Part 1: Bathroom visits and daily routine

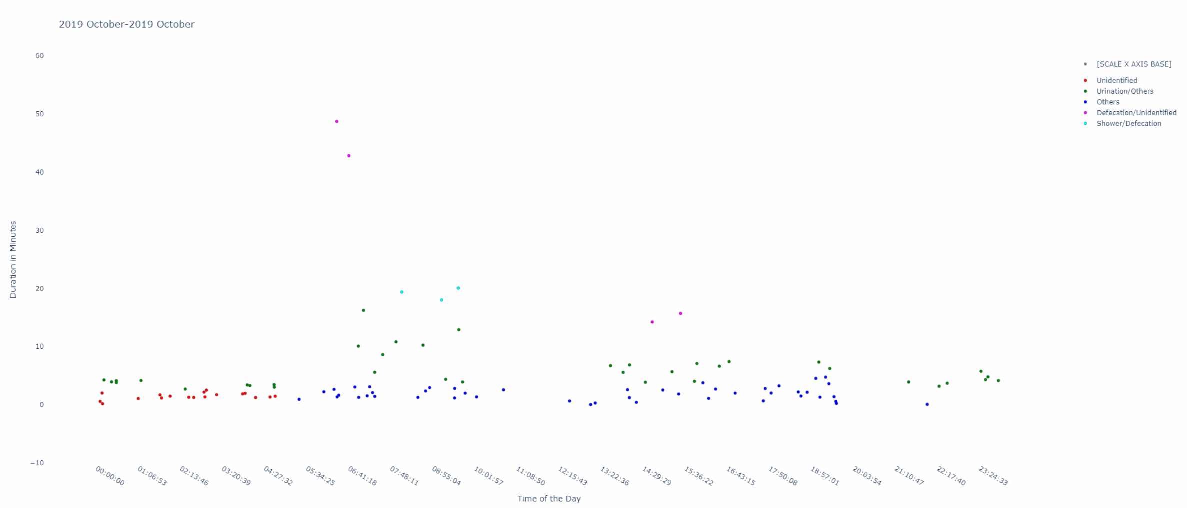

Figure 1. User 5448's activity distribution over 24 hours for "urination" and "others"

[tl;dr]

- Bathroom visits for an elderly is an important piece of medical information for any digital health monitoring system
- In this article, we show how miiCARE leverages only motion data to construct of an understanding of the user's metabolic waste cycles and subtle behaviour changes relevant to bathroom visits
- We were able to assume and validate our models for activity inference in bathroom visits, in the same process
- In the absence of labelled data, we developed an unique approach to classify bathroom visits without needing any machine learning algorithm
- Figure 1 above shows how the approach allowed us to identify and isolate the purpose of each bathroom visit with confidence, over the course of the last 10 months

[Why are bathroom visits important to the miiCUBE?]

Bathroom visits contain crucial bits of information which can inform the miiCUBE about the user's health in several ways. The frequency of visits and the time of the day when the user goes to the bathroom lends direct insight into two categories of medical relevance:

1. **Metabolic Waste Cycles**: Knowing when the user discharges metabolic waste allows us to create a map of their hydration/nutrition intake and track gradual or sudden changes/shifts in the baseline cyclic behaviour. This information can be used to identify development or onset of potential issues affecting the user's metabolic systems.
2. **Subtle Changes in Behaviour**: One can also try to identify whether the user went to the bathroom to wash their hands/face, brush their teeth or take a shower for instance. Such events do not provide direct information unlike the previous category, but they do have relevance as indicators in certain cases. For instance, if the elderly is developing symptoms of forgetfulness as a precursor to dementia, they might start forgetting to brush their teeth in the morning or even delay when they take their shower.

Knowing the probable purpose for which the user went to the bathroom is therefore indispensable in designing the *AI* behind the miiCUBE. In the following sections, we explore how we do this with the minimal amount of information possible and without the use of any visual/acoustic data source medium.

[What does the data look like?]

The miiCUBE collects a variety of information using its IOT network, one of them being motion. When a miiCUBE network is setup at the home of an elderly user, infrared motion sensors are installed in every major room and connecting passage-way. Whenever the user moves in a room or in between rooms, the sensor detects that there has movement and relays this information via the miiCUBE to the Cloud. Every motion event carries with it the information of **the location** and **the timestamp**, among other details. We'll see how this forms the basis of our analyses.

We will be using the data for one of our miiCUBE users to showcase how our approach is applicable across all of them. Due to privacy regulations, we cannot reveal the user's name but we can refer using an anonymized ID `5448`. We will be using the user's motion data from October 2019 to August 2020 to illustrate our process.

