TEAM NULL PRESENTS

An analysis of the Browntail Moth Caterpillar infestation situation in Waterville on distribution pattern

Baron Wang, Ziyan Zhang, Bishal Khadka, Yiheng Su

Introduction

The browntail moth (BTM) is an invasive species found only on the coast of Maine and Cape Cod from April to late June. This moth is a concern for both forest and human health because of its tiny poisonous hairs that can cause dermatitis similar to poison ivy on sensitive individuals.

Given the dataset of browntail caterpillar infestation inspection and treatment, our team aim to provide insights and prediction method through data analysis, machine learning, and model building. We hope to contribute our effort to inhibit the browntail moth caterpillar in Waterville community.

Questions We Want to Ask

How are the BTM infested trees distributed in Waterville? Is there any pattern in the distribution? Is there any factors that determine the distribution?

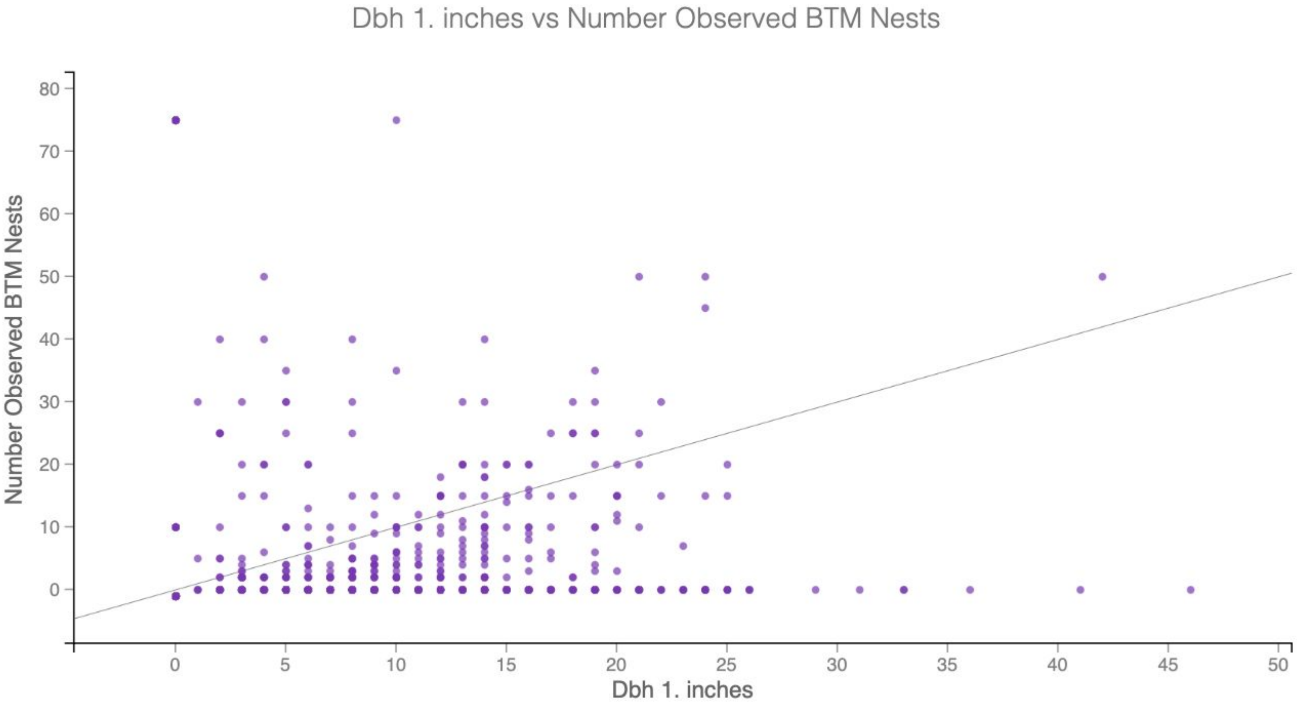
Can we predict the distribution of BTM infestation in Waterville in the future? Can we predict whether a tree will be infested and which treatment will be needed given some information?

How can we improve the treatment and policy to inhabit BTM based on our findings?

