THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

ESTIMATING CROWD SIZE FROM AN ANALYSIS OF MOBILE DEVICE ACTIVITY

Jordi Estivill-Dredge

42875187

The School of Information Technology and Electrical Engineering

The University Of Queensland

2015

Thesis Supervisor:

Dr. Mark Schulz

Jordi Estivill-Dredge
42875187
12A Birdwood Road
Holland Park West
QLD 4121

09/11/2015
Prof Paul Strooper
Head of School
School of Information Technology and Electrical Engineering
The University of Queensland
St Lucia QLD 4072

Dear Professor Strooper,

In accordance with the requirement of the Degree of Bachelor of Engineering in the School of Information Technology and Electrical Engineering, I submit the following thesis entitled"Estimating Crowd Size From An Analysis Of Mobile Device Activity".

The thesis was performed under the supervisor of Dr Mark Schulz. I declare that the work submitted in thesis is my own, except as acknowledged in the text and footnotes, and has not been previously submitted for a degree at The University of Queensland or any other institution.

Yours sincerely

Jordi Estivill-Dredge

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Firstly I would like to express my sincere gratitude and thanks to Dr. Mark Schulz for his continuous support during the preparation of my undergraduate thesis. His guidance helped me to remain focused, on track and achieve my goals.

I would also like to thank my family for everything they have done to support me, from home and abroad. Thank you to my mum, Dianne, for supporting, inspiring and motivating me from afar. Thanks to my dad, Vlad, for helping keep me focused and motivated on achieving great things. Thanks to my sister, Deni, who had to put up with my frustration and helped do all the little things, whether it was bringing a coffee home or helping drive me places to help pick up equipment. Lastly, to the most devoted family member, Millie "the whinger" Dredge, I have to thank you for everything you did. So thank you, little furry one, who didn't mind staying up late to keep me company (as long as there was a treat involved) and always being so happy and enthusiastic to see me, even when things were extremely stressful.

There are a few others I also want to thank for their assistance and support. I would like to thank Dr. Matthew D'Souza, who allowed me to borrow a variety of mobile devices to experiment with. Additionally, I would also like to thank Christine Allsop and the team at Top Nosh Café Aspley for allowing me to use their café to undertake the data gathering stage of this thesis.

ABSTRACT

The analytics of crowd size and movement is an important area of research as managers of public and private spaces seek to manage crowds, and improve the design features, effectiveness and efficiency of space. Having a method to estimate crowd size can be useful to a number of industries such as the retail sector, museums, and public transport. This research investigated whether mobile device activity was a suitable technology to estimate crowd size.

After a review of the available literature, Wi-Fi was selected as the technology to be investigated for its capacity to estimate crowd size. Bluetooth was considerably restricted, including the number of devices per scan and short range.Similarly, cameras do not work well in varying weather and lighting conditions.

A system was developed using Wi-Fi to gather data on the activity of mobile devices. This system used the open source software Kismet running on a Raspberry Pi. It transmitted the data to a laptop, which logged and graphed the information. The data was grouped into experiments and analysed. The data mining and machine learning software Weka was used on the data. This research used decision trees because of their usefulness to identify relationships between variables.Some models were extremely accurate and exceeded competing technologies, whilst others were considered unfeasible as a result of their impracticalities. Consequently, the best model in terms of practicality and accuracy was using RSSI without MAC address filtering.

Some opportunities for future research were also identified. These included better space selection, system design improvements and alternative analysis approaches. Improvements in these areas would significantly increase the practicality and accuracy of models in future development.

In conclusion, Wi-Fi was capable of estimating crowd size through the use of mobile device activity. Compared to other technologies presented in the literature, Wi-Fi was also capable of matching their accuracy. Despite the evidence presented in the current research that suggests estimating crowd size using Wi-Fi is feasible, further research is required.

1.0 INTRODUCTION

1.1 VISION FOR THE FUTURE

Analytics of crowd size is used to identify and describe people in a space, and can be used to understand, for example, queue times and dwell times. These analytics highlight trends in how people use a space over time, although they do not always explain why such conditions exist. Understanding how people interact within their environment can allow for better space optimization. For example, airport managers can identify busy periods and arrange queuing and crowd control mechanisms to better accommodate flows at peak times, retailersmay arrange better staffing to manage crowds during sales, and public transport planners could improve the services offered or the location of entrances, exists and passenger platforms if they better understood the movement of passengers.

The current solutions usedfor estimating crowd size typically require active engagement of the user (e.g. surveys) or are intrusive and detract from the user's experience of the service or space. On one hand, only a minority of users responds to surveys or are willing to provide verbal or digital feedback (e.g. online review) to an operator. Active participation can also be seen as frustrating for users since they do not usually see the impact of their feedback. On the other hand, passive tools such as floor mats or lasers involve expensive infrastructure setupand maintenance compared to active technologies.

An ideal solution to estimating crowd sizewould be accurate, non-invasive and cheap. If a business can receive accurate, real time information on how a crowd is using a space then they will be able to identify trends and provide a better service, or design a more functional space.

1.2 PROJECT DESCRIPTION

A specific technology was chosen for this project that would allow crowd size to be estimated, the technology was then assessed through a series of experiments to determine whether the technology was viable in terms of it being accurate and practical.The specific technology was chosen after reviewing the existing literature and evaluating the strengths and weaknesses of existing technologies. An investigation was then designed and undertaken into whether this technology was viable for estimating crowd size. Multiple experiments were undertaken to collect the required data, which wasthen analysed and conclusions were drawn from it. The conclusions also make recommendations about whether the technology, which was initially chosen as a method to estimate crowd size, is viable compared to competing technologies.

1.3 Project Aim

The objective of this research was to evaluate whether mobile device activity, via a selected technology, could be used to estimate crowd size.Ideally, the solution investigated in this research should provide a system for operatorsor those responsible for the management of spacesto better estimate crowd size.This required two major deliverables. Firstly, a system that was able to gather statistics on crowd mobile device activity was required. Secondly, a series of experiments to gathersomeinitial data was required in order to assess whether the technology developed offered a viablealternative compared to currently available technologies.

1.4 Report Overview

Six major sections describe the method used for the research. Firstly, the Introduction outlines the project need, description and aims. This section provides the context for the project andexplains its relevance in real life scenarios. Secondly, the System Design section includes an overview of existing research and details competing technologies for estimating crowd size. Drawing from this research, one particular technology was chosen and the technical design for the current research experiment is explained. The System Design section also includes an explanation of the experimental setup used to gather data. Thirdly, the Results and Analysis section details the experimental process. This includes how and why the data was gathered and analysed, including a discussion of suitable statistical procedures. Further, it presents the outcomes of the experiments, discussing their significance. Fourthly, the Review section presents a critique of the advantages of limitations of this thesis project. A practical and statistical evaluation is undertaken, which transitions into the fifthsection, Future Directions. The Future Direction section describes what has been identified in this thesis as requiring further investigation or improvement. Lastly, the sixth section, Conclusion, provides a summary of the outcomes of this thesis.

2.0 SYSTEM DESIGN

2.1 ESTIMATING CROWD SIZE

Data on crowdsizeprovides significant information for both operators and users. Allowing an operator the ability to estimate crowd size allows operators the opportunity to provide a better service for the user, it improves efficiencies and lower costs for the operator, and improves the planning of responses to crowd disasters[1]. However crowd estimation can turn into a controversial topic when self-motivated groups with political interests are involved [2]. For example, whether an organisationis supportive of the reason for a crowd (e.g. political protest or a retail sale event) may result in over or under estimation of crowd size[2]. This thesis is not interested in the political motivations behind these scenarios and will assume that the accuracy of crowd estimation is significantly more important than the political or social connotations associated with a larger or smaller crowd size.

2.1.2 PROXIMITY OR LOCATION

Proximity and localization are two methodsused to analyse how a crowd is using a space. However, "proximity and localization of activities and people are two notions that are often mistaken for one another" [3]. On one hand, proximity is simply sensing the presence of an individual. On the other hand, localisation involves identifying the coordinates (in either 2D or 3D space), to allowa systemto preciselyidentify individuals' location compared to proximity. However, despite their differences there remains a semantic wealth of knowledge in the notion of proximity[3]. Further, [4] also suggest that "in many cases, the concept of location can be replaced by that of proximity". Consequently, since indoor localisation has resolutions of a few metres [5, 6] it is difficult to develop a system that is both scalable and accurate.

While the ability to localise individuals can be important, operators tend to require only the number of people present. The number of people in a public transport system or in a shop is sufficient data to satisfy the needs of the majority of operators. As previously stated, localisation of individuals is substantially more difficult, requires a higher start-up cost and is not easily scalable.

2.1.2 Active Approaches

2.1.2.1 Manual Counting

Before the rapid expansion of information technology, person crowd size data was only able to be collected using other humans [7]. Manual counting is expensive, human error is a problem and there are constraints with respect to observable area and time. Further, data such as repeat visitors and length of stay are difficult to gather. [8]found that humans under count by between 8-25%. They noted that the complexity of the scene, the flow rate, width and number of queues could all affect how accurate a count was. However, they also noted that manual observation does not require much planning and is therefore a fast means of obtaining data.

2.1.2.2 Physical Barriers

Person counting has expanded to employ a number of physical interaction barriers such as laser beams, turnstiles and floor mats. [7] investigated the use of switching mats and light barriers in an airport in Vienna, Austria. They used the mat and light barrier to estimate the number of people moving through the airport security line. Their analysisfound that the switching mat over estimated by 12.5%, and the light barrier by 12.6%. The data was more accurate after using a calibration factor, however there was still large error even at low counts. Further, the costs of the laser/light barriers were between US$50 and US$100, while a 75cm by 1m mat costUS$1500.

2.1.2.3 Camera

By adding a camera to a laser beam, [9] also investigated people entering and exiting through a controlled area. Using a cheap video camera and a light barrier, they managed to accurately detect 90% of individuals entering and exiting a space. Similar to [8], accurate data was collected for10-12.6% of events.

Using a video camera is another common method for counting people. This technique is often used to count traffic [10-13]. This is demonstrated by[10], where a satellite camera was used to look down from above to count cars. By identifying individual cars, the researchers were able to gather data about traffic flow. This data did not specifically identify individuals but was still useful to gauge how roads were being used. The system used by [10] managed to correctly identify vehicles 87% of the time, although vehicles were completely identified (i.e. all vehicles correctly identified) only 60% of the time. This is consistent with the findings by[13]. They also used a camera to detect the flow of traffic with an error of 11-16%. [14]developed a blob counting system, whereby image recognition techniques identified people as blobs in order to be counted. When the area had greater

crowd density the data became inaccurate, as the system was not able to distinguish one person from another. An issuefor[14]was that only a single camera was used, which resulted in blind spots. These authors concluded that a "better camera" and "motion tracking" could overcome these issues, although this would add increasing complexity.

2.1.2.4 RFID

Radio frequency tags (also known as RFID) have also been used as a solution to monitor people indoors. Active tags are very cost effective [15] and can be "read despite extreme environmental factors, such as snow or fog" [16]. They use an inbuilt battery that can last 3 to 5 years and allow an RFID scanner to detect 500 tags up to 45 metres away in 7.5 seconds [16]. However, passive tags have a very limited range of approximately a metre and require numerous scanners [15]. Both active and passive tags are application specific and would require users to carry the tags (e.g. in a sports marathon) for data to be gathered.

2.1.3 Passive Approaches

Considering that 90% of adults have a mobile phone, counting a crowd using mobiles appears reasonable [17]. As a result, it seems safe to assume that for every mobile device found, one individual is present. However, this is not the case. Laptops and tablets in addition to mobile phones may broadcast their Bluetooth, Wi-Fi or other wireless signals. Further, a phone is only in arm's reach an average of 58% of the time[18]. Therefore, the assumption that one device equals one person does not always hold. In spite of this, there are still methods to count people using their mobile devices. This typically involves a mathematical model to convert device number to person count. Usually this is dependent on location, the demographic characteristics of the crowd and the system used.

2.1.3.1 Bluetooth

A simple approach forpassiveestimation of crowd size is to measure the number of visible Bluetooth devices. Bluetooth is already being used to estimate crowd density at football stadiums, in retail and at public festivals. However, most mobiles have their Bluetooth status set to undiscoverable and only about 2-7% of mobile phones are visible by Bluetooth [19, 20]. This is still useful however, since the correlation of Bluetooth devices to people is fairly high, R^2=0.89 [19]. Bluetooth has limits of 33 devices per scan, 11 seconds per scan and a range of 10 meters [19-21]. Bluetooth is useful for identifying individual devices because they broadcast their MAC address. This allows a system to identify repeat visitors, or measure the time a person takes between two

points. A three day experiment at Oktoberfest in Munich found an 80% accuracy using Bluetooth [22]. They did notice that using Bluetooth they "had to consider cultural factors" and the model of device to person "may significantly vary depending on who the persons in the crowd are [22]."

2.1.3.2 Wi-Fi

Wi-Fi is the other main wireless system on most mobile phones. Counting the number of people using Wi-Fi has been previously investigated [15, 23, 24]. Over 90% of mobiles are Wi-Fi enabled, making it the most used mode of wireless communication [25]. [23] described the benefits of using mobile devices rather than images and cameras. Cameras "cannot work well in a dim or dark environment", and RFID or tags "require every human being to carry the device, and are inaccurate and unreliable in practice". Additionally, [24] used a system to measure the strength of the Wi-Fi signal from the front of a queue. Judging from the increase in signal strength over time, they were able to see how quickly a queue moved.

2.1.3 Summary

Based on the reviewed literature, a Wi-Fi based passive system that is cost effective, accurate and scalable was chosen for this research. The two main choices are to detect the proximity of people (not localise) and to use Wi-Fi (opposed to competing technologies).

Whether it be to monitor the state of a public transport system, event planning or in retail spaces, knowing the number of people present can improve the efficiency and effectiveness of service and improve the user experience without including users specific location. Systems using Wi-Fi [26, 27] and Google Maps Indoor [28] have already been used to localise people. However these techniques tend to have poor resolution and are only accurate to within a few square meters. Consequently, person count using proximity is a better practical choice for two reasons. Firstly, proximity is significantly easier to setup. It requires only a single device, rather than numerous devices for localization. Secondly, sensing the proximity of individuals is able to gather suitable data on how a crowd uses a space.

The choice to use Wi-Fi was the other major decision that was made based on the above review of literature. The other contender was Bluetooth. Considering the limits of Bluetooth, Wi-Fi was chosen for a number of reasons. Bluetooth has less range, has a smaller limit of devices per scan and has a smaller percentage of discoverable users. Further, the other common passive technology,the video camera, has lighting and blind spot limitations [9, 10, 12-14]. Further, physical

interaction from users on floor mats or through lasers has demonstratedabout the same accuracy as wireless devices [7-9].

The choice was made to develop a Wi-Fi proximity person counter. The accuracy of the systemwastrialled and tested in a number of experiments to determine whether it is a suitable solution.

2.2 Wi-Fi Approach

Wi-Fi is defined as a wireless local area network (WLAN) that complies with the Institute of Electrical and Electronics Engineers (IEEE) 802.11 standards[29]. The 802.11 standards relate to the media access control (MAC) and physical layers (PHY) of the network. Wi-Fi operates on either the 2.4GHz or 5GHz SHF ISM radio bands.

To gather data, the Wi-Fi packets being broadcast in the local vicinity had to be intercepted. Packet sniffing is the process of capturing Wi-Fi transmissions sent over networks and analysing the information contained within [30]. This process is not intrusive in the sense that a computer is being attacked. Conversely, packet sniffing is a passive technique, eavesdropping on the communication that is taking place. This is made possible because the Wi-Fi packets themselves are broadcast to every device, with the destination contained within a header [30]. Subsequently, every nearby device and on the network receives allmessages and evaluates whether it is the intended recipientor whether the message should be discarded. The system designed in this research chose not to reject packets and instead, analysed the publically available information.

Three key hypotheses were made in regard to using Wi-Fi as the technology to estimate crowd size. Firstly, the signal strength of the device could be used to estimate the distance of the device from the scanner. This would allow the size of a space to be defined. Secondly, the number of devices should be larger than the number of people. This is because an individual may have more than one device (e.g. laptop and phone) in addition to static Wi-Fi infrastructure (e.g. routers). Thirdly, the Organisational Unique Identifier (OUI) of a device's MAC address should be related to the device brand. For instance, an Apple iPhone should have an Apple OUI. However, this may not always be the case, since some phone manufacturers may purchase Wi-Fi cards from other companies.

2.3 System Architecture

2.3.1 Scanner

To analyse the packets being broadcast in a space, the packet sniffing software Kismet[31]was chosen. Kismet works with any wireless card or dongle that supports raw monitoring mode.

Kismet was able to detect the presence of both wireless access points and the clients.However, Kismet itself was not sophisticated enough to share the packets it intercepted with anexternal database or logging system. Consequently, a separate program, was required to fetch the data from Kismet. A Python program was created to extract, save and transmit the data (Figure 1).

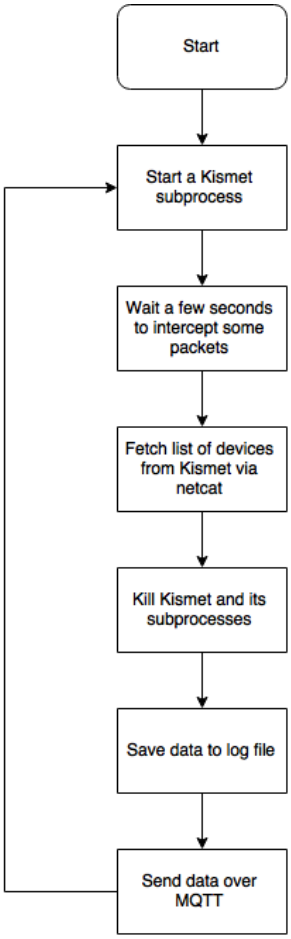


Figure 1: Wi-Fi Scanner Software Flowchart

The software choices had a strong influence on what hardware could be used. A portable device capable of intercepting Wi-Fi packets and running Kismet was required. A Raspberry Pi was chosen for two reasons. Firstly, it is a physically small microcomputer, which compared to desktops or laptops, provides significant portability advantages. The increased portability means that if numerous spaces were to be analysed (e.g. numerous exhibits in a museum), it could be undertaken at a cheaper cost to the operator. Secondly, the Raspberry Pi used the Raspian operating system,whichpermittedmanipulationof Wi-Fi hardware (such as the raw monitoring mode required by Kismet). Network interface manipulation was not possible on some operating systems, for example, Mac OSX Mavericks (10.9) and onwards [32]. The Element 14 Wi-Pi Wi-Fi dongle was used in conjunction with the Raspberry Pi, which supported the raw monitoring mode required by

Kismet.Lastly, since Kismet manipulates the wireless interface, Ethernet was required for shell and Internet access to the Raspberry Pi.

A protocol was required to publish and record the data for analysis, which resulted in the selection of MQ Telemetry Transport (MQTT). The Centre for Educational Innovation and Technology (CEIT) already operates an MQTT server (codenamed Winter). MQTT was chosen for two reasons. Firstly, CEIT already had the existing MQTT infrastructure in place, and with the inclusion of the Paho MQTT Python client library [33], provided a simple method to share the data. Secondly, MQTT allows plain text transmission, not restricting the type of data that can be sent. This allowed the raw data from Kismet to be converted to JavaScript Object Notation (JSON) forpublishing.

The finalised system incorporated the above decisions to create a small, portable microcomputer capable of sensing the proximity of nearby Wi-Fi devices. Figure 2 illustrates the hardware design of the Wi-Fi proximity sensor.

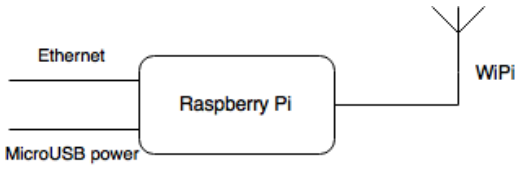


Figure 2: Wi-Fi Scanner Hardware Diagram

The Wi-Fi scanner had a number of software options, allowing future users to customise the software to their situation. These options could facilitate less power and network usage if required. Options available for the scanning software included:

- The MQTT server.
- The MQTT topic.
- Time to listen for devices before publishing to MQTT.
- Time to wait between each scan.
- Name of output files.

2.3.2 Receiver

Once the scanner had been implemented, a separate system was designed to facilitate logging and visualising the data in preparation for analysis. Decisions for the receiver were primarily for the software, since the tasks were not hardware dependent. The receiver software wascompleted using a Python program (Figure 3), which received data from MQTT.The Python script had the option

to set a number of filters on the incoming and existing data. For example, the signal strength (RSSI) threshold and the timeout (number of seconds since a device is last seen before it is removed from the count).

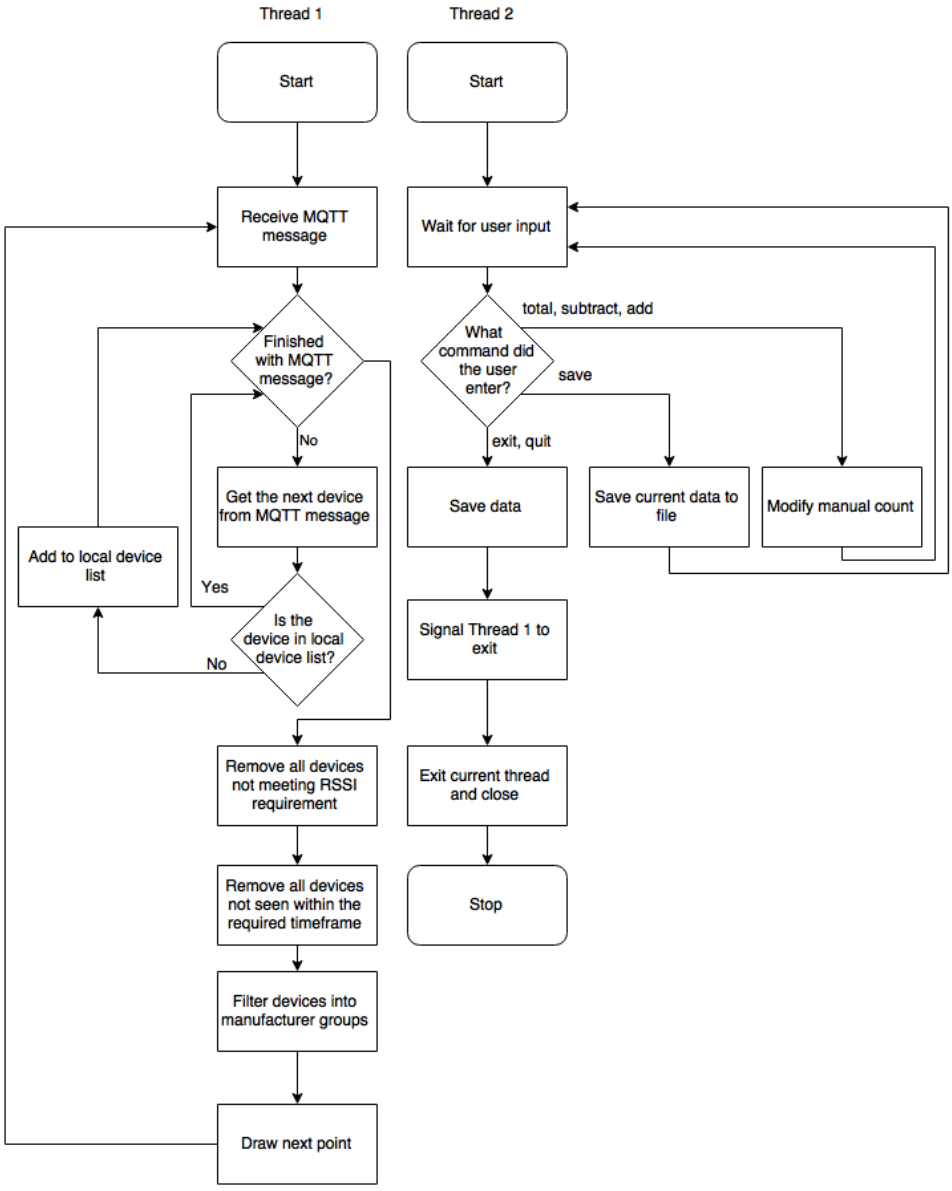


Figure 3: Receiver Software Flowchart

Logs of all the data were stored in comma separated value (csv) files (Figure 4) since they are simple and compatible with a number of applications. Using aprogram specific file format (e.g. Excel) would slow down the logging software and would restrict the potential for others to analyse the same dataset that this research used. In addition to csv files, extensible markup language (xml) files (Figure 5) were created to enable a way to store the multiple levels of data that are created. For example, there are numerous scans, in each scan there are a numerous devices and each device has

a MAC address and a signal strength value. New data was appended to the end of the xml file during runtime, while the csv file was generated in its entirety at the program's exit. This was because it was significantly easier to manipulate rows of a file compared to columns.

```
Timeout,RSSI,Manual_Count,Total_Devs
30,90,0,0
30,85,9,30
120,60,10,3
75,75,11,20
10,80,8,19
```

Figure 4: CSV File Excerpt

```
<root id="1">
 <item>
   <rssi>-61</rssi>
   <MAC>64:D9:89:43:5E:70</MAC>
   <man>CISCO SYSTEMS, INC.</man>
   <pi_id>1</pi_id>
   <time>12:48:03_October_21_2015</time>
 </item>
<item>
   <rssi>-69</rssi>
   <MAC>68:64:4B:59:60:97</MAC>
   <man>Apple</man>
   <pi_id>1</pi_id>
   <time>12:48:03_October_21_2015</time>
 </item>
<item>
   <rssi>-75</rssi>
   <MAC>84:38:38:0A:4C:86</MAC>
   <man>Samsung Electro Mechanics co., LTD.</man>
   <pi_id>1</pi_id>
   <time>12:48:03_October_21_2015</time>
 </item>
</root>
```

Figure 5: XML File Excerpt

Visualising the data was performed using the same Python program running matplotlib, a 2D plotting library. Matplotlib was chosen as the graphing library because of its integration capabilities with the rest of the software. The graphing software would continuously increase the x-axis to accommodate the incoming data. An example screenshot of the software in operation can be seen in Figure 6.

