Using point processes to model categorical data

Shyam Popat[1]   Supervisor: Jess Gillam[2]

[1] MORSE, University of Warwick;   [2] STOR-i CDT, Lancaster University

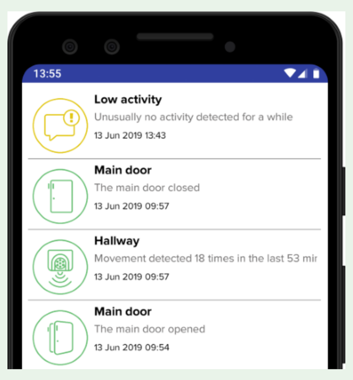
excellence with impact

Background

According to Age UK, as of 2018:

- There are more than 11.9 million people aged 65+ in the UK.
- 3.8 million of these individuals are living alone.
- 50% of those 3.8 million individuals have 3 or more long term health conditions.

Howz is a software company who have designed an award winning home care kit with smart sensors to help people live independently for longer.



With permission, the data from the sensors can be tracked by carers or the NHS via a mobile app.The app should also alert the user when it detects a change in routine.

This project introduces some methodology for modeling Howz sensor readings to form criteria for defining a change in routine.

The data set

Our data comes in the form of pairs of the form (S, t) where S is the sensor that went off and t is the timestamp for when it went off.

We use 4 different sensors- main door opened, main door closed, microwave and landing. The data spans 253 days.

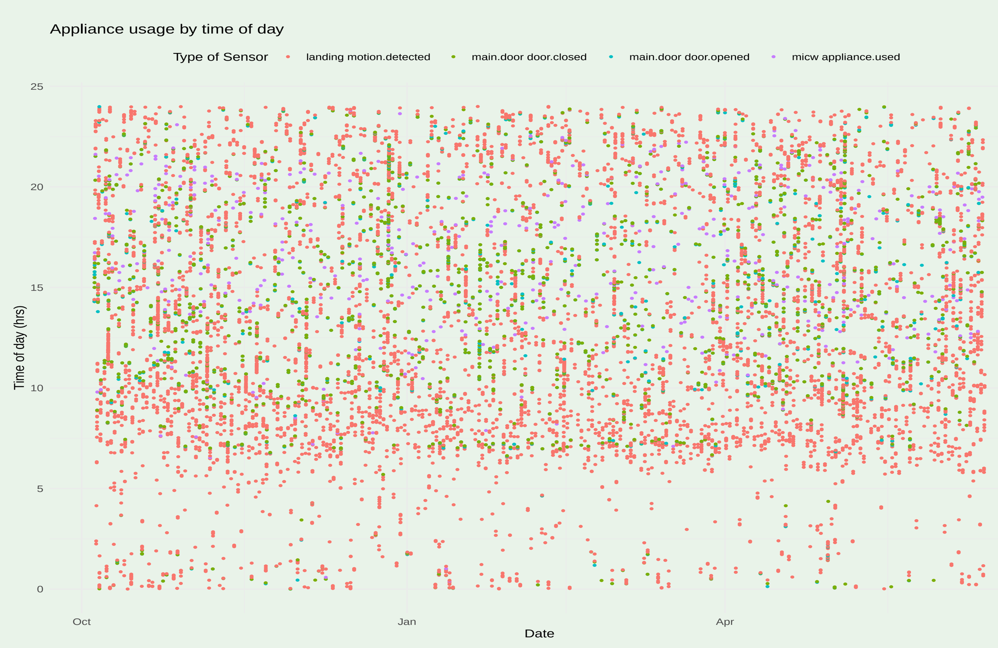


Figure 1: Plot of sensor observations by time of day.

Fitting a basic Poisson model

We first fit a time homogeneous Poisson point process model for each sensor because it is easy to fit:

1. Split the 253 day data into training data (first 127 days) and test data (the final 126).
2. Calculate the mean number of observations per day in the training data.
3. Assume the 253 day data is realisations of a Poisson random variable with mean in step 2.
4. Find a two sided 95% confidence interval for the number of sensor realisations per day.
5. Check the proportion of test data that falls outside this confidence interval.

We also fit models where we separated the data by day, by weekday/weekend and a model where we updated the mean each day.

The results can be found in table 2.

Separating the data by time of day

The basic Poisson model does not account for density of sensor observation throughout a day. Figure 1 shows that sensor observations are not uniformly distributed throughout a typical day.

We tackle this by modeling the density of each sensor by time of day and then using k-means clustering:

