Prediction of Jiuling based on Sample Point Data

VE414 Bayesian Analysis

Yuhao Chen Ruiwei Zhang Chong Hu

VE414 Project, Group 16

August 5, 2019

Contents

1. Introduction
2. Assumptions
3. Solution
4. Result
    - Trees' Location
    - DistributionS
5. Discussion
    - Error Analysis
    - Futre Work

Introduction

Data Interpretation

Raw Data Gives Us:

Sample point coordinates

Which trip are the samples taken

Fruit number within 1m from sample point(100% correct)

Fruit number within 3m from sample point(unknown accuracy)

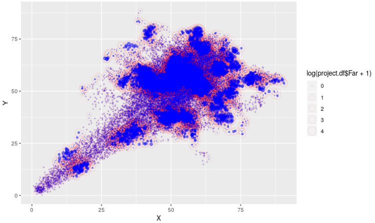

Figure: All coordinates sampled through the 49 trips

Introduction

Fruit Counted (Within 1m) Distribution

Fruits that are counted distribute unevenly

Fruits counted are extremely dense within the center areas

Fruits counted are sparse on the corners

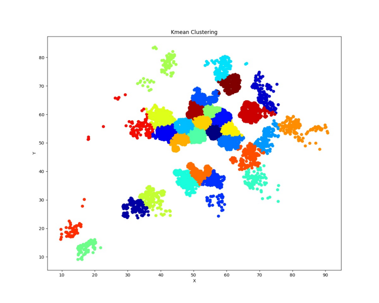

Figure: Sample Point With Fruits Counted within 1m Coordinate Distribution

Introduction

First Inference

Peaks in fruit number counted results from multiple trees in adjacency
Dense sample points result in repeated fruit counting
Sparse fruits area results from few trees in adjacency

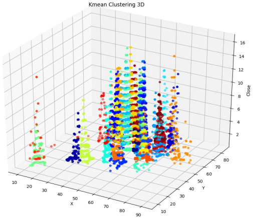

Figure: Distribution of Fruits Counted within 1m

Introduction

Tasks

Parameterize the properties of a Jiuling Tree

- *Tayes generated by a single Jiuling Tree*
- *Distribution of the probability of Tayes dropped on a certain coordinate*

Find isolated single Jiuling tree for property estimation

Estimate amount of Jiuling trees over the sampled area

Make inference on tree coordinates over the sampled area

Predict amount of tree over the area not sampled

Predict coordinates of trees in the area not sampled

Contents

1. Introduction
2. Assumptions
3. Solution
4. Result
   - Trees' Location
   - DistributionS
5. Discussion
   - Error Analysis
   - Futre Work

Assumptions

Claims

- Sample points represents only the points with Tayes Counted not equal to zero
- A_{Tree} represents the area covered by a Jiuling tree and A_{Sample} represents the area covered by the sample points
- F_{Tree} represents Tayes Fruit quantity of a single tree and F_{Sample} represents Tayes Fruit quantity counted by the sample points
- N_{Tree} represents quantity of Jiuling Trees and N_{Sample} represents quantity of Sample Points

Assumptions

Assmptions on Jiuling Tree Properties

Quantity of Tayes counted will be bigger when the sample point is closer to the tree center

Expectation of Tayes quantity will be the same when the distance between sample points from the tree center are the same

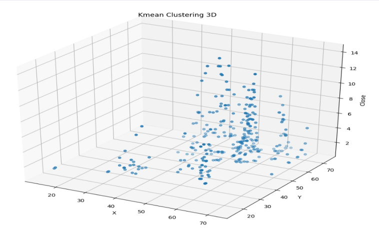

Figure: Tayes Amount Distribution over a Single Tree

Assumptions

Assmptions on Jiuling Tree Properties(Continued)

The Jiuling trees fall the Tayes fruits following the following properties:

The distribution of Tayes quantity on the x-axis direction is independent of the distribution of Tayes quantity on the y-axis direction

Using Bivariate normal distribution as an example:

$$Cov({\sigma_1}^2, {\sigma_2}^2) = 0$$

Assumption

Assumption on Tayes fruits Counted by the Sample Points

In the scenario of no overlapping of two sample points:

More Tayes of a Jiuling tree will be seen given more sample points in the area

Tayes amount per unit area of a single tree is reflected by the average fruit amount in sample points areas

$$\text{Tayes fruit per unit area} = \frac{F_{Tree}}{A_{Tree}}$$

$$\text{Fruits per unit area of a single sample point } = \frac{F_{sample}}{N_{sample}\pi}$$

$$\frac{F_{Tree}}{A_{Tree}} = \frac{F_{sample}}{N_{sample}\pi}$$

Assumptions

Assumption on Tayes fruits Counted by the Sample Points

In the scenario of no overlapping of two sample points:

The more area sample points covered, the more trees will be seen

The quantity of Jiuling trees is proportional to the area covered by the sample points

$$N_{Tree} = \frac{A_{Sample}}{A_{Tree}}$$

Assumptions

Overlap Scenarios

A overlap scenario of two trees is illustrated in the following figure:

We assume that the overlap will never result in two complete overlapped points

$$A_2 < A_1 + A_2$$

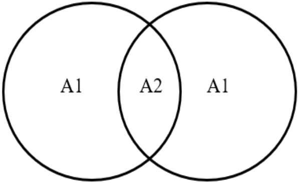

Figure: Overlap Scenario of Two Trees

Assumptions

Overlap Scenarios

We denote the fruits per unit area as ρ for convenience

We assume that the impact of overlap scenario is reflected in the fruits per unit area

Based on previous notations, we have an estimation of Jiuling tree numbers

$$T_N = \frac{\rho_{Sample}}{\rho_{Tree}} * \frac{A_{Sample}}{A_{Tree}} = \frac{\frac{F_{Sample}}{N_{Sample}\pi}}{\frac{F_{Tree}}{A_{Tree}}} * \frac{A_{Sample}}{A_{Tree}}$$

Contents

1. Introduction
2. Assumptions
3. Solution
4. Result
   - Trees' Location
   - DistributionS
5. Discussion
   - Error Analysis
   - Futre Work

Bayesian Network

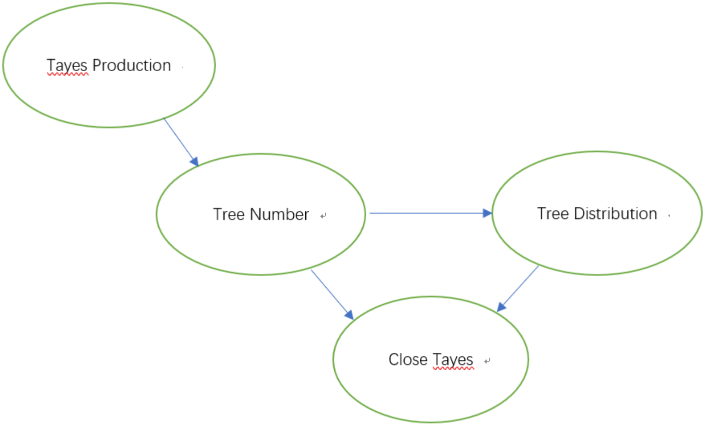

Figure: Bayesian Network

Tayes Production Estimate

Possion Distribution

$$P(x = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

We need to estimate the lambda. We will use MAP to estimate the λ.
Prior:

$$f_\lambda = Gamma(1, 1) \tag{2}$$

Tree Number Estimate

Linear Estimate

The tree number is proportional to the total cover area of trees. The factor is the density of trees, which can be estimated by the density of Tayes.

$$T_N = \frac{\rho_{Sample}}{\rho_{Tree}} * \frac{A_{Sample}}{A_{Tree}} = \frac{\frac{T_{Sample}}{N_{Sample}\pi}}{\frac{T_{Tree}}{T_{Tree}}} * \frac{A_{Sample}}{A_{Tree}}$$

For the total tree number of Jiuling, we use the ratio of total area and sample area to approximately estimate.

$$T_{total} = \frac{Area_{total}}{Area_{sample}} T_N$$

Tree Location Estimate

K-mean Clustering

The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

Other Parameter Estimate

σ estimate

The Tayes distribution for one single tree follows normal distribution.

$$f(x,y) = (2\pi\sigma_1\sigma_2)^{-1/2} exp(-\frac{1}{2}(\frac{(x-\mu_1)^2}{{\sigma_1}^2} + \frac{(y-\mu_2)^2}{{\sigma_2}^2})) \tag{3}$$

The prior of σ distribution is inverse-Gamma.