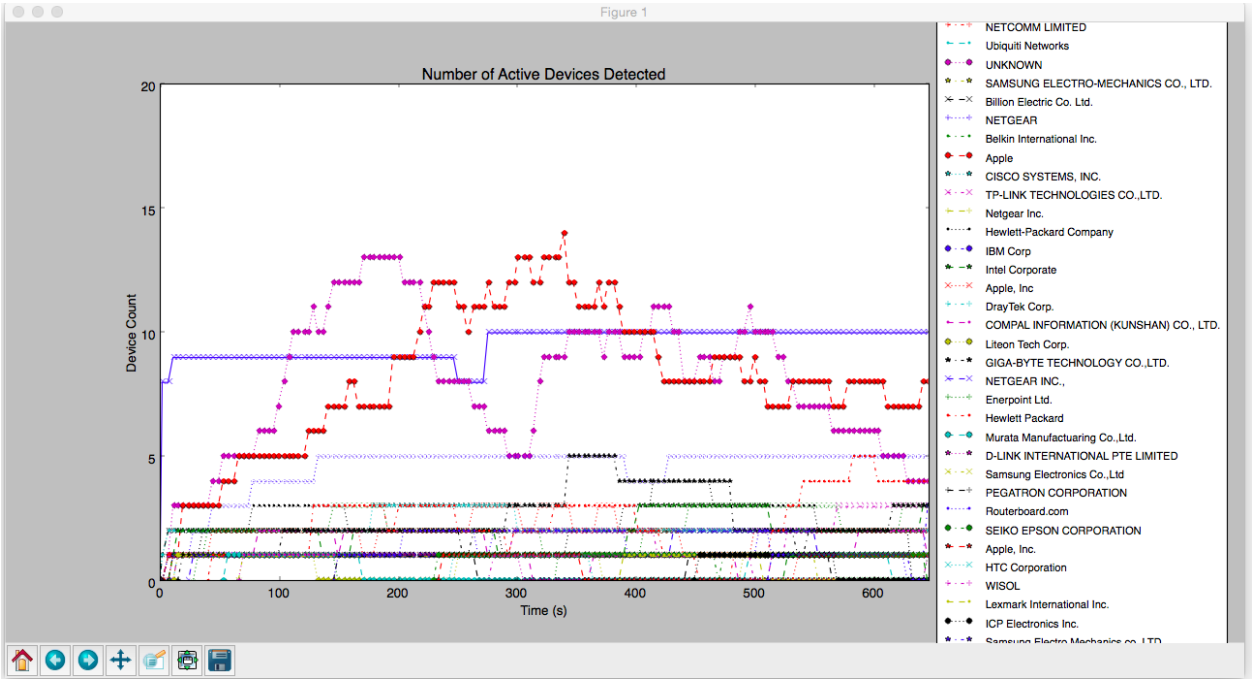


Figure 6: Screenshot of Graphing Software

The software developed for the receiving system had a number of options available to allow future users to gather data relevant to their situation. Additionally, these options were used to analyse the viability of Wi-Fi for estimating crowds. Available options include:

- The MQTT server.
- The MQTT topic.
- Whether to graph the data or not.
- The RSSI threshold.
- How long a device will be remembered before being removed from the count.
- The output filenames.
- The initial number of people.

Hardware decisions for the receiver were not too significant, since the only requirement was compatibility with the Python program and libraries. Consequently, a 2011 MacBook Air running Mac OSX Yosemite (10.10) was used as the receiver.

2.3.4 Experimental Setup

This research involved a number of trials and experiments to determine whether Wi-Fi traffic is a suitable estimator for crowd size. On one hand, a trial is defined as an execution of the system with a number of system parameters remaining constant, such as the RSSI threshold and the timeout.

On the other hand, an experiment is defined as a collection of trials, with differing parameters between them, to investigate whether the parameter is an effective estimator.

The system developed in this research should be suitable for use in a number of environments, for example, classrooms, retail shops and museums. To maintain consistency across trials, Top Nosh Café at Aspley, Brisbane (Figure 7) was chosen as the experiment location. Top Nosh is part a small complex of other shops, which meant that some Wi-Fi traffic from neighbouring businesses was intercepted.The experiments were always undertaken at the same table (red table in Figure 7) within the café. Staff working at the register and kitchen (green area) were not included in the count because only the customer count was important. All customers (except those believed to be younger than 12 years), including those in the café and the gift shop area (blue) were included in the manual count. The crowd was not surveyed because this research aimed to collect the data passively, without active interaction from the crowd. This meant that demographic data was not collected from this research.

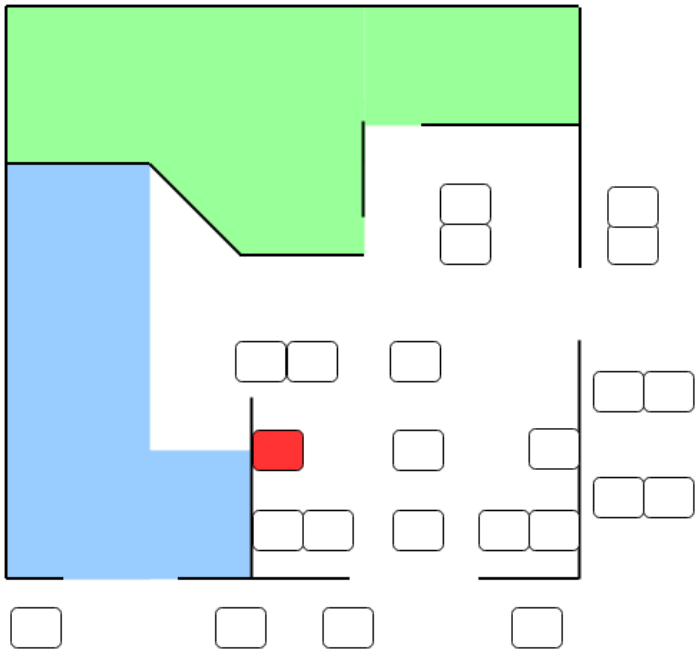


Figure 7: Floor Map of Top Nosh Café

Once the scanning device was setup and operating, the trials were run using a variety of different parameters. The two variables that were modified between trials included the RSSI threshold and the timeout. A variety of combinations of these two variables were recorded, with each trial being run for approximately five minutes.

3.0 RESULTS AND ANALYSIS

To evaluate whether mobile device activity is a reliable estimator for the size of a crowd, several trials were performed and grouped into experiments. The data gathered from those experiments were then analysed and evaluated in comparison to other technologies. One similarity across all experiments was the changing timeout variable. The timeout variable was different between trials and did not remain constant for any experiment. Timeout was intended to vary across trials since there was no reason to support a particular value to be effective across all mobile devices. Further, the variable could not be removed from an experiment because it is impractical to measure the number of mobile devices at an instant.

To investigate the accuracy of mobile device activity as an estimator for crowd size, decision trees were selected as the primary analysis approach. Decision trees are a simple and effective technique for predicting, investigating and explaining relationships between variables [34]. A decision tree consists of nodes that all link together to form a rooted tree. The root node has no incoming edges and is the starting point of the tree. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as a test node, while all other nodes (with no outgoing edges) are called leaves. Decision trees were chosen as the primary analysis tool because they are self-explanatory, easy to follow and nonparametric (no assumptions are maderegarding the relationships of variables) [35]. This meant that a decision tree could be implemented for a business fairly easily.

The primary decision tree algorithm chosen was M5P for two reasons in regard to its practicality to implement and analyse. Firstly, the M5P algorithm generates multiple linear regression models at the tree's leaves[36]. Consequently, once at a leaf node of the tree, a linear function can be used to estimate the crowd size. Since it could be more accurate, it was significantly more desirable than a single figure at the leaf node. Secondly, the M5P algorithm can handle missing variables[36]. This is critical, because it meant that while the experiment parameters were different between trials, it was not crucial to cover every possibility. The data mining software Weka [37] was used to execute the decision tree algorithm. Weka is a collection of machine learning and data mining algorithms implemented in Java that allows pre-processing, classification and regression. The Weka software included the M5P decision tree algorithm. This research used the algorithm with cross validation to give an indication of how well the model would perform when asked to make new predictions for data it has not already seen [38]. Using only some of the data for the training and the rest to evaluate the performance,gave an indication of how the model would perform in the future.

3.1 Device MAC Address and RSSI Investigation

To verify the impact of RSSI and MAC address, a preliminary investigation was undertaken to evaluate the impactof RSSI and MAC address filtering. Firstly, it was important to verify what RSSI measurements implied. For example, when a device was detected with an RSSI of 67, could the distance between the sensor and the device be estimated? This meant that an operator could set a RSSI filter level suitable for the space they wished to examine. Secondly, the MAC addresses of a number of known devices were recorded and their OUI was obtained from the IEEE register. This was to verify whether or not the OUI mightmatch the brand of phone. For example, a phone manufacturer mightpurchase Wi-Fi and Bluetooth chips from one company, a CPU from another and so on, limiting the ability to obtain the mobile device's actual manufacturer.

3.1.1 Signal Strength Analysis

To measure the impact of RSSI filtering, a number of different devices, from different manufacturers were placed at specific distances away from the scanner and their RSSI was recorded. The devices used were an Apple iPhone 6, Apple iPad 3, Samsung Pocket Neo and Asus PadPhone. During the testing, the devices played videos from YouTube to ensure they were actively using Wi-Fi. The measurements, outlined in Table 1 and graphed in Figure 8, had lines of best fit added to analyse the relationship.

Table 1: RSSI Measurements Over Distances

Distance (cm)	**RSSI Reading (dBm)**			
	Apple iPhone 6	Asus PadPhone	Samsung Pocket Neo	Apple iPad
10	31	27	33	23
50	35	41	45	33
100	43	51	57	37
150	47	55	57	41
200	51	61	59	47
250	63	63	61	51
500	67	65	65	61

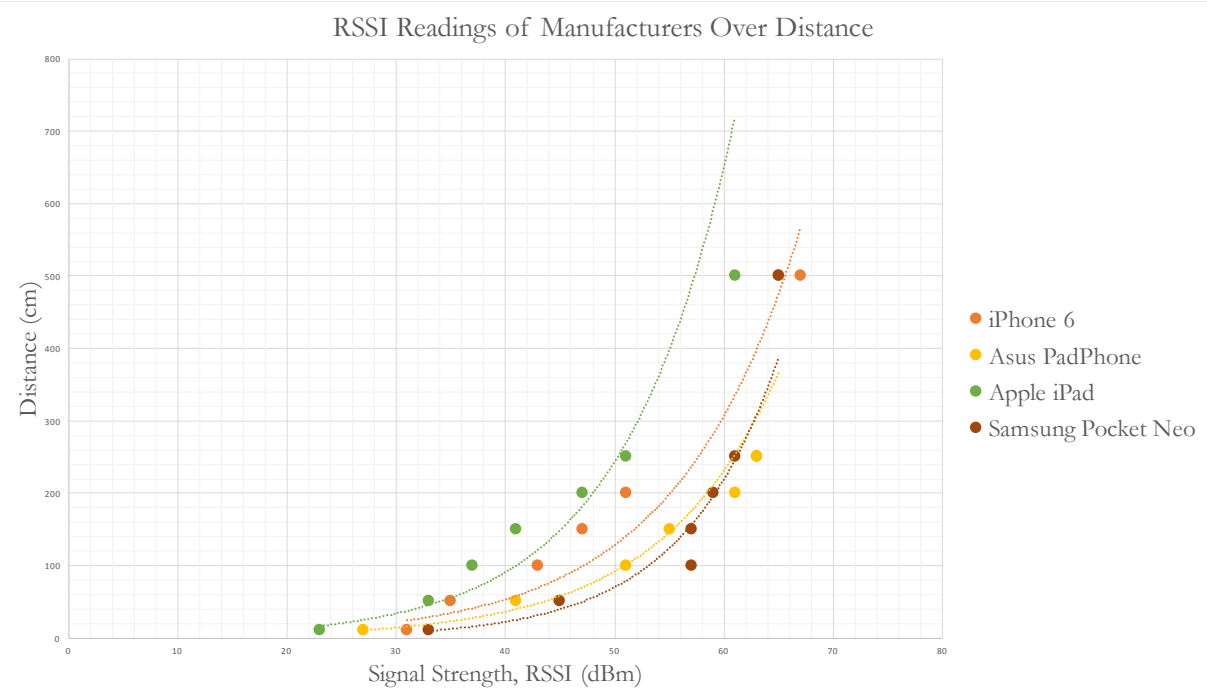


Figure 8: RSSI Readings of Manufacturers Over Distance

Since an unknown number of devices would be encountered, the above RSSI readings were averaged(Table 2 and Figure 9). The aim of this step was to develop a model that would work across a variety of manufacturers, not just the manufacturersshown in Figure 8.

Table 2: Averaged RSSI Measurements

Distance (cm)	**Average RSSI Reading (dBm)**
10	29
50	39
100	47
150	50
200	55
250	60
500	65

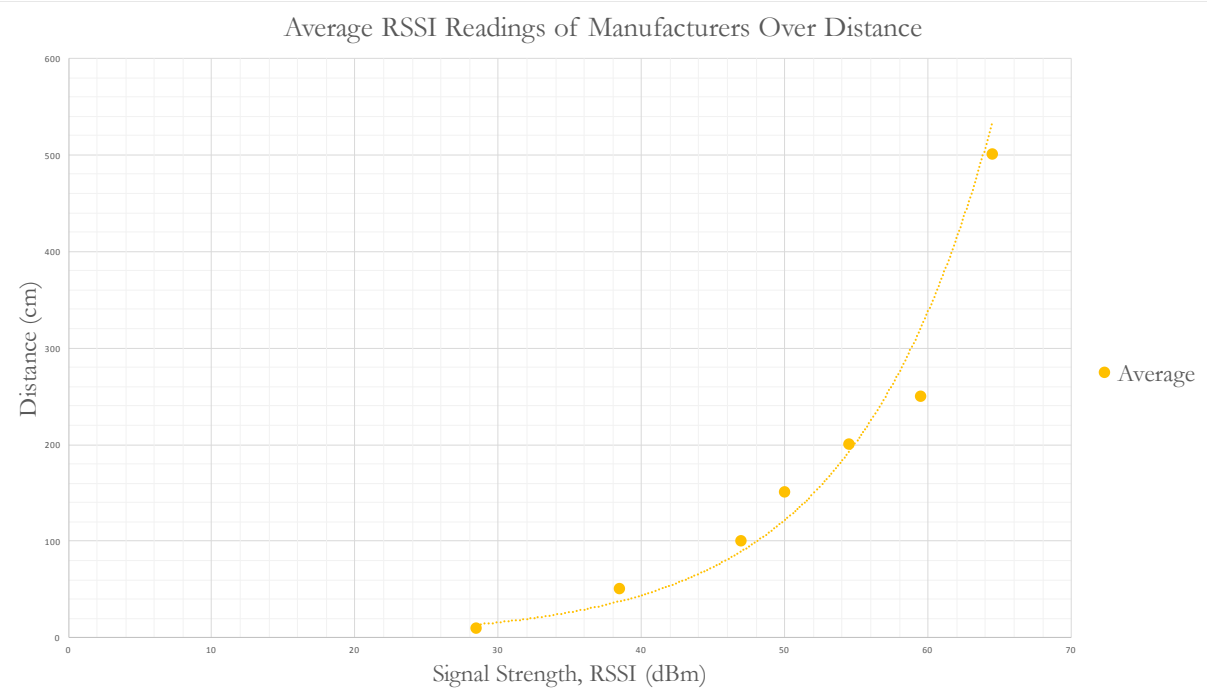


Figure 9: Average RSSI Readings of Manufacturers Over Distance

The averaged RSSI function had two advantages. Firstly, it had a good fit (R^2=0.9) and secondly, it should work across a larger varietyof devices and manufacturers. The averaged function was:

$$y = 0.75e^{0.1018x}$$

Top Nosh Café had a width and length of approximately five metres either direction of the scanning location. Therefore, an RSSI filter of 65 may have been appropriate. However, the trials undertaken to gather this data had different conditions to the later experiments. This was because the later experiments were practical evaluations of the system, while this experiment was to verify that the metadata was reliable. Consequently, there were two key differences between this experiment and the others. Firstly, the data was gathered in a straight line with no obstacles such as chairs, tables and walls. Each barrier may significantly impact the measured signal strength. Secondly, the data was gathered in an environment with significantly less wireless interference than the later experiments. Top Nosh Café had a large number of neighbouringbusinesses with Wi-Fi networks and other electronic appliances (e.g. microwave), which would have impacted measured signal strength.

3.1.2 MAC Address Filtering

While a certain company might manufacturer or assemble a model of mobile phone, not all components might be from the same company. A number of consumer electronics manufacturers do not manufacturer all the components themselves, instead choosing to purchase components from other businesses. This is particularly the case for mobile device CPUs, GPUs and wireless cards (Bluetooth and Wi-Fi). Consequently, referring a MAC address to the IEEE register and obtaining an OUI may not always be reliable. To investigatethis issue, a number of mobile phones had their Wi-Fi MAC address examined (Table 3).

Table 3: Device Wi-Fi Organisational Unique Identifier

MAC Address	**Device**	**Organizational Unique Identifier**
48:43:7C:63:54:4F	Apple iPhone 6	Apple
30:19:66:55:B6:78	Samsung Pocket Neo	Samsung Electronics Co.,Ltd
CC:78:5F:B3:7B:AE	Apple iPad	Apple
5C:FF:35:77:45:5F	Asus PadPhone	Wistron Corporation

As hypothesised, while some device's OUI matched the brand, this did not always occur. This is the result of components from different companies being used to manufacturer mobile devices. Consequently, it is not appropriate to assume the brand of the phone and the OUI to be similar. A component manufacturer may sell parts to a number of phone brands. For example Wistron Corporation components might be present in numerous different phone brands. This preliminary investigationprovided evidence that a manufacturer breakdown of mobile devices may not be suitable for estimating crowd size. However, a practical application in an experiment of manufacturer breakdown was required to further verify this.

3.2 TOTAL DEVICES

Analysing the total number of mobile devices was the first experiment to investigate the accuracy of Wi-Fi device activity as an estimator of crowd size. This experiment had no restrictions on the data gathered during the trials. Themain purpose of this experiment was to act as a comparison for the following experiments. The variables included in this experiment are listed in Table 4.

Table 4: Total Devices Experiment Variables

Variable Name	Type	Classification
Manual Count	Dependent	Numerical
Total Devices	Independent	Numerical
Timeout	Independent	Numerical

The spread of these variables shown in Figure 10 shows that trials with different parameters were performed. Shorter timeout values (t < 60s) were usually used, but there was still a suitable amount of data for larger values. As hypothesised, the device count's mean (41) was higher than the manual count (15). This was expected since an individual may have more than one device and the areashould have static Wi-Fi infrastructure.