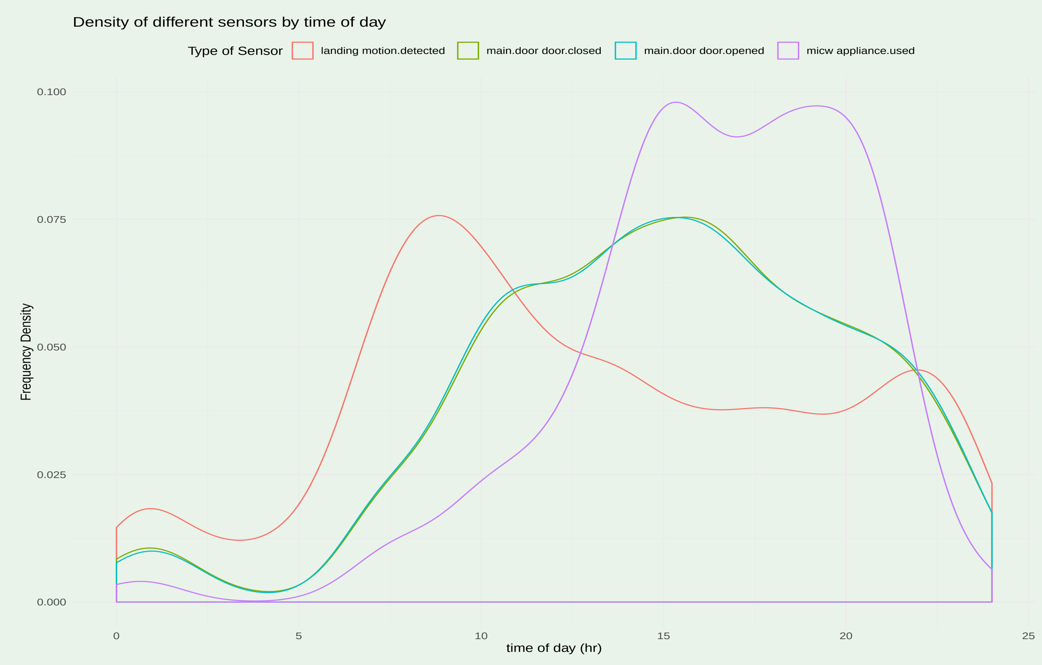


Figure 2: Density of each sensor by using first 127 days.

Determining number of clusters

The Elbow method looks at the total within sum of squares (WSS) as a function of the number of clusters. We choose a number of clusters so that adding another cluster doesn't improve the total WSS by much. Applying this idea gives:

	Sensor			
	Landing	Door close	Door Open	Microwave
#clusters	4	5	5	3

Table 1: Number of clusters we should fit on each sensor using the Elbow method.

Maximum likelihood estimation

We now fit a time in-homogeneous Poisson process by fitting a different constant to each cluster. The constant is determined using Maximum Likelihood Estimation:

Suppose we have a sample of size n, $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, and 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.

Then if we let

$$\theta : [0 : T] \to \mathbb{R}_{\geq 0}$$

be the intensity function, the log likelihood is

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

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

.

By solving the system

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

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

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

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

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

Results

The parameter estimates are as we would expect. Fitting this to the door opening sensor gives:

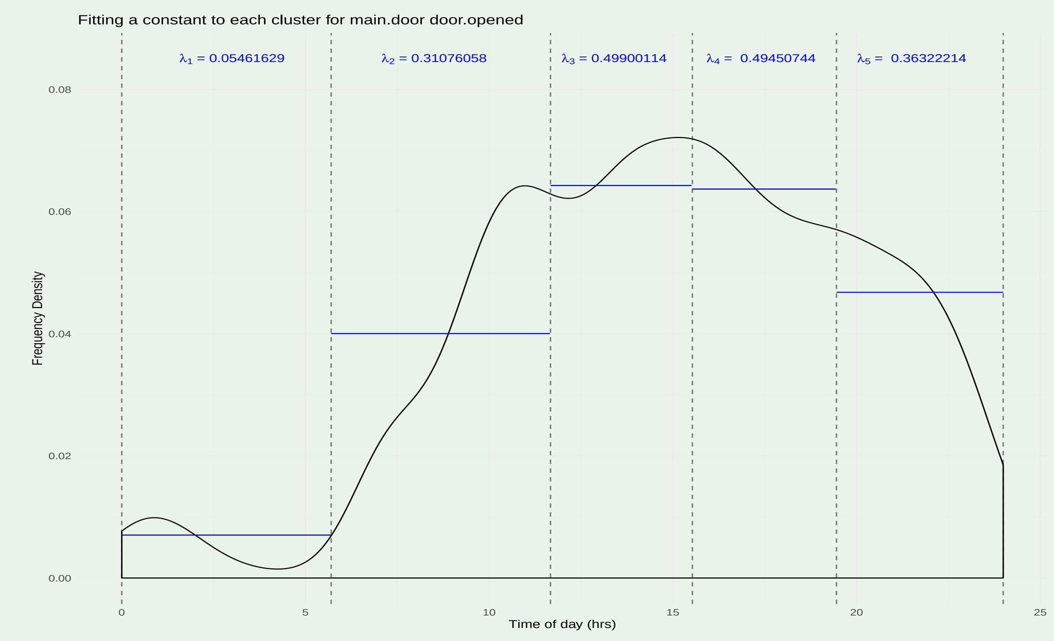


Figure 3: Density and rates obtained the first 127 days.

To model the data using an in-homogeneous Poisson process,

- Split the 253 day data into training data (first 127 days) and test data (last 126 days).
- Use k-means on the training data and fit a constant to each cluster (as in figure 3).
- For each element in the test data, first assign it a cluster based on the time of day.
- Apply an analogous algorithm to the basic model.

The table below summarises the results of applying all of the algorithms discussed thus-far:

		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 2: Percentage of observations outside the confidence interval (rounded to nearest %). Black represents homogeneous models and red represents the in-homogeneous model.

Conclusion

- We have come up with a fairly accurate model that allows us to predict what we expect to see on a cluster by cluster basis.
- Our method flags days with observations either too high or too low. In reality we should not be alarmed if there is a lot of movement- indeed this can be a good sign. Removing this condition nearly halves the error on every sensor.
- The algorithm runs very quickly (few seconds) so can be used in real time.

Further work

Some ideas to improve our current method include:

- Fitting higher degree polynomials or parametric functions to each cluster. This may improve the accuracy of our models and reduce errors.
- Looking at consecutive sensor readings and how we can incorporate them into the model.

References

- Age UK (2019) *Later Life in the United Kingdom*
- Alboukadel Kassambara (2018) *Determining The Optimal Number Of Clusters: 3 Must Know Methods*
- Ludwik Czeslaw Drazek (2013) *Intensity estimation for Poisson processes* pp.32-33

s.popat@lancaster.ac.uk