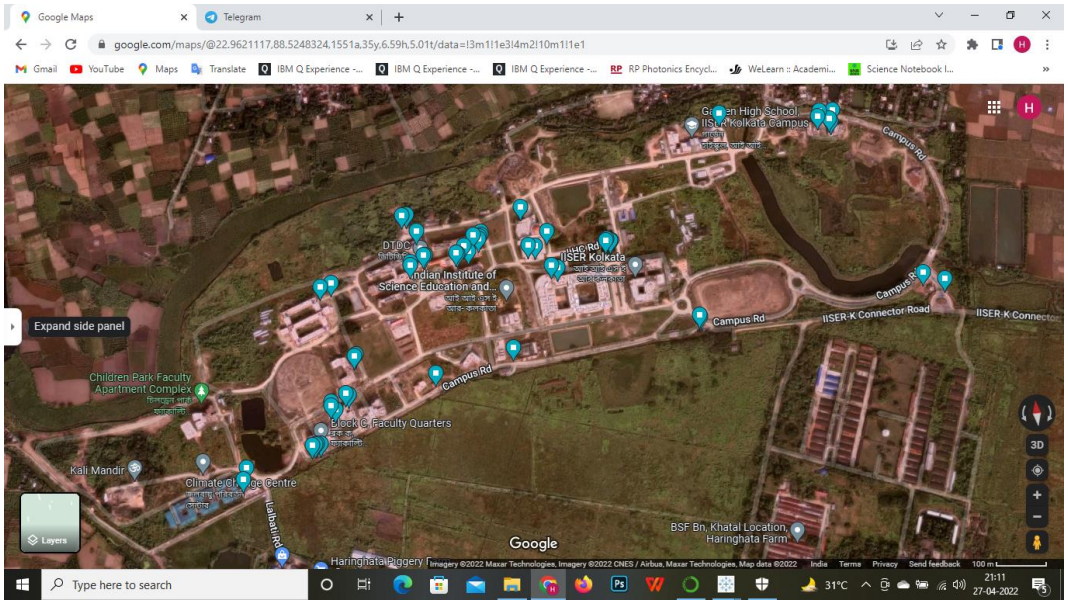
Silhouette coefficient :

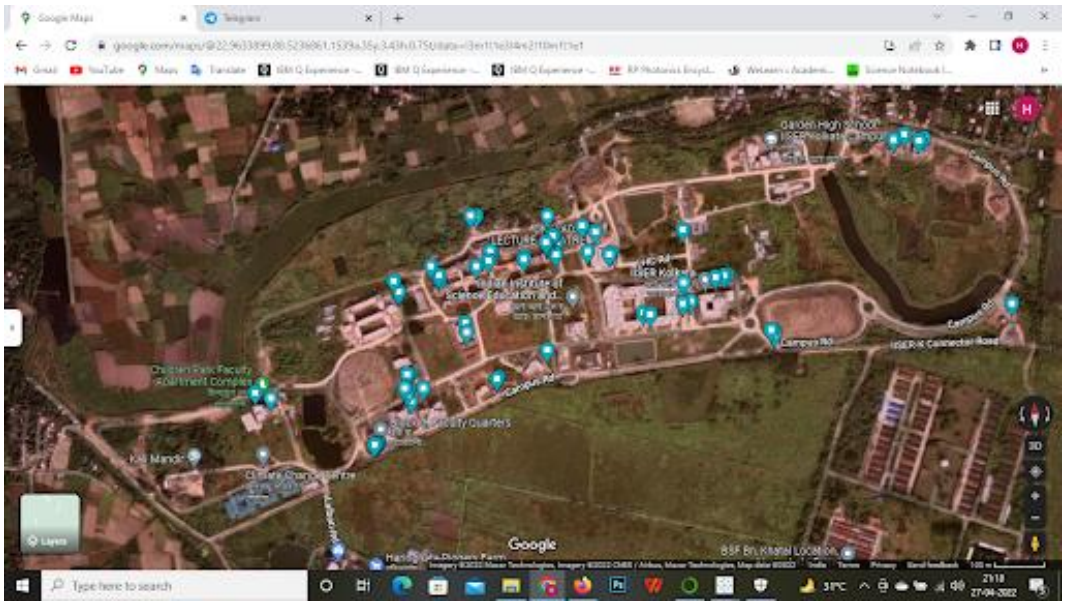
Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