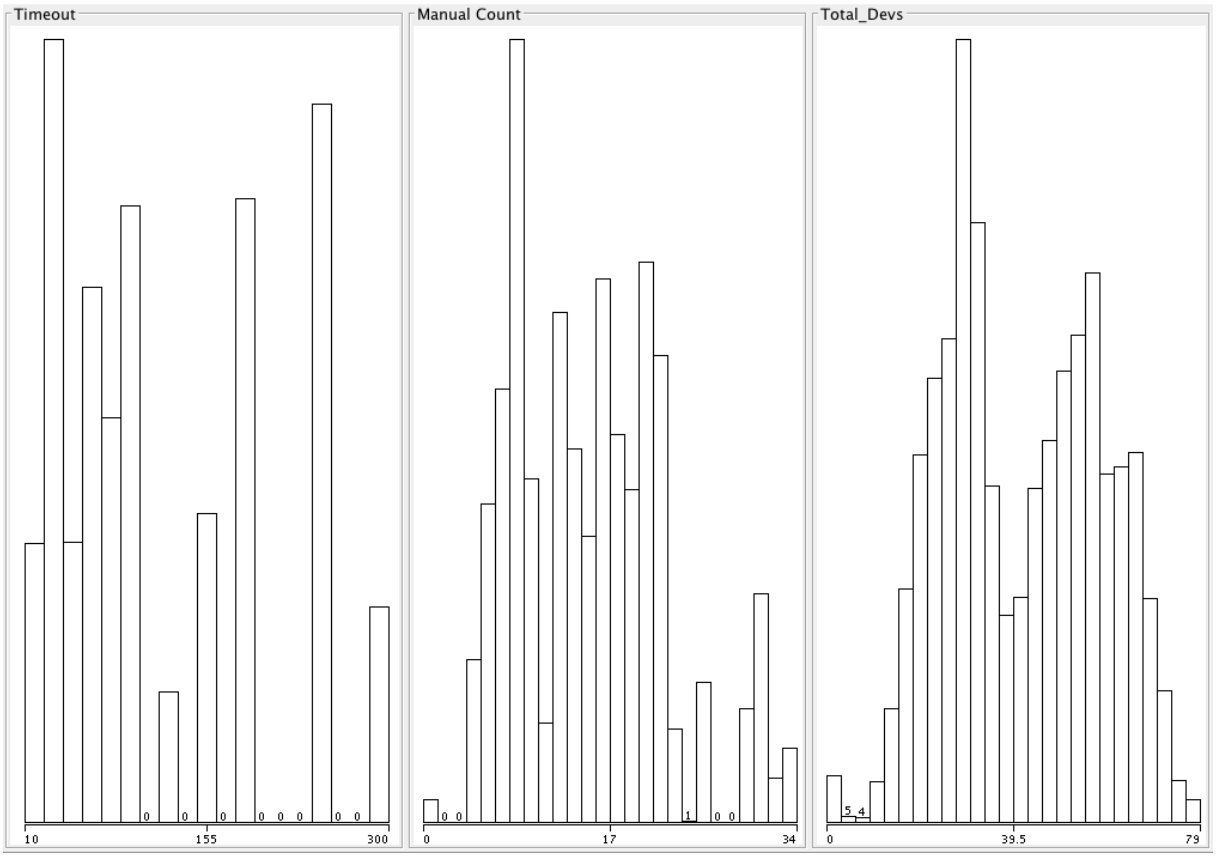


Figure 10: Total Devices Experiment Variable Spread

The Weka M5P decision tree algorithm was run with the data and resulted in the 45 leaf decision tree shown in Figure 11 and summarised in Table 5 (full output can be found in Appendix 2):

Table 5: Total Devices Decision Tree Model

Parameter	**Value**
Correlation coefficient, r	0.7899
Coefficient of determination, R^2	0.6239
Mean absolute error	2.9851
Root mean squared error	4.2683
Relative absolute error	52.6306 %
Root relative squared error	61.3398 %
Total Number of Instances	6505
Number of Leaves	45

The M5P decision tree for this experiment wasnot as accurate compared to other technologies that were reviewed in the literature,but would be practical to implement. On one hand, the coefficient of determination indicated that it was not as accurate as competing technologies. For example, Bluetooth has demonstrated R^2=0.80 and cameras have R^2=0.90. Therefore, looking at only the total number of active Wi-Fi devices was not as accurate as other solutions. On the other hand, the decision tree implemented in Weka had 45 nodes, and would be fairly practical to implement. Since the number of nodes was not too large, it would be practical to deploy this decision tree.

The decision tree did however highlight some unusual scenarios and relationships between the variables. It was expected that there would always be a positive coefficient between device count and the manual count. In other words, as the number of devices increased, so should the number of people that were counted. This however was not the case in the leaves listed in Table 6. This seemed to occur when the device counts were large, but the timeout was moderate. This may be due to Wi-Fi devices from neighbouring businesses becoming active and saturating the count. However, it is still unusual for these leaves to have negative coefficients. Further research may be required to identify the conditions that promote a negative coefficient between the device count and the crowd size.

Leaves with coefficients greater than one also provide insight into a variety of scenarios since it was hypothesised that there should be more devices than people. For instance, individuals may havephones, laptops and tablets, in addition to static Wi-Fi infrastructure such as routers and cash registers. Therefore, it was unanticipated thatthe two leaves in Table 6 had positive coefficients greater than one. Firstly, leaf nine was when less than 5.5 devices were detected, which may have been caused by the crowd not having many Wi-Fi devices, perhaps due to demographic factors.

Additionally, leaf nine could be considered an outlier since only 0.29% of the data fell into this leaf. Secondly, leaf 13 had a coefficient of 1.9749, however it had a negative constant of 35. Further, since the device count was between 24.5 and 27.5, it resulted in a crowd size range of 13 to 19, the M5P regression algorithm may have determined that the linear model, while not consistent with the hypothesis, fit the data in this category more accurately. However, further investigation is required into scenarios that create a coefficient greater than one.

Table 6: Total Device Decision Tree Notable Leaves

Leaf	Note	Timeout (t) values	Device count (x) values
9	Coefficient over one	52.5 < t <= 165	x <= 5.5
13	Coefficient over one	t <= 37.5	24.5 < x < 27.5
16	Negative coefficient	37.5 < t <= 82.5	24.5 < x <= 32.5
17	Negative coefficient	37.5 < t <= 82.5	32.5 < x < 38.5
20	Negative coefficient	t <= 60	38.5 < x <= 40.5
21	Negative coefficient	t <= 60	40.5 < x
22	Negative coefficient	60 < t <= 82.5	38.5 < x <= 42.5
23	Negative coefficient	82.5 < t <= 135	38.5 < x <= 42.5
25	Negative coefficient	82.5 < t <= 135	42.5 < x <= 52.5
26	Negative coefficient	t <= 82.5	52.5 < x
27	Negative coefficient	82.5 < t <= 105	52.5 < x
29	Negative coefficient	135 < t <= 165	38.5 < x <= 43.5
30	Negative coefficient	135 < t <= 165	43.5 < x <= 45.5
31	Negative coefficient	135 < t <= 165	45.5 < x <= 53.5
32	Negative coefficient	135 < t <= 165	53.5< x <= 59.5
39	Negative coefficient	165 < t <= 210	68.5 < x

Ultimately, the use of total device count did not serve as an accurate predictor for crowd size compared to other technologies. When compared to Bluetooth or cameras, a decision tree using the total number of Wi-Fi devices was not as accurate based on the reviewed literature. However, the decision tree did highlight a number of interesting scenarios, when there was a negative or large coefficient. In an attempt to improve the accuracy, the software then restricted the RSSI of devices in the next experiment.

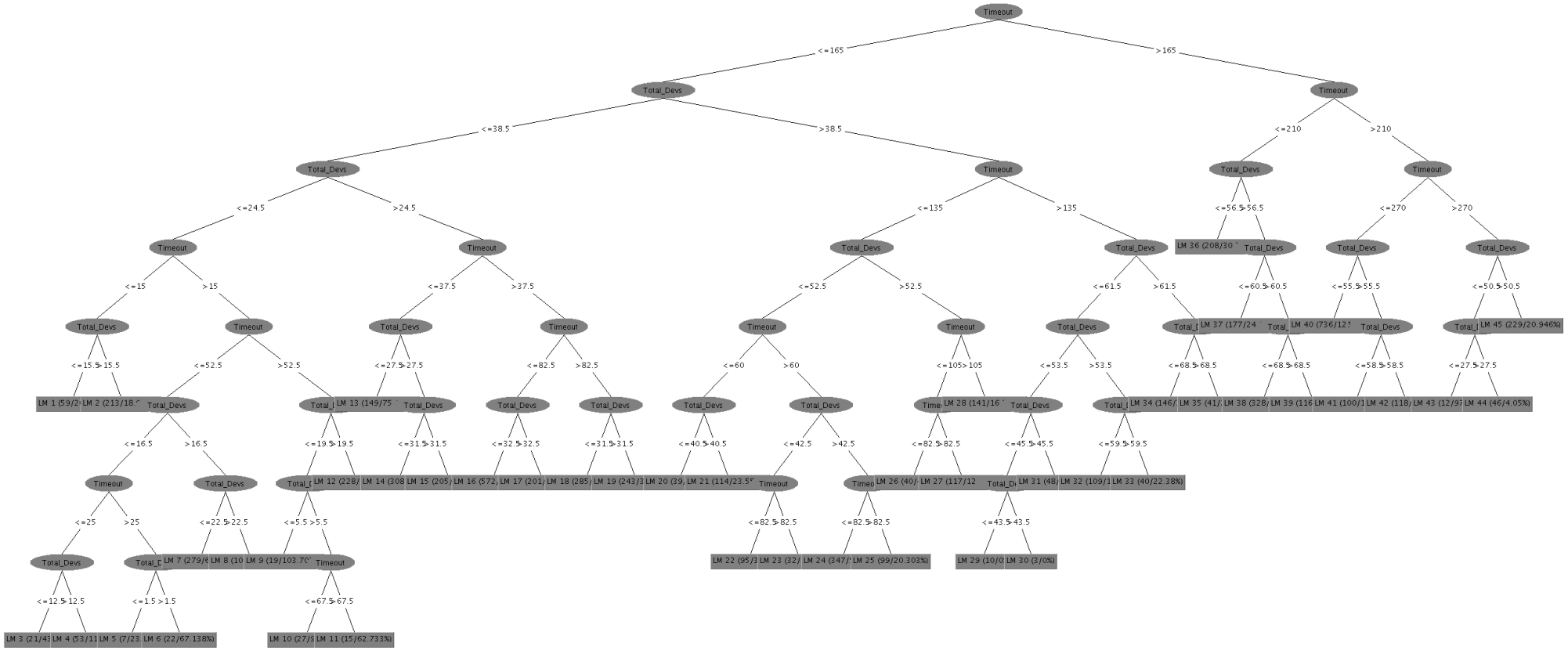


Figure 11: Total Devices Decision Tree

3.3 Total Devices With RSSI Filtering

The total number of mobile devices with RSSI filtering was the second experiment used to investigate the accuracy of Wi-Fi device activity to estimate crowd size. This experiment had only one filter, removing devices that had signals weakerthan a certain threshold value. Since lower RSSI values indicated a stronger signal, a higher threshold is considered less restrictive, while a lower threshold will limit more devices. The main purpose of the second experiment was to measure the impact of only RSSI filtering compared to the previous experiment (no RSSI filtering). The variables included in this experiment were:

Table 7: Total Devices with RSSI Variables

Variable Name	Type	Classification
Manual Count	Dependent	Numerical
Total Devices	Independent	Numerical
Timeout	Independent	Numerical
RSSI	Independent	Numerical

The spread of the variables shown in Figure 12contrasted starkly to the previous experiment. Most notably, the average number of devices had significantly decreased from 41 to 16 while the manual count had remained similar (13). This could be attributed to the RSSI filtering which would reduce the number of devices eligible to be counted. In particular, a number of trials with low RSSI thresholds prevented any devices from being counted at all.

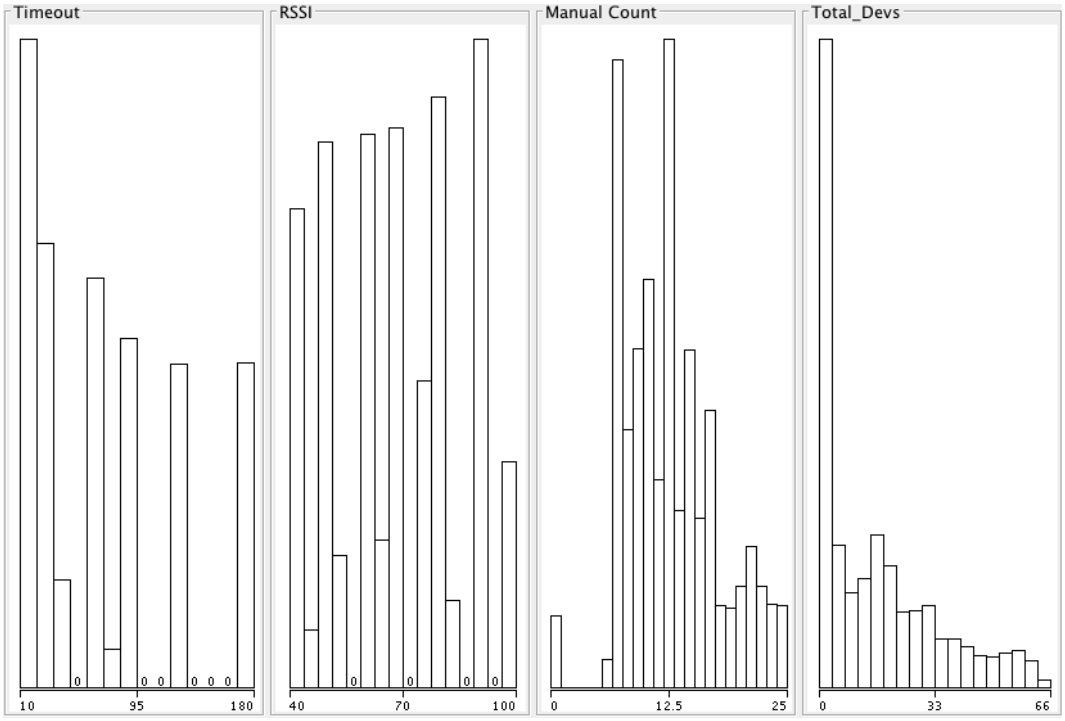


Figure 12: Total Devices with RSSI Variable Spread

The Weka M5P decision tree algorithm was run with the data and resulted in the tree shown in Figure 13 and summarised in Table 8 (full details can be found in Appendix 3):

Table 8: Total Devices With RSSI Decision Tree Model

Parameter	**Value**
Correlation coefficient, r	0.8953
Coefficient of determination, R^2	0.8015
Mean absolute error	1.2259
Root mean squared error	2.2455
Relative absolute error	29.9499%
Root relative squared error	44.864%
Total Number of Instances	4018
Number of Leaves	53

The M5P decision tree for this experiment had significantly improved accuracy at only a slight hindrance to its practicality. Compared to the previous experiment with no RSSI filtering, the RSSI filteredmodel had a 20% increase in explained variance. Compared to the literature, this brought the model to similar accuracy as Bluetooth (R^2=0.80), however it did not reach the accuracy of cameras (R^2=0.90). This improvement came at the cost of eight more leaves on the decision tree. The increased accuracy however, significantly outweighed the minor decline in practicality.

Further, compared to the previous experiment, the RSSI filtering had reduced the number of unusual leaves in the decision tree (Table 9).However, there still remained a few leaves with negative and large coefficients for the device count. The only two leaves with coefficients greater than one, had device counts of less than 3.5 and 1.5. These trials may have occurredin two possible scenarios. Firstly, a significant number of café patrons had their mobile device's Wi-Fidisabled, which would have yielded a lower device count with minimal change to the manual count. Secondly, the unusual coefficient may have been caused by large fluctuations inpatronage. For example, since the timeout values were fairly low for these scenarios, the manual count may have increased without any more mobile devices detected. The negative coefficients tended to occur when one or more of three variables were relatively high. In other words, these situations occurred when the timeout was very long, there was a large number of devices, or the RSSI threshold was high. Additionally, uncommon scenarios saturating certain trials may have caused the negative coefficients. For example, if the timeout was large, sudden changes in manual count may have resulted in difficulties for the algorithm to identify a suitable linear model.However, compared to the total devices experiment, the inclusion of RSSI filtering has reduced the number of unusual tree leaves while significantly improving the accuracy.

Table 9: Total Device with RSSI Decision Tree Notable Leaves

Leaf	Note	Timeout (t) values	Device (x) values	RSSI (r) values
7	Negative Coefficient	t <= 15	2 < x	77.5 < r <=85
10	Coefficient over one	15 < t <= 52.5	x <= 3.5	47.5 < r <=55
11	Coefficient over one	15 < t <= 52.5	x <= 1.5	55 < r < 87.5
18	Negative Coefficient	15 < t <= 52.5	3.5 < x	77.5 < r <= 87.5
22	Negative Coefficient	37.5 < t <= 52.5	37.5 < x	87.5 < r
43	Negative Coefficient	105 < t	50.5 < x	85 < r
48	Negative Coefficient	150 < t	17.5 < x <= 35	r <= 85
49	Negative Coefficient	150 < t	35 < x <= 45.5	r <= 85

The inclusion of RSSI filtering had significantly improved the accuracy of crowd size estimation without creating a decision tree that was impractical. Further, when compared to other competing technologies, it has shown similar accuracy to Bluetooth but not cameras. The RSSI filtered decision tree also highlighted a number of interesting relationships between the variables. In particular, trails with high timeout or RSSI threshold values had generated coefficients that were negative or greater than one. To further improve the accuracy, the devices were broken down into manufacturers by looking up the MAC address's OUI in the IEEE Register. This allowed an even more sophisticated model, however, it came at the cost of reduced practicality.

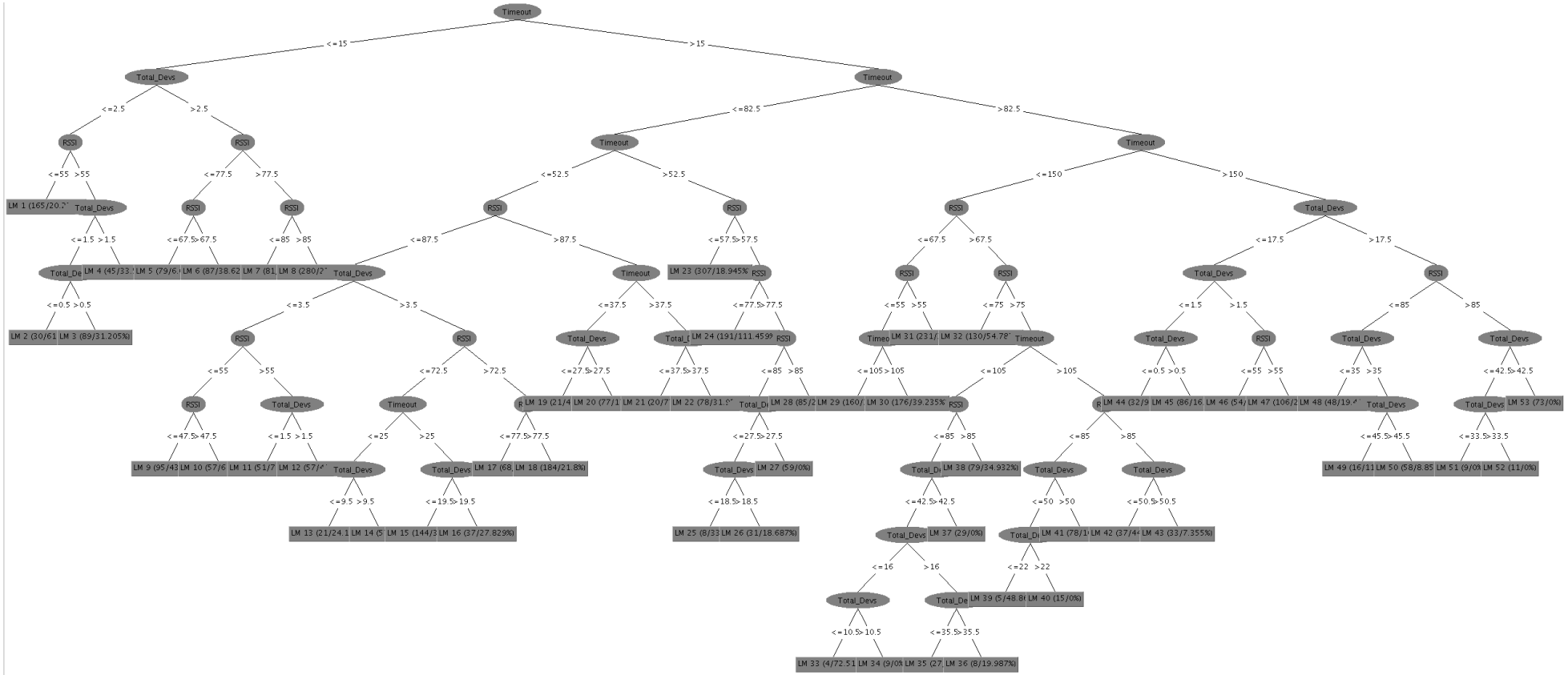
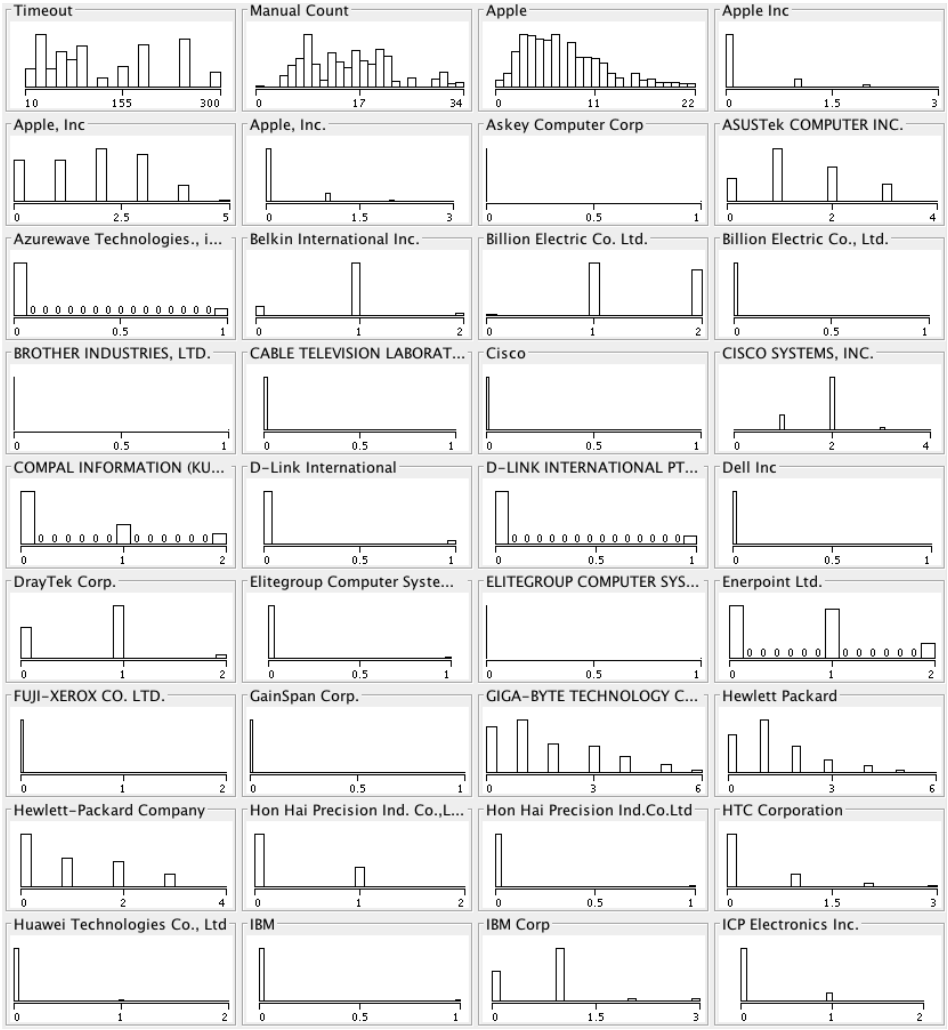


Figure 13: Total Devices With RSSI Decision Tree

3.4 MAC Filtering Without RSSI

The third experiment undertaken to evaluate whether Wi-Fi device activity could estimate crowd size added a manufacturer breakdown of devices, while removing the RSSI threshold restrictions. Once a device was detected, the netaddr Python library[39] was used to obtain the OUI of the MAC address. The main purpose of this experiment was to evaluate the impact of manufacturer breakdown on both the practicality and accuracy of the decision tree. The full variable list for this experiment can be found in the Git repository at the file location: `data/other/NOrssi_man_breakdown_variables.csv`.

The spread of the variables shown in Figure 14revealed significant detail of the potential devices in the space compared to the previous experiments. On one hand, a significant fraction of manufacturers did not have more than a couple of devices detected at any time. On the other hand, a few manufacturers such as Apple, Netgear and Giga-Byte had a noticeably larger variance. A device was listed as having an Unknown manufacturer when the netaddr Python library encountered a "NotRegisteredError" exception.