Data Cleaning-Preparation

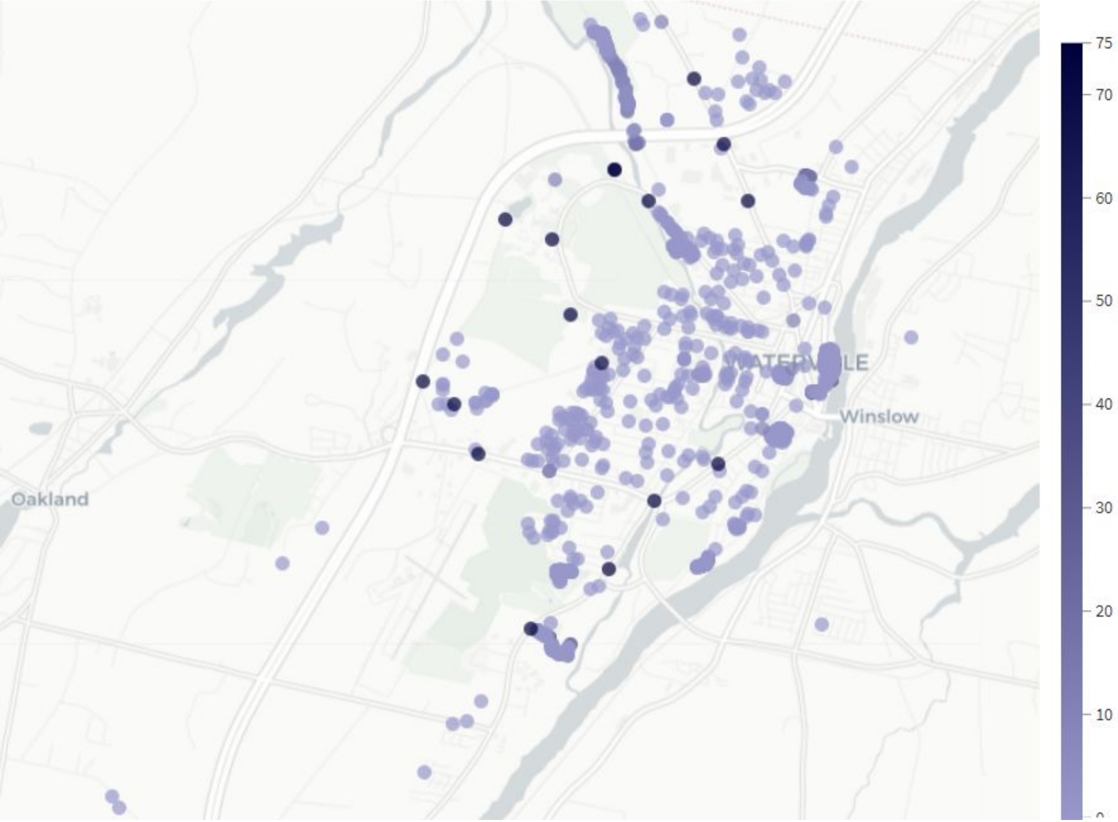
- Remove row and column whose values are all null;
- Identify the correct semantic type for each column;
  - E.g. tree ID→String;
  - Longitude→Double;
  - Number of BTM positive Tree Species on property→Integer
- Fill in the missing data /convert data based on the semantic type of each columns;
  - E.g. replace “” to N/A in tree ID, Condition Class / Infestation Pattern, etc
  - “” to Other in Common Name
  - “” to 0 in Dbh 1.inches
  - Null to -1 if 0 is already a part the data
- Create a new column Geopoint from Longitude and Latitude

DBH of Tree vs. Number of Observed BTM Nests



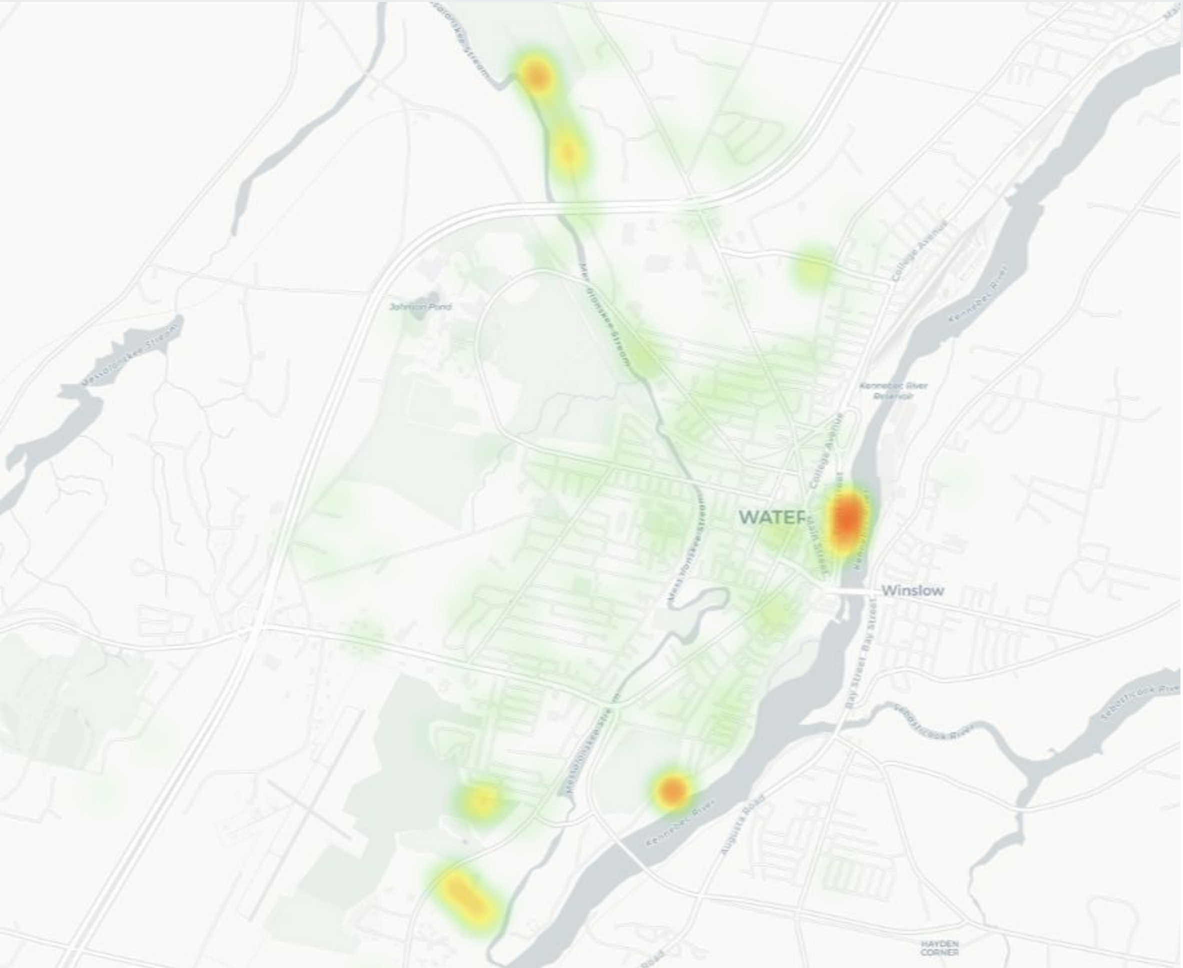
There is a positive correlation between the DBH of a tree and number of observed BTM nests. So, we conclude that **larger trunked trees would have higher chances of having BTM nests** and consequently more BTMs.

