Redwood

Xinyi Sheng, Yucan Zeng

2022-10-11

Abstract

While the original 2005 paper was a preliminary study of the macroclimate of Sequoia trees along the California coast using wireless sensor networks, this report thoroughly examines the Sequoia dataset in a different way, focusing mainly on the preparation and cleaning before starting the data analysis. Five sections are included: raw data collection, data cleaning, data exploration, interesting findings, and conclusions.

1. Data Collection

a.Paper Summary

The thesis focuses on the ecophysiology of Sonoma redwood trees in California using wireless sensor networks to explore the complex spatial variability and temporal dynamics of microclimate around coastal redwood trees. Air temperature, relative humidity, and photosynthetically active solar radiation (PAR) were monitored every 5 min for 2 m for 44 days during the early summer of the redwood tree's life by placing magnetic suction sensors on a 70 m high redwood tree. The paper describes the placement of magnetic suction sensors, one sensor every 2 m from 15 m to 70 m above ground. Since the canopy is thicker on the western side of the redwood tree, the sensors this to be placed on the western side. Sensors were also placed at 0.1-1.0 m from the trunk to ensure that microclimate trends affecting the trees could be captured directly. And TASK and TinySQL software were used as the framework for collecting and processing the data. The first phase investigates the extent and distribution of the data independent of time and space by transforming a 3-dimensional dataset with a temporal dimension, a spatial dimension, and a dimension related to the sensor values themselves into a 1-dimensional dataset, the second phase adds a temporal dimension to investigate the temporal trend of the data can be seen in the weather over the time scale with a temporal gradient, and the third phase investigates the spatial trend of the data on the basis of the 1-dimensional dataset. The third stage examines the spatial trend of the data based on a 1-dimensional dataset, where the trend in height helps to derive a theoretical model of canopy density and the existence of a spatial gradient in tree height, and the last stage examines the three-dimensional data, where the sensor values are contrasted by shades of color in different times and spaces. Through humidity graphs, the dense canopy can have a buffering effect on the lower part of the tree.

b.

i.Sensor Deployment Mode

Time: Early summer for one month, with all sensors sampled every 5 minutes. Early summer contains the most dynamic microclimate changes. We decided that sampling every 5 minutes would be sufficient to capture this variability. Vertical distance: from 15 meters from the ground to 70 meters from the ground,

with a spacing of about 2 meters between nodes. This spatial density ensured that we could cap the gradients in sufficient detail to interpolate with accuracy. envolop starts at 15 m because most of the foliage is in the upper part of the tree. Angular position: west side of the tree. The western side of the tree has a thicker canopy and provides the greatest buffer against direct environmental impacts. Radial distance: 0.1-1.0 m from the trunk. These nodes were placed very close to the trunk to ensure that we capture microclimate trends that directly affect the tree, rather than the broader climate. Figure 1 shows the final location of each mote in the tree. We also placed several nodes outside of the angular and radial envelopes to monitor microclimate near other previously installed biosensing devices.

ii.Duration of Data

From 4/27/2004 at 5:10 p.m. (epoch 1) to 6/10/2004 at 2 p.m., a little over 44 days' worth of data were collected (epoch 12635). Every five minutes, measurements were obtained, and battery-operated nodes were duty cycled to save energy when not in use. They were switched on for four seconds to take measurements before being turned off until the next reading.

iii.Main Variable

Traditional climatic factors such as temperature, humidity, and light levels were of interest to researchers. These variables were monitored using photosynthetically active radiation (PAR). Two measurements were made of light between the wavelengths of 350 nm and 700 nm: one was incident (direct), which gives data on the energy available for photosynthesis, and the other was reflected (ambient), which was used to validate observations by satellites.

iv.Difference between the Dataset

Log is retrieved from flash log after the deployment. net is retrieved from wireless internet. to provide backup in case of network failure and to provide a basis for analyzing the performance of network work, we extended the TASK framework to include a local data log system. The data logger records every read performed by each query before the reads are passed to the multi-hop routing layer and stops recording when the 512kB flash chip is full. After deployment, we connected each mote to a serial connection and then installed a new program to transfer the contents of the flash memory over the serial link. We chose to include a full data logger because we knew the capacity of the flash would be sufficient for the duration of our deployment. Longer deployments should consider including a storage system that supports multiple resolutions [6] or one that can be partially retrieved over the network. Each reading received from the multi-hop network is stored in a database management system running on the Stargate gateway node as a staging area until we retrieve the results. We retested some results in real time directly over a GPRS cellular modem connection and periodically connected the laptop to the Stargate in order to download the rest. the GPRS connection was also used to remotely restart the gateway node in the event of a failure.

2.Data Cleaning

a.Check Histogram

Before plotting the histogram, we initially filtered the variables according to the article table1 range for each sensor and range of voltage, removed the NA, controlled the locallog voltage at (2.4, 3), network voltage at (200, 250), then controlled the temperature at (6.6, 32.6) and humidity at (16.4, 100.2). Also, because the time is a trash, we also corrected the time.