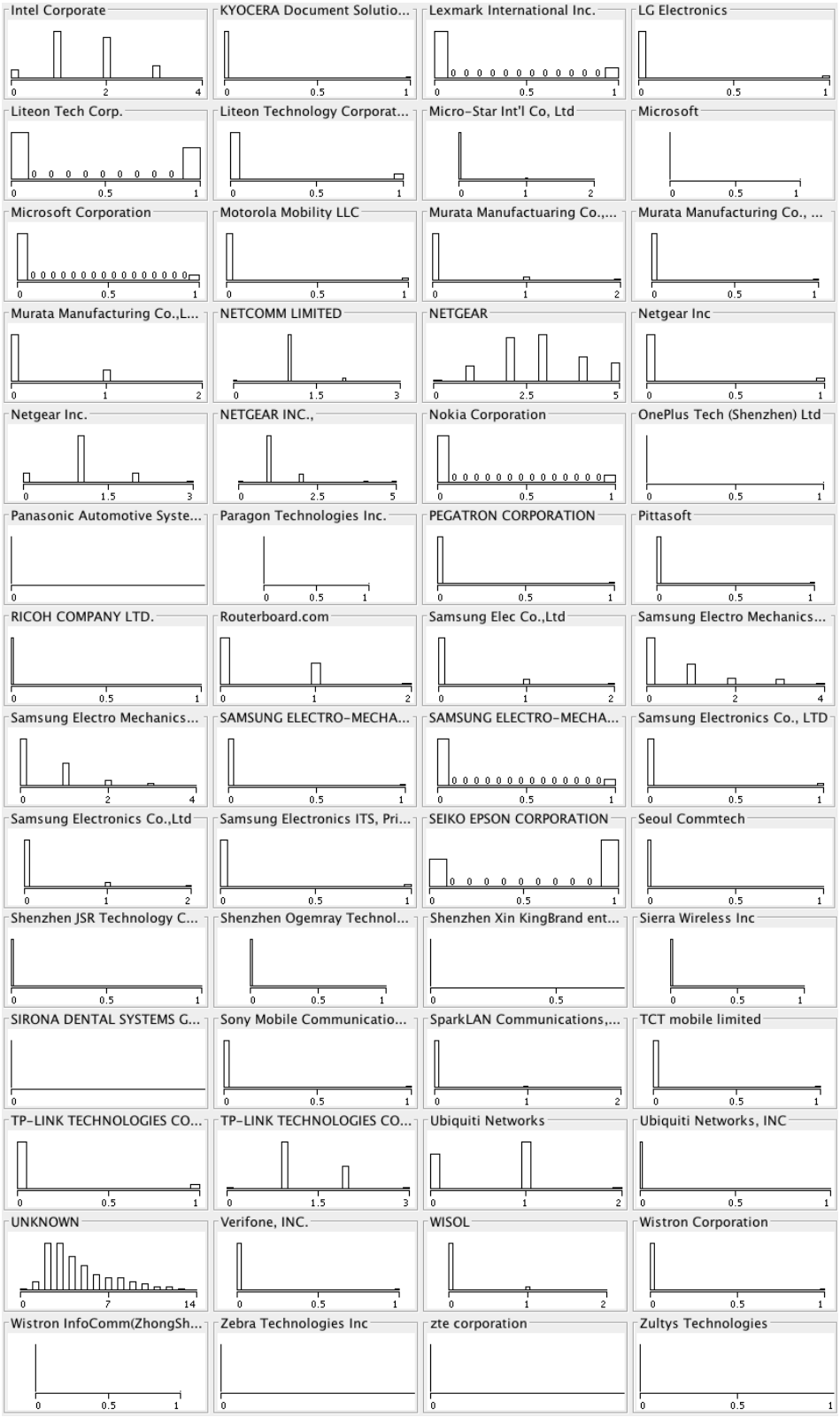


Figure 14: MAC Filtering Without RSSI Variable Spread

3.4.1 M5P Decision Tree

The Weka M5P decision tree algorithm was run with the data and resulted in a tree with 275 leaves, which is summarised in Table 10 (full output can be found in `data/weka_models/NOrssi_man_breakdown.txt`):

Table 10: MAC Filtering Without RSSI M5P Decision Tree Model

Parameter	**Value**
Correlation coefficient, r	0.9772
Coefficient of determination, R^2	0.9549
Mean absolute error	0.7758
Root mean squared error	1.4814
Relative absolute error	13.6786%
Root relative squared error	21.2896%
Total Number of Instances	6505
Number of Leaves	275

The M5P decision tree for this experiment had significantly reduced practicality while also improving the accuracy. Compared to the previous experiment, with only RSSI filtering, this experiment had a 519% increase in leaves for only a 19% improvement in explained variance. Compared to competing technologies in the reviewed literature, this experiment exceeded the accuracy of both cameras (R^2=0.90) and Bluetooth (R^2=0.80), but had significantly reduced practicality. Consequently, the M5P decision tree for this experiment was not suitable for use because the high number of leaves increased the complexity beyond what was practical.

Due to the severe impracticalities of the M5P decision tree, two alternative approaches were considered. Firstly, using theREPTree algorithm, which was a classification decision tree. Compared to the M5P algorithm, a maximum depth was able to be set and it generated a single figure at each leaf rather than a linear function. Secondly, rather than generate numerous linear functions, a single linear regression analysis was performed.

3.4.2 REPTree Decision Tree

The REPTreedecision tree approach allowed the configuration to limit the number of leaves, improving the practicality of the decision tree. However, the REPTree algorithm produced a single figure for each leaf rather than a linear function. This meant that a REPTree decision tree with the same number of leaves as a M5P tree would be less accurate. Nevertheless, the option to set the maximum depth was required to create a practical decision tree. The Weka REPTree decision tree

algorithm was run with the data and resulted in a tree shown in Figure 15with 48 leaves, which is summarised in Table 11 (full details can be found in Appendix3).

Table 11: MAC Filtering Without RSSI REPTree
Decision Tree Model

Parameter	**Value**
Correlation coefficient, r	0.8745
Coefficient of determination, R^2	0.7569
Mean absolute error	2.3385
Root mean squared error	3.374
Relative absolute error	41.2305%
Root relative squared error	48.4882%
Total Number of Instances	6505
Maximum Depth	6
Maximum Number of Leaves	64
Number of Leaves	48

The REPTree decision tree is substantially more practical than the 275 leaves M5P tree that was previously generated. However, when compared to RSSI filtering, it did not perform as accurately. This was attributed to the single figure leaves, rather than the linear function leaves of the M5P tree. However, it still performed better than no RSSI filtering. Compared to the accuracy of competing technologies in the reviewed literature, the system including manufacturer breakdown was the least accurate.

Unlike the M5P decision tree, a REPTree tree did not highlight unique scenarios that had negative or large coefficients. However, it did highlight the scenarios in which there were the largest (over 30 people) or smallest crowds (less than 10) (Tables 12 and 13). Further, the REPTree highlighted the manufacturers that had the most information gain. Notably, a number of mobile phone manufacturers (e.g Apple, Samsung), electronic component manufacturers (e.g. Murata Manufacturing, Hon Hai Precision) and network equipment manufacturers (Billion, Netgear) had the highest information gain. While mobile phone manufacturers were hypothesised to have the largest influence, the mobile electronics manufacturers may have high influence because their components are used in other company's devices. For example, since Hon Hai Precision (also known as Foxconn) is a primary manufacturer of Apple devices, the OUI of the Wi-Fi chip might not match the brand of the device. Therefore, the REPTree decision tree provided evidence to support the hypothesis that mobile device activity could be used to estimate crowd size.

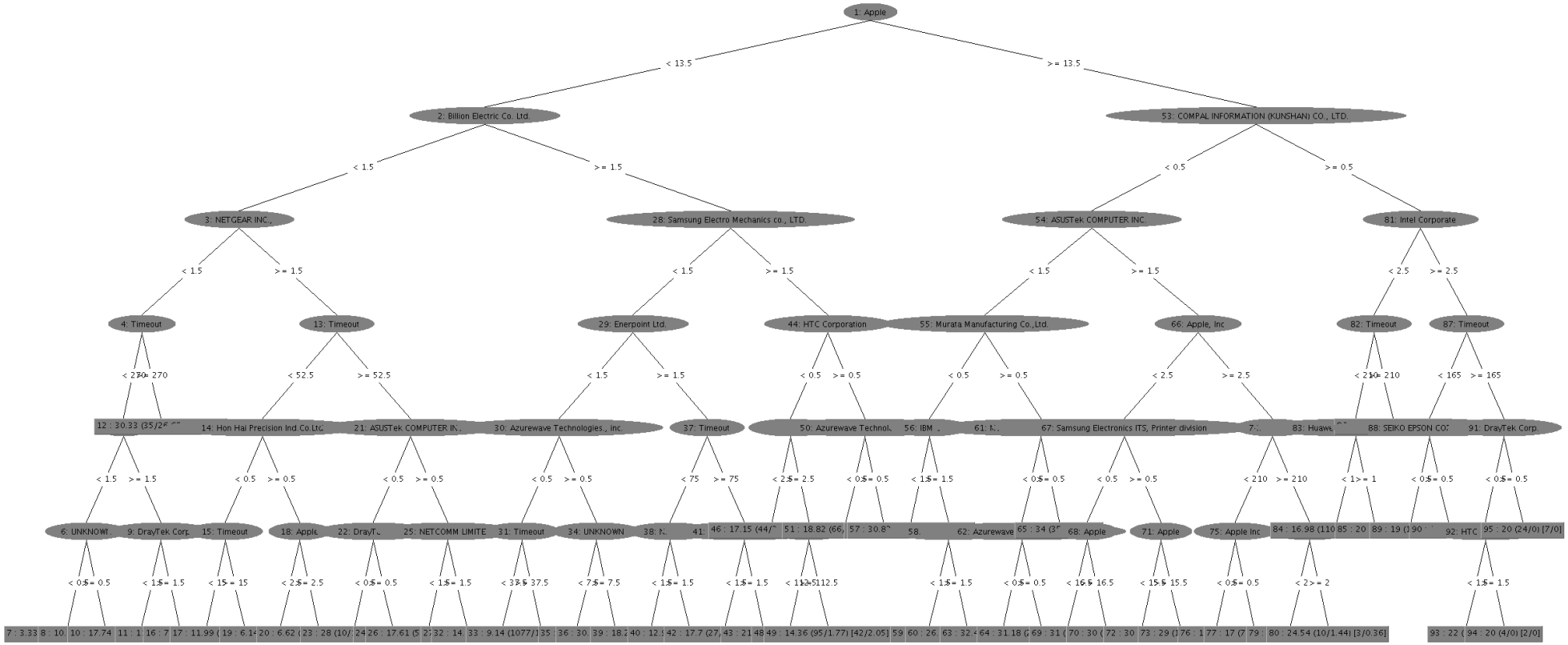


Figure 15: MAC Filtering Without RSSI REPTree

Table 12: MAC Filtering Without RSSI Largest Crowds

Timeout (t)	Manufacturer
Any	Azurewave Technologies., inc.>= 0.5 Enerpoint Ltd. < 1.5 Samsung Electro Mechanics co., LTD. < 1.5 Billion Electric Co. Ltd. >= 1.5 Apple < 13.5
Any	Azurewave Technologies., inc.>= 0.5 HTC Corporation >= 0.5 Samsung Electro Mechanics co., LTD. >= 1.5 Billion Electric Co. Ltd. >= 1.5 Apple < 13.5
Any	IBM Corp < 1.5 Murata Manufacturing Co.,Ltd. < 0.5 ASUSTek COMPUTER INC. < 1 COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5 Apple >= 13.5
Any	HTC Corporation < 1.5 IBM Corp >= 1.5 Murata Manufacturing Co.,Ltd. < 0.5 ASUSTek COMPUTER INC. < 1.5 COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5 Apple >= 13.5
Any	ASUSTek COMPUTER INC. < 1.5 COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5 Apple >= 13.5
Any	Apple, Inc< 2.5 ASUSTek COMPUTER INC. >= 1.5 COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5 Apple >= 13.5

Table 13: MAC Filtering Without RSSI Smallest Crowds

Timeout (t)	Manufacturer
Any	UNKNOWN < 0.5 Apple < 1.5 Timeout < 270 NETGEAR INC., < 1.5 Billion Electric Co. Ltd. < 1.5 Apple < 13.5
t < 15	Hon Hai Precision Ind.Co.Ltd< 0.5 NETGEAR INC., >= 1.5 Billion Electric Co. Ltd. < 1.5 Apple < 13.5
t < 52.5	Hon Hai Precision Ind.Co.Ltd>= 0.5 NETGEAR INC., >= 1.5 Billion Electric Co. Ltd. < 1.5 Apple < 13.5

3.4.3 Linear Regression

The second analysis tool that was used on the manufacturer breakdown waslinear regression to generate a linear function to estimate crowd size. The Weka software included a linear regression classifier, which was used to generate the function. The model's summary is detailed in Table 14 (full output in Appendix 5).

Table 14: MAC Filtering Without RSSI Linear Regression Summary

Parameter	**Value**
Correlation coefficient, r	0.8650
Coefficient of determination, R^2	0.7482
Mean absolute error	2.6166
Root mean squared error	3.4913
Relative absolute error	46.133%
Root relative squared error	50.1736%
Total Number of Instances	6505

The linear regression analysis resulted in a slight decline in explained variance compared to the REPTree, anddid not reach the level accuracy of the M5P decision tree. However, this is expected since the M5P tree had the opportunity to undertake numerous linear regressions. Further, the single linear regression failed to achieve the accuracy of other technologies reviewed in the literature. However, the linear function still had a fairly good fit to the data, and could be practically used to estimate the crowd. Analysing the coefficients for the linear regression was not feasible since only 13 of the 91 manufacturers had positive coefficients less than one. This was likely caused by a large number of data points saturating the regression algorithm, resulting in a function inconsistent with the original hypothesises.

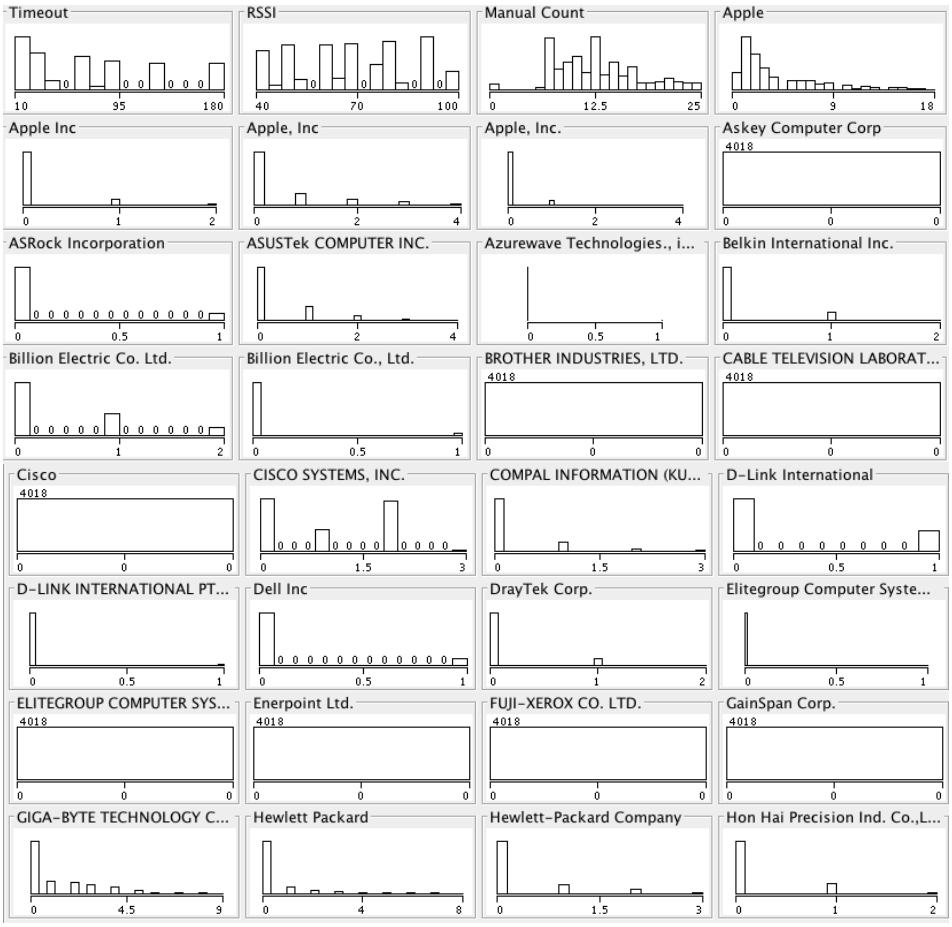
3.4.4 Summary

Of the three methods considered for manufacturer breakdown without RSSI filtering, none proved to be both accurate and practical. The M5P decision tree, while extremely accurate, was far too impractical for use. Subsequently, REPTree algorithm was used to restrict the size and improve practicality. However, the REPTree decision tree lacked the accuracy to compete with other technologies. Since the REPTree only placed a single figure on each leaf, linear regression was considered as the final analysis technique in an attempt to generate a mathematical function for crowd size. The linear regression function failed to reach the accuracy of both the M5P decision tree, Bluetooth and cameras. These results led to the final experiment, to estimate the crowd size using both MAC address and RSSI filtering.

3.5 MAC Address and RSSI Filtering

The fourth and final experiment undertaken to evaluate whether Wi-Fi device activity could estimate crowd size included both a manufacturer breakdown of devices in addition to RSSI filtering. Similar to the previous experiment, the netaddr python library [39] was used to obtain the OUI of the MAC address. This experiment evaluated whether RSSI and MAC address filtering would further improve the accuracy of estimating crowds using Wi-Fi device activity. The full variable list for this experiment can be found in the Git repository at the file location: `data/other/rssi_man_breakdown_variables.csv`.

The spread of the variables (Figure 16)revealed similar information as the previous experiment. Similar to the previous experiment, a significant portion of manufacturers had no more than a few devices at any time. However, major manufacturers such as Apple and Netgear still recorded a significant variance in the device numbers reordered. As with the previous experiment, if the netaddr python extension encountered a "NotRegisteredError" then the device was listed as Unknown.



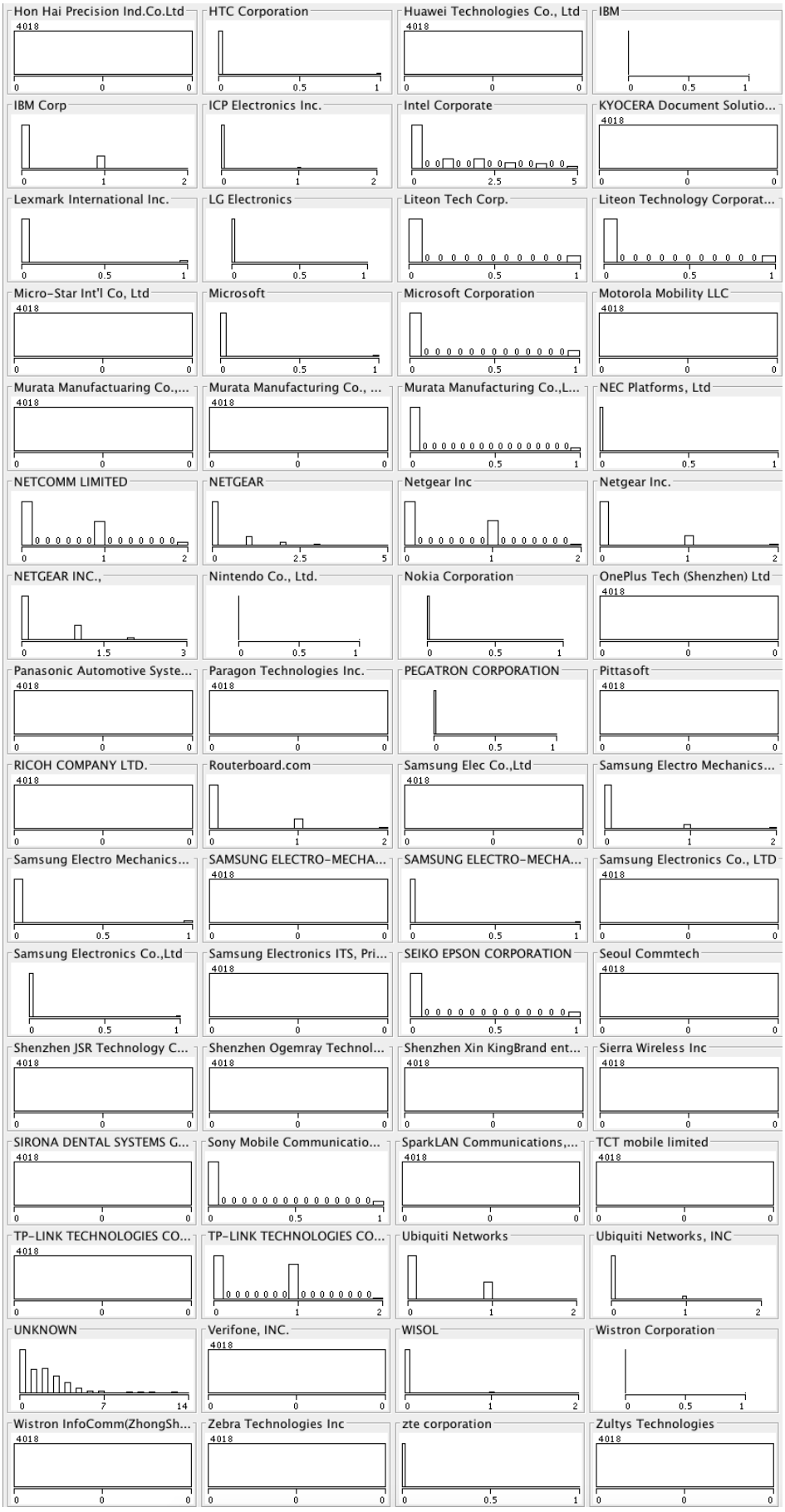


Figure 16: MAC and RSSI Filtering Variable Spread

3.5.1 M5P Decision Tree

The Weka M5P decision tree algorithm was run with the data and resulted in a tree with 100 leaves, which is summarised in Table 15 (full details can be found in `data/weka_models/rssi_man_breakdown.txt`):

Table 15: MAC Filtering Without RSSI M5P Decision Tree Model

Parameter	**Value**
Correlation coefficient, r	0.9235
Coefficient of determination, R^2	0.8528
Mean absolute error	0.9698
Root mean squared error	1.9238
Relative absolute error	23.6916%
Root relative squared error	38.4362%
Total Number of Instances	4018
Number of Leaves	100

The M5P decision tree for this experiment had a minor improvement in practicality over the previous experiment but with some reduced accuracy. Compared to technologies in the reviewed literature, this experiment exceeded the accuracy of Bluetooth (R^2=0.80) but not cameras (R^2=0.90). However, it was considered impractical to implement because the decision tree had 100 leaves. Similar to the previous experiment, other analysis techniques were considered.

Due to the severe impracticalities of the M5P decision tree, the same two approaches as the previous experiment were used. Theses techniques were the REPTreedecision tree and linear regression. The two techniques were chosen to improve the practicality of the model because the M5P decision tree was not practically feasible.

3.5.2 REPTree Decision Tree

The REPTree algorithm was chosen since it had the ability to limit the number of leaves, improving the practicality of the decision tree. However, the accuracy of the tree would be diminished since the REPTree algorithm produced a single figure for each leaf rather than a linear function. The maximum depth was set to the same value of the previous experiment to remain consistent. The Weka REPTree decision tree algorithm was run with the data and resulted in a tree shown in Figure 17 with 38 leaves, which is summarised in Table 16 (full details can be found in Appendix 6).