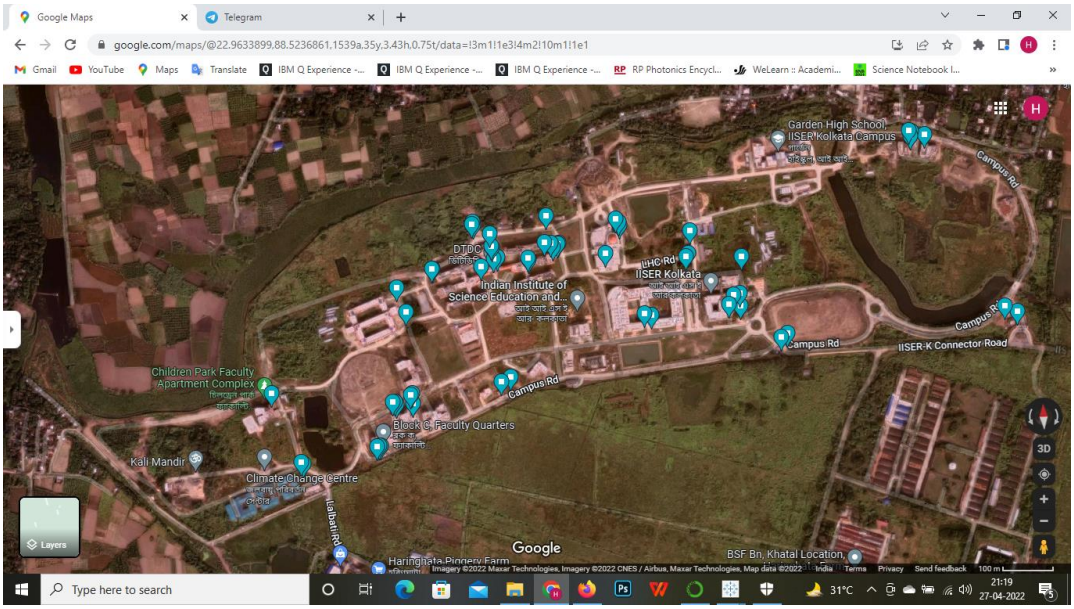
The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

Protocol

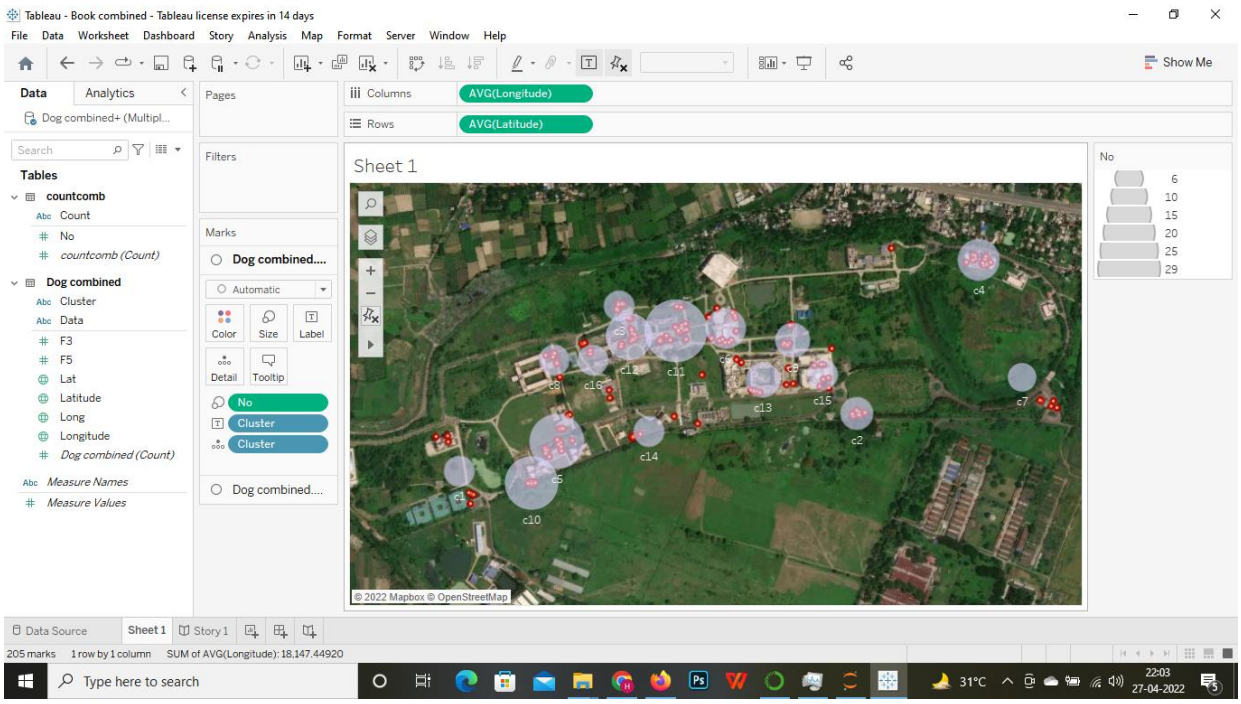
1. Firstly, we specified the area within which we want to observe the organisms (IISER Kolkata campus).
2. We take a specified time duration within which we take the observation. We took all our observations during 4pm to 6pm.
3. During the specific interval, we walk around the whole campus.
4. We avoid walking through the same area twice.
5. During walking whenever we see a dog within 25m distance in our eyesight, we go to the location, stand on near the dog and pin the location on Google maps, save the location on a list which we create for the particular day.
6. During pinning the location the map should be zoomed to the maximum to pin most accurate location.
7. We repeat our observation on 3 different days, so we get 3 set of location pins to get the general location of the dog.