Data Visualization–Distribution of BTM Nests



This diagram shows the distribution of BTM Nests on trees in Waterville area. **The darker dot represents the larger amount of nests.**

To visualize this, we first created Geopoints from the given Latitude and Longitude of trees. We associated the ***Geopoint*** with ***Number Observed BTM Nest*** via the Scatter Map in Dataiku.



This chart shows the density of BSM infestation. **Area in red represents a denser BSM presence**. Waterville residents are able to avoid the severely afflicted area based on this map.

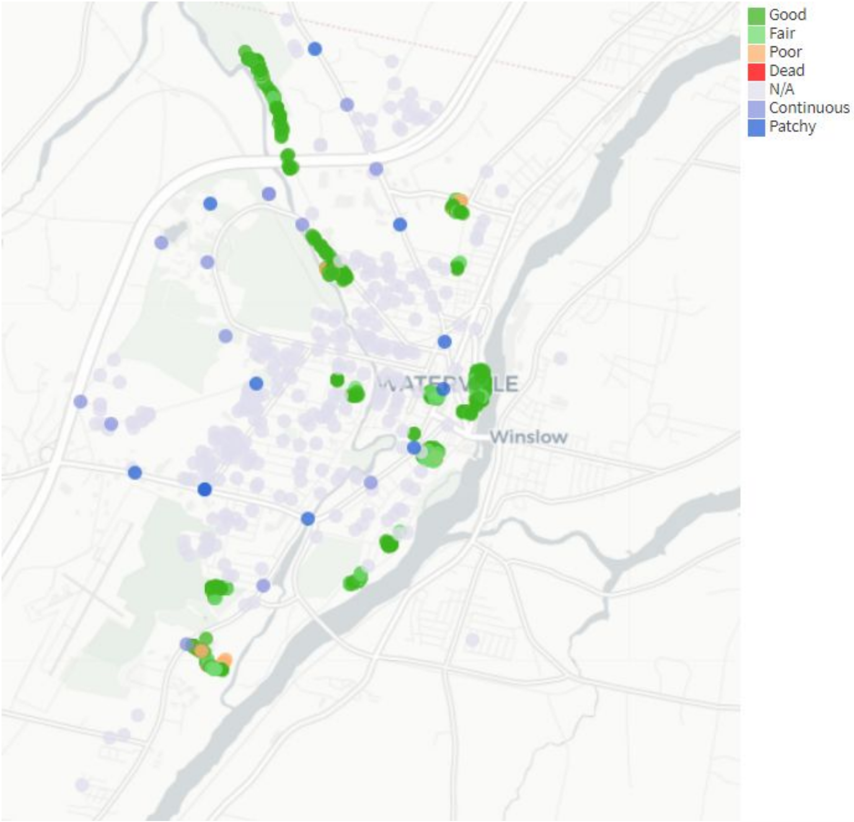
We observed that **the denser spots are distributed close to the rivers.**

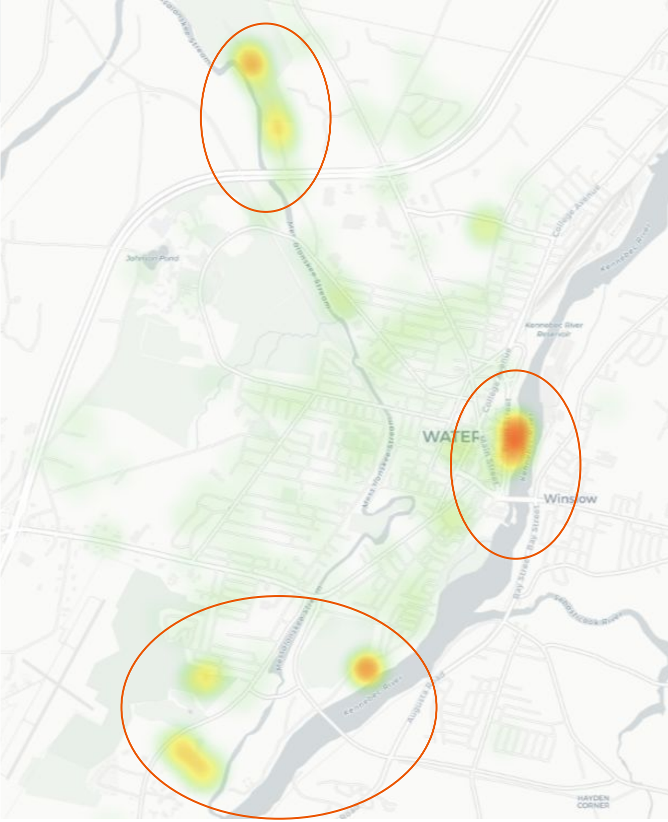
This is visualized by using the ***Geopoint*** with ***Number Observed BTM Nest*** via the Density Map in Dataiku.