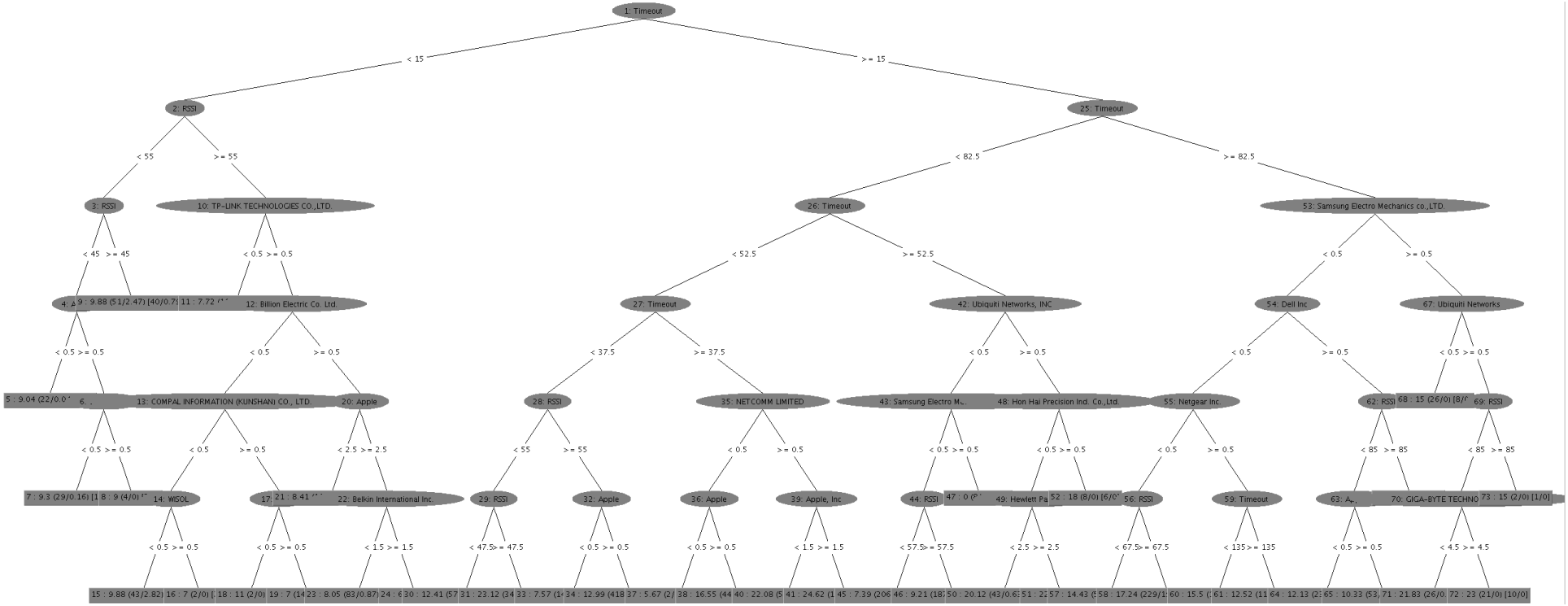


Figure 17: MAC and RSSI Filtering REPTree

Table 16: MAC and RSSI Filtering REPTree Decision Tree Model

Parameter	Value
Correlation coefficient, r	0.7684
Coefficient of determination, R^2	0.5904
Mean absolute error	2.0525
Root mean squared error	3.2024
Relative absolute error	50.1418%
Root relative squared error	63.9819%
Total Number of Instances	4018
Maximum Depth	6
Maximum Number of Leaves	64
Number of Leaves	38

Compared to the previous M5P decision tree, the REPTree is substantially more practical, but not as accurate. It has reduced the number of leaves from 100 to 38, but has reduced the explained variance by 31%. However, compared to all the analysis techniques across all experiments, it is the least accurate. Consequently, it is does not perform as accurately when compared to the technologies in the reviewed literature. The reduction in accuracy was attributed to the single figure leaves, which would have less accuracy than a linear function. However, the magnitude of the difference in accuracy between the REPTree and the analysis of previous experiments was not expected.

Similar to the previous REPTree analysis, the algorithm did highlight scenarios where the crowd would be large (larger than 20 people) or small (less than 10 people). However, it did not highlight scenarios that resulted in coefficients that were negative or greater than one, like the M5P algorithm could. In contrast to the previous experiment'sREPTree decision tree, a significant number of variables had no influence. As seen in Table 17 and 18, a significant number of leaves that had high or low crowd sizes had no rules for the RSSI or manufacturers. However, similar to the previous REPTree, the manufacturers with high influence are mobile phone companies (e.g. Samsung, Apple) and electronic components manufacturers (e.g. UbiquitiNetworks, Hon Hai Precision).The REPTree decision tree supported the hypothesis that mobile device activity could be used to estimate crowd.

Table 17: MAC and RSSI Filtering Largest Crowds

Timeout (t)	RSSI (r)	Manufacturer
15 <= t <37.5	47.5 <= r < 55	*Any*
37.5<= t <52	*Any*	NETCOMM LIMITED >= 0.5
52.5<= t <82.5	*Any*	Hon Hai Precision Ind. Co.,Ltd. < 0.5 Ubiquiti Networks, INC >= 0.5
82.5<= t	r < 85	Ubiquiti Networks >= 0.5 Samsung Electro Mechanics co.,LTD. >= 0.5

Table 18: MAC and RSSI Filtering Smallest Crowds

Timeout (t)	RSSI (r)	Manufacturer
15 <= t <37.5	55 <= r	Apple < 0.5
37.5 <= t < 52.5	*Any*	Apple < 0.5 NETCOMM LIMITED < 0.5
52.5 <= t < 82.5	*Any*	Samsung Electro Mechanics co.,LTD. < 0.5 Ubiquiti Networks, INC < 0.5

3.4.3 Linear Regression

The second analysis tool used in the final experiment was linear regression. Similar to the previous experiment, Weka was used to perform the regression and generate the linear function. The model's summary is detailed in Table 19 (full details in Appendix 7).

Table 19: MAC and RSSI Filtering Linear Regression Summary

Parameter	Value
Correlation coefficient, r	0.6369
Coefficient of determination, R^2	0.4056
Mean absolute error	2.7798
Root mean squared error	3.8575
Relative absolute error	67.9115%
Root relative squared error	77.0718%
Total Number of Instances	4018

The linear regression analysis resulted in severely reduced accuracy in the model, making the linear regression analysis of RSSI and MAC address filtering the worst performing model. While it was more practical than the decision trees, the coefficient of determination suggested it was not accurate enough to use. Similar to the linear regression analysis of the previous experiment, 6 of the 95 manufacturers had coefficients with values that were hypothesised, lying inside the positive unit interval (0 to 1). This may have been caused by the large volume of data saturating the algorithm

and resulting in a function inconsistent with the hypothesis. Regardless, the linear regression analysis was too inaccurate to be suitable to estimate crowd size.

3.6 Summary of Experiments

All the experiments provide evidence to support the hypothesis that mobile device activity could be used to estimate crowd size. However, a vast number of experimental and analytical approaches resulted in differing levels of practicality and accuracy. For example, a M5P decision tree which used MAC, but not RSSI filtering was the most accurate model (R^2=0.9549), but was not practical because the decision tree had 275 leaves. Therefore, it was important to evaluate the practicality to the same extent as the accuracy. In this regard, the most balanced, in terms of accuracy and practicality, was the M5P decision tree that did have RSSI but not MAC address filtering. On one hand, the analysis of MAC address filtering did support the hypothesis that a sensed device with an OUI belonging to a mobile device manufacturer had high information gain. On the other hand, the increased complexity made the model tooimpractical.

Compared to technologies that have been reviewed in the literature, Wi-Fi appeared to be a feasible alternative. All M5P decision trees, except total devices without RSSI filtering, exceeded the accuracy of Bluetooth. However, none exceeded the accuracy of cameras while being practical at the same time. Cameras continue to be the best crowd estimation technology, so long as the situation practically permits their implementation. For example, a space with different weather and lighting throughout the day would not be practical for a camera to the same extent as the Wi-Fi solution proposed in this research. Ultimately, accurately and practically estimating crowd size was a difficult task for the system developed in this research, and while its performance exceeds competing technologies in some regards, further research is required to evaluate is position in the crowd estimation landscape.

4.0 REVIEW

4.1 TECHNOLOGY SELECTION

This research evaluated whether Wi-Fi in mobile devices could serve as an estimator for crowd size. Reviewed literature demonstrated that Bluetooth [22] and cameras [10] had previously worked as technologies to estimate crowd size. Further, the previous studies had shown how these technologies could be used to estimate density and flow rate. However, the literature did list some difficulties with these technologies. Firstly, a significant portion of the population had their Bluetooth disabled and Bluetooth had proximity detection limitations. This meant for small crowds, no Bluetooth devices might be detected, while large crowds might saturate a scan. Secondly, cameras worked well in certain setups, but lacked the flexibility to work in different lighting and weather conditions. Consequently, Wi-Fi was chosen to take minimise the impracticalities of the other technologies. Wi-Fi is used substantially more in mobile devices and works in a larger variety of setups.

There were some benefits of choosing Wi-Fi as the technology to estimate crowds using mobile device activity. Firstly, since Wi-Fi was a broadcast technology, transmissions could be passively intercepted and analysed. This meant that no individualshad to actively participate in the research. Secondly, the Wi-Fi transmission headers resulted in some metadata being extracted. Being able to analyse the metadata was a unique approach this research took, which allowed the estimation to be more accurate. The signal strength and MAC address of a device were analysed to improve the estimation techniques.

However, the choice to use Wi-Fi hadsome weaknesses. Firstly, there was a significant number of static Wi-Fi devices. Routers and modems near the space added extra devices the scanner would pick up. This is in contrast to Bluetooth, which would not have any static infrastructure it might accidentally detect. Secondly, Wi-Fi had significantly more range than other technologies. This meant that devices from neighbouring businesses were detecteddespitethe fact thatthese devices were outside the café space,which was the subject of this research. The RSSI filtering experiments attempted to remove this, however since signal strength was not very reliable, this did not always work.

The choice to use Wi-Fi did provide a solution that was comparable to other technologies. While it may have had some practicality issues associated with its use, it provided a significant number of

benefits over competing technologies. Wi-Fi was an appropriate technology choice to estimate crowd size.

4.2 System Design

A system was developed to gather data to evaluate Wi-Fi as the technology choice to estimate crowd size. The system had two parts, the scanner and the receiver. The scanner was a Raspberry Pi running Kismet though a Python script. The receiver logged and graphed the data using a Python script. The two systems communicated over MQTT.

The scanner was able to passively scan for active Wi-Fi devices in the nearby space. Further, Kismet and Python softwarewere able to extract the metadata from the Wi-Fi messages. The Kismet software did not have difficulty detecting any devices, even when the device was not connected to a nearby Wi-Fi network. However, while Kismet had a graphical interface, the Python scriptwas restricted toextracting data from Ksimet'slocal network server. Additionally, poor documentation for Kismet restricted the ability to interact with the local Kismet server. This had two impacts on the software. Firstly, the device list obtained from the Kismet server could not be forgotten internally by Kismet. This meant that once a device was detected, it would remain in the list permanently. Consequently, a Kismet process had to be started and killed for each scan to be run. Secondly, Kismet did not provide the time a device was seen. Similar to the first issue, a Kismet process was started, executed for a short period of time, and then each device was associated with the Raspberry Pi's local system time. These two issues with the Kismet server required a loop of continuously creating and killing Kismet processes. Ultimately however,this loop worked and Kismet was able to sense nearby Wi-Fi devices while extracting metadata from them.

The receiver's main purpose was to log and graph the data so that it could be analysed. Similar to the scanner, a Python script was created to do this task. The Python script performed its task well through the use of sub-processes and threads. This was due to performance issues because when the graph had numerous variables, a single Python processes was not capable of handling the workload. However, having numerous threads and processes resulted in difficulties in handling the interactions between tasks. For instance, queues had to be used to pass data between MQTT, logging, graphing and user input tasks. All of these complex interactions, in addition to Python being strongly typed, but having no type checking until runtime, meant that the system was vulnerable to numerous programming issues. Consequently, the receiver system was not thoroughly tested, meaning it may not be suitable for actual deployment to estimate crowd size.

The systems designed to estimate crowd size did have some issues that prevented them being suitable for real world use. However, this research was aimed at evaluating whether Wi-Fi was suitable as a technology to estimate crowd size. In this regard, the systems were able to gather appropriate data to evaluate Wi-Fi as a tool for estimating crowd size.

4.3 EXPERIMENTS

A series of experiments was undertaken to evaluate Wi-Fi as a technology for estimating crowd size. Further, the system design allowed metadata about the devices to be gathered and it was expected this extra information would improve the model. A series of experiments were undertaken to evaluate whether metadata did improve the estimation.

The preliminary investigation evaluated whether the metadata was reliable. This was done for the two different types of metadata that could be gathered, signal strength and the OUI of the MAC address. As expected, the distance over RSSI model was exponential, as the signal strength was weaker, smaller differences in RSSI would correlate with larger differences in distance. This would have positive implications for small spaces, but would not be suitable in larger spaces and crowds. Additionally, signal strength figures would have been distorted by a number of factors, such as the different power of antenna between devices and obstacles restricting the propagation of a signal (e.g. tables, chairs, walls). The second metadata gathered was the OUI of the MAC address. This investigation demonstrated that the OUI might not match the device's brand. Consequently, it may not have been a reliable indicator for the device that was detected. However, both types of metadata did prove to have some practicality. Consequently, the metadata was included separately and then together in experiments to evaluate whether they did indeed improve the accuracy of the crowd estimation.

The first experiment to evaluate using Wi-Fi to estimate crowd size did not use any of the metadata. This experiment used the M5P decision tree in Weka which resulted in a R^2=0.6239. Compared to technologies examined in the literature, Wi-Fi without the metadata was not as accurate. However, a fairly practical decision tree was generated. Compared to the other experiments, this served, as a good comparison to whether the metadata was useful.

The second experiment included threshold value on the RSSI of devices, preventing devices with weaker signals being included in the count. Compared to the previous experiment, the M5P decision tree had 53 leaves, 8 more than without RSSI. However, it had significantly improvedaccuracy(R^2=0.8015). Consequently, this M5P decision tree exceeded the accuracy of

Bluetooth but not cameras. Therefore, the inclusion of RSSI filtering did have a positive effect on being able to estimate crowd size.

The third experiment included MAC address but not RSSI filtering on devices to estimate crowd size. This experiment allowed a comparison of the two metadata types, whether just RSSI or MAC address filtering were more accurate. In contrast to the slight increase in decision tree leaves in the previous experiment, MAC address filtering increased the decision tree by222. Even though it was significantly more accurate (R^2=0.9549), the M5P decision tree was not practical. Therefore, the REPTree algorithm was chosen as an alternative to limit the size of the tree. The resulting decision tree supported the hypothesis that mobile phone and electronics manufacturers had the most information gain. However, it failed to achieve the accuracy of the M5P tree, Bluetooth or cameras, since it had R^2=0.7569. Linear regression was chosen as an alternative analysis technique to contrast the single figure leaves of the REPTree decision tree. While it proved to be more accurate (R^2=0.7482), it did not highlight relationships between variables. This experiment demonstrated that MAC address filtering simply saturated the algorithm with too many variables, whichsuggested that MAC address filtering was not feasible.

The final experiment used both types of metadata, filtering both the signal strength and the RSSI reading of devices. Initially a M5P decision tree was used, which had less accuracy than the previous experiment (R^2=0.8528). Further, while it had a smaller tree than previously, it was still too impractical to use. Therefore, the REPTree and linear regression algorithms were used as alternatives. These two alternatives were the two worst performing models (R^2=0.5904 and R^2=0.4056) across all experiments and analysis techniques. The MAC filtering, in addition to the RSSI filtering has severely saturated the algorithm with too many variables, resulting in a model for estimation crowd size that was neither accurate nor practical.

The preliminary investigation provided the first hint that the OUI of a mac address is not always reliable, instead identifying the device as belonging to a completely different company. This was confirmed in the third and fourth experiments, which showed that MAC address filtering overwhelmed the algorithm with different variables and resulted in models that were neither practical nor accurate. While MAC filtering proved to be an unfeasible addition to estimate crowd size, the experiments provided strong support for RSSI filtering.The improvement in accuracy between experiments one and two, in addition to the preliminary investigation into RSSI over distance, showed that RSSI had some significant improvements to the accuracy of a model. Further, the addition of RSSI into a decision tree did not inflate the size of the decision tree beyond

a usable level. Therefore, the most appropriate estimation model to estimate crowds was to use thedevice total with RSSI filtering.

4.4 Summary

This research investigated whether using Wi-Fi activity was a suitable method to estimate crowd size. Further, it took advantage of two forms of metadata, the signal strength and the MAC address's OUI to improve the crowd size estimation. A system was designed to scan a space for active environments. While the system was appropriate for evaluating whether Wi-Fi was a good tool to estimate crowd size, the complexity of software would hinder real life deployment. Consequently, refactoring and testing the software is required before it can be used to estimate crowd size. While the RSSI filtering improved the accuracy of the model, MAC filtering resulted in models too complex and impractical to use.

5.0 FUTURE DIRECTION

This research into estimating crowd size using Wi-Fi device activity has identified numerous improvements that could be implemented in future investigations. There are a number of ways in which both the system architecture as well as the analysis could be enhanced. These improvements could significantly improve the accuracy of a crowd size estimate.

Firstly, the system could be redesigned to use multiple technologies, rather than just one. For instance, a mobile device scanner that has both Bluetooth and Wi-Fi scanning could be used. Since Bluetooth has already demonstrated that it has similar accuracy to the Wi-Fi approach, this would improve the accuracy of the crowd size estimate. This design approach could reincorporate the RSSI and MAC address filtering. Additionally, OUI of the Bluetooth MAC address could be matched against the OUI obtained from Wi-Fi. While this improvement would increase the complexity of both the system and analysis techniques, it might significantly improve the accuracy.

Alternatively, surveying users of the space to obtain demographic data may allow additional relationships between variables to be discovered. This research did not survey patrons at Top Nosh Café, meaning that if someone had one or more mobile devices, it was not explicitly known. If the crowd was surveyed, then the explicit device count could be compared against the detected device count creating a baseline for further analysis. The information gathered from surveys could help understand how crowds use their mobile device and could help improve the model and generate better crowd size estimates. However this approach would make the estimation technique a more active approach, as opposed to the passive technique favoured in this research, since it would need to explicitly engage the crowd in information gathering.

While surveying the crowd would gather demographic data and change the solution to being active engagement, some additional information could be passively gathered without surveying the crowd. For example, the time of day, day of the week, or the weather could all be included as passive data. This data could reveal relationships that the total number of devices alone could not. For instance, the data gathered in this research was not able to identify if Top Nosh Cafés crowd size might change based on whether it was breakfast, lunch or in-between. In future research it would be useful to examine if additional environment data, which could be obtained passively, assists modelling algorithms in identifying a model to estimate crowd size.

Estimates on crowd size are useful to a variety of businesses and spaces, and investigating how an estimate could change depending on location could be a topic for future research. For instance, libraries, classrooms or museums are alternative locations that could be examined. The café used in this research had a number of different entrances and exits, in addition to a pedestrian footpath. This meant it was occasionally difficult to determine if an individual was entering the café, or passing by. Therefore, future research should pick a location that has clearly defined entrances, exits and would not accidentally include passers by.

While RSSI filtering was performed in this research, an investigation of the variation of RSSI over time might also be worthwhile. Since individuals in the crowd move around the space, the signal strength will be constantly changing. Meanwhile, the static Wi-Fi infrastructure remains stationary and would have fairly constant signal strength to the scanner. Therefore, if the variance in a single device was measured over time, it may provide a method to verify if individuals are moving around or stationary in the space. However, as discussed in the preliminary investigation, RSSI is unpredictable and therefore, may not be practical. This would assist algorithms such as the decision tree to determine whether the device can be associated with an individual or not. However, this would require a decision tree for each device to verify if it should be counted. This would increase the complexity of the estimation model but would lead to a more accurate result.

Lastly, the packet sniffing technology could be expanded to examine the contents of Wi-Fi messages. This would have some security and privacy concerns, which would trigger an ethical evaluation of such. Individuals would have to be satisfied that their personal mobile device activity was being examined only for the purpose of crowd estimation. However, if the system was able to see what a transmission contains, it might be able to better estimate the number of people. For example, whether a transmission was a device establishing a connection to a Wi-Fi network or a HTTP message may assist in estimating crowd size. While this might be a possible venture for future research, care should be taken to ensure the security and privacy of individuals is not compromised as a result.

This research's aim was to examine whether Wi-Fi device activity could be used to estimate crowd size. A series of experiments was undertaken to evaluate whether Wi-Fi was a suitable choice. However, there remain several areas with potential for future research. In particular, analysis techniques could be enhanced to include multiple technologies or Wi-Fi metadata could be more closely analysed. The suggested improvements could allow Wi-Fi to exceed the accuracy of Bluetooth and cameras to estimate crowd size in the future.

6.0 CONCLUSION

The analytics of crowd size and movement is an increasingly important area of research as managers of public and private spaces seek to manage crowds, and improve the design features, effectiveness and efficiency of space. Knowledge about crowd size can be useful to a number of industries such as retail spaces, museums, and public transport. This research investigated whether mobile device activity, in particular Wi-Fi, could be used to estimate crowd size. The literature revealed a number of competing technologies such as camera and Bluetooth. Wi-Fi was selected because Bluetooth had technological limitations and video cameras had practical restraints. In this regard, the Wi-Fi approach used in this research took advantage of the weaknesses in other technologies. The Wi-Fi system was able to work in any environmental condition and did not have limitations on how many devices could be detected.

After the technology had been chosen, a series of experiments was undertaken to evaluate whether Wi-Fi could estimate crowd size. Most experiments were accurate when compared to the technologies reported in the literature, for some models were too impractical to consider feasible. Consequently, the best model in terms of practicality and accuracy was using RSSI but not MAC address filtering. This RSSI filtered model was more accurate than Bluetooth, but not cameras.

This investigation looked at a number of factors to estimate crowd size and identified improvements that can still be made in the future. Most notably, a number of other analytical techniques, such as the variance in a device's signal strength, could be investigated. Moreover, the passive approach adopted in this research provided advantages over active approaches. For example, the crowd was not interrupted or made aware they were being counted.

In conclusion, using mobile device activity, in particular Wi-Fi can be used to estimate a crowd in a given area. Compared to other technologies presented in the literature, Wi-Fi was capable of matching their accuracy. Additionally, the decision tree approach highlighted that mobile devices had the highest information gain when estimating crowd size. While there are still improvements that could be made to both the system developed and the analysis techniques used, the solution developed in this research provided strong evidence that Wi-Fi device activity correlates with the number of people in a space.

REFERENCES

[1] M. Jiang, J. Huang, X. Wang, J. Tang, and C. Wu, "An Approach for Crowd Density and Crowd Size Estimation," *Journal of Software,* vol. 9, pp. 757-762, 2014.

[2] P. F. Yip, R. Watson, K. S. Chan, E. H. Y. Lau, F. Chen, Y. Xu*, et al.*, "Estimation of the number of people in a demonstration," *Australian & New Zealand Journal of Statistics,* vol. 52, pp. 17-26, 2010.

[3] A. Torre and A. Rallet, "Proximity and Localization," *Regional Studies,* vol. 39, pp. 47-59, 2005.

[4] D. Namiot and M. Sneps-Sneppe, "Mobile Services and Network Proximity," 2013.

[5] Navizon, "Navizon Indoor Triangulation System," ed: Navizon, 2013.

