ECE PROJECT

GRIDWATCH NIGERIA DATA RELIABILITY ANALYSIS

Name	Sushmita Das
Department	ECE
Email	sushmitd@andrew.cmu.edu
Date	May 12, 2022

Carnegie Mellon University Africa

Acknowledgement

I would like to thank my advisor Prof. Barry Rawn for giving me the opportunity to work on this interesting project. I am thankful to him for his guidance, support and feedback on my work.

Executive Summary

Keywords: Intermittent Power, Wall-plug sensors

Intermittent power is a problem in many developing countries. It occurs due to irregular or frequent interruption of electric power. In recent years, there are many sensors developed which can be used to collect power availability data which can then be analysed to conduct research for solving the problem of intermittent power to make the power system reliable.

In this project, the main objective is to analyse the time series data of power availability to draw insights on the cause of intermittent power. The data on power availability used in this project was collected by a one year campaign in the University of Lagos by using a handful of wall-plug sensors in different houses in a limited geographical scope. The sensors were manufactured by nLine which is a spin-off company from the University of California, Berkeley. The findings of this project can help to draw insights into the patterns of power interruption in different households which in the future can potentially help find and eliminate the route of unreliability in the power system.

Contents

1. Background

1.1 Motivation

Dealing with intermittent power is very crucial to ensure a reliable electrical power system. Data on power availability collected from sensors can be used to understand the cause and to reduce the problem of intermittent power. nLine is a spin off company from the University of California, Berkeley which manufactures sensors that can be used to solve the problem of intermittent power. These type of sensor are deployed in large amounts in Africa and Nigeria Campaign. The vast amount of data generated from these sensors can be very useful and the motivation behind this project is to use the data on power availability collected from sensors for drawing useful insights.

1.2 Prior Work

The authors of [2] introduced PowerWatch, an agile methodology to directly measure customer experience and aggregated grid performance without relying on the utility for deployment or management. PowerWatch employs a system of distributed sensors coupled with cloud-based analytics. This paper evaluated the PowerWatch methodology by deploying 462 sensors in homes and businesses in Accra, Ghana for over a year, yielding the largest open-source data set on electricity reliability at the customer-level in the region.
The authors of [1] provided a detailed mapping of voltage quality in a sub-Saharan African (SSA) community and connected power systems data to household surveys and interviews for a comprehensive understanding of life under an unreliable grid.

1.3 This project

The Gridwatch Nigeria Data Reliability Analysis project was started in the locality of the University of Lagos using wall-plug sensors in a limited geographical area. In this project, the aim is to used the data collected by the one year campaign in University of Lagos to draw insights on power interruption which in the future can potentially help mitigate the issue.

2. Data Quality Evaluation

2.1 Data Overview & Preprocessing

- Data Overview:
  There are three datasets used for this project.

  i. The 'powered_data.csv' dataset has a total of 10 columns where 9 columns correspond to 9 devices and 1 column corresponds to timestamp. At each timestamp, the entries in the 9 columns denote whether the device was turned on or turned off. '0' denotes that the device was turned off and '1' denotes that the device was on. However, there are NaN values in these columns which denote that the device has lost connection and so there is no data about it's turn on/turn off status.

  ii. The 'device-latitudes-jan2019.csv' and 'device-longitudes-jan2019.csv' dataset contains the latitude and longitude values for all of the 9 devices. The columns corresponding to devices 1,8 and 9 have all NaN values. The other devices have both valid data and NaN values.

- Data Pre-Processing:
  I have performed data-preprocessing on the three available datasets mentioned above to create two new datasets which will used for solving the objectives of this project.

  I have created a dataset named 'Nigeria_DA' by doing the following changes to the 'powered_data.csv'.

  i. I have created 9 new columns which denotes the 9 devices named as 'Device 1', 'Device 2',...'Device9'. These new columns have -1 where the value is NaN. Thus, these columns can take 3 values : 1(On), 0(Off) and -1 (NaN).

  ii. I have also created 3 new columns using the timestamp. The three new columns are 'Date' (YYYY-MM-DD), 'Day' (Name of weekday) and 'Month' (Name of the month).