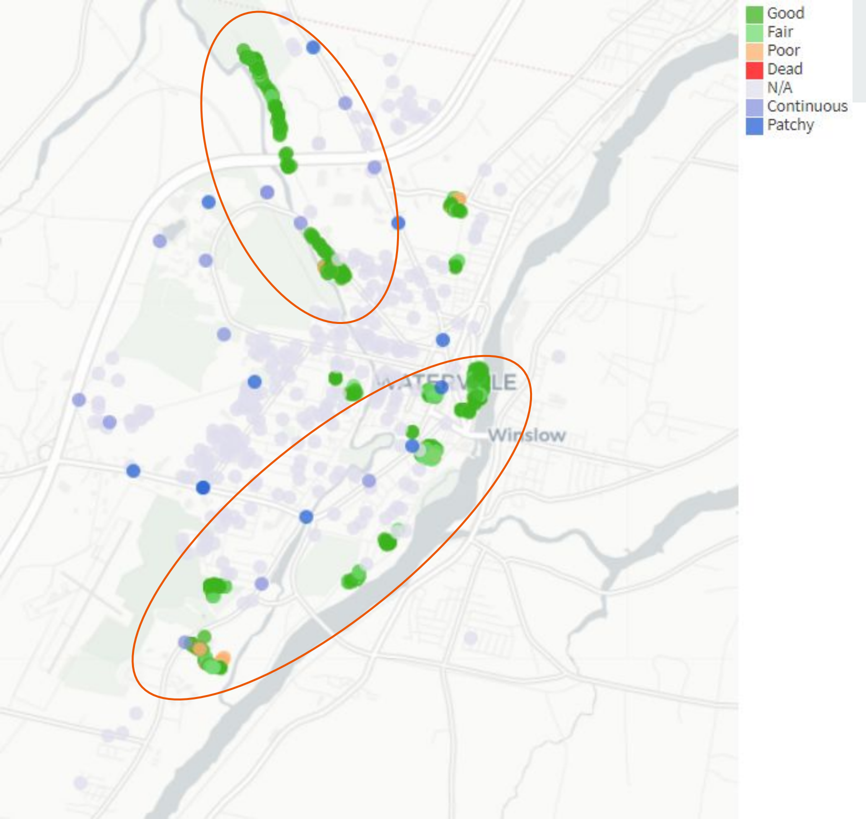
Data Visualization–Distribution of Infestation Pattern

This diagram shows the infestation condition of trees in Waterville area. The color light purple represents no response in the survey. Notice that the chart shows **a green line along the rivers.**

To visualize this, We associated the ***Geopoint*** with ***Condition Class/Infestation Pattern*** via the Scatter Map in Dataiku.

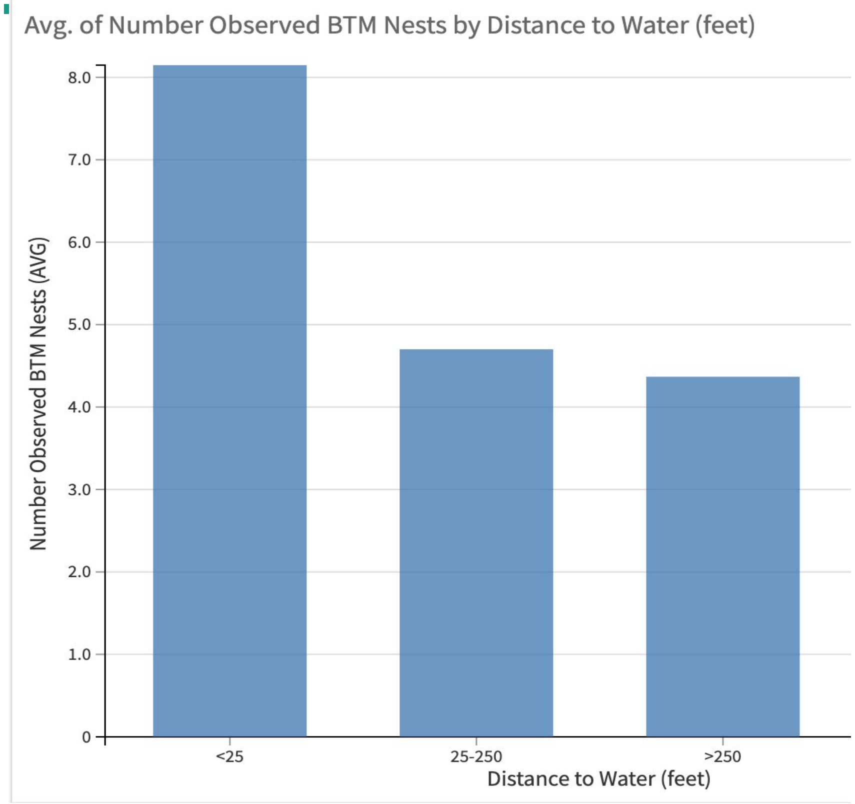






From our visualization, it seems that the closer to water, the denser the BTM. This attracts our attention. Is **the distance to water** an important factor to consider in this context?

Number of BTM Nest vs. Distance to Water



Here we first modify the data. For each category (e.g. <25), we **take the average # of BTM nest of the trees** in this category.