[Formulating the hypothesis]

In this analysis, we are concerned with motion events associated with the bathroom(s) in an elderly's residence. Given the timestamp of when motion is detected in the bathroom, we can assign a new piece of information to it; namely, the `part_of_the_day` Let's assume the following demarcations for a day:

- `early morning`: Between 12 AM (midnight) and 5 AM
- `morning`: Between 5 AM and 1 PM
- `afternoon`: Between 1 PM and 3 PM
- `evening`: Between 3 PM and 8 PM
- `night`: Between 8 PM and 12 AM (midnight)

We now claim that there exists a function f such that f(`duration`, `part_of_the_day`) = `activity` where `duration` is the the amount of time spent in the bathroom and `part_of_the_day` is our aforementioned demarcation based on the timestamp. We posit that only the following five classes of bathroom `activity` (visit purpose) can be inferred with confidence:

- `shower`
- `defecation`
- `urination`
- `others`: this class comprise small duration events like washing hands/face, brushing teeth, etc.
- `unidentified`: this class houses anomalies which should raise flags because the time taken is too long given the normal ranges of duration the user has been observed to spend in the bathroom for a `part_of_the_day`.

[Identifying the right classification approach]

We know that there are no pre-existing labels that can tell us what activity happened at a certain time of the day in the bathroom. So we have to resort to unsupervised data mining approaches. In an unsupervised approach, the unlabelled data is assumed to contain hidden patterns which we can help us to assign custom labels. In our case, we set the assumption that the various types of `activity` are distinguishable using only `duration` and `part_of_the_day` and the separation can be identified by the said pre-existing patterns.

Existing data mining literature and machine learning approaches are predominantly built for datasets with multiple fields/variables/inputs (i.e. multivariate datasets). We have one numeric variable - `duration` - and one categorical variable `part_of_the_day`. We could have converted `part_of_the_day` into mutually exclusive binary variables (i.e. through One-Hot Encoding[1]) like `is_it_morning_or_not`, `is_it_afternoon_or_not`, `is_it_evening_or_not` and other such features before moving ahead with one of the contemporary unsupervised machine learning algorithms (e.g. DBSCAN[2] or KMeans Clustering[3]). However, there is a caveat with that: clustering approaches generally do not work well when majority of the variables in the data set are either categorical or derivatives of it. This is because they are discrete.

> *Clustering algorithms use distance between data-points as a classification criteria and it does not make sense to define "distance" between discrete classes*.

[A novel way of classification: 1D segmentation]

We utilized an algorithm called Fischer-Jenks Natural Breaks Optimization[4] which performs "segmentation" with only an array/list of numeric data. This segmentation extends the idea of quartiles into dynamically generated unequal intervals which takes after the numeric progression of the given list of numbers.

Let's consider an example to understand how the Breaks Optimization algorithm works:

- Let 4, 4.1, 4.2, -50, 200.2, 200.4, 200.9, 80, 100, 102 be a list of numbers
- We can intuit that this list can be "segmented" into 4 classes: (4, 4.1, 4.2), (-50), (200.2, 200.4, 200.9), (80, 100, 102)
- The algorithm finds these "breaks" as intervals for us to put the numbers in them. *Please note that we are required to specify the interval count to the algorithm beforehand.*
- The breaks found by the algorithm are -50, 4, 4.2, 102, 200.9
- The breaks can be interpreted as follows:
  - Interval 1 : [-50, 4) (-50 is the lower limit)
  - Interval 2 : [4, 4.2]
  - Interval 3 : (4.2, 102]
  - Interval 4 : (102, 200.9] (200.9 is the upper limit)
  - We considered a fifth interval as well, which we defined as: Interval 5 : (200.9, \infty).

When we applied this to the `duration` values for bathroom visits by User `5448`, we obtained the following intervals. The `duration` values are in minutes here and will be for the other two users as well.

- For `early morning`, breaks = 0, 2, 4, 10, 28.0, INFINITY
- For `morning`, breaks = 0, 3, 9, 17, 29, INFINITY
- For `afternoon`, breaks = 0, 2, 9, 19, 43, INFINITY
- For `evening`, breaks = 0, 2, 5, 10, 21, INFINITY
- For `night`, breaks = 0, 2, 6, 10, 26, INFINITY

The algorithm has essentially reduced univariate (i.e. with one variable) classification into an `if-else` conditional where you can compare lower and upper limits given a new number and allocate that interval (or class) to the new number. For example, taking the breaks obtained for User `5448` as example, if the user goes to the bathroom in the afternoon and spends 5 minutes there, we can put this value in the bin (9, 19] for `afternoon`.