  Further, I have created another dataset named 'Nigeria_Clean_Lat_Long - Sheet1.csv' by doing the following changes to the 'devicelatitudesjan2019.csv' and 'device-longitudes-jan2019.csv' . It does not contain information for devices 1,8 and 9 because the entire column corresponding to these three devices have NaN values in the 'device-latitudes-jan2019.csv' and 'device-longitudes-jan2019.csv' dataset.

  i. Created three columns corresponding to the device name, latitudes and longitudes.

ii. The ‘Device Name’ column has the name of the devices. The ‘Latitudes’ column has the latitude of each device which is calculated by taking an average of all non-NaN latitude values in the ‘device-latitudes-jan2019.csv’ dataset. Similarly, the ‘Longitudes’ column has the longitude of each device which is calculated by taking an average of all non-NaN longitude values in the ‘device-longitudes-jan2019.csv’ dataset.

Both the datasets ‘Nigeria_Clean_Lat_Long - Sheet1.csv’ and ‘Nigeria_DA.xlsx’ are used for the purpose of this project.

2.2 Visualising the Location of Devices

I have visualized the location of the devices using ’Nigeria_Clean_Lat_Long - Sheet1.csv’. From Figure 2.1, we can see that the location of Device 6 is probably incorrect as Device 6 seems to be in the middle of an ocean. The other devices are in the same locality and so we need to zoom in the map to see the location more clearly as they are overlapped in this figure. The other devices excluding Device 6 can be seen more clearly in Figure 2.2 where we can see that the location of all the devices except Device 6 is in the University of Lagos.

We can see the locality of all the devices except Device 6 in Figure 2.3. We can see that the all the devices are in an urban environment. We can also see a zoomed in version of the urban environment in Figure 2.4

2.3 Computing the Fraction of Null Values

For each device, I have computed the fraction of null values by dividing the number of times there is NaN in the column to the total number of entries in the column.

From the results (i.e Figure 2.5), we can see that Device 3 has the highest fraction of null values and Device 8 has the lowest fraction of null values.

2.4 Computing the Power Availability

- Overall Power Availability:

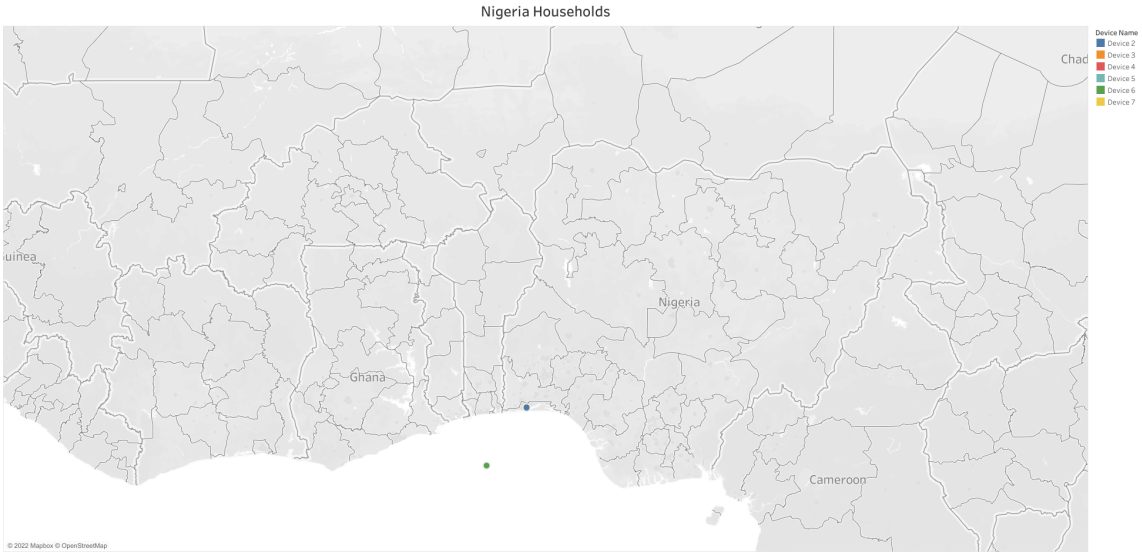
  For each device, I have computed the overall power availability of the device by dividing the number of times the device is turned on (i.e 1) to the total number of entries in the column.