In this histogram, we do see a relationship between the Average Number of Observed BTM Nests and Distance to Water: t**he closer to water, the larger is the average number of BTM nests observed.**

Data Visualization–Violin Plot

1. *When the distance to water is greater than 250 ft, we have two tree types that has BTM nests. The fruiting/flowering type has dense data at points around 10 but has some outliers of 90. The ornamental type has more values closer to 1 and some outlines that goes as high as 55.*
2. *There are 3 tree types for the distance to water less than 25 feet. The bush type has more uniform data without any outliers. There is one data point for the fruiting/flowering type.*
3. *There is no bush at 25-250 ft from water source that has BTM nests.*
4. *Other tree types do not have any BTM nests.*
5. *The negative values are NULL values.*

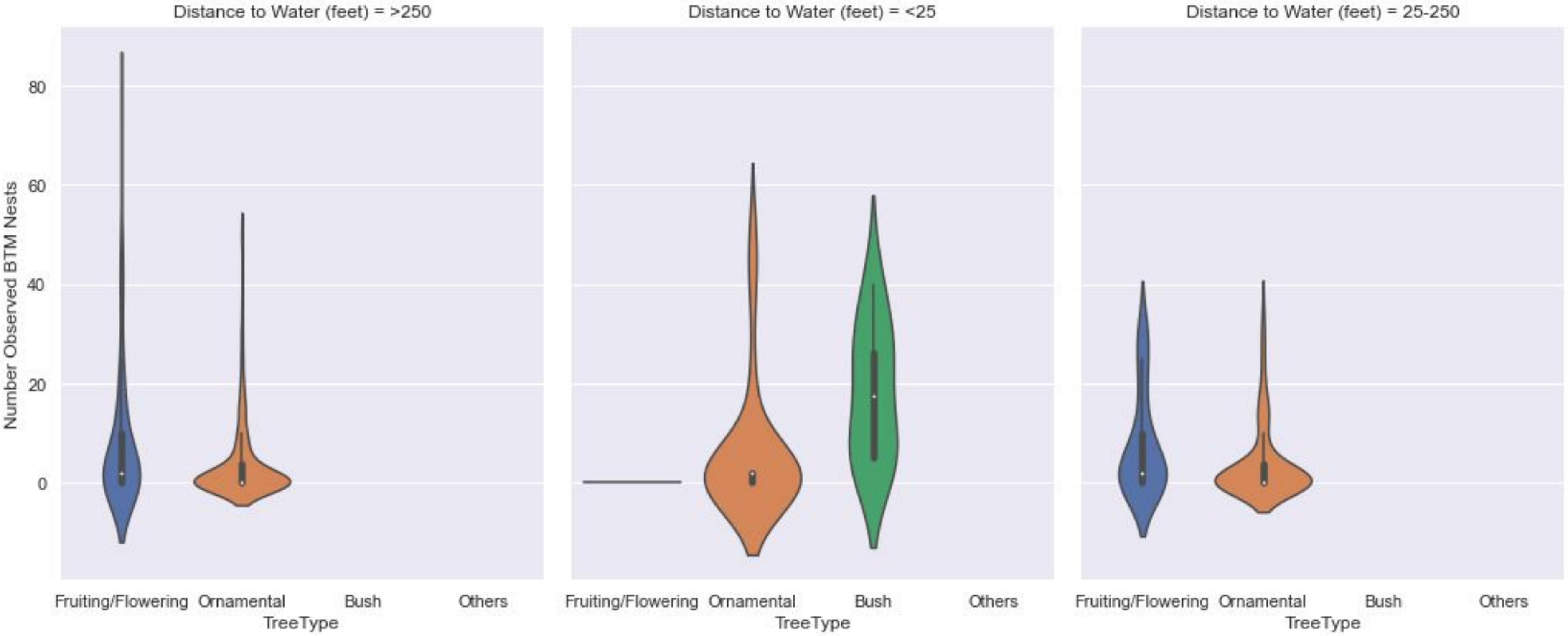
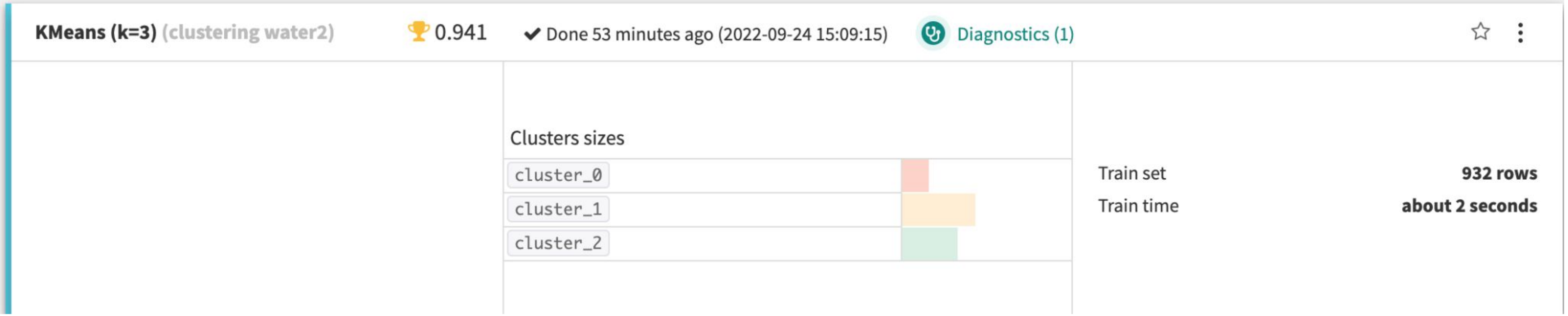


Fig. Violin Plot

Cluster the Trees

From the density and distribution map, we can see that the infested trees are roughly located along the rivers. We can actually use some clustering methods to separate the trees based on the distance from the river and the number of BTMs variables. We used k-means clustering method with k = 3. Here is the resulting of the graph aftering running the k-means clustering on it. 0.941 implies the clusters fitted our model really well!