We use the raw data to plot the number of measurements recorded by loggers and networks for each period. Figure 1 shows that the network data are missing and only appear from May 7 to June 2. locallog data

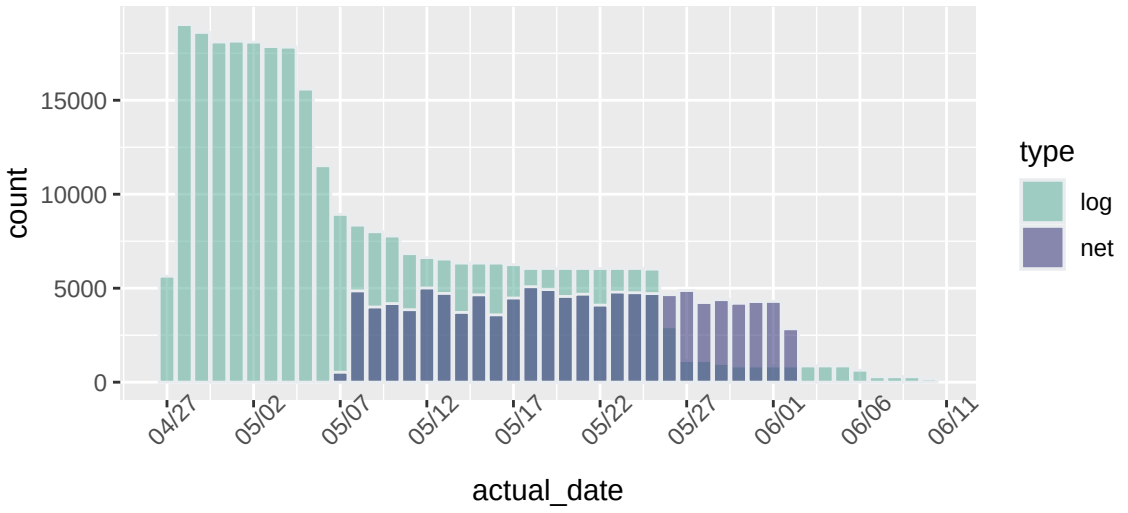

Figure 1: Data Distribution over Time by Log and Net

decreased on May 5 and May 6, and on May 26. Next, we made histograms of the number of measurements recorded by the logger and the network for each period for voltage, humidity, temperature, hamatop, and hamabot, respectively. It is obvious that in the voltage distribution, the network data is very problematic, and the locallog data is concentrated around 0. And we find that the value 200 to 300 is taken when filtering the network because if we go beyond this range we can only see the locallog and the network value is not useful. Also, we can clearly see from the locallog and network two graphs that the units of voltage are not the same, so we performed the operation of converting the voltage. In the humidity histogram, we can see that locallog records data evenly distributed at each humidity value, but network logger records more data at humidity greater than 50, and we speculate that network logger may not be sensitive enough for smaller humidity. In the temperature histogram, both locallog and network data are well recorded. Since hamatop and hamabot only record data when there is light, it is reasonable that most of the data of hamatop and hamabot are recorded near 0.

b.Remove Missing Value

After viewing the dataset we ensured the best analysis of the data by removing NA, outluers, duplicates and filtering the ranges of the variables. First, we drop the rows containing NA, and then we filter the data for duplicate rows of epoc and nodeid, keeping only one row. Since the network voltages were converted to adc, we divided the network voltage data by the conversion factor. Next, we replicated the voltage filtering steps outlined in Tolle's paper, removing entries with voltages greater than 3V or less than 2.4V, and filtering the data according to the data range in table1 in the paper. Finally we filtered out the anomalous sensors using moto-location-data.We use table 1 to show the data filtering process.

c.Identify Outliers

Tolle's article mentions the problems associated with outliers corresponding to the measurement range of humidity, and we also found when looking at the histogram data that humidity less than 0 and temperature up to 100 degrees are obvious outliers that do not correspond to reality, so these data should be excluded from the analysis.

In the box line plot, firstly for humidity, we filtered only the readings with quantile at (0, 0.02). Secondly,