[6] M. D'Souza, M. Ros, and M. Karunanithi, "An Indoor Localisation and Motion Monitoring System to Determine Behavioural Activity in Dementia Afflicted Patients in Aged Care," *electronic Journal of Health Informatics,* vol. 7(2):e14, 2012.

[7] D. Bauer, M. Ray, and S. Seer, "Simple sensors used for measuring service times and counting pedestrians: Strengths and weaknesses," *Transportation Research Record,* vol. 2214, pp. 77-84, 2011.

[8] D. Bauer, N. Brandle, S. Seer, M. Ray, and K. Kitzzawa, "Measurement of Pedestrian Movements: A Comparative Study on Various Existing Systems," *Pedestrian Behaviour: Models, Data Collection and Applications,* 2009.

[9] D. Hernández-Sosa, M. Castrillón-Santana, and J. Lorenzo-Navarro, "Multi-sensor People Counting." vol. 6669, ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 321-328.

[10] U. Stilla, E. Michaelsen, U. Soergal, S. Hinz, and J. Ender, "Airborne monitoring of vehicle activity in urban areas ".

[11] ACT Government. (2014, 3 April). *Use of Bluetooth technology for traffic studies*. Available: http://www.tams.act.gov.au/roads-transport/traffic/use-of-bluetooth-technology-for-traffic-studies

[12] J. Segen and S. Pingali, "A camera-based system for tracking people in real time," Vienna, 1996, pp. 63-67.

[13] E. Bas, E. Bas, M. Tekalp, M. Tekalp, F. S. Salman, and F. S. Salman, "Automatic Vehicle Counting from Video for Traffic Flow Analysis," pp. 392-397.

[14] D. Lefloch, F. A. Cheikh, J. Y. Hardeberg, P. Gouton, and R. Picot-Clemente, "Real-time people counting system using a single video camera," in *Real-Time Image Processing 2008*, 2008.

[15] Ekahau, "Comparison of Wireless Indoor Positioning Technologies " Ekahau2005.

[16] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "LANDMARC: Indoor Location Sensing Using Active RFID," *Wireless Networks,* vol. 10, pp. 701-710, 2004.

[17] P. R. Center, "Mobile Technology Fact Sheet," in *Pew Research Center*, ed, 2014.

[18] S. N. Patel, J. A. Kientz, G. R. Hayes, S. Bhat, and G. D. Abowd, "Farther than you may think: An empirical investigation of the proximity of users to their mobile phones," in *UbiComp 2006: Ubiquitous Computing*, 2006, pp. 123-140.

[19] T. Nicolai and H. Kenn, "About the relationship between people and discoverable Bluetooth devices in urban environments," in *4th international conference on mobile technology, applications, and systems* 2007, pp. 72-78.

[20] D. M. Bullock, R. Haseman, J. S. Wasson, and R. Spitler, "Automated measurement of wait times at airport security," *Transportation Research Record,* pp. 60-68, 2010.

[21] A. Morrison, M. Bell, and M. Chalmers, "Visualisation of Spectator Activity at Stadium Events," in *Information Visualisation, 2009 13th International Conference*, Barcelona, 2009, pp. 219-226.

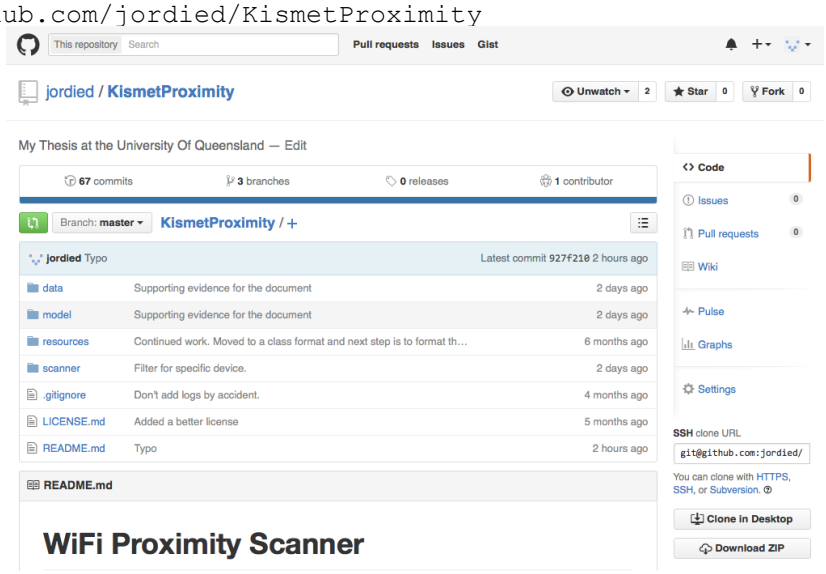
[22] J. Weppner and P. Lukowicz, "Bluetooth based collaborative crowd density estimation with mobile phones," in *Pervasive Computing and Communications (PerCom), 2013 IEEE International Conference*, ed: IEEE, 2013, pp. 193-200.
[23] W. Xi, J. Zhao, X.-Y. Li, K. Zhao, S. Tang, X. Liu*, et al.*, "Electronic frog eye: Counting crowd using WiFi," pp. 361-369.
[24] Y. Wang, J. Yang, H. Liu, Y. Chen, M. Gruteser, and R. Martin, "Measuring human queues using WiFi signals," in *19th Annual International Conference on Mobile Computing & Networking*, 2013, pp. 235-238.
[25] V. Afshar, "50 Incredible WiFi Tech Statistics That Businesses Must Know " in *Huffington Post*, ed, 2014.
[26] N. Lane, Y. Chon, L. Zhou, Y. Zhang, F. Li, D. Kim*, et al.*, "Piggyback CrowdSensing (PCS): energy efficient crowdsourcing of mobile sensor data by exploiting smartphone app opportunities," in *11th ACM Conference on Embedded Networked Sensor Systems*, 2013, pp. 1-14.
[27] J. Sturmer, "Use of phone-tracking technology in shopping centres set to increase," in *ABC news*, ed. Adelaide, 2013.
[28] C. Zhang, K. P. Subbu, J. Luo, and J. Wu, "GROPING: Geomagnetism and cROwdsensing Powered Indoor NaviGation," *IEEE Transactions on Mobile Computing,* vol. 14, pp. 387-400, 2015.
[29] Tektronix, "Wi-Fi: Overview of the 802.11 Physical Layer and Transmitter Measurements," 2013.
[30] A. Hannah, "Packet Sniffing Basics," in *Linux Journal*, ed, 2011.
[31] M. Kershaw. (2013, 18 April). *Kismet*. Available: http://www.kismetwireless.net/
[32] M. Kershaw. (2013, 15 April). *Trying to compile kismet on OS X 10.9*. Available: https://www.kismetwireless.net/Forum/General/Messages/1382365265.01709
[33] The Eclipse Foundation, "paho-mqtt 1.1," ed, 2015.
[34] H. Dahan, *Proactive data mining with decision trees* vol. 1. New York, NY: Springer, 2014.
[35] L. Rokach, O. Z. Maimon, and S. World, *Data mining with Decision Trees: Theory and Applications* vol. 69. Singapore: World Scientific Pub, 2008.
[36] C. Zhan, A. Gan, and M. Hadi, "Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm," *IEEE Transactions on Intelligent Transportation Systems,* vol. 12, pp. 1549-1557, 2011.
[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter,* vol. 11, pp. 10-18, 2009.
[38] J. Schneider. (1997). *Cross Validation*. Available: https://www.cs.cmu.edu/~schneide/tut5/node42.html
[39] D. Moss. (2015, 5 May). *netaddr*. Available: https://github.com/drkjam/netaddr

APPENDICIES

APPENDIX 1: PROJECT REPOSITORY

The source code, raw data and Weka models are all located in a GitHub repository at:

`https://github.com/jordied/KismetProximity`



The key files and directories in the repository are:

Filename or Directory	Description
./LICENSE.md	The license for the source code, the GNU general public license.
./README.md	The project Readme which describes the contents of the repository. Including options for the Python programs.
./data	Folder which contains all the data and analysis. Has an additional Readme to elaborate on the subfolders.
./data/arff	The native Weka format files.
./data/csv	The raw csv files of the experiments. Each csv file contains numerous trials.
./data/Log_Files	Contains numerous subfolders, each containing a trial.
./data/other	The variable lists for the third and fourth experiment.
./data/weka_models	Contains the Java byte file from the Weka model in addition to the raw text output from Weka.
./model	Contains the receiver and analysis python scripts.
./model/manlookup.py	Lookup a MAC address's OUI.
./model/receiver.py	Main receiver Python program. Run with the '-h' flag for more information.
./model/transpose.py	Swaps rows and columns of a csv file.
./resources	Contains a short text file with a Kismet response on netcat.
./scanner/wifiScanner.py	Main Python program used on the Raspberry Pi to scan for local Wi-Fi devices. Execute with `-h` flag for more information.

APPENDIX 2: TOTAL DEVICES ANALYSIS

Option	Selection
Data File	data/csv/NOrssi_total.csv
Weka File	data/arff/NOrssi_total.arff
Classifier	weka.classifiers.trees.M5P
Classifier Options	buildRegressionTree: False debug: False minNumInstances: 4.0 saveInstances: False unPruned: False useUnsmoothed: False
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka Output:

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     rssi_NOrssi_total
Instances:    6505
Attributes:   3
              Timeout
              Manual Count
Total_Devs
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

Timeout <= 165 :
|   Total_Devs<= 38.5 :
|   |   Total_Devs<= 24.5 :
|   |   |   Timeout <= 15 :
|   |   |   |   Total_Devs<= 15.5 : LM1 (59/20.942%)
|   |   |   |   Total_Devs>  15.5 : LM2 (213/18.063%)
|   |   |   Timeout >  15 :
|   |   |   |   Timeout <= 52.5 :
|   |   |   |   |   Total_Devs<= 16.5 :
|   |   |   |   |   |   Timeout <= 25 :
|   |   |   |   |   |   |   Total_Devs<= 12.5 : LM3 (21/43.28%)
|   |   |   |   |   |   |   Total_Devs>  12.5 : LM4 (53/11.39%)
|   |   |   |   |   |   Timeout >  25 :
|   |   |   |   |   |   |   Total_Devs<= 1.5 : LM5 (7/23.049%)
|   |   |   |   |   |   |   Total_Devs>  1.5 : LM6 (22/67.138%)
|   |   |   |   |   Total_Devs>  16.5 :
|   |   |   |   |   |   Total_Devs<= 22.5 : LM7 (279/69.359%)
|   |   |   |   |   |   Total_Devs>  22.5 : LM8 (108/82.599%)
|   |   |   |   Timeout >  52.5 :
|   |   |   |   |   Total_Devs<= 19.5 :
|   |   |   |   |   |   Total_Devs<= 5.5 : LM9 (19/103.702%)
|   |   |   |   |   |   Total_Devs>  5.5 :
|   |   |   |   |   |   |   Timeout <= 67.5 : LM10 (27/93.673%)
|   |   |   |   |   |   |   Timeout >  67.5 : LM11 (15/62.733%)
|   |   |   |   |   Total_Devs>  19.5 : LM12 (228/64.061%)
|   |   Total_Devs>  24.5 :
|   |   |   Timeout <= 37.5 :
|   |   |   |   Total_Devs<= 27.5 : LM13 (149/75.054%)
|   |   |   |   Total_Devs>  27.5 :
|   |   |   |   |   Total_Devs<= 31.5 : LM14 (308/43.42%)
|   |   |   |   |   Total_Devs>  31.5 : LM15 (205/38.142%)
|   |   |   Timeout >  37.5 :
|   |   |   |   Timeout <= 82.5 :
```

```
|   |   |   |   |   Total_Devs<= 32.5 : LM16 (572/52.401%)
|   |   |   |   |   Total_Devs>  32.5 : LM17 (201/30.877%)
|   |   |   |   Timeout >  82.5 :
|   |   |   |   |   Total_Devs<= 31.5 : LM18 (285/34.827%)
|   |   |   |   |   Total_Devs>  31.5 : LM19 (243/39.503%)
|   Total_Devs>  38.5 :
|   |   Timeout <= 135 :
|   |   |   Total_Devs<= 52.5 :
|   |   |   |   Timeout <= 60 :
|   |   |   |   |   Total_Devs<= 40.5 : LM20 (39/53.24%)
|   |   |   |   |   Total_Devs>  40.5 : LM21 (114/23.559%)
|   |   |   |   Timeout >  60 :
|   |   |   |   |   Total_Devs<= 42.5 :
|   |   |   |   |   |   Timeout <= 82.5 : LM22 (95/37.911%)
|   |   |   |   |   |   Timeout >  82.5 : LM23 (32/78.519%)
|   |   |   |   |   Total_Devs>  42.5 :
|   |   |   |   |   |   Timeout <= 82.5 : LM24 (347/51.397%)
|   |   |   |   |   |   Timeout >  82.5 : LM25 (99/20.303%)
|   |   |   Total_Devs>  52.5 :
|   |   |   |   Timeout <= 105 :
|   |   |   |   |   Timeout <= 82.5 : LM26 (40/44.571%)
|   |   |   |   |   Timeout >  82.5 : LM27 (117/12.165%)
|   |   |   |   Timeout >  105 : LM28 (141/16.522%)
|   |   Timeout >  135 :
|   |   |   Total_Devs<= 61.5 :
|   |   |   |   Total_Devs<= 53.5 :
|   |   |   |   |   Total_Devs<= 45.5 :
|   |   |   |   |   |   Total_Devs<= 43.5 : LM29 (10/0%)
|   |   |   |   |   |   Total_Devs>  43.5 : LM30 (3/0%)
|   |   |   |   |   Total_Devs>  45.5 : LM31 (48/14.42%)
|   |   |   |   Total_Devs>  53.5 :
|   |   |   |   |   Total_Devs<= 59.5 : LM32 (109/15.876%)
|   |   |   |   |   Total_Devs>  59.5 : LM33 (40/22.38%)
|   |   |   Total_Devs>  61.5 :
|   |   |   |   Total_Devs<= 68.5 : LM34 (146/30.537%)
|   |   |   |   Total_Devs>  68.5 : LM35 (41/26.187%)
Timeout >  165 :
|   Timeout <= 210 :
|   |   Total_Devs<= 56.5 : LM36 (208/30.771%)
|   |   Total_Devs>  56.5 :
|   |   |   Total_Devs<= 60.5 : LM37 (177/24.983%)
|   |   |   Total_Devs>  60.5 :
|   |   |   |   Total_Devs<= 68.5 : LM38 (328/26.035%)
|   |   |   |   Total_Devs>  68.5 : LM39 (116/24.427%)
|   Timeout >  210 :
|   |   Timeout <= 270 :
|   |   |   Total_Devs<= 55.5 : LM40 (736/121.149%)
|   |   |   Total_Devs>  55.5 :
|   |   |   |   Total_Devs<= 58.5 : LM41 (100/132.523%)
|   |   |   |   Total_Devs>  58.5 : LM42 (118/50.17%)
|   |   Timeout >  270 :
|   |   |   Total_Devs<= 50.5 :
|   |   |   |   Total_Devs<= 27.5 : LM43 (12/97.966%)
|   |   |   |   Total_Devs>  27.5 : LM44 (46/4.05%)
|   |   |   Total_Devs>  50.5 : LM45 (229/20.946%)

LM num: 1
Manual Count = 0.0056 * Timeout + 0.2838 * Total_Devs + 4.1691
LM num: 2
Manual Count = 0.0056 * Timeout + 0.0718 * Total_Devs + 6.6892
LM num: 3
Manual Count = -0.2077 * Timeout + 0.7198 * Total_Devs + 16.7136
LM num: 4
Manual Count = -0.2077 * Timeout + 0.1659 * Total_Devs + 23.3189
LM num: 5
Manual Count = -0.3177 * Timeout + 0.789 * Total_Devs + 15.1294
LM num: 6
Manual Count = -0.3177 * Timeout + 0.2183 * Total_Devs + 17.7224
LM num: 7
```

```
Manual Count = -0.0265 * Timeout + 0.0207 * Total_Devs + 9.9982
LM num: 8
Manual Count = -0.0265 * Timeout + 0.04 * Total_Devs + 11.5154
LM num: 9
Manual Count = -0.0221 * Timeout + 1.4928 * Total_Devs + 7.4259
LM num: 10
Manual Count = -0.0376 * Timeout + 0.2338 * Total_Devs + 18.2708
LM num: 11
Manual Count = -0.0471 * Timeout + 0.2338 * Total_Devs + 16.31
LM num: 12
Manual Count = 0.003 * Timeout + 0.0328 * Total_Devs + 16.7007
LM num: 13
Manual Count = 0.0003 * Timeout + 1.9749 * Total_Devs - 35.0662
LM num: 14
Manual Count = 0.0003 * Timeout + 0.0274 * Total_Devs + 18.243
LM num: 15
Manual Count = 0.0003 * Timeout + 0.0323 * Total_Devs + 19.3081
LM num: 16
Manual Count = 0.0007 * Timeout - 0.1086 * Total_Devs + 16.6076
LM num: 17
Manual Count = -0.1319 * Timeout - 0.2406 * Total_Devs + 27.2801
LM num: 18
Manual Count = 0.0036 * Timeout + 0.0133 * Total_Devs + 16.8968
LM num: 19
Manual Count = 0.0036 * Timeout + 0.0151 * Total_Devs + 18.3941
LM num: 20
Manual Count = -0.0008 * Timeout - 0.0401 * Total_Devs + 13.378
LM num: 21
Manual Count = -0.0008 * Timeout - 0.0192 * Total_Devs + 11.2449
LM num: 22
Manual Count = 0.0195 * Timeout - 0.0041 * Total_Devs + 7.4006
LM num: 23
Manual Count = -0.0902 * Timeout - 0.9607 * Total_Devs + 60.869
LM num: 24
Manual Count = -0.0019 * Timeout + 0.0944 * Total_Devs + 6.0044
LM num: 25
Manual Count = 0.0233 * Timeout - 0.0041 * Total_Devs + 6.4242
LM num: 26
Manual Count = 0.0162 * Timeout - 0.4467 * Total_Devs + 30.3311
LM num: 27
Manual Count = 0.0105 * Timeout - 0.029 * Total_Devs + 8.1154
LM num: 28
Manual Count = 0.0069 * Timeout + 0.114 * Total_Devs  + 1.8077
LM num: 29
Manual Count = 0.0027 * Timeout - 0.205 * Total_Devs + 24.0092
LM num: 30
Manual Count = 0.0027 * Timeout - 0.2598 * Total_Devs + 26.0451
LM num: 31
Manual Count = 0.0027 * Timeout - 0.0336 * Total_Devs + 15.9078
LM num: 32
Manual Count = 0.0027 * Timeout - 0.0721 * Total_Devs + 17.4067
LM num: 33
Manual Count = 0.0027 * Timeout + 0.0325 * Total_Devs + 12.2235
LM num: 34
Manual Count = 0.0027 * Timeout + 0.0259 * Total_Devs + 13.7943
LM num: 35
Manual Count = 0.0027 * Timeout + 0.3999 * Total_Devs - 10.9758
LM num: 36
Manual Count = 0.002 * Timeout + 0.1132 * Total_Devs + 9.0371
LM num: 37
Manual Count = 0.002 * Timeout + 0.0107 * Total_Devs + 16.0682
LM num: 38
Manual Count = 0.002 * Timeout + 0.0054 * Total_Devs + 17.7834
LM num: 39
Manual Count = 0.002 * Timeout - 0.1723 * Total_Devs + 29.6875
LM num: 40
Manual Count = 0.0049 * Timeout + 0.016 * Total_Devs + 11.4649
LM num: 41
Manual Count = 0.0049 * Timeout + 0.1845 * Total_Devs + 7.5273
```

```
LM num: 42
Manual Count = 0.0049 * Timeout + 0.3602 * Total_Devs + 3.2899
LM num: 43
Manual Count = 0.0127 * Timeout + 0.2933 * Total_Devs + 19.1066
LM num: 44
Manual Count = 0.0127 * Timeout + 0.0541 * Total_Devs + 24.5948
LM num: 45
Manual Count = 0.0127 * Timeout + 0.0192 * Total_Devs + 26.7176