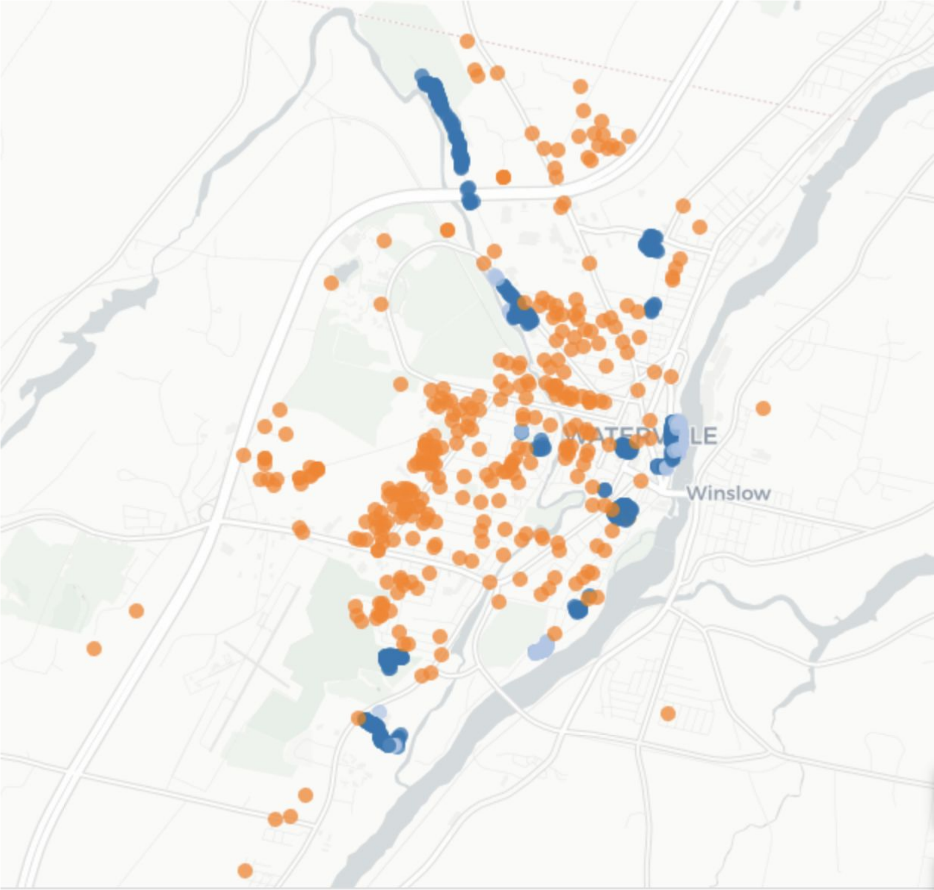


KMeans (k=3) (clustering water2) 🏆 0.941 ✓ Done 53 minutes ago (2022-09-24 15:09:15) Diagnostics (1)

Clusters sizes

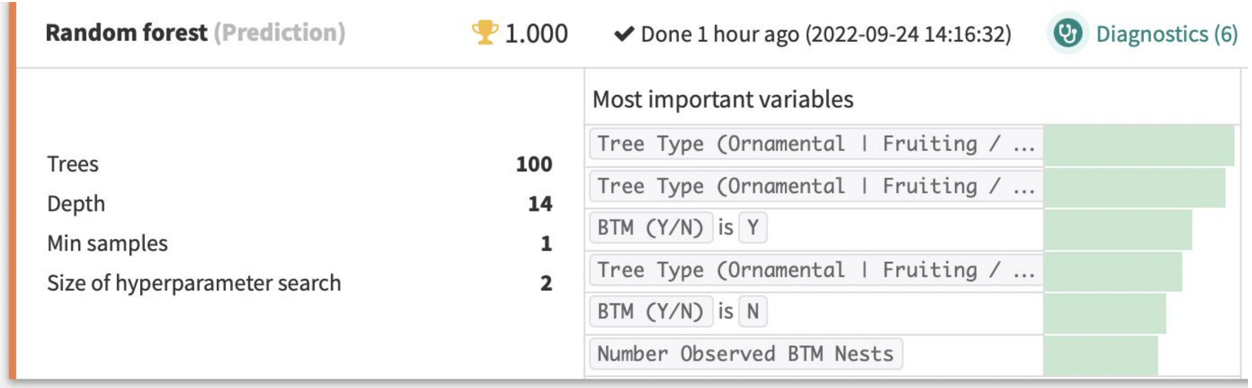
cluster_0	🟥
cluster_1	🟧
cluster_2	🟩

Train set	**932 rows**
Train time	**about 2 seconds**



From the map, we can see that the orange points are scattered at Waterville while the blue points are following the rivers. Notably, the trees align the two rivers are clustered together, indicating a potential tendency of BTM infestation with close proximity to water.

Treatment Type Prediction Model



Random forest (Prediction) 🏆 1.000 ✓ Done 1 hour ago (2022-09-24 14:16:32) Diagnostics (6)

Trees	**100**
Depth	**14**
Min samples	**1**
Size of hyperparameter search	**2**

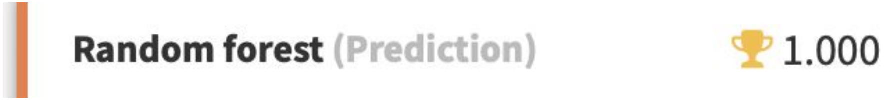
Most important variables

```
Tree Type (Ornamental | Fruiting / ...
Tree Type (Ornamental | Fruiting / ...
BTM (Y/N) is Y
Tree Type (Ornamental | Fruiting / ...
BTM (Y/N) is N
Number Observed BTM Nests
```

Using the Random Forest modelling technique, we are successful at building a model that predicts the type of treatment a tree needs to receive, based on other variables.

