Using Point Processes to model Categorical Data

Shyam Popat
Supervisor: Jess Gillam
Meme Lord Sibling: Aastha Popat

September 5, 2019


Hiya, Jenny, I'm getting old!
Okay.


How often will you visit me to take care of me?


With these bad boyz, I'll always be with you.

Background Information-Context

- Howz is a start-up software company based in Manchester.
- They have designed an award winning home care kit to help people live independently for longer.
- The kit comes with smart sensors which will learn routines within the home by spotting patterns.
- With the app, the household or a carer can track the sensor activity in the house. But the app should also tell you when it detects a change in routine!

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.
- Fitting a homogeneous Poisson model.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.
- Fitting a homogeneous Poisson model.
- Improving the model - separating by time of day.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.
- Fitting a homogeneous Poisson model.
- Improving the model - separating by time of day.
- K-means clustering.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.
- Fitting a homogeneous Poisson model.
- Improving the model - separating by time of day.
- K-means clustering.
- Maximum Likelihood Estimation.

Background Information- Outline of work.

The goal of this project is to create accurate models for real life sensor data.

Some of the approaches include:

- Preliminary work - approaches and challenges in visualising the data.
- Fitting a homogeneous Poisson model.
- Improving the model - separating by time of day.
- K-means clustering.
- Maximum Likelihood Estimation.
- Future work.

The problem in visualising the data

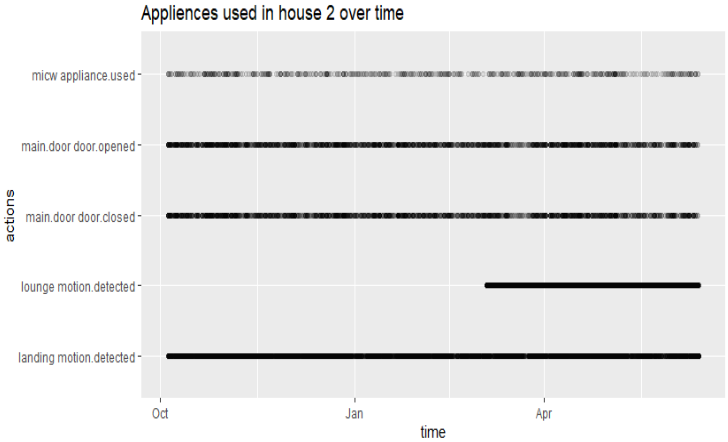


Figure: Data can become difficult to visualise on a larger scale.

One approach- Violin Plot

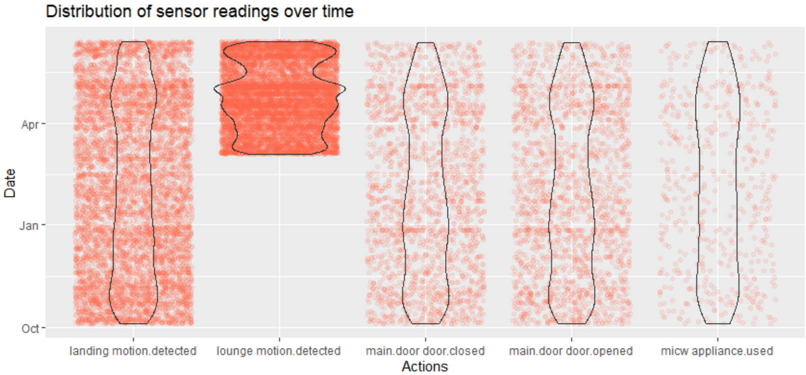


Figure: The violin plot makes it easy to visualise the distribution over time.

A slide for the change point people

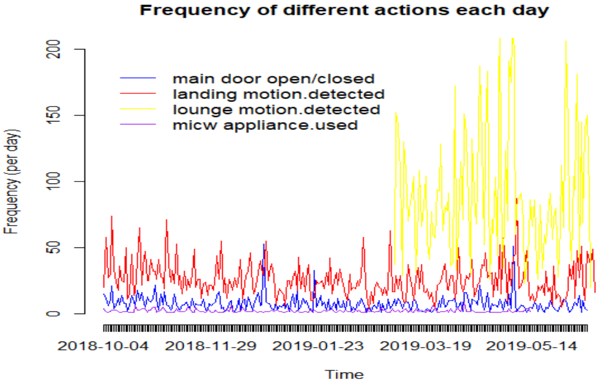


Figure: It is precisely these frequency per day's that our first model will try and estimate.