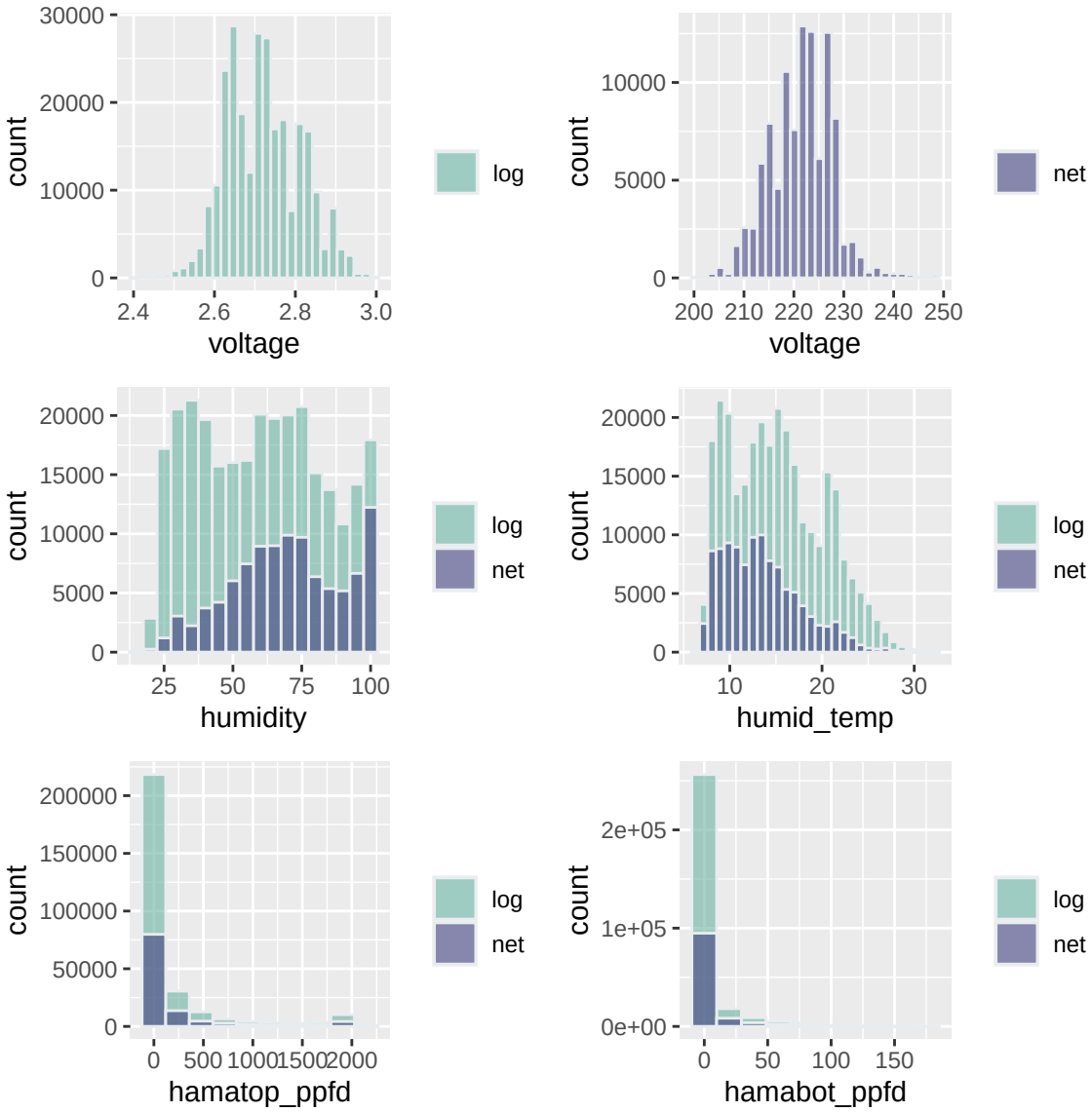

Figure 2: Count of Data

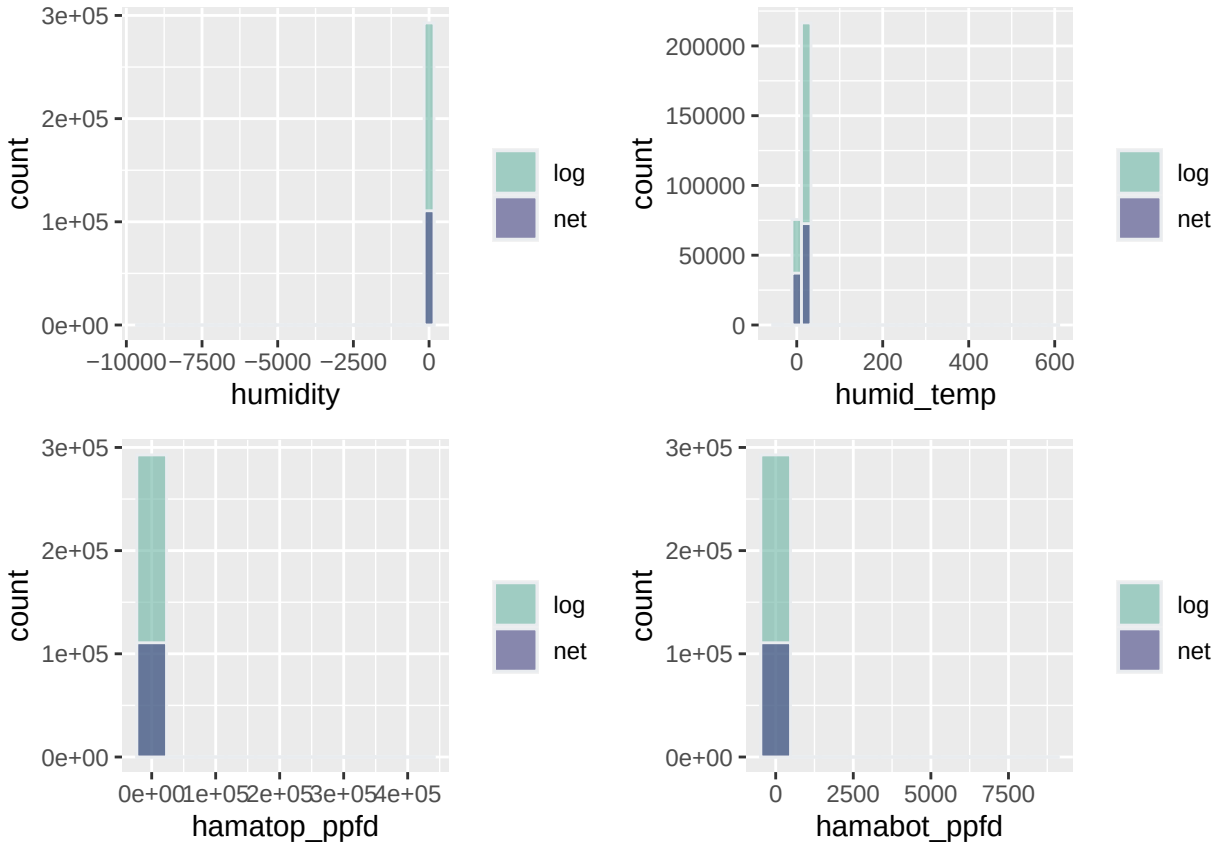

Figure 3: Histograms

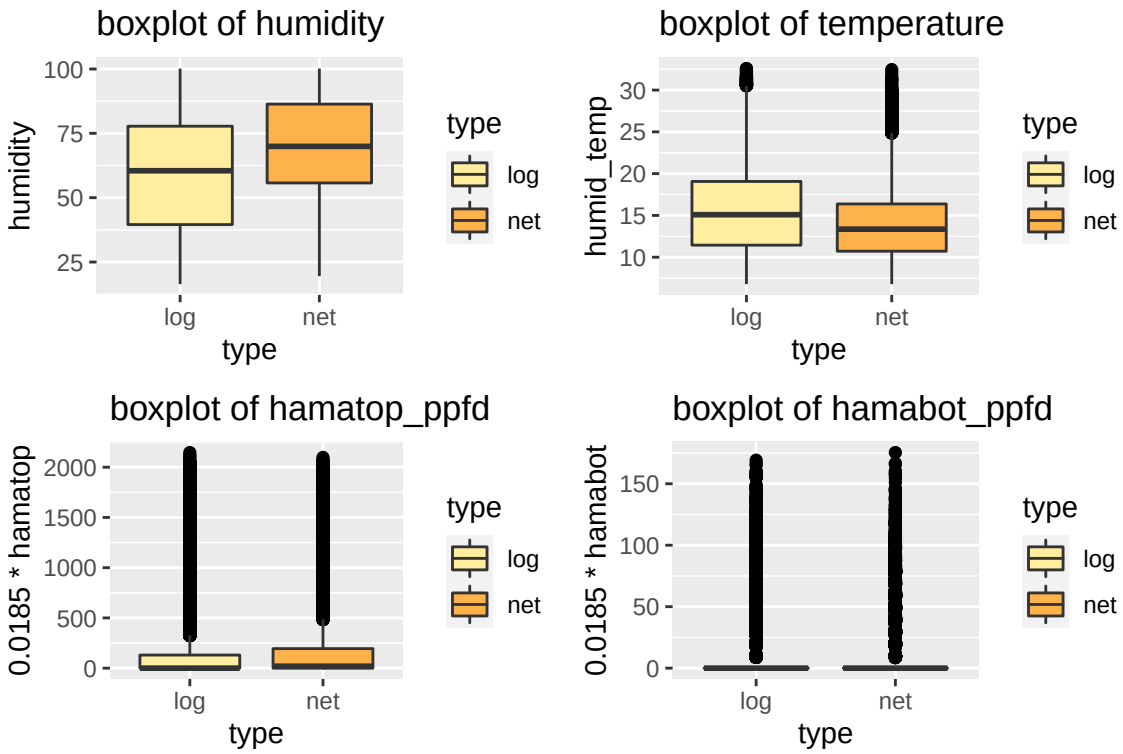

Figure 4: Boxplots

for the temperature boxplot we filtered the quantile readings at (0.98, 1). And we also screened the data for hamatop and hamabot at (0.98, 1). In the humidity boxplot, the mean value of network is larger than the mean value of locallog exactly in line with our guess that the network sensor is not sensitive to lower humidity and may be missing a large number of readings. In the box line plots of hamatop and hamabot it is obvious that the readings are heavily concentrated around 0, so the box line plots would have more outliers. Because there is a large amount of time without sunlight, resulting in 0 readings, we do not remove a large number of outliers considering the specificity of the data.

In addition to using the filtering conditions given in the article, we also converted the battery voltage of the network sensor by a factor during data processing, thus further removing the outliers instead of simply discarding that part of the data or leaving it unprocessed.

3.Data Exploration

a.Pairwise Analysis

Three different time periods were chosen: 6 am to 9 am (sunrise time), 10 pm to 18 pm, and 17 pm to 20 pm (sunset time). These three time periods were chosen because it is clear in Figure 4 of the Tolle's article that there are spatial and temporal trends in humidity and temperature during these three time periods, and intuitively, these two variables change dramatically at sunrise and sunset, and we wanted to find and quantify the relationship between temperature and humidity. At the same time, humidity and temperature show small and dramatic fluctuations at midday, so we chose to explore the potential correlation between humidity and temperature during the three time periods.