  From the results (i.e Figure 2.6a), we can see that Device 7 has highest overall power availability and Device 3 has the lowest overall power availability.

- Weekly Power Availability

  For each device, I have computed the weekly power availability by finding the power availability for each day of the week. For example : The power availability for Sunday is computed by finding the total number of times the device was turned on (i.e 1) on all Sundays divided by the total number of entries when it was a Sunday.

  From the results (i.e figure 2.6b), we can see that the weekly power availability was highest on Wednesday and the lowest weekly power availability was on Friday.



(a) Tableau Dashboard

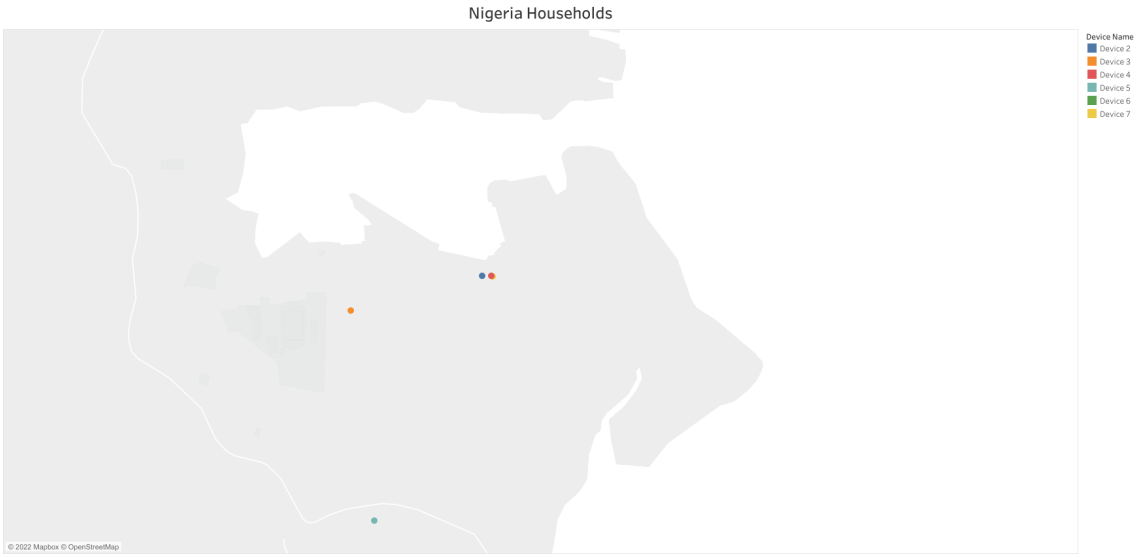
(b) Google Map - Satellite Layer

Figure 2.1: Figures showing all the devices

(a) Tableau Dashboard

(b) Google Map - Satellite Layer

Figure 2.2: Figures showing the location of all the devices excluding Device 6



(a) Tableau Dashboard

(b) Google Map - Satellite Layer

Figure 2.3: Figures showing the locality of all the devices excluding Device 6

Figure 2.4: Zoomed In Version of showing the locality of the Devices.