A homogeneous Poisson model to predict frequency per day by sensor

1. The data spans 253 days. We will use the first 127 days as our training data and the last 126 days as our test data.

A homogeneous Poisson model to predict frequency per day by sensor

1. The data spans 253 days. We will use the first 127 days as our training data and the last 126 days as our test data.
2. Let θ_i be the mean number of realisations of sensor i per day in the training data.

A homogeneous Poisson model to predict frequency per day by sensor

1. The data spans 253 days. We will use the first 127 days as our training data and the last 126 days as our test data.
2. Let θ_i be the mean number of realisations of sensor i per day in the training data.
3. Model 253 days as realisations of a $Poisson(\theta_i)$ random variable.

A homogeneous Poisson model to predict frequency per day by sensor

1. The data spans 253 days. We will use the first 127 days as our training data and the last 126 days as our test data.
2. Let θ_i be the mean number of realisations of sensor i per day in the training data.
3. Model 253 days as realisations of a $Poisson(\theta_i)$ random variable.
4. Find a 2 sided 95% confidence interval for $Poisson(\theta_i)$.

A homogeneous Poisson model to predict frequency per day by sensor

1. The data spans 253 days. We will use the first 127 days as our training data and the last 126 days as our test data.
2. Let θ_i be the mean number of realisations of sensor i per day in the training data.
3. Model 253 days as realisations of a $Poisson(\theta_i)$ random variable.
4. Find a 2 sided 95% confidence interval for $Poisson(\theta_i)$.
5. Check the proportion of test data that falls outside this interval.

Results

		Sensor			
		door close	door open	microwave	landing
Method used	Simple mean	19%	19%	2%	48%
	Moving mean	20%	20%	2%	42%
	separating weekend/weekday	22%	19%	2%	44%
	separating by day	24%	24%	2%	52%

Table: Percentage of observations outside the confidence interval (rounded to nearest %).

Another approach- by time of day.

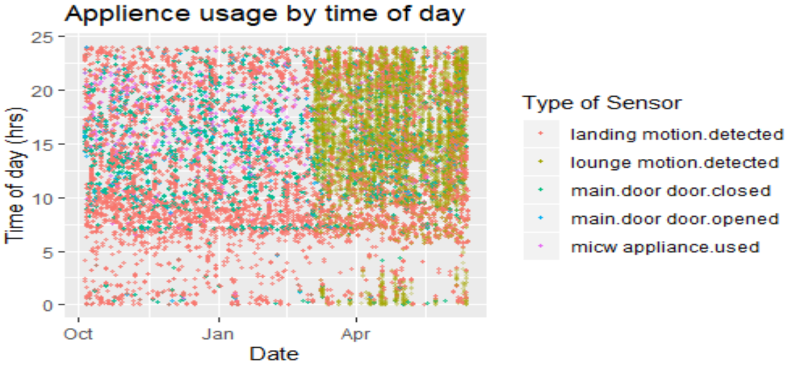


Figure: We can see that the distribution of sensor observations is not uniform over a day.

Patterns by sensor

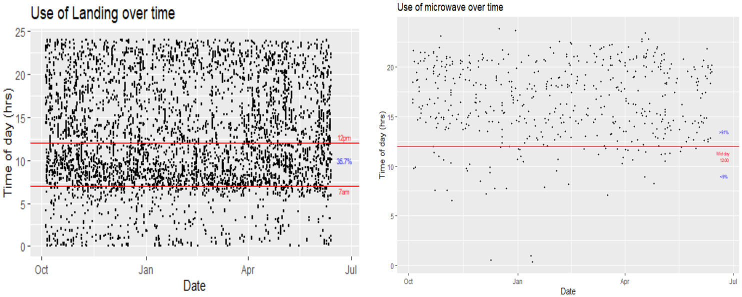
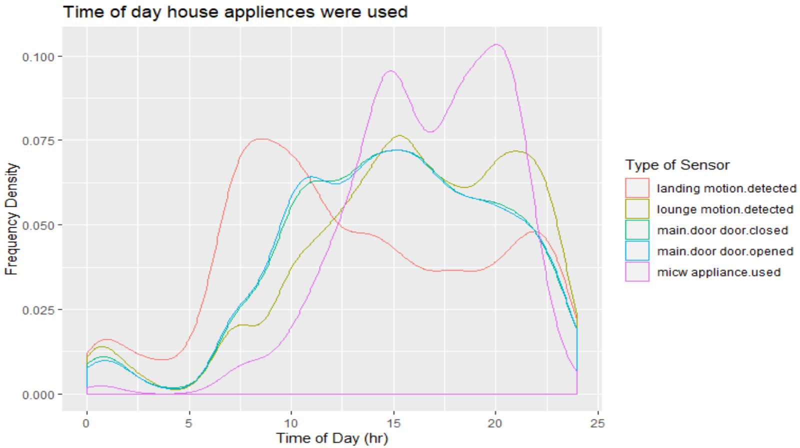


Figure: It is clear that different sensors have different distributions. We should try to take this into account in future models.