The pairwise scatter plot of humidity and temperature shows a relationship between temperature and humidity. We can see that as the temperature rises, the humidity decreases. There is a linear relationship

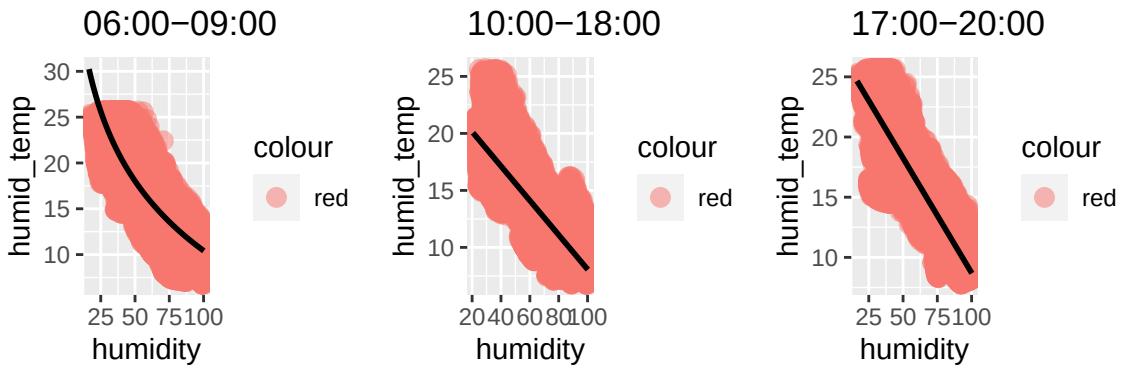

Figure 5: Scatterplots

between temperature and humidity at sunrise and noon. However, the trend is logarithmic at sunset.

b.Correlation

We created correlation plots based on our three time periods to recognize any of the predictors associated with Incident PAR. It is evident that Incident PAR has a medium correlation with Reflect PAR, and the closer they are to sunset, the more robust the correlation is. In addition, Incident PAR has a weak correlation with temperature, and the correlation coefficient is more significant during the day than at night.

