Defeating the Digital Divide

1 Executive Summary

As the world becomes increasingly reliant on the internet, from online schooling to working from home, broader and higher quality access has never been more important. However, expanding internet infrastructure presents a unique challenge in terms of cost, economic efficiency, and capacity requirements. Our team aims to optimize the process of improving connectivity by predicting the price of bandwidth over the next 10 years, calculating bandwidth needs for a variety of household scenarios, and determining the best distribution of cellular nodes over a given region.

We first predicted future internet costs in the United States(US) and the United Kingdom(UK) by using random forest regression, a supervised machine learning algorithm that combines the predictions of multiple decision trees to give accurate forecasts. The model was trained on a dataset of 48 mainland US states and their respective population densities, cost of living, average download speeds, and average price per Mbps. After achieving a reasonable root mean square error of 0.660, the random forest was then applied to the entirety of the US and UK. The population densities and cost of living were adjusted based on the expected percent change over 10 years, and a regressed exponential function was used to determine future average download speeds. These features served as input into the random forest which then predicted the future price. The model ultimately predicted a decrease of \$0.23 per Mbps in the US and \$0.57 in the UK over the next decade, which proved to be a sensible result.

Next, we calculated the bandwidth demands for various household scenarios and determined the bandwidth required to satisfy 90% and 99% of predicted demand. Each scenario involved a household with a hypothetical weekly schedule of activities and information based on age, education, and job status. We created a simulation of a typical week with 112 waking hours and determined the bandwidth usage for each hour where the probability of a person performing the activity was proportional to the percent of total time needed for the task. For each scenario, the maximum bandwidth was recorded over 1000 trials, and the 90th and 99th percentile of the distribution was found. 14.5 Mbps and 15.5 Mbps were sufficient for 90% and 99% coverage for scenario 1, 20.8 Mbps and 21.8 Mbps for scenario 2, and 20.6 Mbps and 21.9 Mbps for scenario 3.

Finally, we developed a model that optimally distributes 4G and 5G cellular nodes in arbitrary regions. We studied the comparative advantages and disadvantages of each network type—surmising that while 4G networks are often discounted, its 5G counterparts offer fast speeds at shorter signal ranges. To calculate the locations for placing the nodes, we calculated the center of mass by integrating mass with respect to position for each region. We also determined that cities eligible for 5G networks should have population densities of 777 people per square mile or higher and an annual household income of at least \$103,689.32. Combining these two techniques yielded an efficient distribution plan for cellular nodes.

Contents

2 Part I: The Cost of Connectivity

2.1 Restatement of the Problem

In this problem, we are tasked with predicting the cost per unit of bandwidth(Mbps) over the next decade for consumers in the United States and the United Kingdom.

2.2 Assumptions

1. *There will be no significant policy changes regarding standards for internet speed, cost, and availability in the next 10 years.* Due to the unpredictable nature of new legislation and government initiatives, we are unable to account for these potential changes in our modeling.

2. *The amount of internet provided in Mbps per US dollar will be constant regardless of how much money is spent.* The majority of internet users will not spend enough money such that they begin experiencing diminishing returns for each dollar spent. Economics of scale are therefore not applicable to individual households.

3. *Future internet usage and cost trends in the UK are comparable to US states and depend on the same factors.* The United States and the United Kingdom have similar internet infrastructure and geographical trends such as population density (e.g. the two countries contain both high and low population density areas).

4. *The population density and cost of living index in a given area increase by 1.31% and 2.20% respectively, which are sensible rates for the United States.* The percent increase in population density in California was found to be 1.31% according to the PPIC[1]. California has a large land area and contains a relatively uniform mix of urban and rural areas. As a result, its overall increase in population density can likely be generalized to the United States. The average increase in the cost of living index can be modeled by the inflation rate[2]. Since we are looking at the change of the index over a long period of time (10 years), it is reasonable to use a constant average inflation rate per year in order to model the growth of the cost of living over a decade. .