Separating by sensor

We use all of the data for each sensor to form the densities:



Method for in-homogeneous Poisson modeling.

1. Split the data into training data and test data as before.

Method for in-homogeneous Poisson modeling.

1. Split the data into training data and test data as before.
2. Use K-means clustering on the training data and fit a constant to each cluster.

Method for in-homogeneous Poisson modeling.

1. Split the data into training data and test data as before.
2. Use K-means clustering on the training data and fit a constant to each cluster.
3. Get a confidence interval for each cluster.

Method for in-homogeneous Poisson modeling.

1. Split the data into training data and test data as before.
2. Use K-means clustering on the training data and fit a constant to each cluster.
3. Get a confidence interval for each cluster.
4. For each element in the test data, assign it a cluster based on the time of day.

Method for in-homogeneous Poisson modeling.

1. Split the data into training data and test data as before.
2. Use K-means clustering on the training data and fit a constant to each cluster.
3. Get a confidence interval for each cluster.
4. For each element in the test data, assign it a cluster based on the time of day.
5. Repeat a process analogous to homogeneous process.

Elbow Method

The basic idea behind partitioning methods is to define clusters so that the total intra-cluster variation- within sum of squares WSS- is minimised.

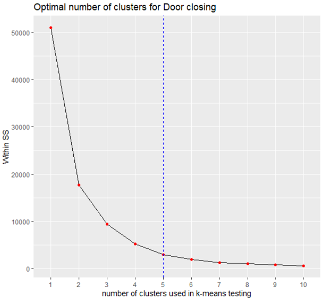


Figure: The data used for WSS is the training data.

Result of clustering

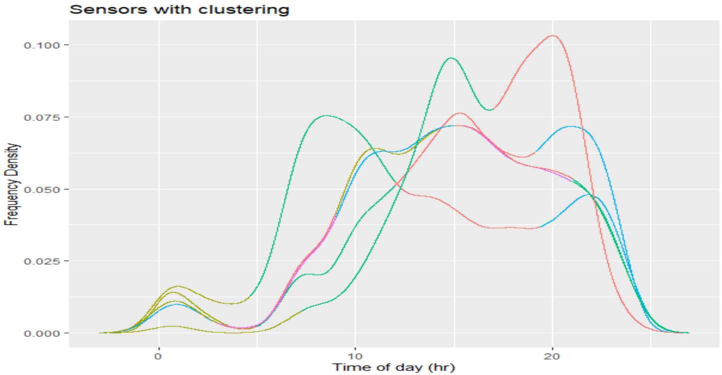


Figure: We can see that usually different peaks have their own clusters.

Modeling each cluster.

For each sensor, we want to fit a constant to each of the clusters.

We determine what constant to use by using Maximum Likelihood Estimation.

MLE

The log likelihood for a time in-homogeneous Poisson Process is:

$$\ell(\theta|\mathbf{x}) = -\int_0^T \lambda(x)dx + \sum_{i=1}^{n} log(\lambda(x_i)).$$

Suppose we are trying to fit m clusters, where cluster i ranges from time t_{i-1} *to* t_i, has k_i observations and an intensity λ_i.

$$= -\lambda_1 t_1 - \lambda_2(t_2 - t_1) - ... - \lambda_m(t_m - t_{m-1})$$
$$+k_1 \; log(\lambda_1) + k_2 \; log(\lambda_2) + ... + k_m \; log(\lambda_m).$$

.

Differentiating wrt $(\lambda_1, \lambda_2, ...\lambda_n)$, we get for each cluster:

$$\frac{\#\text{ of observations in the time interval}}{\text{length of time interval}}.$$

Fitting the constant model

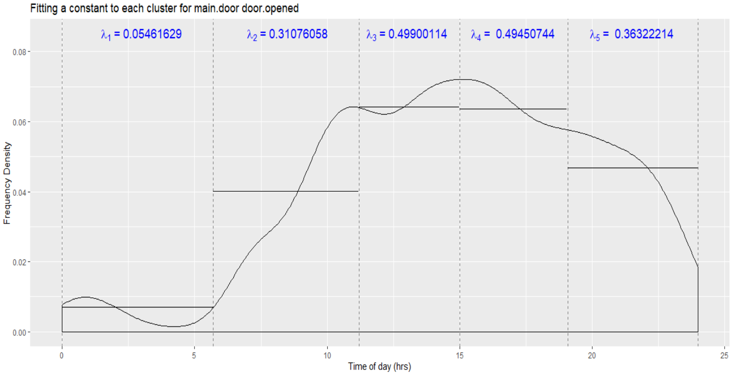


Figure: λ_i represents the per hour frequency in cluster i

Results

		Sensor			
		Door close	Door open	Microwave	Landing
Method	Simple mean	19%	19%	2%	48%
	Moving mean	20%	20%	2%	42%
	weekend/weekday	22%	19%	2%	44%
	Separating by day	24%	24%	2%	52%
	K-means clustering	5%	5%	1%	21%