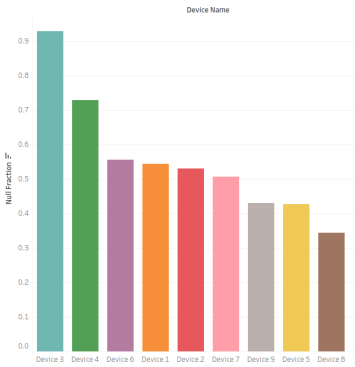


Figure 2.5: Fraction of Null Values

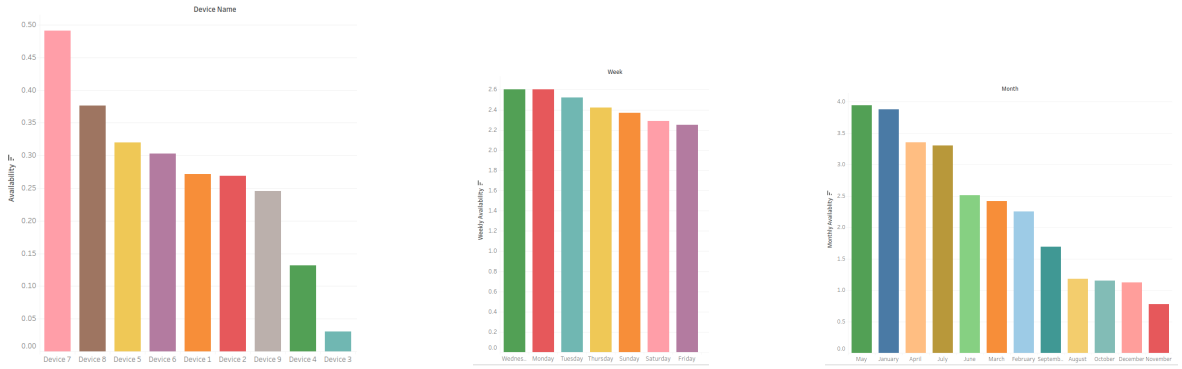
- Monthly Power Availability

  For each device, I have computed the monthly power availability by finding the power availability for each month of the year. For example : The power availability for January is computed by finding the total number of times the device was turned on (i.e 1) in the month of January divided by the total number of entries when it was January.

  From the results (i.e figure 2.6c), we can see that the monthly power availability was highest on May and the lowest monthly power availability was on November.

2.5 Finding & Aggregating Valid Data Ranges

There is a high fraction of null values in the columns of all the devices and so I have found out all possible combination of intervals in the dataset where there are no null values. Further, I have created a valid data by concatenating all those intervals without null values. The valid data contains discontinuous data equivalent of one month duration and we will use the valid data for further analysis.



(a) Overall (b) Weekly (c) Monthly

Figure 2.6: Power Availability

2.6 Visualizing Valid Data

I have created a time series plot for each of the devices using the valid data shown in Figure 2.7. I have made the following observations after visualizing the plots :

i. Devices 2 and 4 have almost identical patterns and so must be highly correlated.

ii. Devices 2 and 4 have similar patterns with device 3.

iii. Devices 1 and 8 also have similar patterns.

Based on these observations, I hypothesize that the correlation might be related to inter-house distance. That is, the houses that are closer have higher correlation between their power availability data.

I have then computed the correlation matrix and the distance matrix to test this hypothesis using the valid data. These matrices are discussed in the future sections.