8. We extracted latitude, longitude data from csv file exported using google takeout.
9. After getting the location data points of the dogs, we imported the data in python to perform k-mean clustering.
10. We used the Elbow algorithm and Silhouette coefficient method, to find the optimum K (or number of clusters for the K-mean algorithm).
11. We define the minimum of a cluster as 3 data points.
12. After getting the optimum K-value, we used the K-mean constrained algorithm, to form the cluster of points using the longitude and latitude of the location points.
13. After getting the longitude and latitude of centroid of clusters, we use “tableau” software to visualize the clusters on the google map. We set the size of cluster proportional to data points, they have.



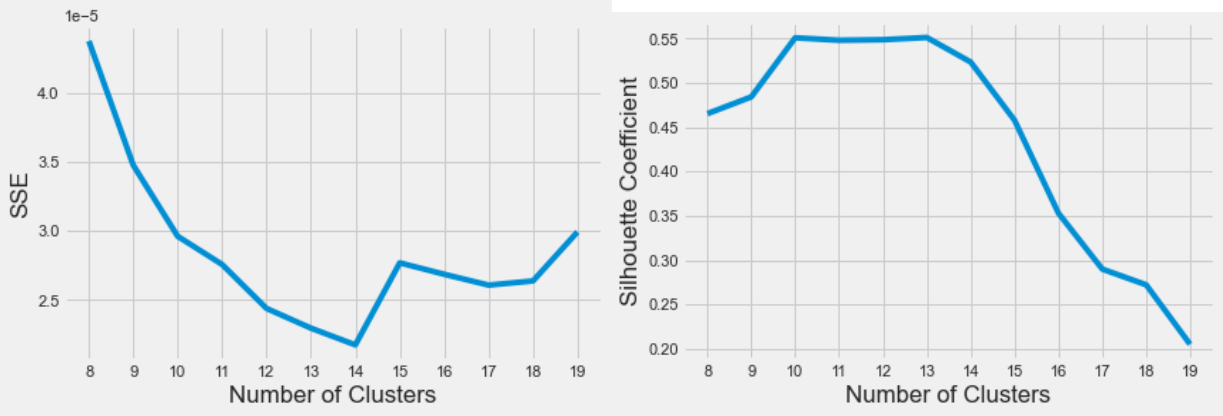
14. We repeated the process separately for 3 days, and then we combined the data points for 3 days, and repeated the analysis process and make the cluster visualizing maps for each.
15. We observe the location of the cluster and the human establishments in the campus. And try to observe if we see any relation.
16. Lastly, we selected 8 points, which we observed to be possible as a food source for dogs, and pin down the locations for that.
17. Firstly we calculated minimum distance of each data points from the all possible food sources. For this, distance of all the food source from the point was calculated, and the values were put into a matrix.
18. Then we analyzed the frequency of occurrence dogs with distance from food source, plotted the histogram for the same.

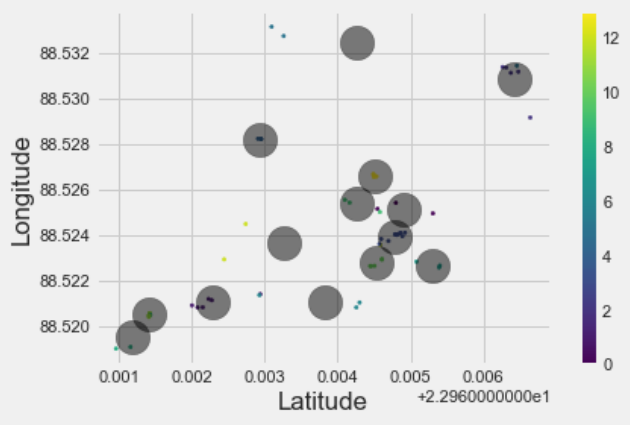
RESULTS

Day 1 results:

Elbow curve and Silhouette coefficient:

Minimum number of points in the cluster=3, K-value=14



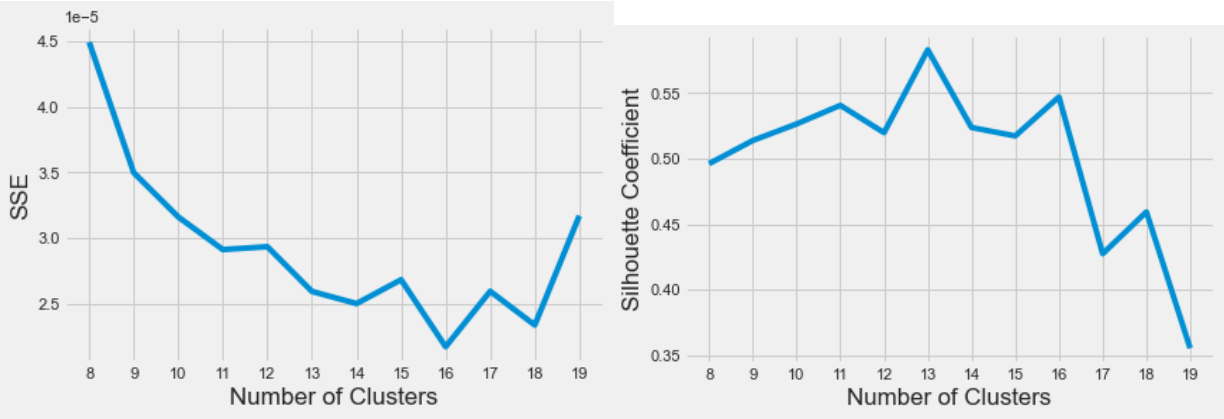


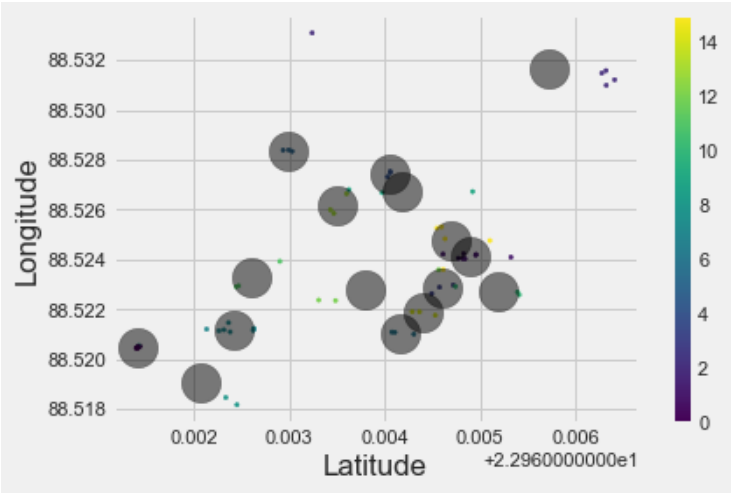
Sheet 1

Map based on average of Longitude and average of Latitude. For marks layer Dog day1.Latitude: Details are shown for Data. For marks layer Dog day1.Latitude (2): Size shows sum of No. The marks are labeled by Cluster. Details are shown for Cluster.