Number of Rules : 45

Time taken to build model: 0.66 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7899
Mean absolute error                      2.9851
Root mean squared error                  4.2683
Relative absolute error                 52.6306 %
Root relative squared error             61.3398 %
Total Number of Instances             6505
```

APPENDIX 3: TOTAL DEVICES WITH RSSI FILTERING ANALYSIS

Option	Selection
Data File	data/csv/rssi_total.csv
Weka File	data/arff/rssi_total.arff
Classifier	weka.classifiers.trees.M5P
Classifier Options	buildRegressionTree: False debug: False minNumInstances: 4.0 saveInstances: False unPruned: False useUnsmoothed: False
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka Output:

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     rssi_rssi_total
Instances:    4018
Attributes:   4
              Timeout
              RSSI
              Manual Count
Total_Devs
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

Timeout <= 15 :
|   Total_Devs<= 2.5 :
|   |   RSSI <= 55 : LM1 (165/20.229%)
|   |   RSSI >  55 :
|   |   |   Total_Devs<= 1.5 :
|   |   |   |   Total_Devs<= 0.5 : LM2 (30/61.997%)
|   |   |   |   Total_Devs>  0.5 : LM3 (89/31.205%)
|   |   |   Total_Devs>  1.5 : LM4 (45/33.521%)
|   Total_Devs>  2.5 :
|   |   RSSI <= 77.5 :
|   |   |   RSSI <= 67.5 : LM5 (79/6.017%)
|   |   |   RSSI >  67.5 : LM6 (87/38.622%)
|   |   RSSI >  77.5 :
|   |   |   RSSI <= 85 : LM7 (81/0%)
|   |   |   RSSI >  85 : LM8 (280/22.941%)
Timeout >  15 :
|   Timeout <= 82.5 :
|   |   Timeout <= 52.5 :
|   |   |   RSSI <= 87.5 :
|   |   |   |   Total_Devs<= 3.5 :
|   |   |   |   |   RSSI <= 55 :
|   |   |   |   |   |   RSSI <= 47.5 : LM9 (95/43.721%)
|   |   |   |   |   |   RSSI >  47.5 : LM10 (57/61.292%)
|   |   |   |   |   RSSI >  55 :
|   |   |   |   |   |   Total_Devs<= 1.5 : LM11 (51/71.559%)
|   |   |   |   |   |   Total_Devs>  1.5 : LM12 (57/46.408%)
|   |   |   |   Total_Devs>  3.5 :
|   |   |   |   |   RSSI <= 72.5 :
|   |   |   |   |   |   Timeout <= 25 :
|   |   |   |   |   |   |   Total_Devs<= 9.5 : LM13 (21/24.153%)
|   |   |   |   |   |   |   Total_Devs>  9.5 : LM14 (57/0%)
|   |   |   |   |   |   Timeout >  25 :
|   |   |   |   |   |   |   Total_Devs<= 19.5 : LM15 (144/31.583%)
|   |   |   |   |   |   |   Total_Devs>  19.5 : LM16 (37/27.829%)
```

```
|   |   |   |   |   RSSI >  72.5 :
|   |   |   |   |   |   RSSI <= 77.5 : LM17 (68/0%)
|   |   |   |   |   |   RSSI >  77.5 : LM18 (184/21.8%)
|   |   |   RSSI >  87.5 :
|   |   |   |   Timeout <= 37.5 :
|   |   |   |   |   Total_Devs<= 27.5 : LM19 (21/46.145%)
|   |   |   |   |   Total_Devs>  27.5 : LM20 (77/17.756%)
|   |   |   |   Timeout >  37.5 :
|   |   |   |   |   Total_Devs<= 37.5 : LM21 (20/77.312%)
|   |   |   |   |   Total_Devs>  37.5 : LM22 (78/31.955%)
|   |   Timeout >  52.5 :
|   |   |   RSSI <= 57.5 : LM23 (307/18.945%)
|   |   |   RSSI >  57.5 :
|   |   |   |   RSSI <= 77.5 : LM24 (191/111.459%)
|   |   |   |   RSSI >  77.5 :
|   |   |   |   |   RSSI <= 85 :
|   |   |   |   |   |   Total_Devs<= 27.5 :
|   |   |   |   |   |   |   Total_Devs<= 18.5 : LM25 (8/33.698%)
|   |   |   |   |   |   |   Total_Devs>  18.5 : LM26 (31/18.687%)
|   |   |   |   |   |   Total_Devs>  27.5 : LM27 (59/0%)
|   |   |   |   |   RSSI >  85 : LM28 (85/25.328%)
|   Timeout >  82.5 :
|   |   Timeout <= 150 :
|   |   |   RSSI <= 67.5 :
|   |   |   |   RSSI <= 55 :
|   |   |   |   |   Timeout <= 105 : LM29 (160/39.197%)
|   |   |   |   |   Timeout >  105 : LM30 (176/39.235%)
|   |   |   |   RSSI >  55 : LM31 (231/32.358%)
|   |   |   RSSI >  67.5 :
|   |   |   |   RSSI <= 75 : LM32 (130/54.782%)
|   |   |   |   RSSI >  75 :
|   |   |   |   |   Timeout <= 105 :
|   |   |   |   |   |   RSSI <= 85 :
|   |   |   |   |   |   |   Total_Devs<= 42.5 :
|   |   |   |   |   |   |   |   Total_Devs<= 16 :
|   |   |   |   |   |   |   |   |   Total_Devs<= 10.5 : LM33 (4/72.518%)
|   |   |   |   |   |   |   |   |   Total_Devs>  10.5 : LM34 (9/0%)
|   |   |   |   |   |   |   |   Total_Devs>  16 :
|   |   |   |   |   |   |   |   |   Total_Devs<= 35.5 : LM35 (27/0%)
|   |   |   |   |   |   |   |   |   Total_Devs>  35.5 : LM36 (8/19.987%)
|   |   |   |   |   |   |   Total_Devs>  42.5 : LM37 (29/0%)
|   |   |   |   |   |   RSSI >  85 : LM38 (79/34.932%)
|   |   |   |   |   Timeout >  105 :
|   |   |   |   |   |   RSSI <= 85 :
|   |   |   |   |   |   |   Total_Devs<= 50 :
|   |   |   |   |   |   |   |   Total_Devs<= 22 : LM39 (5/48.861%)
|   |   |   |   |   |   |   |   Total_Devs>  22 : LM40 (15/0%)
|   |   |   |   |   |   |   Total_Devs>  50 : LM41 (78/16.267%)
|   |   |   |   |   |   RSSI >  85 :
|   |   |   |   |   |   |   Total_Devs<= 50.5 : LM42 (37/44.619%)
|   |   |   |   |   |   |   Total_Devs>  50.5 : LM43 (33/7.355%)
|   |   Timeout >  150 :
|   |   |   Total_Devs<= 17.5 :
|   |   |   |   Total_Devs<= 1.5 :
|   |   |   |   |   Total_Devs<= 0.5 : LM44 (32/90.667%)
|   |   |   |   |   Total_Devs>  0.5 : LM45 (86/16.644%)
|   |   |   |   Total_Devs>  1.5 :
|   |   |   |   |   RSSI <= 55 : LM46 (54/7.77%)
|   |   |   |   |   RSSI >  55 : LM47 (106/29.394%)
|   |   |   Total_Devs>  17.5 :
|   |   |   |   RSSI <= 85 :
|   |   |   |   |   Total_Devs<= 35 : LM48 (48/19.462%)
|   |   |   |   |   Total_Devs>  35 :
|   |   |   |   |   |   Total_Devs<= 45.5 : LM49 (16/11.912%)
|   |   |   |   |   |   Total_Devs>  45.5 : LM50 (58/8.853%)
|   |   |   |   RSSI >  85 :
|   |   |   |   |   Total_Devs<= 42.5 :
|   |   |   |   |   |   Total_Devs<= 33.5 : LM51 (9/0%)
|   |   |   |   |   |   Total_Devs>  33.5 : LM52 (11/0%)
```

```
|   |   |   |   |   Total_Devs>  42.5 : LM53 (73/0%)

LM num: 1
Manual Count = 0.0004*Timeout + 0.0512*RSSI + 0.0615*Total_Devs + 7.1718

LM num: 2
Manual Count = 0.0004*Timeout - 0.1571*RSSI + 0.6089*Total_Devs + 17.0094

LM num: 3
Manual Count =0.0004*Timeout - 0.0316*RSSI + 0.3575*Total_Devs + 9.9018

LM num: 4
Manual Count = 0.0004*Timeout - 0.0301*RSSI + 0.2938*Total_Devs + 10.0962

LM num: 5
Manual Count = 0.0004*Timeout - 0.0474*RSSI + 0.0024*Total_Devs + 10.2897

LM num: 6
Manual Count = 0.0004*Timeout + 0.016 *RSSI + 0.0024*Total_Devs + 7.4482

LM num: 7
Manual Count = 0.0004*Timeout - 0.0011*RSSI - 0.0028*Total_Devs + 8.2321

LM num: 8
Manual Count = 0.0004*Timeout - 0.0498*RSSI + 0.0007*Total_Devs + 13.1562

LM num: 9
Manual Count = 0.0344*Timeout - 1.7857*RSSI + 0.1386*Total_Devs + 88.6991

LM num: 10
Manual Count = 0.0344*Timeout + 0.0624*RSSI + 1.03*Total_Devs + 15.49

LM num: 11
Manual Count = 0.1909*Timeout - 0.2782*RSSI + 3.7793*Total_Devs + 20.984

LM num: 12
Manual Count = 0.3406*Timeout - 0.0958*RSSI + 0.4587*Total_Devs + 10.4704

LM num: 13
Manual Count = 0.0411*Timeout + 0.1137*RSSI - 0.2484*Total_Devs + 5.7333

LM num: 14
Manual Count = 0.0411*Timeout - 0.0308*RSSI + 0.0113*Total_Devs + 13.2333

LM num: 15
Manual Count = 0.167*Timeout - 0.1501*RSSI + 0.0265*Total_Devs + 18.8561

LM num: 16
Manual Count = -0.0213*Timeout - 0.0834*RSSI + 0.2172*Total_Devs + 18.6564


LM num: 17
Manual Count = 0.0194*Timeout - 0.018*RSSI + 0.0007*Total_Devs + 11.3846

LM num: 18
Manual Count = 0.0194*Timeout - 0.5147*RSSI - 0.0827*Total_Devs + 55.9635

LM num: 19
Manual Count = 0.0719*Timeout - 0.0124*RSSI + 0.4219*Total_Devs + 1.0075

LM num: 20
Manual Count = 0.0719*Timeout - 0.0124*RSSI + 0.0988*Total_Devs + 10.5965

LM num: 21
Manual Count = 0.0719*Timeout - 0.0124*RSSI + 0.1908*Total_Devs + 12.9244

LM num: 22
Manual Count = 0.0719*Timeout - 0.0124*RSSI - 0.0844*Total_Devs + 23.9392
```

```
LM num: 23
Manual Count = 0.0023*Timeout - 0.0059*RSSI + 0.0587*Total_Devs + 7.441

LM num: 24
Manual Count = -0.5069*Timeout + 0.3082*RSSI + 0.0308*Total_Devs + 25.142

LM num: 25
Manual Count = 0.0015*Timeout - 0.0196*RSSI + 0.1364*Total_Devs + 6.9396

LM num: 26
Manual Count = 0.0015*Timeout - 0.0196*RSSI + 0.0536*Total_Devs + 8.1138

LM num: 27
Manual Count = 0.0015*Timeout - 0.0196*RSSI + 0.0392*Total_Devs + 8.0817

LM num: 28
Manual Count = 0.0015*Timeout - 0.0175*RSSI + 0.1375*Total_Devs + 6.1988

LM num: 29
Manual Count = 0.0118*Timeout + 0.1669*RSSI + 0.5177*Total_Devs + 4.3689

LM num: 30
Manual Count = 0.011*Timeout - 0.0452*RSSI + 0.2863*Total_Devs + 17.4948

LM num: 31
Manual Count = -0.098*Timeout - 0.4128*RSSI + 0.053*Total_Devs + 46.6168

LM num: 32
Manual Count = -0.0125*Timeout - 0.0243*RSSI + 0.0609*Total_Devs + 23.2299

LM num: 33
Manual Count = -0.021*Timeout - 0.0973*RSSI + 0.3663*Total_Devs + 24.7871

LM num: 34
Manual Count = -0.021*Timeout - 0.0973*RSSI + 0.2209*Total_Devs + 26.2666

LM num: 35
Manual Count = -0.021*Timeout - 0.0973*RSSI + 0.0543*Total_Devs + 28.5611

LM num: 36
Manual Count = -0.021*Timeout - 0.0973*RSSI + 0.06*Total_Devs + 28.5483

LM num: 37
Manual Count = -0.021*Timeout - 0.0973*RSSI + 0.0324*Total_Devs+30.0154

LM num: 38
Manual Count = -0.021*Timeout - 0.0954*RSSI + 0.0727*Total_Devs + 23.2189

LM num: 39
Manual Count = -0.02*Timeout + 0.0609*RSSI + 0.096*Total_Devs + 5.34

LM num: 40
Manual Count = -0.02*Timeout + 0.0609*RSSI + 0.0445*Total_Devs + 6.3231

LM num: 41
Manual Count = -0.02*Timeout + 0.0609*RSSI + 0.0552*Total_Devs + 5.5975

LM num: 42
Manual Count = -0.02*Timeout + 0.0839*RSSI + 0.0981*Total_Devs + 6.9383

LM num: 43
Manual Count = -0.02*Timeout + 0.0839*RSSI - 0.0054*Total_Devs + 11.291

LM num: 44
Manual Count = -0.0003*Timeout - 0.2144*RSSI + 0.7926*Total_Devs + 22.8403

LM num: 45
Manual Count = -0.0003*Timeout - 0.0434*RSSI + 0.3874*Total_Devs + 17.2519
```

```
LM num: 46
Manual Count = -0.0003*Timeout + 0.0284*RSSI + 0.2343*Total_Devs + 13.8563

LM num: 47
Manual Count = -0.0003*Timeout - 0.1193*RSSI + 0.158*Total_Devs + 21.006

LM num: 48
Manual Count = -0.0003*Timeout - 0.0733*RSSI - 0.002*Total_Devs + 18.8518

LM num: 49
Manual Count = -0.0003*Timeout - 0.0259*RSSI - 0.0252*Total_Devs + 14.9193

LM num: 50
Manual Count = -0.0003*Timeout - 0.0259*RSSI + 0.0037*Total_Devs + 14.1946

LM num: 51
Manual Count = -0.0003*Timeout - 0.0041*RSSI + 0.0218*Total_Devs + 12.2948

LM num: 52
Manual Count = -0.0003*Timeout - 0.0041*RSSI + 0.0205*Total_Devs + 12.5108

LM num: 53
Manual Count = -0.0003*Timeout - 0.0041*RSSI + 0.0004*Total_Devs + 13.3602

Number of Rules : 53

Time taken to build model: 0.18 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8953
Mean absolute error                      1.2259
Root mean squared error                  2.2455
Relative absolute error                 29.9499 %
Root relative squared error             44.864  %
Total Number of Instances             4018
```

APPENDIX 4: MAC FILTERING WITHOUT RSSI REPTREE

Option	Selection
Data File	data/csv/NOrssi_man_breakdown.csv
Weka File	data/arff/NOrssi_man_breakdown.arff
Classifier	weka.classifiers.trees.REPTree
Classifier Options	debug: False maxDepth: 6 minNum: 2.0 minVarianceProp: 0.001 noPruning: False numFolds: 3 seed: 1
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka output:

```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 6
Relation:     rssi_NOrssi_man_breakdown
Instances:    6505
Attributes:   92
              Timeout
              Manual Count
              Apple
              Apple Inc
              Apple, Inc
              Apple, Inc.
Askey Computer Corp
ASUSTek COMPUTER INC.
Azurewave Technologies., inc.
Belkin International Inc.
              Billion Electric Co. Ltd.
              Billion Electric Co., Ltd.
              BROTHER INDUSTRIES, LTD.
              CABLE TELEVISION LABORATORIES, INC.
              Cisco
              CISCO SYSTEMS, INC.
              COMPAL INFORMATION (KUNSHAN) CO., LTD.
              D-Link International
              D-LINK INTERNATIONAL PTE LIMITED
              Dell Inc
DrayTek Corp.
Elitegroup Computer System Co.
              ELITEGROUP COMPUTER SYSTEM CO., LTD.
Enerpoint Ltd.
              FUJI-XEROX CO. LTD.
GainSpan Corp.
              GIGA-BYTE TECHNOLOGY CO.,LTD.
              Hewlett Packard
              Hewlett-Packard Company
              Hon Hai Precision Ind. Co.,Ltd.
              Hon Hai Precision Ind.Co.Ltd
              HTC Corporation
              Huawei Technologies Co., Ltd
              IBM
              IBM Corp
              ICP Electronics Inc.
              Intel Corporate
              KYOCERA Document Solutions Inc.
              Lexmark International Inc.
              LG Electronics
Liteon Tech Corp.
Liteon Technology Corporation
              Micro-Star Int'l Co, Ltd
```

```
              Microsoft
              Microsoft Corporation
              Motorola Mobility LLC
              Murata ManufactuaringCo.,Ltd.
              Murata Manufacturing Co., Ltd.
              Murata Manufacturing Co.,Ltd.
              NETCOMM LIMITED
              NETGEAR
NetgearInc
Netgear Inc.
              NETGEAR INC.,
              Nokia Corporation
OnePlus Tech (Shenzhen) Ltd
              Panasonic Automotive Systems Company of America
              Paragon Technologies Inc.
              PEGATRON CORPORATION
Pittasoft
              RICOH COMPANY LTD.
              Routerboard.com
              Samsung ElecCo.,Ltd
              Samsung Electro Mechanics co., LTD.
              Samsung Electro Mechanics co.,LTD.
              SAMSUNG ELECTRO-MECHANICS
              SAMSUNG ELECTRO-MECHANICS CO., LTD.
              Samsung Electronics Co., LTD
              Samsung Electronics Co.,Ltd
              Samsung Electronics ITS, Printer division
              SEIKO EPSON CORPORATION
              Seoul Commtech
              Shenzhen JSR Technology Co.,Ltd.
              Shenzhen Ogemray Technology Co., Ltd.
              Shenzhen XinKingBrand enterprises Co.,Ltd
              Sierra Wireless Inc
              SIRONA DENTAL SYSTEMS GmbH & Co. KG
              Sony Mobile Communications AB
SparkLAN Communications, Inc.
              TCT mobile limited
              TP-LINK TECHNOLOGIES CO., LTD.
              TP-LINK TECHNOLOGIES CO.,LTD.
Ubiquiti Networks
Ubiquiti Networks, INC
              UNKNOWN
Verifone, INC.
              WISOL
Wistron Corporation
WistronInfoComm(ZhongShan) Corporation
              Zebra Technologies Inc
zte corporation
Zultys Technologies
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

Apple < 13.5
|   Billion Electric Co. Ltd. < 1.5
|   |   NETGEAR INC., < 1.5
|   |   |   Timeout < 270
|   |   |   |   Apple < 1.5
|   |   |   |   |   UNKNOWN < 0.5 : 3.33 (19/14.76) [11/88.08]
|   |   |   |   |   UNKNOWN >= 0.5 : 10.71 (27/24.89) [18/34.46]
|   |   |   |   Apple >= 1.5
|   |   |   |   |   DrayTek Corp. < 1.5 : 17.74 (1492/13.5) [745/11.35]
|   |   |   |   |   DrayTek Corp. >= 1.5 : 11.5 (118/2.44) [62/2.36]
|   |   |   Timeout >= 270 : 30.33 (35/26.65) [14/0.78]
|   |   NETGEAR INC., >= 1.5
```

```
|   |   |   Timeout < 52.5
|   |   |   |   Hon Hai Precision Ind.Co.Ltd< 0.5
|   |   |   |   |   Timeout < 15 : 7.83 (121/1.62) [70/1.58]
|   |   |   |   |   Timeout >= 15 : 11.99 (256/30.07) [138/30.47]
|   |   |   |   Hon Hai Precision Ind.Co.Ltd>= 0.5
|   |   |   |   |   Apple < 2.5 : 6.14 (33/0.42) [17/0.65]
|   |   |   |   |   Apple >= 2.5 : 6.62 (51/0.32) [36/0.34]
|   |   |   Timeout >= 52.5
|   |   |   |   ASUSTek COMPUTER INC. < 0.5
|   |   |   |   |   DrayTek Corp. < 0.5 : 28 (10/1.89) [5/2.25]
|   |   |   |   |   DrayTek Corp. >= 0.5 : 26 (8/0) [0/0]
|   |   |   |   ASUSTek COMPUTER INC. >= 0.5
|   |   |   |   |   NETCOMM LIMITED < 1.5 : 17.61 (55/9.5) [34/10.21]
|   |   |   |   |   NETCOMM LIMITED >= 1.5 : 25.82 (7/0) [4/1]
|   Billion Electric Co. Ltd. >= 1.5
|   |   Samsung Electro Mechanics co., LTD. < 1.5
|   |   |   Enerpoint Ltd. < 1.5
|   |   |   |   Azurewave Technologies., inc.< 0.5
|   |   |   |   |   Timeout < 37.5 : 14.52 (125/36.23) [44/30.21]
|   |   |   |   |   Timeout >= 37.5 : 9.14 (1077/12.75) [536/14.86]
|   |   |   |   Azurewave Technologies., inc.>= 0.5
|   |   |   |   |   UNKNOWN < 7.5 : 30 (4/0) [3/0]
|   |   |   |   |   UNKNOWN >= 7.5 : 30.55 (6/0.22) [5/0.31]
|   |   |   Enerpoint Ltd. >= 1.5
|   |   |   |   Timeout < 75
|   |   |   |   |   NETGEAR < 1.5 : 18.21 (38/2.36) [20/1.5]
|   |   |   |   |   NETGEAR >= 1.5 : 12.93 (99/2.18) [51/2.6]
|   |   |   |   Timeout >= 75
|   |   |   |   |   Belkin International Inc. < 1.5 : 17.7 (27/0.89)
[16/0.94]
|   |   |   |   |   Belkin International Inc. >= 1.5 : 21 (17/0) [8/0]
|   |   Samsung Electro Mechanics co., LTD. >= 1.5
|   |   |   HTC Corporation < 0.5
|   |   |   |   NETGEAR < 2.5 : 17.15 (44/2.54) [21/2.91]
|   |   |   |   NETGEAR >= 2.5
|   |   |   |   |   Timeout < 112.5 : 10 (5/0) [2/0]
|   |   |   |   |   Timeout >= 112.5 : 14.36 (95/1.77) [42/2.05]
|   |   |   HTC Corporation >= 0.5
|   |   |   |   Azurewave Technologies., inc.< 0.5 : 18.82 (66/2.95)
[37/0.95]
|   |   |   |   Azurewave Technologies., inc.>= 0.5 : 31 (6/0) [2/0]
Apple >= 13.5
|   COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5
|   |   ASUSTek COMPUTER INC. < 1.5
|   |   |   Murata Manufacturing Co.,Ltd. < 0.5
|   |   |   |   IBM Corp < 1.5 : 30.82 (134/0.3) [64/7.62]
|   |   |   |   IBM Corp >= 1.5
|   |   |   |   |   HTC Corporation < 1.5 : 32 (5/0) [2/0]
|   |   |   |   |   HTC Corporation >= 1.5 : 26.27 (26/6.38) [18/8.7]
|   |   |   Murata Manufacturing Co.,Ltd. >= 0.5
|   |   |   |   Murata ManufactuaringCo.,Ltd. < 0.5
|   |   |   |   |   Azurewave Technologies., inc.< 0.5 : 32.43 (21/0.54)
[14/0.53]
|   |   |   |   |   Azurewave Technologies., inc.>= 0.5 : 31.18 (23/0.17)
[11/0.67]
|   |   |   |   Murata ManufactuaringCo.,Ltd. >= 0.5 : 34 (35/0) [15/0]
|   |   ASUSTek COMPUTER INC. >= 1.5
|   |   |   Apple, Inc< 2.5
|   |   |   |   Samsung Electronics ITS, Printer division < 0.5
|   |   |   |   |   Apple < 16.5 : 31 (8/0) [3/0]
|   |   |   |   |   Apple >= 16.5 : 30 (2/1) [0/0]
|   |   |   |   Samsung Electronics ITS, Printer division >= 0.5
|   |   |   |   |   Apple < 15.5 : 30 (6/0) [1/0]
|   |   |   |   |   Apple >= 15.5 : 29 (13/0) [7/0]
|   |   |   Apple, Inc>= 2.5
|   |   |   |   Timeout < 210
|   |   |   |   |   Apple Inc< 0.5 : 14 (2/0) [0/0]
|   |   |   |   |   Apple Inc>= 0.5 : 17 (7/1.96) [5/4.9]
|   |   |   |   Timeout >= 210
```

```
|   |   |   |   |   IBM Corp < 2 : 22 (27/0) [11/0]
|   |   |   |   |   IBM Corp >= 2 : 24.54 (10/1.44) [3/0.36]
|   COMPAL INFORMATION (KUNSHAN) CO., LTD. >= 0.5
|   |   Intel Corporate < 2.5
|   |   |   Timeout < 210
|   |   |   |   Huawei Technologies Co., Ltd < 1 : 16.98 (110/1.47)
[50/1.91]
|   |   |   |   Huawei Technologies Co., Ltd >= 1 : 20 (5/0) [2/0]
|   |   |   Timeout >= 210 : 20.3 (15/0.22) [8/0.19]
|   |   Intel Corporate >= 2.5
|   |   |   Timeout < 165
|   |   |   |   SEIKO EPSON CORPORATION < 0.5 : 19 (12/0) [3/0]
|   |   |   |   SEIKO EPSON CORPORATION >= 0.5 : 18 (3/0) [1/0]
|   |   |   Timeout >= 165
|   |   |   |   DrayTek Corp. < 0.5
|   |   |   |   |   HTC Corporation < 1.5 : 22 (3/0) [1/0]
|   |   |   |   |   HTC Corporation >= 1.5 : 20 (4/0) [2/0]
|   |   |   |   DrayTek Corp. >= 0.5 : 20 (24/0) [7/0]

Size of the tree : 95

Time taken to build model: 0.31 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8745
Mean absolute error                      2.3385
Root mean squared error                  3.374
Relative absolute error                 41.2305 %
Root relative squared error             48.4882 %
Total Number of Instances             6505
```

APPENDIX 5: MAC FILTERING WITHOUT RSSI LINEAR REGRESSION

Option	Selection
Data File	data/csv/NOrssi_man_breakdown.csv
Weka File	data/arff/NOrssi_man_breakdown.arff
Classifier	weka.classifiers.functions.LinearRegression
Classifier Options	attributeSelectionMethod: M5 Method debug: False eliminateColinearAttributes: True ridge: 1.0E-8
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka output:

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     rssi_NOrssi_man_breakdown
Instances:    6505
Attributes:   92
              Timeout
              Manual Count
              Apple
              Apple Inc
              Apple, Inc
              Apple, Inc.
Askey Computer Corp
ASUSTek COMPUTER INC.
Azurewave Technologies., inc.
Belkin International Inc.
              Billion Electric Co. Ltd.
              Billion Electric Co., Ltd.
              BROTHER INDUSTRIES, LTD.
              CABLE TELEVISION LABORATORIES, INC.
              Cisco
              CISCO SYSTEMS, INC.
              COMPAL INFORMATION (KUNSHAN) CO., LTD.
              D-Link International
              D-LINK INTERNATIONAL PTE LIMITED
              Dell Inc
DrayTek Corp.
Elitegroup Computer System Co.
              ELITEGROUP COMPUTER SYSTEM CO., LTD.
Enerpoint Ltd.
              FUJI-XEROX CO. LTD.
GainSpan Corp.
              GIGA-BYTE TECHNOLOGY CO.,LTD.
              Hewlett Packard
              Hewlett-Packard Company
              Hon Hai Precision Ind. Co.,Ltd.
              Hon Hai Precision Ind.Co.Ltd
              HTC Corporation
              Huawei Technologies Co., Ltd
              IBM
              IBM Corp
              ICP Electronics Inc.
              Intel Corporate
              KYOCERA Document Solutions Inc.
              Lexmark International Inc.
              LG Electronics
Liteon Tech Corp.
Liteon Technology Corporation
              Micro-Star Int'l Co, Ltd
              Microsoft
              Microsoft Corporation
              Motorola Mobility LLC
```

```
              Murata ManufactuaringCo.,Ltd.
              Murata Manufacturing Co., Ltd.
              Murata Manufacturing Co.,Ltd.
              NETCOMM LIMITED
              NETGEAR
NetgearInc
Netgear Inc.
              NETGEAR INC.,
              Nokia Corporation
OnePlus Tech (Shenzhen) Ltd
              Panasonic Automotive Systems Company of America
              Paragon Technologies Inc.
              PEGATRON CORPORATION
Pittasoft
              RICOH COMPANY LTD.
              Routerboard.com
              Samsung ElecCo.,Ltd
              Samsung Electro Mechanics co., LTD.
              Samsung Electro Mechanics co.,LTD.
              SAMSUNG ELECTRO-MECHANICS
              SAMSUNG ELECTRO-MECHANICS CO., LTD.
              Samsung Electronics Co., LTD
              Samsung Electronics Co.,Ltd
              Samsung Electronics ITS, Printer division
              SEIKO EPSON CORPORATION
              Seoul Commtech
              Shenzhen JSR Technology Co.,Ltd.
              Shenzhen Ogemray Technology Co., Ltd.
              Shenzhen XinKingBrand enterprises Co.,Ltd
              Sierra Wireless Inc
              SIRONA DENTAL SYSTEMS GmbH & Co. KG
              Sony Mobile Communications AB
SparkLAN Communications, Inc.
              TCT mobile limited
              TP-LINK TECHNOLOGIES CO., LTD.
              TP-LINK TECHNOLOGIES CO.,LTD.
Ubiquiti Networks
Ubiquiti Networks, INC
              UNKNOWN
Verifone, INC.
              WISOL
Wistron Corporation
WistronInfoComm(ZhongShan) Corporation
              Zebra Technologies Inc
zte corporation
Zultys Technologies
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

Manual Count =

      0.0213 * Timeout +
      0.4192 * Apple +
     -0.1845 * Apple, Inc +
     -0.2756 * Apple, Inc. +
     -0.63   * ASUSTek COMPUTER INC. +
      3.067  * Azurewave Technologies., inc. +
      0.6106 * Belkin International Inc. +
     -1.5596 * Billion Electric Co. Ltd. +
     -1.677  * Billion Electric Co., Ltd. +
     -2.0262 * CABLE TELEVISION LABORATORIES, INC. +
      0.1598 * CISCO SYSTEMS, INC. +
     -0.7869 * COMPAL INFORMATION (KUNSHAN) CO., LTD. +
     -5.9152 * D-Link International +
      0.3058 * D-LINK INTERNATIONAL PTE LIMITED +
```

```
      -1.3297 * DrayTek Corp. +
      -3.0105 * Elitegroup Computer System Co. +
       1.2444 * Enerpoint Ltd. +
      -0.5526 * FUJI-XEROX CO. LTD. +
      -5.6587 * GainSpan Corp. +
      -0.3072 * GIGA-BYTE TECHNOLOGY CO.,LTD. +
       0.1588 * Hewlett Packard +
      -0.1817 * Hewlett-Packard Company +
      -1.5455 * Hon Hai Precision Ind. Co.,Ltd. +
      -8.7087 * Hon Hai Precision Ind.Co.Ltd +
       0.797  * HTC Corporation +
      -0.7436 * IBM Corp +
      -0.3998 * ICP Electronics Inc. +
      -0.3778 * Intel Corporate +
       1.1606 * KYOCERA Document Solutions Inc. +
      -0.4476 * Lexmark International Inc. +
       1.9445 * LG Electronics +
      -0.2242 * Liteon Tech Corp. +
       0.4552 * Liteon Technology Corporation +
       3.0383 * Microsoft +
       0.684  * Microsoft Corporation +
      -0.996  * Motorola Mobility LLC +
       1.3765 * Murata ManufactuaringCo.,Ltd. +
       5.0672 * Murata Manufacturing Co., Ltd. +
       1.0073 * Murata Manufacturing Co.,Ltd. +
       3.0664 * NETCOMM LIMITED +
      -1.0237 * NETGEAR +
      -2.0779 * NetgearInc +
       1.4819 * Netgear Inc. +
      -0.1108 * NETGEAR INC., +
      -1.6923 * Nokia Corporation +
      -7.616  * OnePlus Tech (Shenzhen) Ltd +
      -3.8433 * Paragon Technologies Inc. +
      -1.3266 * PEGATRON CORPORATION +
      -0.5364 * Pittasoft +
      -2.0037 * Samsung ElecCo.,Ltd +
      -0.4367 * Samsung Electro Mechanics co., LTD. +
      -0.4729 * Samsung Electro Mechanics co.,LTD. +
      -2.8635 * SAMSUNG ELECTRO-MECHANICS +
      -2.5956 * SAMSUNG ELECTRO-MECHANICS CO., LTD. +
      -4.3559 * Samsung Electronics Co., LTD +
      -0.8508 * Samsung Electronics Co.,Ltd +
      -0.8579 * Samsung Electronics ITS, Printer division +
       0.3427 * SEIKO EPSON CORPORATION +
      -4.7025 * Seoul Commtech +
       2.2399 * Shenzhen JSR Technology Co.,Ltd. +
       1.6489 * Shenzhen Ogemray Technology Co., Ltd. +
      -6.0489 * SIRONA DENTAL SYSTEMS GmbH & Co. KG +
       3.1828 * Sony Mobile Communications AB +
       0.8365 * SparkLAN Communications, Inc. +
      -3.8369 * TCT mobile limited +
      -2.4783 * TP-LINK TECHNOLOGIES CO., LTD. +
      -1.2001 * TP-LINK TECHNOLOGIES CO.,LTD. +
       0.5265 * Ubiquiti Networks +
       4.0048 * Ubiquiti Networks, INC +
       0.3884 * UNKNOWN +
      -1.7403 * Verifone, INC. +
       0.5142 * WISOL +
       1.9541 * Wistron Corporation +
       9.0821 * Zultys Technologies +
      14.0161

Time taken to build model: 0.64 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.865
Mean absolute error                      2.6166
```

```
Root mean squared error                  3.4913
Relative absolute error                 46.133  %
Root relative squared error             50.1736 %
Total Number of Instances             6505
```

APPENDIX 6: MAC AND RSSI FILTERING REPTREE

Option	Selection
Data File	data/csv/rssi_man_breakdown.csv
Weka File	data/arff/rssi_man_breakdown.arff
Classifier	weka.classifiers.trees.REPTree
Classifier Options	debug: False maxDepth: 6 minNum: 2.0 minVarianceProp: 0.001 noPruning: False numFolds: 3 seed: 1
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka output:

```
=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L 6
Relation:     rssi_rssi_man_breakdown
Instances:    4018
Attributes:   96
              Timeout
              RSSI
              Manual Count
              Apple
              Apple Inc
              Apple, Inc
              Apple, Inc.
Askey Computer Corp
ASRock Incorporation
ASUSTek COMPUTER INC.
Azurewave Technologies., inc.
Belkin International Inc.
              Billion Electric Co. Ltd.
              Billion Electric Co., Ltd.
              BROTHER INDUSTRIES, LTD.
              CABLE TELEVISION LABORATORIES, INC.
              Cisco
              CISCO SYSTEMS, INC.
              COMPAL INFORMATION (KUNSHAN) CO., LTD.
              D-Link International
              D-LINK INTERNATIONAL PTE LIMITED
              Dell Inc
DrayTek Corp.
Elitegroup Computer System Co.
              ELITEGROUP COMPUTER SYSTEM CO., LTD.
Enerpoint Ltd.
              FUJI-XEROX CO. LTD.
GainSpan Corp.
              GIGA-BYTE TECHNOLOGY CO.,LTD.
              Hewlett Packard
              Hewlett-Packard Company
              Hon Hai Precision Ind. Co.,Ltd.
              Hon Hai Precision Ind.Co.Ltd
              HTC Corporation
              Huawei Technologies Co., Ltd
              IBM
              IBM Corp
              ICP Electronics Inc.
              Intel Corporate
              KYOCERA Document Solutions Inc.
              Lexmark International Inc.
              LG Electronics
Liteon Tech Corp.
```

```
Liteon Technology Corporation
              Micro-Star Int'l Co, Ltd
              Microsoft
              Microsoft Corporation
              Motorola Mobility LLC
              Murata ManufactuaringCo.,Ltd.
              Murata Manufacturing Co., Ltd.
              Murata Manufacturing Co.,Ltd.
              NEC Platforms, Ltd
              NETCOMM LIMITED
              NETGEAR
NetgearInc
Netgear Inc.
              NETGEAR INC.,
              Nintendo Co., Ltd.
              Nokia Corporation
OnePlus Tech (Shenzhen) Ltd
              Panasonic Automotive Systems Company of America
              Paragon Technologies Inc.
              PEGATRON CORPORATION
Pittasoft
              RICOH COMPANY LTD.
              Routerboard.com
              Samsung ElecCo.,Ltd
              Samsung Electro Mechanics co., LTD.
              Samsung Electro Mechanics co.,LTD.
              SAMSUNG ELECTRO-MECHANICS
              SAMSUNG ELECTRO-MECHANICS CO., LTD.
              Samsung Electronics Co., LTD
              Samsung Electronics Co.,Ltd
              Samsung Electronics ITS, Printer division
              SEIKO EPSON CORPORATION
              Seoul Commtech
              Shenzhen JSR Technology Co.,Ltd.
              Shenzhen Ogemray Technology Co., Ltd.
              Shenzhen XinKingBrand enterprises Co.,Ltd
              Sierra Wireless Inc
              SIRONA DENTAL SYSTEMS GmbH & Co. KG
              Sony Mobile Communications AB
SparkLAN Communications, Inc.
              TCT mobile limited
              TP-LINK TECHNOLOGIES CO., LTD.
              TP-LINK TECHNOLOGIES CO.,LTD.
Ubiquiti Networks
Ubiquiti Networks, INC
              UNKNOWN
Verifone, INC.
              WISOL
Wistron Corporation
WistronInfoComm(ZhongShan) Corporation
              Zebra Technologies Inc
zte corporation
Zultys Technologies
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


REPTree
============

Timeout < 15
|   RSSI < 55
|   |   RSSI < 45
|   |   |   Apple < 0.5 : 9.04 (22/0.04) [4/0]
|   |   |   Apple >= 0.5
|   |   |   |   UNKNOWN < 0.5 : 9.3 (29/0.16) [14/0.34]
|   |   |   |   UNKNOWN >= 0.5 : 9 (4/0) [2/0]
|   |   RSSI >= 45 : 9.88 (51/2.47) [40/0.79]
```

```
|   RSSI >= 55
|   |   TP-LINK TECHNOLOGIES CO.,LTD. < 0.5 : 7.72 (195/3.18) [101/3.8]
|   |   TP-LINK TECHNOLOGIES CO.,LTD. >= 0.5
|   |   |   Billion Electric Co. Ltd. < 0.5
|   |   |   |   COMPAL INFORMATION (KUNSHAN) CO., LTD. < 0.5
|   |   |   |   |   WISOL < 0.5 : 9.88 (43/2.82) [9/2.24]
|   |   |   |   |   WISOL >= 0.5 : 7 (2/0) [2/0]
|   |   |   |   COMPAL INFORMATION (KUNSHAN) CO., LTD. >= 0.5
|   |   |   |   |   Apple < 0.5 : 11 (2/0) [0/0]
|   |   |   |   |   Apple >= 0.5 : 7 (14/0) [10/0]
|   |   |   Billion Electric Co. Ltd. >= 0.5
|   |   |   |   Apple < 2.5 : 8.41 (119/0.99) [69/1.05]
|   |   |   |   Apple >= 2.5
|   |   |   |   |   Belkin International Inc. < 1.5 : 8.05 (83/0.87)
[39/1.1]
|   |   |   |   |   Belkin International Inc. >= 1.5 : 6 (2/0) [0/0]
Timeout >= 15
|   Timeout < 82.5
|   |   Timeout < 52.5
|   |   |   Timeout < 37.5
|   |   |   |   RSSI < 55
|   |   |   |   |   RSSI < 47.5 : 12.41 (57/42.74) [38/49.25]
|   |   |   |   |   RSSI >= 47.5 : 23.12 (34/0.58) [23/25.22]
|   |   |   |   RSSI >= 55
|   |   |   |   |   Apple < 0.5 : 7.57 (14/34.41) [7/37.59]
|   |   |   |   |   Apple >= 0.5 : 12.99 (418/7.94) [213/5.94]
|   |   |   Timeout >= 37.5
|   |   |   |   NETCOMM LIMITED < 0.5
|   |   |   |   |   Apple < 0.5 : 5.67 (2/72.25) [1/72.25]
|   |   |   |   |   Apple >= 0.5 : 16.55 (44/3.16) [20/1.85]
|   |   |   |   NETCOMM LIMITED >= 0.5
|   |   |   |   |   Apple, Inc< 1.5 : 22.08 (51/1.6) [24/3.01]
|   |   |   |   |   Apple, Inc>= 1.5 : 24.62 (11/0.79) [10/0.23]
|   |   Timeout >= 52.5
|   |   |   Ubiquiti Networks, INC < 0.5
|   |   |   |   Samsung Electro Mechanics co.,LTD. < 0.5
|   |   |   |   |   RSSI < 57.5 : 7.39 (206/0.62) [101/1.63]
|   |   |   |   |   RSSI >= 57.5 : 9.21 (187/14.7) [79/9.53]
|   |   |   |   Samsung Electro Mechanics co.,LTD. >= 0.5 : 0 (8/0) [5/0]
|   |   |   Ubiquiti Networks, INC >= 0.5
|   |   |   |   Hon Hai Precision Ind. Co.,Ltd. < 0.5
|   |   |   |   |   Hewlett Packard < 2.5 : 20.12 (43/0.63) [22/0.47]
|   |   |   |   |   Hewlett Packard >= 2.5 : 22 (6/0) [10/0]
|   |   |   |   Hon Hai Precision Ind. Co.,Ltd. >= 0.5 : 18 (8/0) [6/0]
|   Timeout >= 82.5
|   |   Samsung Electro Mechanics co.,LTD. < 0.5
|   |   |   Dell Inc< 0.5
|   |   |   |   Netgear Inc. < 0.5
|   |   |   |   |   RSSI < 67.5 : 14.43 (506/9.49) [254/9.93]
|   |   |   |   |   RSSI >= 67.5 : 17.24 (229/19.79) [101/26.59]
|   |   |   |   Netgear Inc. >= 0.5
|   |   |   |   |   Timeout < 135 : 15.5 (18/0.8) [18/0.7]
|   |   |   |   |   Timeout >= 135 : 12.52 (114/0.37) [45/0.39]
|   |   |   Dell Inc>= 0.5
|   |   |   |   RSSI < 85
|   |   |   |   |   Apple Inc< 0.5 : 12.13 (23/0.14) [16/6.54]
|   |   |   |   |   Apple Inc>= 0.5 : 10.33 (53/0.42) [27/0.65]
|   |   |   |   RSSI >= 85 : 17 (5/0) [2/0]
|   |   Samsung Electro Mechanics co.,LTD. >= 0.5
|   |   |   Ubiquiti Networks < 0.5 : 15 (26/0) [8/0]
|   |   |   Ubiquiti Networks >= 0.5
|   |   |   |   RSSI < 85
|   |   |   |   |   GIGA-BYTE TECHNOLOGY CO.,LTD. < 4.5 : 21.83 (26/0.21)
[9/0.4]
|   |   |   |   |   GIGA-BYTE TECHNOLOGY CO.,LTD. >= 4.5 : 23 (21/0) [10/0]
|   |   |   |   RSSI >= 85 : 15 (2/0) [1/0]

Size of the tree : 73
```

```
Time taken to build model: 0.17 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.7684
Mean absolute error                      2.0525
Root mean squared error                  3.2024
Relative absolute error                 50.1418 %
Root relative squared error             63.9819 %
Total Number of Instances             4018
```

APPENDIX 7: MAC AND RSSI FILTERING LINEAR REGRESSION

Option	Selection
Data File	data/csv/rssi_man_breakdown.csv
Weka File	data/arff/rssi_man_breakdown.arff
Classifier	weka.classifiers.functions.LinearRegression
Classifier Options	attributeSelectionMethod: M5 Method debug: False eliminateColinearAttributes: True ridge: 1.0E-8
Test Options	Cross-Validation: 10 Fold
Dependent Variable:	Manual Count

Weka output:

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     rssi_rssi_man_breakdown
Instances:    4018
Attributes:   96
              Timeout
              RSSI
              Manual Count
              Apple
              Apple Inc
              Apple, Inc
              Apple, Inc.
Askey Computer Corp
ASRock Incorporation
ASUSTek COMPUTER INC.
Azurewave Technologies., inc.
Belkin International Inc.
              Billion Electric Co. Ltd.
              Billion Electric Co., Ltd.
              BROTHER INDUSTRIES, LTD.
              CABLE TELEVISION LABORATORIES, INC.
              Cisco
              CISCO SYSTEMS, INC.
              COMPAL INFORMATION (KUNSHAN) CO., LTD.
              D-Link International
              D-LINK INTERNATIONAL PTE LIMITED
              Dell Inc
DrayTek Corp.
Elitegroup Computer System Co.
              ELITEGROUP COMPUTER SYSTEM CO., LTD.
Enerpoint Ltd.
              FUJI-XEROX CO. LTD.
GainSpan Corp.
              GIGA-BYTE TECHNOLOGY CO.,LTD.
              Hewlett Packard
              Hewlett-Packard Company
              Hon Hai Precision Ind. Co.,Ltd.
              Hon Hai Precision Ind.Co.Ltd
              HTC Corporation
              Huawei Technologies Co., Ltd
              IBM
              IBM Corp
              ICP Electronics Inc.
              Intel Corporate
              KYOCERA Document Solutions Inc.
              Lexmark International Inc.
              LG Electronics
Liteon Tech Corp.
Liteon Technology Corporation
              Micro-Star Int'l Co, Ltd
              Microsoft
```

```
              Microsoft Corporation
              Motorola Mobility LLC
              Murata ManufactuaringCo.,Ltd.
              Murata Manufacturing Co., Ltd.
              Murata Manufacturing Co.,Ltd.
              NEC Platforms, Ltd
              NETCOMM LIMITED
              NETGEAR
NetgearInc
Netgear Inc.
              NETGEAR INC.,
              Nintendo Co., Ltd.
              Nokia Corporation
OnePlus Tech (Shenzhen) Ltd
              Panasonic Automotive Systems Company of America
              Paragon Technologies Inc.
              PEGATRON CORPORATION
Pittasoft
              RICOH COMPANY LTD.
              Routerboard.com
              Samsung ElecCo.,Ltd
              Samsung Electro Mechanics co., LTD.
              Samsung Electro Mechanics co.,LTD.
              SAMSUNG ELECTRO-MECHANICS
              SAMSUNG ELECTRO-MECHANICS CO., LTD.
              Samsung Electronics Co., LTD
              Samsung Electronics Co.,Ltd
              Samsung Electronics ITS, Printer division
              SEIKO EPSON CORPORATION
              Seoul Commtech
              Shenzhen JSR Technology Co.,Ltd.
              Shenzhen Ogemray Technology Co., Ltd.
              Shenzhen XinKingBrand enterprises Co.,Ltd
              Sierra Wireless Inc
              SIRONA DENTAL SYSTEMS GmbH & Co. KG
              Sony Mobile Communications AB
SparkLAN Communications, Inc.
              TCT mobile limited
              TP-LINK TECHNOLOGIES CO., LTD.
              TP-LINK TECHNOLOGIES CO.,LTD.
Ubiquiti Networks
Ubiquiti Networks, INC
              UNKNOWN
Verifone, INC.
              WISOL
Wistron Corporation
WistronInfoComm(ZhongShan) Corporation
              Zebra Technologies Inc
zte corporation
Zultys Technologies
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

Manual Count =

      0.0357 * Timeout +
     -1.1726 * Apple Inc +
      0.5328 * Apple, Inc +
      2.0039 * Apple, Inc. +
      0.4588 * ASUSTek COMPUTER INC. +
     -2.688  * Azurewave Technologies., inc. +
      1.8054 * Belkin International Inc. +
     -0.2973 * Billion Electric Co. Ltd. +
     -2.042  * Billion Electric Co., Ltd. +
     -1.6698 * CISCO SYSTEMS, INC. +
```

```
      1.4392 * D-LINK INTERNATIONAL PTE LIMITED +
      1.812  * Elitegroup Computer System Co. +
      0.1736 * GIGA-BYTE TECHNOLOGY CO.,LTD. +
     -0.2118 * Hewlett Packard +
      0.2677 * Hewlett-Packard Company +
     -0.7838 * Hon Hai Precision Ind. Co.,Ltd. +
     -2.8332 * IBM +
      0.6611 * IBM Corp +
     -0.6311 * ICP Electronics Inc. +
     -0.1858 * Intel Corporate +
     -1.1797 * Lexmark International Inc. +
     -3.4918 * LG Electronics +
      1.2702 * Liteon Tech Corp. +
     -1.287  * Microsoft +
      3.4156 * NEC Platforms, Ltd +
     -0.2686 * NETCOMM LIMITED +
      0.4618 * NETGEAR +
      2.5543 * NetgearInc +
     -0.7195 * NETGEAR INC., +
     -2.8579 * PEGATRON CORPORATION +
      1.5933 * Routerboard.com +
     -0.3589 * Samsung Electro Mechanics co., LTD. +
      1.2937 * Samsung Electro Mechanics co.,LTD. +
     -0.9735 * Samsung Electronics Co.,Ltd +
      1.3626 * SEIKO EPSON CORPORATION +
     -0.4762 * TP-LINK TECHNOLOGIES CO.,LTD. +
      1.8804 * Ubiquiti Networks +
      5.2147 * Ubiquiti Networks, INC +
     -0.3773 * UNKNOWN +
     -2.5007 * WISOL +
      6.0872 * zte corporation +
     10.1069

Time taken to build model: 0.28 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6369
Mean absolute error                      2.7798
Root mean squared error                  3.8575
Relative absolute error                 67.9115 %
Root relative squared error             77.0718 %
Total Number of Instances             4018
```