The discovery of this algorithm and understanding how to interpret its results with ease gave us the idea to implement f by breaking it down to its arguments:

- We redefine f as 5 simpler functions: one for every `part_of_the_day`
- For every `part_of_the_day`, we generate 5 Fischer-Jenks Breaks for the bathroom visit durations available for an elderly user during the specific hours of the day. (We want 4 + 1 breaks because we have 4 + 1 possible activity classes to infer for)
- The breaks generated can be used to create an `if-else` conditional ladder to bin values into.

[Naming Intervals for Activity Inference]

So now, we have a `if-else` based classifier function f which, given a `part_of_the_day` and the `duration` spent in the bathroom, provides us with the interval number where it was binned to. This, by itself, is not useful to us in its current form. ***We need to give meaning to what these intervals mean in our consequent analyses.***

We do so by assigning the available activities as labels to the interval bins. There is a simple rationale behind this step:

- We can use our existing experiences to make an initial assumption about what the user could have been doing given the `duration` and `part_of_the_day`.
- For example, we can claim that it takes about 1 ~ 5 minutes to wash our face/hands or brush our teeth. So we can name the bin "Washing Hands/Face" when a bathroom visit is in the afternoon or evening and takes less than 5 minutes OR "Brushing Teeth" when the bathroom visit is in the morning and also takes less than 5 minutes.
- In our case, we have grouped activities like brushing teeth, washing hands/faces, etc. taking less than 5 minutes for an elderly user as `Others`
- Similarly, we have named interval bins for durations between 5 to 15 minutes as for `urination` and for durations between 15 to 35 minutes as `defecation` or `shower`.
- For durations beyond an hour (which is almost non-existent and usually means either a computation error or an emergency), we named the intervals as `unidentified`.
- It is important to note that the naming convention for the intervals is highly subjective and has to be specific to the user.
- It is apparent that we cannot claim that the user was performing a *predicted* bathroom activity with a high degree of confidence. So we paired up reasonable activities together that could be done in succession in one bathroom visit. For example, we have paired up `urination` with `others` as a distinct group. For times when we wish to state that there is a possibility that we have no idea what activity is being performed, we paired up the second most likely activity with `unidentified`.

Let's take up the running example for User `5448` and assign some labels for the intervals obtained in the previous section. We proceed to assign an activity label to every interval range for one specific `part_of_the_day`:

For `early morning`:

- [0, 2] = `unidentified` (We don't what kind of activity would take so less time when it's the middle of the night)
- (2, 4] = `urination/others` (If the user wakes up and goes to the bathroom, a combination of these activities is likely)
- (4, 10) = `urination/others` (same reason as the previous interval's assignment)

- (10, 28] = `defecation/unidentified` (If the user is in the bathroom for more than 10 minutes, one could say that the user is most probably pooping. But there could be instances where we actually have no idea)
- (28, INFINITY) = `defecation/unidentified` (same reason as the previous interval's assignment)

For the sake of brevity, we have shown how we rationalized assigning activity labels to each interval bin obtained for `early morning` for the user. The same process applies to all other `part_of_the_day` for any other user. Although it would be better to do this with prior knowledge about the user's behaviours, one can use common reasoning to do it as well without being too far off from the ground truth.

[Visualization]

When we formulated the approach of using interval binning and classifying events by their `duration` and `part_of_the_day`, we also had the notion that this would help us visualizing how the bathroom events are spread out across the duration of a typical day in the life of the elderly. What we have obtained as of now are assumed label or "inferred activities" for every bathroom visit for every user. To visualize this, we intuited that we should plot events across a 24-hour duration which could help us obtain an idea of the distribution of the various activities across a day. In terms of choosing the plot type, we wanted to stick to flat `xy` plots instead of going about a polar plot for the sake for familiarity and ease of interpretation. So we defined a flat `xy` plot with:

- 24 hours on the `x-axis`
- the `duration` of the event on the `y-axis`
- an unique color for every assumed `activity` label for every event

[Visualizing activity classes : `urination/others` and `others`]

We first observe how the two categories: `urination/others` and `others` are distributed across a typical day for User `5448` in Figure 1 below.