Day 2

Elbow curve and silhouette coefficient: K=16



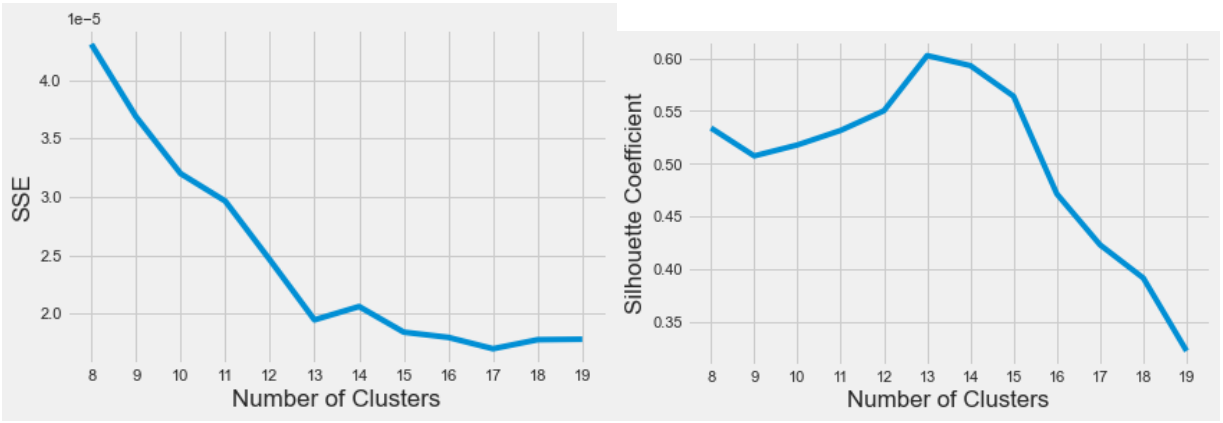


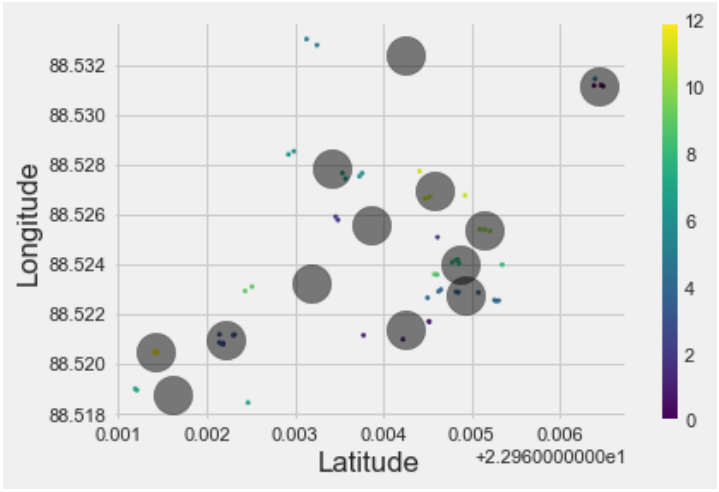
Sheet 1

Map based on average of Longitude and average of Latitude. For marks layer Dog day2.Latitude: Details are shown for Data. For marks layer Dog day2.Latitude (2): Size shows sum of No. The marks are labeled by Cluster. Details are shown for Cluster.

Day 3

Elbow curve and Silhouette coefficient: K=13



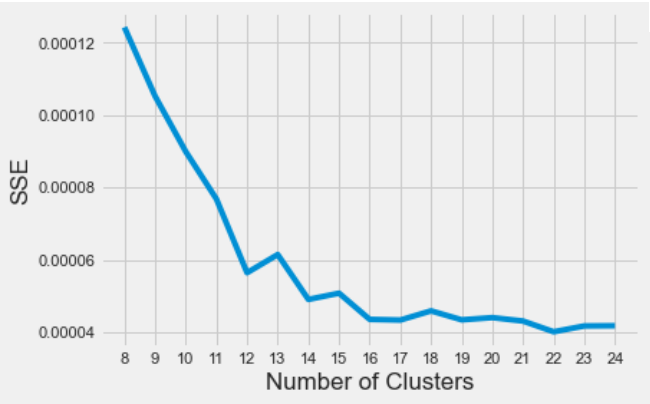


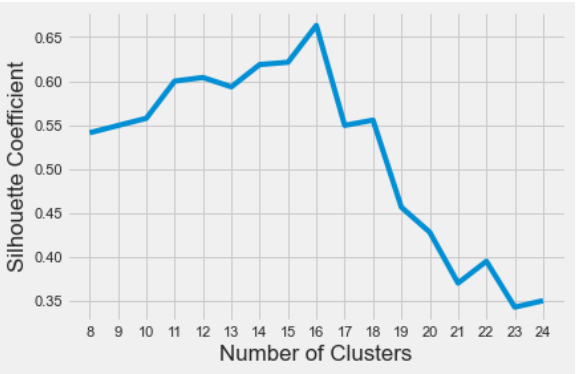
Sheet 1

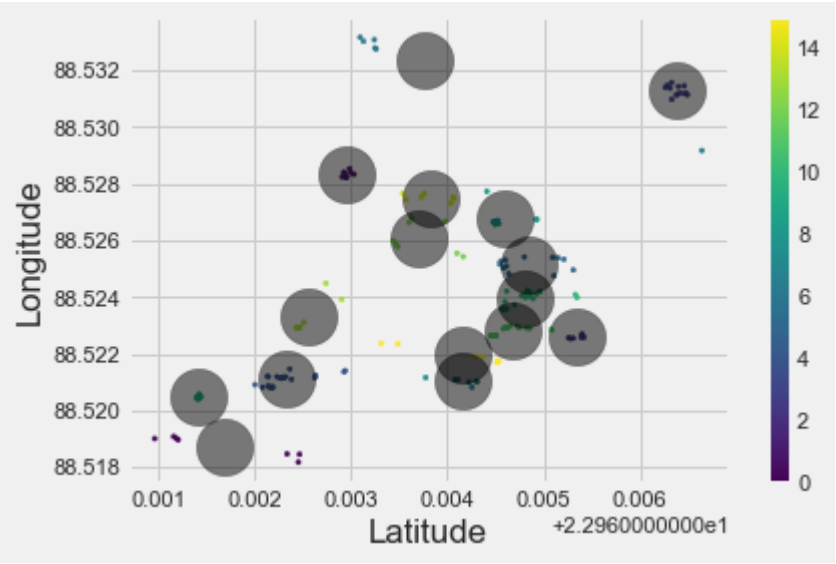
Map based on average of Longitude and average of Latitude. For marks layer Dog day3.Latitude: Details are shown for Data. For marks layer Dog day3.Latitude (2): Size shows sum of No. The marks are labeled by Cluster. Details are shown for Cluster.