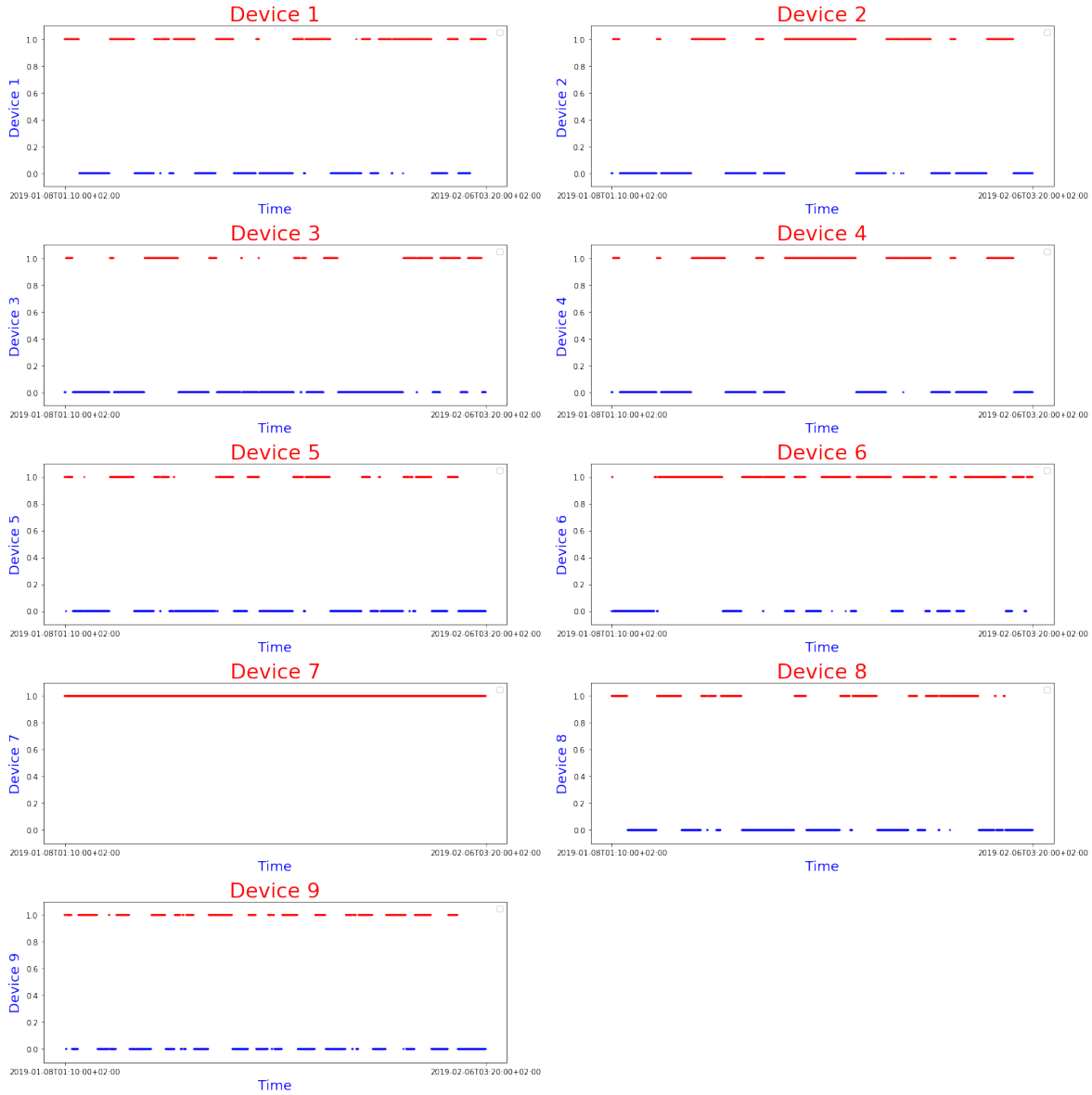


Figure 2.7: Time Series Plot for each of the devices using the valid data.

3. Data Analysis

3.1 Computing Correlation Matrix

Correlation matrix is computed to find the correlation of each device with the other devices. This helps to understand if the devices have similar power pattern. The 'Nigeria_DA.xlsx' dataset is used for computing the correlation matrix. The formula used to calculate the correlation between any device i and j can be written as:

$$corr(i,j) = \frac{cov(i,j)}{\sigma_x \sigma_y}$$

Device 7 is omitted from the correlation matrix. This is because we have seen in Section 2.6 before that device 7 was turned on in the entire duration of the valid data. Since the column corresponding to device 7 contains only '1', the correlation of device 7 is undefined. As we can see from the formula above that to find correlation we use standard deviation and the standard deviation of a constant is zero which makes the correlation of device 7 undefined.