$$f_{\sigma^2} = Inverse - Gamma(\frac{1}{2}, \frac{1}{2}) \tag{4}$$

Contents

1 Introduction

2 Assumptions

3 Solution

4 Result
- Trees' Location
- DistributionS

5 Discussion
- Error Analysis
- Futre Work

Trees' Location

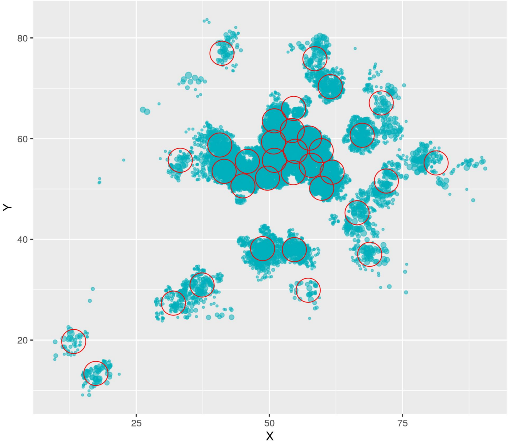

Figure: Fruits location with trees' location

Distribution

Fruit number for each tree

$$F_{Tree} = 59$$

$$\text{numbers of fruits} \sim Possion(F_{Tree})$$

Fruit location Distribution For a tree

$$\text{Tree's location: } \mu = \begin{bmatrix} \alpha_X \\ \alpha_Y \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma 2 \end{bmatrix}$$

where $\sigma_1 = 4.86$ and $\sigma_2 = 2.82$

Tree number

$$N_{Tree} = 38$$

Distribution

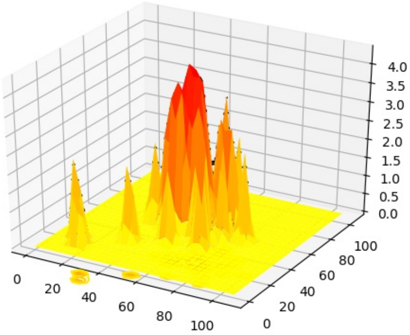

Figure: Distribution Simulation Result

Contents

1. Introduction
2. Assumptions
3. Solution
4. Result
   - Trees' Location
   - DistributionS
5. Discussion
   - Error Analysis
   - Futre Work

Error Analysis

- In this project, we split the dataset randomly into two parts, training part and test part. The proportion is 0.9:0.1
- After we get our model on training set, we use test set to check our result. Basically we use two standard to evaluate our model.
  - The first is `Mean Squared Error` between predicted fruits and ground truth fruits.
  - The second is a kind of soft accuracy. To mark the error within 3 fruits as correct.

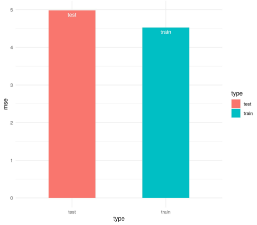

Figure: MSE

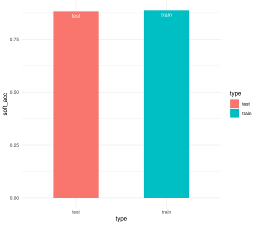

Figure: soft accuracy

Future Work

As you can see, our model still need to be improved. Both the training loss and prediction error are not satisfying.

- The clustering method we used has a large scale of randomness, which couldn't guarantee the performance. Switch to other clustering method or combining with Bayesian method should be one solution.
- Our prediction is much smooth than the ground truth, since we have to sacrifice some extreme value to get an indifferent result. Selecting more specific distribution would be helpful.

What if Jiuling actually can move

We need more data to address the main task and more actuate time-stamp. For example, when did they count fruits"; How long did fruits disappear? We also need to propose appriopate assumption to simplify the problem. Applying some time series or HMM model would be helpful to solve the question. There might be a lot of hidden variables, so we'd better first observe data, then propose model.

Contribution

- Poster: Chen Yuhao, Ruiwei Zhang, Chong Hu
- PPT: Chen Yuhao, Ruiwei Zhang, Chong Hu
- model: Chen Yuhao, Ruiwei Zhang
- visualizaiton: Chong Hu

Reference

[1] Carlin, J., et al. Bayesian Data Analysis. Chapman and Hall/CRC, 2003.
[2] Pearl, Judea. Bayesian Networks. Computer Science Dept., University of California, 1998.
[3] Gelman, Andrew. Bayesian Data Analysis. CRC Press, 2014.