All 3 days combined: K= 16, Minimum no. Of points in cluster= 6

Elbow curve & Silhouette coefficient:







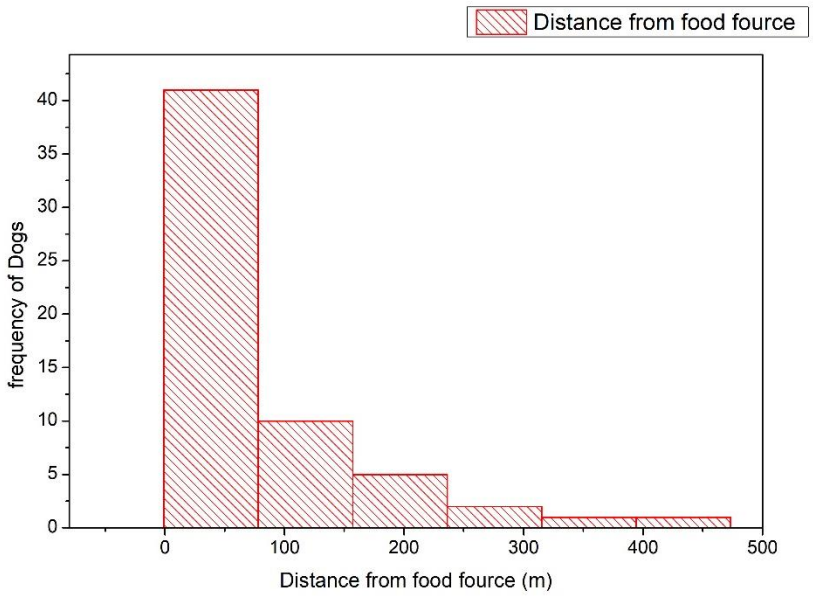
Sheet 1

Map based on average of Longitude and average of Latitude. For marks layer Dog combined.Latitude: Details are shown for Data. For marks layer Dog combined.Latitude (2): Size shows sum of No. The marks are labeled by Cluster. Details are shown for Cluster.

We plotted the frequency of dogs against the distance from the food sources.

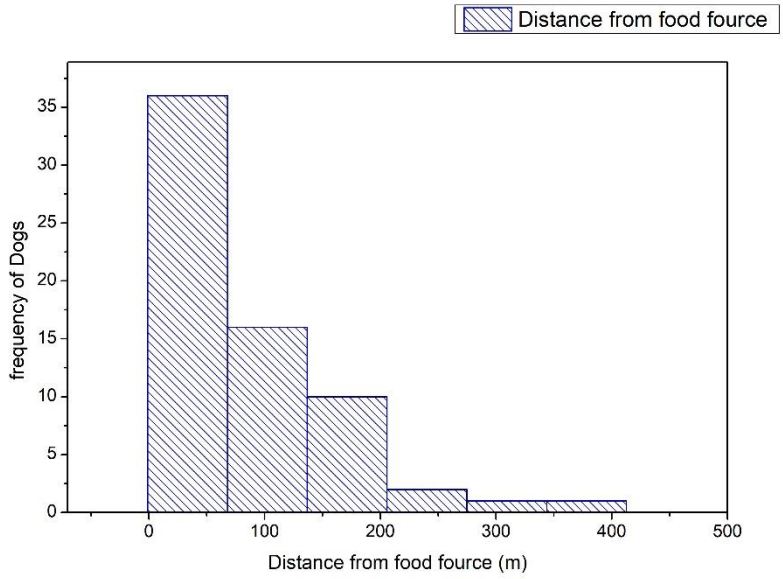
For the histogram, we determined bin width by Scott's rule. Scott's rule to choose bin sizes is based on the standard deviation (σ) of the data. The formula is: $3.49\sigma n^{-1/3}$.