3.2 Computing Distance Matrix

Distance matrix is computed to find the distance of a particular device with other devices. I have calculated the distance measured in km between the pair of devices using the haversine distance metric from sklearn which takes the latitude and longitude in radians as input. The 'Nigeria_Clean_Lat_Long - Sheet1.csv' is used to compute the distance matrix. It also outputs the result in radians and so I have multiplied by 6371 which is the earth's radius to get the distance in km. The result is a 6x6 matrix showing the distance between the pair of devices. The resultant matrix is a 6x6 matrix instead of a 9x9 matrix because we have seen in Section 2.1 that devices 1,8 and 9 do not have valid longitude and latitude values.

3.3 Pairwise Distance Versus Correlation

I have created a pairwise distance versus correlation plot to test our hypothesis that some devices are highly correlated with each other and it is probably the distance between them that drives the high correlation. It is expected to have high correlation for devices closer to each other. I have computed the pairwise distance versus correlation only for devices 2,3,4 and 5. This is because as we have seen in Section 2.1 that there is no information about the location for devices 1,8 and 9. Also, device 6 is omitted because we have seen in Section 2.2 that the location of device 6 is probably reported incorrectly.

4. Results

4.1 Correlation Matrix and Heatmap

	Device 1	Device 2	Device 3	Device 4	Device 5	Device 6	Device 8	Device 9
Device 1	1.000000	-0.235148	0.233583	-0.241801	0.657004	0.071464	0.762440	0.249066
Device 2	-0.235148	1.000000	0.240854	0.993497	0.046390	0.013823	-0.091859	-0.011056
Device 3	0.233583	0.240854	1.000000	0.236115	0.249268	0.212592	0.136730	-0.046953
Device 4	-0.241801	0.993497	0.236115	1.000000	0.041348	0.009316	-0.097607	-0.008596
Device 5	0.657004	0.046390	0.249268	0.041348	1.000000	0.335215	0.621252	0.293890
Device 6	0.071464	0.013823	0.212592	0.009316	0.335215	1.000000	-0.019246	-0.028839
Device 8	0.762440	-0.091859	0.136730	-0.097607	0.621252	-0.019246	1.000000	0.257205
Device 9	0.249066	-0.011056	-0.046953	-0.008596	0.293890	-0.028839	0.257205	1.000000

Figure 4.1: Correlation Matrix

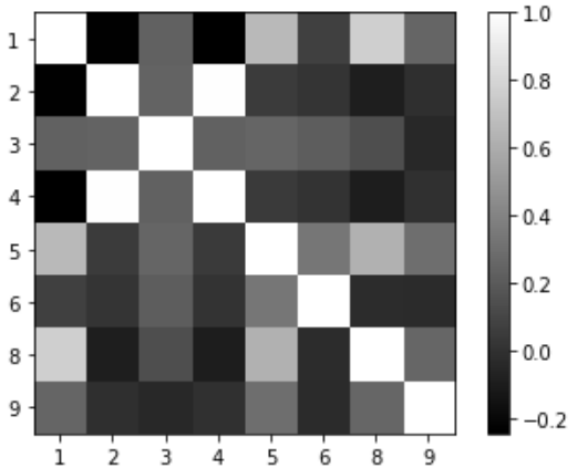


Figure 4.2: Heatmap

From the correlation matrix (i.e Figure 4.1) and the heatmap (i.e Figure 4.2), we can see that the correlation values are in agreement with the observed similarity pattern in availability data. This is because device 2 has the highest correlation with 4 and the correlation is very close to 1 (i.e 0.992497). We had seen earlier that these 2 devices had almost identical availability pattern which explains the high correlation. Similarly, device 2 has high correlation with device 3. Device 1, which had similar availability pattern with device 8, has the highest correlation with device 8 and the correlation value is also high i.e 0.762440.