2.3 Model Development

2.3.1 Factor Identification

The population density is a relevant factor given the cost discrepancies between urban and rural areas. For a sparsely populated region, the price of internet is significantly higher for an individual consumer because fewer people are available to recuperate the fixed infrastructure cost. On the other hand, the utility obtained per unit of internet infrastructure is greater in densely populated areas where more consumers are available to handle the expenses. Thus,

we incorporate the population density data of mainland US states[3] into our model along with the expected annual change of 1.31

Next, the cost of living indicates the amount of money required to achieve a certain standard of living in a given area. With a higher cost of living, the dollar price for goods and services, including the internet, will also be greater. Therefore, we utilize a cost of living index[4] that compares a state's cost of living with the average for the United States. For example, a state with an index value of 110 compared to the mean of 100 in the US has a 10% higher cost of living. We utilize the average change in the index to model the internet costs for the next 10 years while accounting for inflation.

Lastly, we include the average download speed for each state[5] to ensure that the model accounts for any potential trends between the factor and internet cost. For example, it is possible that areas with high download speed have well-established infrastructure and therefore have cheaper internet. We also created two exponential regression models based on US and the UK download speed trends[6] to predict the values over the next 10 years as shown in Figure 1. These predictions are then used as part of the overall regression model.

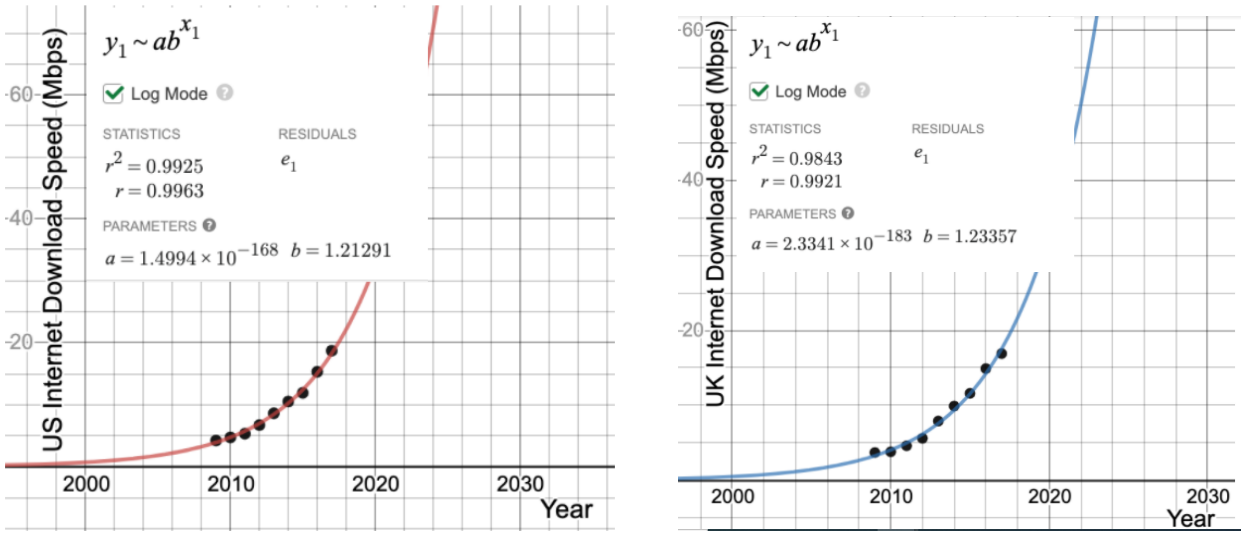

Figure 1: Exponential Regression Models

2.3.2 Random Forest Regression

To predict internet cost over time, we used the Scikit-learn Python library to train a random forest regression model on the factors described above. This supervised machine learning model employs an ensemble of decision trees to make an accurate and robust prediction. Each decision tree is trained on a random subset of the dataset, shown in Figure 2, to reduce intercorrelation and utilizes algorithms to determine optimal logical splits. Following the created splits, data can be categorized and a regression output is produced. The outputs of the individual decision trees are then averaged for the final ensemble prediction. Our

random forest model consists of 20 decision trees, each limited to a max depth of five nodes to improve generalization to unseen data.