Day 1:



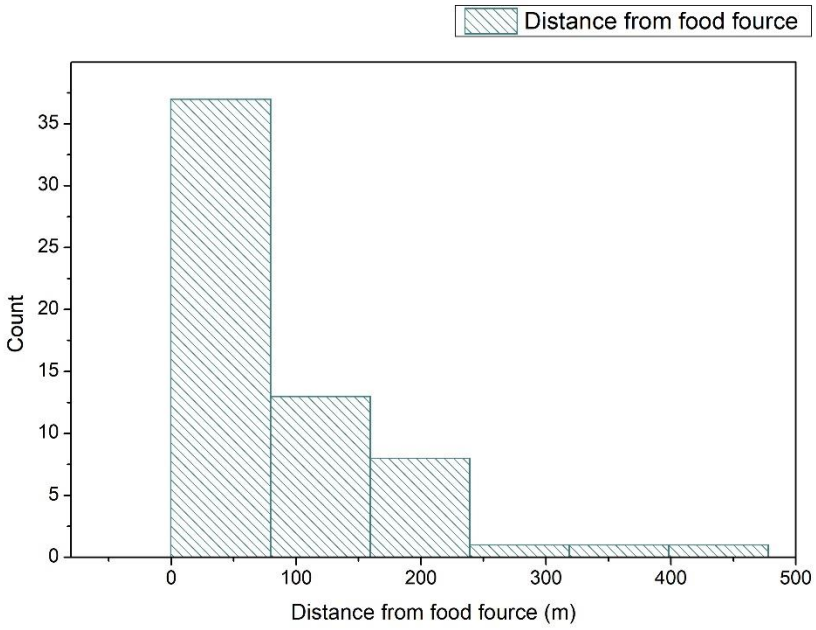
Bin width= 79.457m

Day 2:



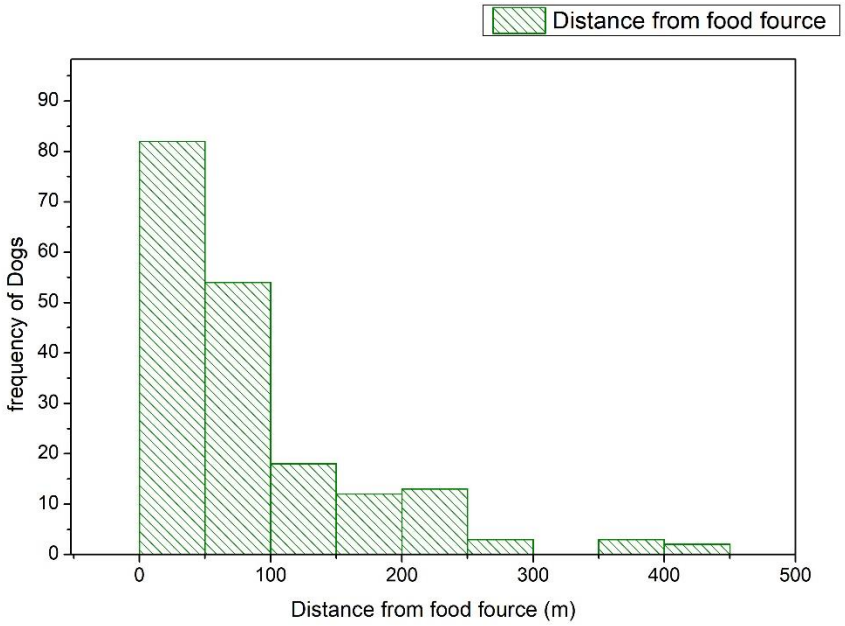
Bin width: 68.949

Day 3:

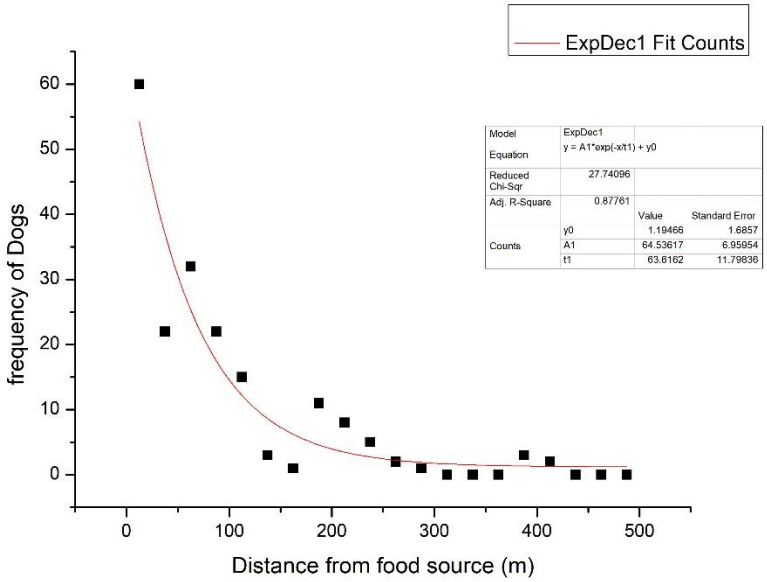


Bin width= 79.702

All 3 days combined:



Bin width= 55.442



We see the trend of exponential decrease of number of dogs as the distance from the food source increases.

We performed for ANOVA for 3 dataset of distance of the points from the food source on 3 different days.