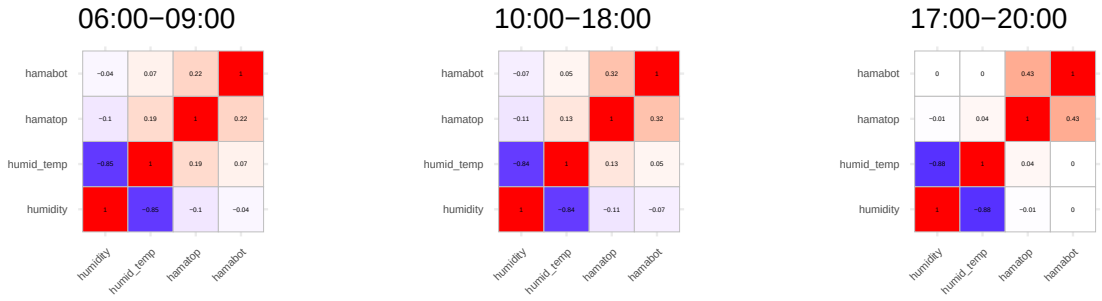

Figure 6: Correlation plots

c.Time Series

d.PCA

4.Interesting Findings

finding 1

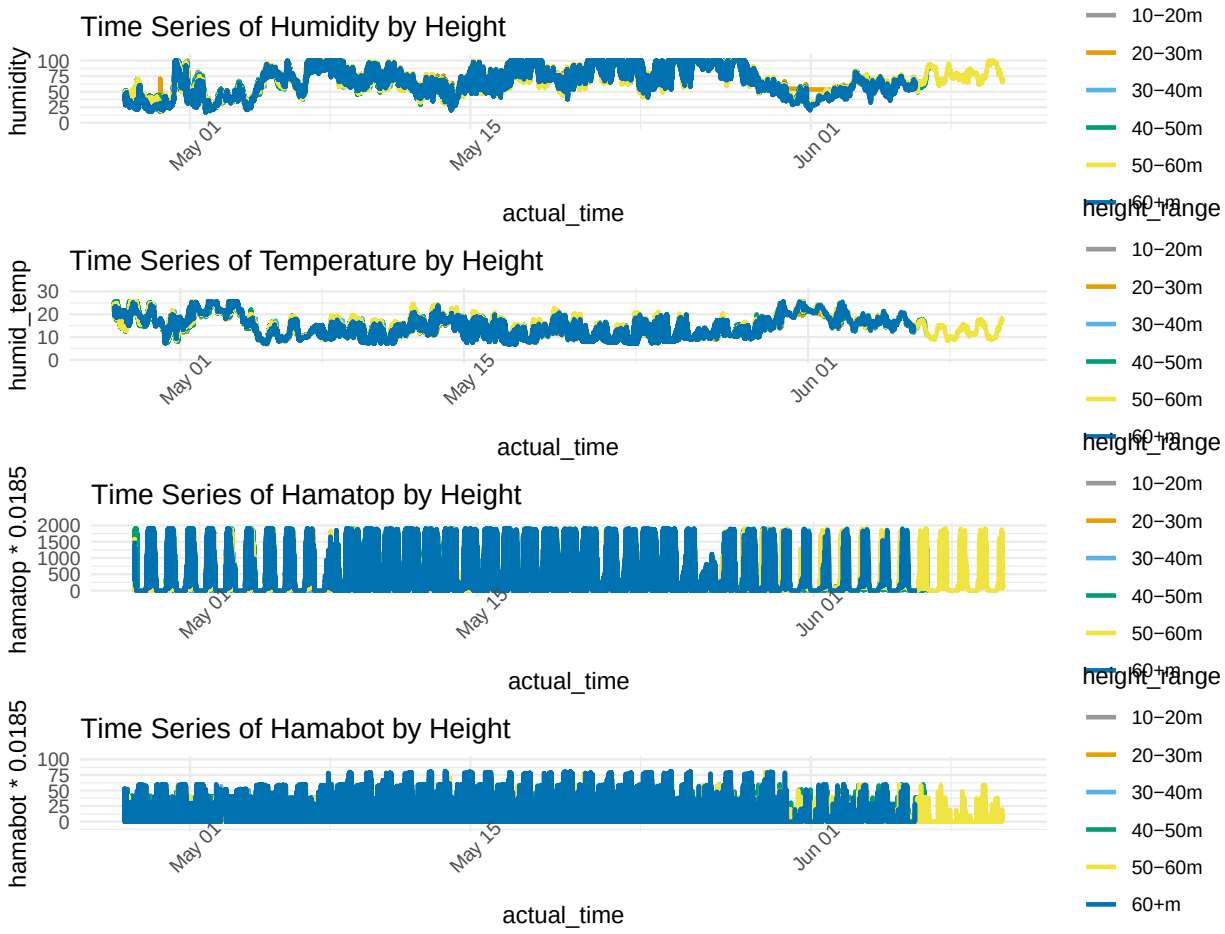