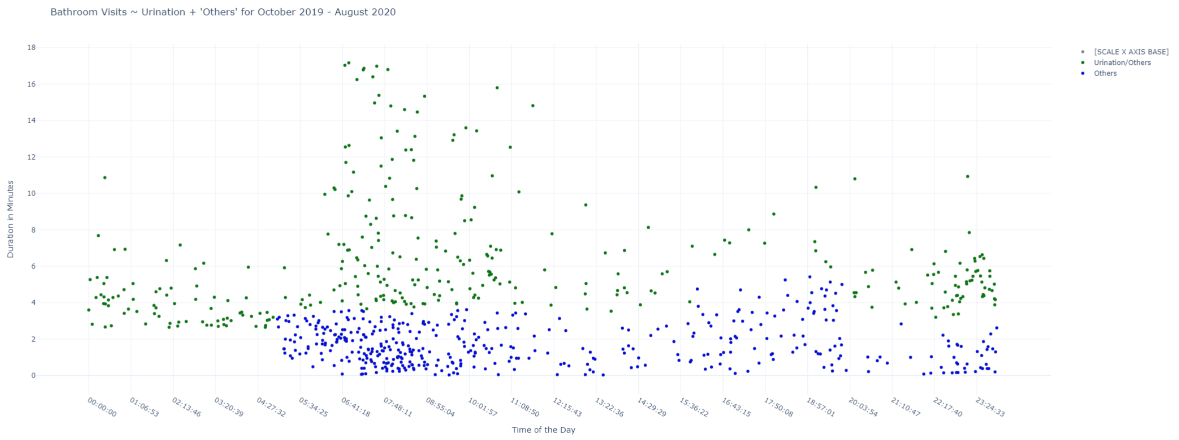

Figure 1. User 5448's activity distribution over 24 hours for "urination" and "others"

1. We see presence of `urination/others` activity during the early morning hours between midnight and 5 AM. This is in fact true to our existing knowledge about this user: that urination visits at this time is not unusual for this person and actually is a symptom of an underlying medical condition (We will have an article on this kind of extrapolative inference in the future).

2. The visits tend to take more time during the morning after waking up. We can attribute this to the fact that she is actually doing multiple activities in the bathroom. Most of these morning visits for `urination/others` typically last to about 12 minutes to our knowledge. The events above the 12 minute mark for `urination/others` can be considered instances when the user took a bit longer at times, hence the increasing sparsity for longer durations in that `part_of_the_day`. This could be up for a better explanation because it may be possible that there are `shower` or `defecation` events among these and they have wrongly classified as `urination/others`. (See Figure 2 below)
3. For `afternoon`, `evening` and `night` the sparsity for `urination/others` decrease a lot when compared to the morning. This tells us that the probability the user visits the bathroom for this class of activity is low during this period. Maybe this hints at an interesting pattern in this user's hydration intake? (Something to explore in a later article)
4. Events for `others` are concentrated in the `morning` when the user does the usual freshening up routine. Although less in density, it is still present all throughout the day. This usually is because of either the user washes is cleaning up or there is a visitor doing the same.
5. We see a slight increase in density of `urination/others` and `others` in the `night` between 10 PM and 12 AM. This is when the user prepares to go to bed for the day. To our knowledge, the user always visits the bathroom before going to bed and these observations conform to that existing information.

[Visualizing activity classes: `shower/defecation` and `defecation/others`]

We now observe the distribution of bathroom visits for `defecation/unidentified` and `shower/defecation` in Figure 2. There are certain things to note here:

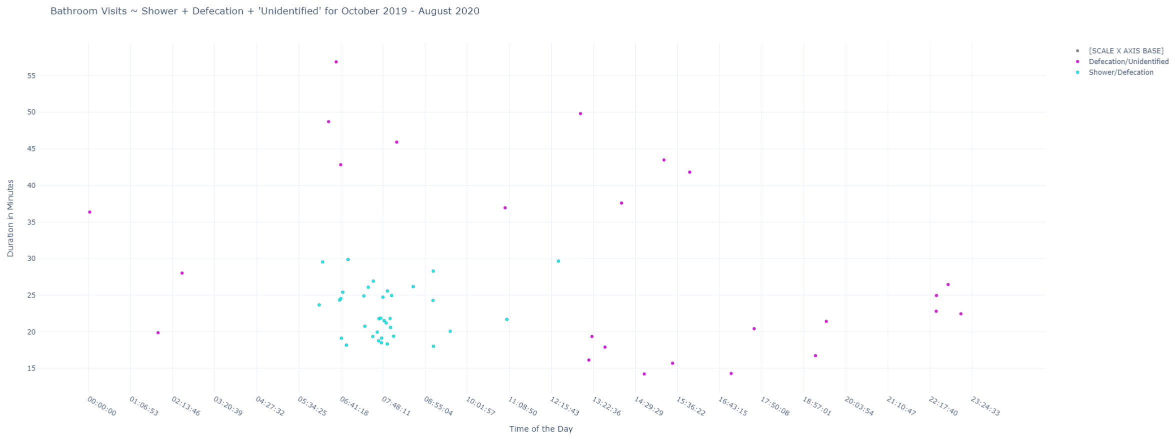

Figure 2. User 5448's activity distribution over 24 hours for "shower", "defecation" and "unidentified"