Summary of Data						
	Day					
	1	2	3	4	5	Total
N	60	66	61			187
∑X	4533.8905	5846.0725	5118.1945			15498.1575
Mean	75.5648	88.5769	83.9048			82.878
∑X²	811316.2011	932166.619	914354.4311			2657837.2512
Std.Dev.	89.1308	79.8402	89.8993			85.9289

Result Details				
Source	***SS***	***df***	***MS***	
Between-Samples	5416.7333	2	2708.3667	$F = 0.36429$
Within-Samples	1367966.5761	184	7434.601	
Total	1373383.3095	186		

We get the F-value of 0.36429, which is way below the critical value for df = 3 & 370 after using the table. So, we conclude that the 3 Samples were from the same Population.

Conclusion:

After clustering the data points from all three days and also clustering the combined data points of the 3 days, we can clearly observe that most of the clusters are formed at the location, where there is some kind of human establishment, also we see rare occurrence of any cluster outside the human establishment or occurring very far from it. So, we can state that dog microhabitats are formed mostly co-occurring with human establishment. The reason for that might be the higher probability of getting food, establishing relationship with humans with time. Or as the study was during summer and day time, the microhabitat formation may be related to availability of shade, which is efficiently provided by the human establishments. For food availability, after analyzing all the data points over the course of 3 days, we saw from the plot that the occurrence of dog decreases, as we move farther away from any food source (8 food sources in our study). So, we can suggest some of the determining factors that could lead to small microhabitat formations inside the campus, which are: Shade availability, which leads to more comfort in resting, human-dog relation, and Food source.

Discussion:

In our study, the weather is hot summer, which can impact the actual distribution of the dogs in campus, so if a study is done over the course of a year which include all types of weather, then we can observe the microhabitats better and in more accurate way. During our study there is a chance that the dogs of a particular area could have moved to another place for some time. There is a huge probability of another kind of human interference, like dogs moving with any familiar person from one location to another.

Code1 – for clustering

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import seaborn as sns; sns.set()
import csv

df = pd.read_csv('Dog combined.csv')
features = df[['Latitude', 'Longitude']]
X = np.array(features)
print(X[:10])

from k_means_constrained import KMeansConstrained
sse = []
for k in range(8, 25):
    clf = KMeansConstrained(
    n_clusters=k,
    size_min=6,
    random_state=0
    )
    clf.fit(X)
    sse.append(clf.inertia_)

plt.style.use("fivethirtyeight")
plt.plot(range(8, 25), sse)
plt.xticks(range(8, 25))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
plt.show()

from k_means_constrained import KMeansConstrained
silhouette_coefficients = []
for k in range(8, 25):
    clf = KMeansConstrained(
    n_clusters=k,
    size_min=6,
    random_state=0
    )
    clf.fit(X)
    score = silhouette_score(X, clf.labels_)
    silhouette_coefficients.append(score)

plt.style.use("fivethirtyeight")
plt.plot(range(8, 25), silhouette_coefficients)
plt.xticks(range(8, 25))
plt.xlabel("Number of Clusters")
plt.ylabel("Silhouette Coefficient")
plt.show()


clf = KMeansConstrained(
    n_clusters=16,
    size_min=6,
    random_state=0
    )
clf.fit_predict(X)
# save results
```

```
labels = clf.labels_
centers = clf.cluster_centers_ # Coordinates of cluster centers.
# send back into dataframe and display it
df['cluster'] = labels
# display the number of mamber each clustering
_clusters = df.groupby('cluster').count()
print(_clusters)

df.plot.scatter(x = 'Latitude', y = 'Longitude', c=labels, s=5,cmap='viridis' )
plt.scatter(centers[:, 0], centers[:, 1], c='black', s=800, alpha=0.5)

print(centers)
pd.DataFrame(centers).to_csv('samplecomb.csv')
clusterCount = np.bincount(labels)
pd.DataFrame(clusterCount).to_csv('countcomb.csv')
```

Code2 – for minimum distance from food source

```
import haversine as hs
import pandas as pd
import numpy as np
from haversine import Unit

df = pd.read_csv('Dog day3.csv')
features = df[['Latitude', 'Longitude']]
X = np.array(features)
print(X[:10])

df2 = pd.read_csv('Food_Source.csv')
features = df2[['Latitude', 'Longitude']]
Y = np.array(features)


dis_mat = [ [0]* len(Y) for i in range(len(X))]
for i in range(len(X)):
    for j in range(len(Y)):
        dis= hs.haversine(X[i],Y[j],unit=Unit.METERS)
        dis_mat[i][j]=dis
print(dis_mat[:10])

dis_list= []
for i in range(len(X)):
    dis_list.append(min(dis_mat[i]))
dis_list

pd.DataFrame(dis_list).to_csv('dis_meas_day3.csv')
```