Figure 7: Time series

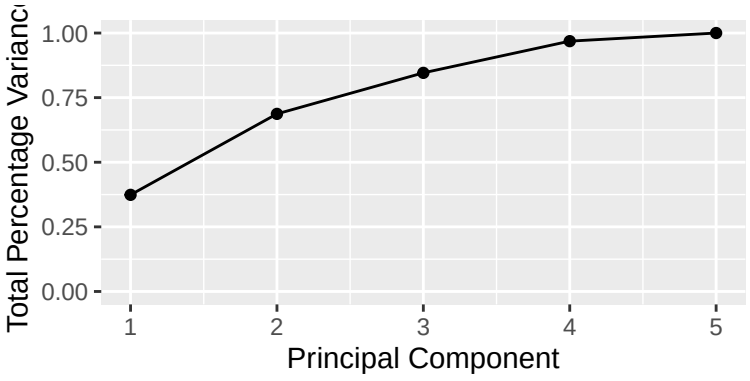

Figure 8: Scree Plot

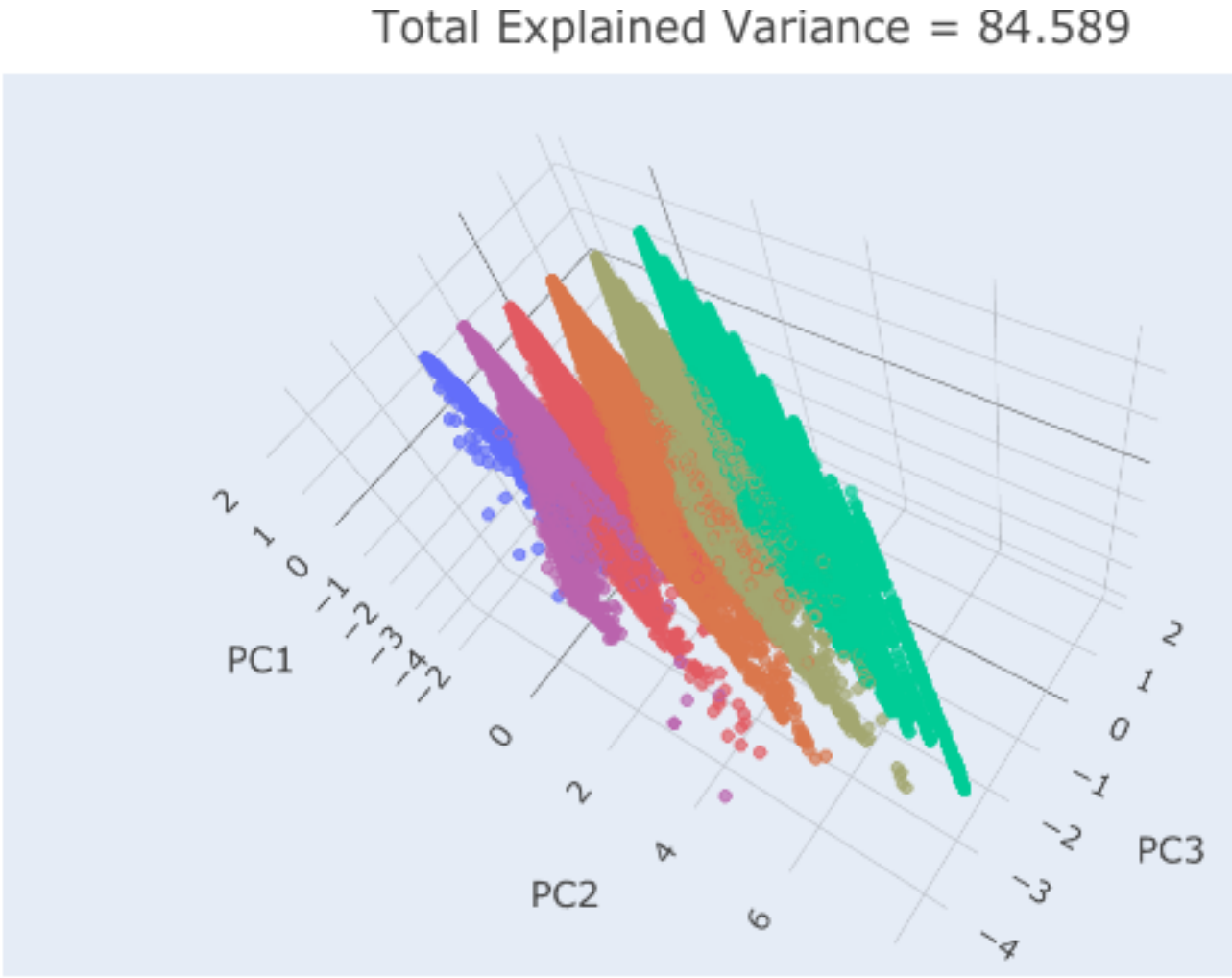

Figure 9: PCA

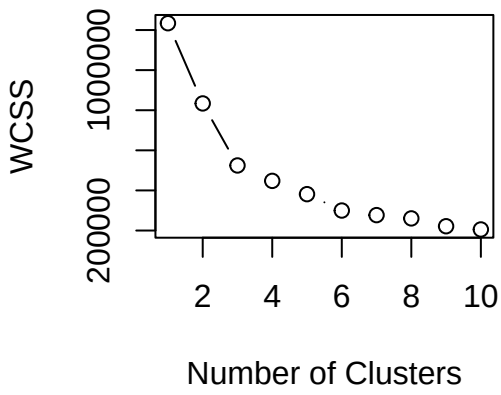