4.2 Distance Matrix

	Device 2	Device 3	Device 4	Device 5	Device 6	Device 7
Device 2	0.000000	0.504982	0.032866	0.990255	218.198497	0.037845
Device 3	0.504982	0.000000	0.536795	0.780954	217.813419	0.540590
Device 4	0.032866	0.536795	0.000000	1.003932	218.217225	0.006025
Device 5	0.990255	0.780954	1.003932	0.000000	217.226530	1.002731
Device 6	218.198497	217.813419	218.217225	217.226530	0.000000	218.216975
Device 7	0.037845	0.540590	0.006025	1.002731	218.216975	0.000000

Figure 4.3: Distance Matrix

From the results (i.e figure 4.3), we can see that the hypothesis is in alignment with the distance between devices 2 and 4. Amongst all the devices, device 2 and 4 have the smallest inter-device distance and the highest correlation in availability data. Device 2 has a small distance with device 3 and there is high correlation between the availability data of these 2 devices. This also aligns well with our hypothesis. We cannot comment on the relationship between device 1 and 8. Their inter-device distances could not be computed because there is no available information about the location of device 8.
Also, we have seen in Section 2.2 that the location of device 6 is probably incorrect as it shows the device to be in middle of a ocean. This is the reason that the row and the column corresponding to device 6 have high values in the distance matrix.

4.3 Pairwise Distance Versus Correlation

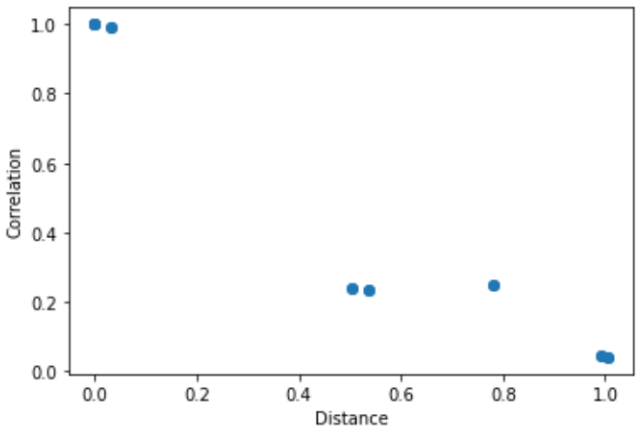


Figure 4.4: Dependence of correlation on distance - devices 2,3,4 and 5

The results (i.e Figure 4.4) show that the correlation decreases as distance increases. Thus, it provides a strong evidence in support of our hypothesis that some devices are highly correlated with each other and the correlation is probably driven by the distance between them.

5. Conclusion

In this project, I analyzed the power availability data of 9 devices deployed in 9 different houses in the locality of the University of Lagos. The power availability varies across the days of the week and across different months.

The available data had a lot of Nan values and I performed data cleaning and aggregated valid intervals. Upon analyzing this clean data, I saw that some devices have highly correlated power availability data. Based on this observation, I hypothesized that the devices with higher correlation are located closes to each other. I plotted the correlation vs inter-device distance to test this hypothesis. I saw a clearly decreasing trend in correlation with increasing distance. This strongly supports my hypothesis.

References

[1] Veronica Jacome, Noah Klugman, Catherine Wolfram, Belinda Grunfeld, Duncan Callaway, and Isha Ray. Power quality and modern energy for all. *Proceedings of the National Academy of Sciences*, 116(33):16308–16313, 2019. [Cited on page 1]

[2] Noah Klugman, Joshua Adkins, Emily Paszkiewicz, Molly G Hickman, Matthew Podolsky, Jay Taneja, and Prabal Dutta. Watching the grid: Utility-independent measurements of electricity reliability in accra, ghana. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*, pages 341–356, 2021. [Cited on page 1]