1. We paired up `shower/defecation` as a single class just by noting `duration` and `part_of_the_day`. This was in essence an assumption just like the other paired up activity classes. Now when we see where those specific events happened during the day, we see them localized densely only in the morning hours. This is a validation of our assumption, of sorts: that we were correct into assigning a label to a specific interval bin.
2. For 10 months worth of data, the cluster for `shower/defecation` looks very sparse here. This can be explained as an error caused by hard margins and `if-else` conditions. In all probability, some of the "classified" `urination/others` events taking more than 2 minutes were most probably `shower/defecation` events in reality .

3. It makes more sense to pair up `unidentified` events than with `defecation` than with `showers` because `unidentifed` events are assumed to take an unusually long amount of time (more than 45-50 minutes) irrespective of `part_of_the_day`. `defecation` events may take a long time, but not that long. If any event takes longer than 50 minutes, it is highly likely that it is an anomaly or an emergency (for example, the user has fallen and become unconsciousness in the bathroom).
4. `defecation/unidentified` events are rare and very sparse as seen in Figure 1. We posit that the events taking over 50 minutes were most likely caused by computational issues (hence "outliers") since they are so few in number.

[Putting it all together]

We now put Figures 1 and 2 together to obtain a more holistic outlook on the distribution of the various classified bathroom activities over the last 10 months worth of data for this user in Figure 3 below. This is actually Figure 1 above with the consolidated data for 10 months.

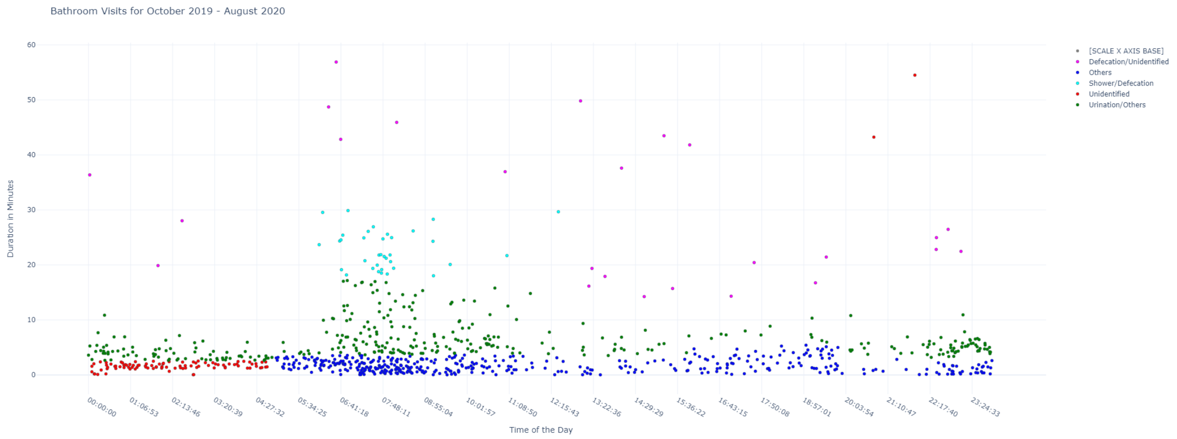

Figure 3. User 5448's bathroom activity distribution over 24 hours

1. We have labelled bathroom visits taking less than 3 minutes between midnight and 5 AM as `unidentified` (red). This is because we weren't quite sure what to name that interval bin initially. Now that we see it on a plot, we believe that those should most likely be `urination/others` events instead.
2. The morning hours from 5 AM to noon are important hours when it comes to the bathroom activities that happen during the hours. Our assumptions in labelling some intervals as `urination/others` and some as `shower/defecation` could be refined and improved, but the plot more or less agrees to our pre-existing knowledge about the user's behaviours.
3. The bathroom activity becomes more sparse in the `afternoon` and `evening`.
4. The user always visits the bathroom before going to the bed and this is indicated from the dense cluster of events for `others` and `urination/others`.

[References]

[1] Draper, N. R.; Smith, H. (1998)."'Dummy' Variables". Applied Regression Analysis. Wiley. pp. 299–326. ISBN 0-471-17082-8.
[2] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).
[3] Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.
[4] Jenks, G. F. (1967). The data model concept in statistical mapping. International yearbook of cartography, 7, 186-190.