Figure 10: Clusters of Clients

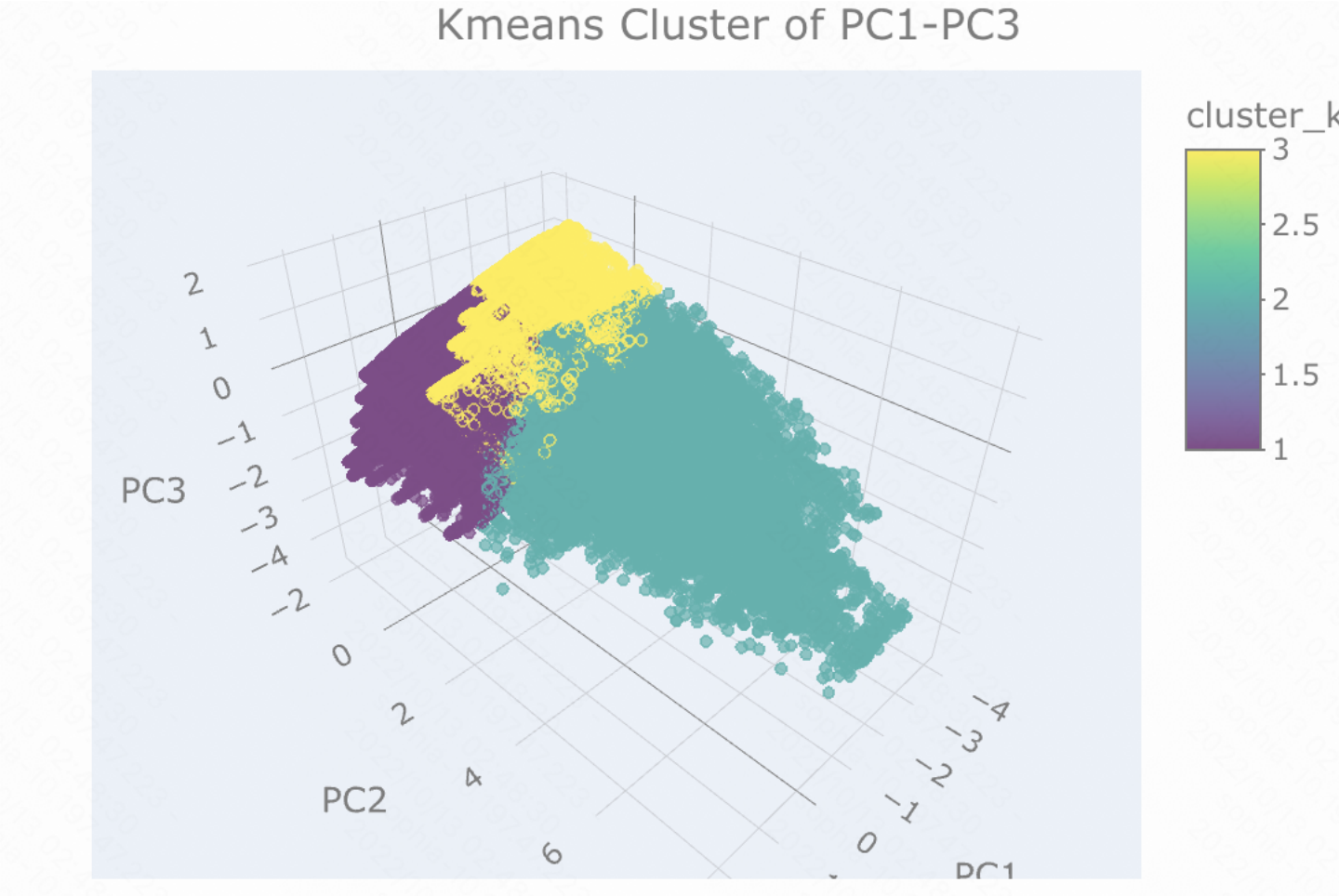

Figure 11: Kmeans

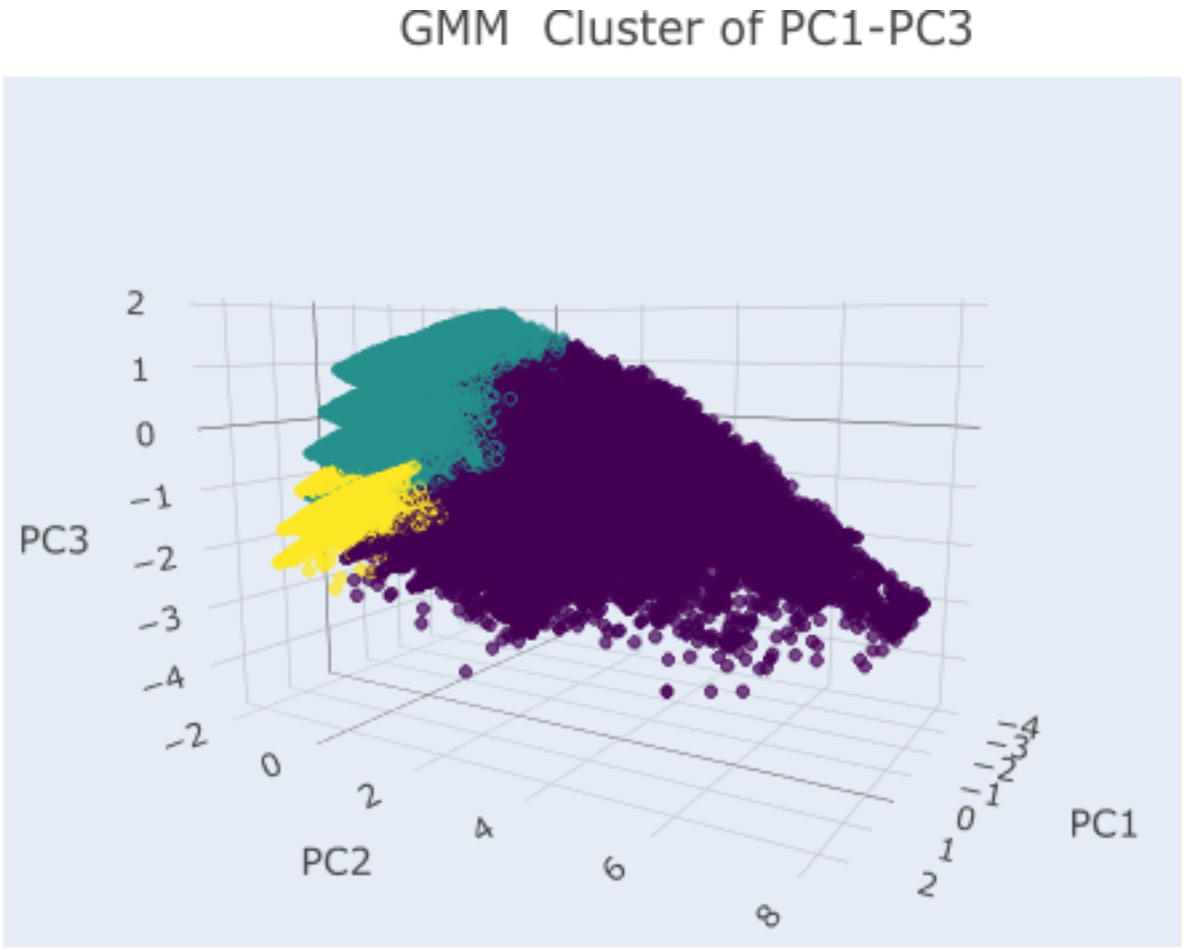

Figure 12: GMM