Table: Percentage of observations outside CI

Further work

There are many avenues I would have liked to explore given more time:

- Modeling clusters with more complicated functions.Hypothesis tests.
- Looking at waiting times between consecutive sensor readings.
- looking at consecutive readings- if the door opens, what can we expect to see next?
- Modeling clusters with more complicated functions. Making a mixture model where we predict different function with different functions.

References

Burr, J. (2019).
https://howz.com/.

Drazek, L. C. (2013).
Intensity estimation for poisson processes.
pages 32–33.

Kassambara, A. (2018).
Determining the optimal number of clusters: 3 must know methods.

Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. (1996).
Stochastic processes, volume 2.
Wiley New York.

Thank you for listening

Any questions?

Appendix 1: Deriving the MLE:

Suppose that we have a sample $\mathbf{x} = \{x_1, x_2, .., x_n\}$ that comes from a non-homogeneous Poisson process. The likelihood $L(\lambda|\mathbf{x}) = f(\mathbf{x}|\lambda)$ is the probability of getting the sample $\mathbf{x} = \{x_1, x_2, .., x_n\}$.

Let $\lambda : [0, T] \to \mathbb{R}_{\geq 0}$ be the intensity function.

$$\Lambda = \int_0^T \lambda(x)dx$$

Probability of observing n points$= e^{-\Lambda}\frac{\Lambda^n}{n!}$

Probability density function of observation $x_i = \frac{\lambda(x_i)}{\Lambda}$

Deriving the MLE for a time in-homogeneous Poisson Process Cont.

Independent observations imply

$$P(\mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda(x_i)}{\Lambda}.$$

Likelihood of getting the sample $\mathbf{x} = P(n)P(\mathbf{x}|n)$, i.e.

$$L(\lambda|x) = e^{-\Lambda}\frac{\Lambda^n}{n!}.\prod_{i=1}^{n} \frac{\lambda(x_i)}{\Lambda}.$$

Our sample $x_1 < x_2 < ... < x_n$ is ordered. Then the likelihood of the ordered sample is the above likelihood multiplied by n!.

Deriving the MLE for a time in-homogeneous Poisson Process Cont.

$$L(\lambda) = e^{-\Lambda}.\prod_{i=1}^{n}\lambda(x_i) = e^{-\int_0^T \lambda(x)dx}.\prod_{i=1}^{n}\lambda(x_i)$$

Therefore the log likelihood is:

$$\ell(\theta|\mathbf{x}) = -\int_0^T \lambda(x)dx + \sum_{i=1}^{n} log(\lambda(x_i)).$$

Appendix 2: Solving for the constant model

Suppose we are trying to fit m clusters, where cluster i ranges from time t_{i-1} *to* t_i, has k_i observations and an intensity λ_i.

$$-\int_0^T \lambda(x)dx + \sum_{i=1}^{n} log(\lambda(x_i))$$

$$= -\left(\int_0^{t_1} \lambda_1 dx + \int_{t_1}^{t_2} \lambda_2 dx + ... + \int_{t_{m-1}}^{t_m} \lambda_m dx\right)$$
$$+k_1\ log(\lambda_1) + k_2\ log(\lambda_2) + ... + k_m\ log(\lambda_m).$$

$$= -\lambda_1 t_1 - \lambda_2(t_2 - t_1) - ... - \lambda_m(t_m - t_{m-1})$$
$$+k_1\ log(\lambda_1) + k_2\ log(\lambda_2) + ... + k_m\ log(\lambda_m).$$

.

Solving for the constant model Cont.

By solving the system

$$\frac{\partial \ell}{\partial \lambda_1} = \frac{\partial \ell}{\partial \lambda_2} = \dots = \frac{\partial \ell}{\partial \lambda_m} = 0$$

We get that for $i \in \{1, \dots, m\}$,

$$\hat{\lambda}_i = \begin{cases} \frac{k_i}{t_i} & \text{if} \quad i = 1, \\ \frac{k_i}{t_i - t_{i-1}} & \text{if} \quad i \neq 1, \end{cases}$$

In each case, this is telling us to set the parameter to

$$\frac{\#\text{ of observations in the time interval}}{\text{length of time interval}}.$$