The most important variables for the prediction are the tree type variables, i.e. whether the tree is fruiting/flowering, ornamental, or others.

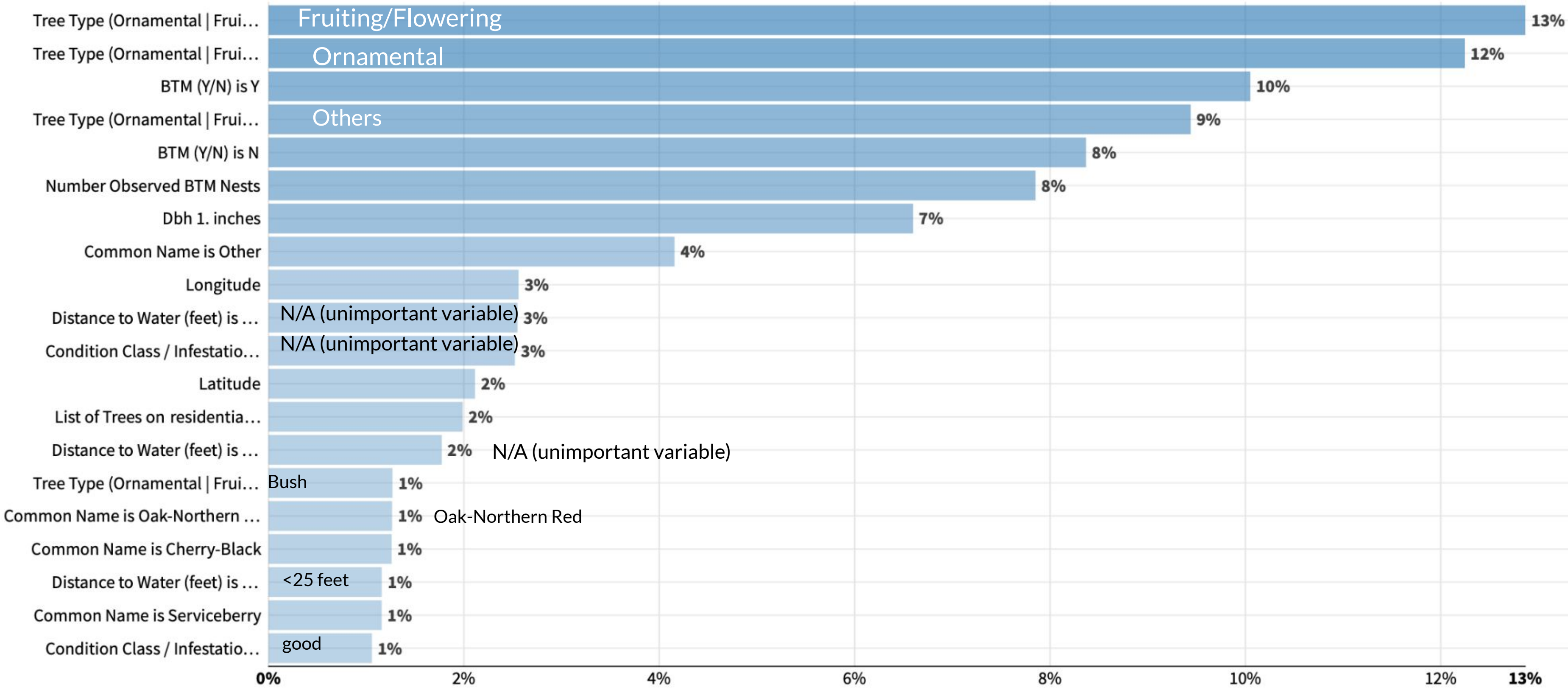
Perfect Model?



Random forest (Prediction) 🏆 1.000

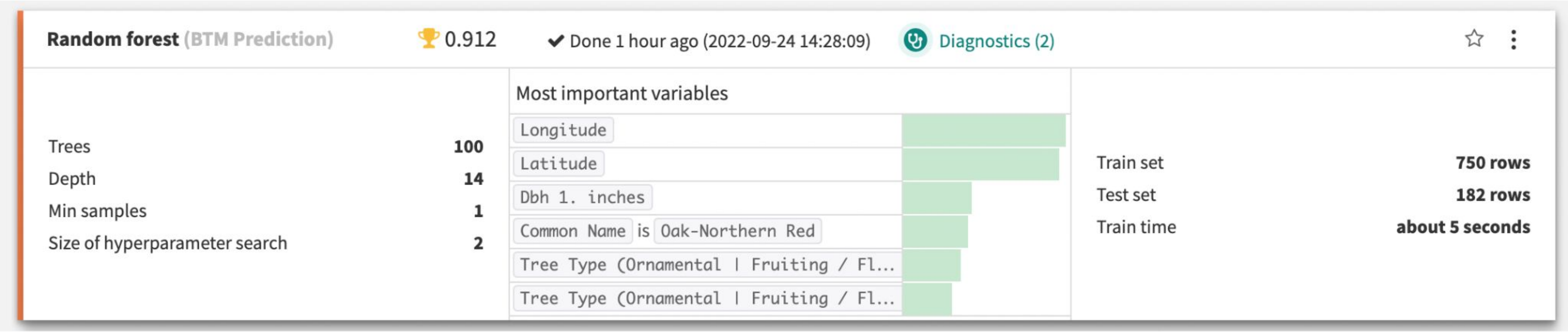
The most interesting thing is that the accuracy of our model is 1! It means no matter what basic information of a tree given to this model, it will produce the most suitable treatment to the tree. In the real life, the perfect model rarely happens. When it happens, it usually implies we did something wrong with the train and test data or made some mistakes building the model. However, Dr. Thomas Klepach,the owner of the dataset, told us that the treatments of the tree in the data were generated by some algorithms based on the basic information of the tree. It means that the prediction model is similar to the algorithms generating the treatments. Hence, we arrive at the perfect accuracy. Though the perfect model confused us a lot at first, it is really interesting to learn how they develop algorithm for choosing which types of the treatment for a tree.

All Variables Used in the Treatment Type Prediction Model and Their Importance



Tree Infestation Prediction Model

We have built a perfect model to predict what kinds of treatment to use on a tree given the tree’s basic information, such as distance from water, number of BTM nests on the tree, and etc. Now, we want to build a model to predict whether a tree get infestation by BTMs given some information of a tree. We use the Random Forest model as our prediction model. The basis information of the tree we are interested are tree types, common name of the tree, distance from the river, and diameter at breast height. Here is the result of model.



Random forest (BTM Prediction) 🏆 0.912 ✓ Done 1 hour ago (2022-09-24 14:28:09) Diagnostics (2)

Trees	100
Depth	14
Min samples	1
Size of hyperparameter search	2

Most important variables

- Longitude
- Latitude
- Dbh 1. inches
- Common Name is Oak-Northern Red
- Tree Type (Ornamental | Fruiting / Fl...
- Tree Type (Ornamental | Fruiting / Fl...

Train set	750 rows
Test set	182 rows
Train time	about 5 seconds

The accuracy of our model is 0.912 which is very high! The most important variables in this prediction model is longitude and latitude. It implies the geolocation of the infested area is really important. Hence, it also suggests that the infested trees tend to cluster together because the prediction is heavily depend on the location. This result agrees with the map we have shown. The infested trees are clustered along the river or in the forests.

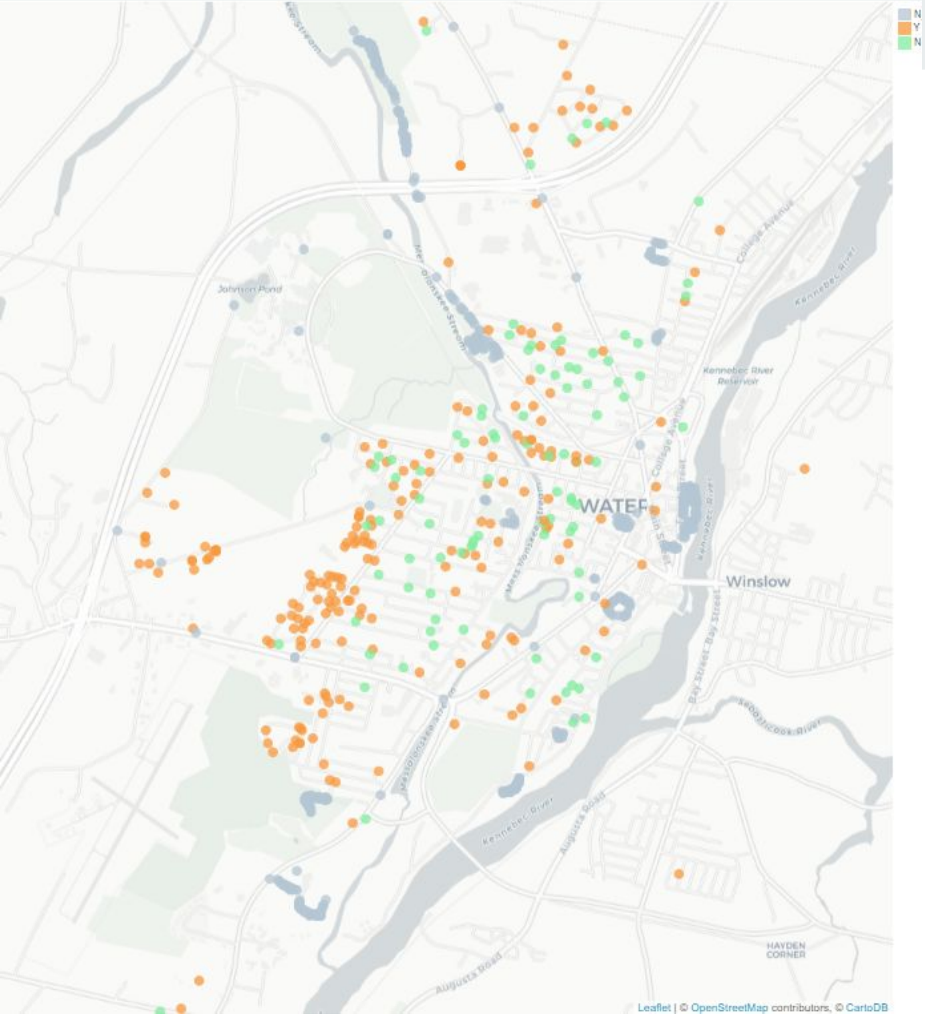
Policies

We have done a lot of data analysis. Let’s give some policies for inhibiting the browntail moth caterpillar infestation in the Waterville community.

- **Insect control campaign with a focus on areas along the river**
  - A lot of of graphs show that the BMT nests is higher in number nearby water resources.
- **Development of better algorithms for offering treatment**
  - We have develop a random tree prediction model to identify whether a tree has been infested. We can also use our prediction model to assign a proper treatments for a tree.

- **Advisory for residents with susceptible types of trees in their property**

This graph shows the location of people who are seeking help and who are not seeking help . We can make policies keeping this in mind, targeting people who need help more.



Leaflet | © OpenStreetMap contributors, © CartoDB

Thanks!