	State	Population Density	Cost of Living Index	Average Download Speed	Cost per Mbps
0	Alabama	95.8	89.3	33.7	5.82
1	Arizona	60.1	97.0	33.9	4.35
2	Arkansas	57.2	86.9	25.0	5.13
3	California	251.0	151.7	29.0	1.86
4	Colorado	52.6	105.6	40.9	3.90

Figure 2: First few rows of the training data

The random forest regressor was first trained on the set of 48 US states using the features of population density, cost of living index, and average download speed. The target column consisted of the state's average internet cost[7] in dollars per Mbps. After training, the model was used to predict the cost of the internet for the US and UK over the next 10 years. This was accomplished by making predictions for each country's average feature values and then applying the appropriate annual change to each input for the next prediction. The population density was increased by 1.31%, the cost of living index by 2.20%, and the average download speed by the amount indicated by the exponential regressions. By performing this process over 10 iterations, we were able to find the predicted future internet costs for users in the two countries.

2.4 Results

The random forest regression achieved a root mean square error (RMSE) of 0.660 when predicting the internet costs for US states. This value represents a typical difference between the predicted and actual internet prices in US dollars, which is an acceptable error given the magnitude of the prices.

The 10-year US and UK regressions produce the trends shown in Figure 3. Although there are some deviations, the model predicts a decrease of $0.57 (from $4.33 to $3.76) in the average price per Mbps in the US and a decrease of $0.57 (from $3.18 to $2.61) in the UK. As expected, the price of internet access decreases over time which can be attributed to increased population density and improved infrastructure. Overall, the predictions are consistent with the expected results.

2.5 Strengths and Weaknesses

The random forest model allows us to effectively combine multiple factors to make a prediction. The algorithm is robust to skewed, non-linear data which makes it ideal for processing features such as population density. Furthermore, the ensembling process minimizes the risk of overfitting, a prevalent issue where a machine learning model learns the training data

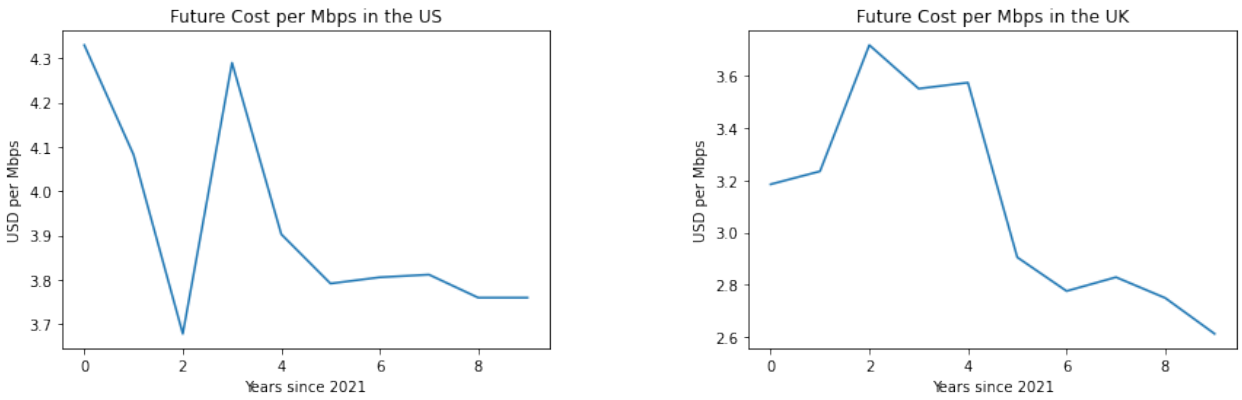

Figure 3: Projected cost of bandwidth in Mbps for the US and the UK

too closely and later fails to accurately predict unseen examples. However, a weakness of random forest regression is the fact that it is not easily interpretable since it consists of a large number of decision trees. It is difficult to pinpoint the exact decision process that the algorithm undergoes to arrive at its prediction.

3 Part II: Bit by Bit

3.1 Restatement of the Problem

In this problem, we are tasked with finding what minimum bandwidth households would need to cover their internet usage 90% and 99% of the time. Three scenarios were given:

1. A couple in their early 30's (one is looking for work and the other is a teacher) with a 3-year-old child.
2. A retired woman in her 70's who cares for two school-aged grandchildren twice a week.
3. Three former M3 Challenge participants sharing an off-campus apartment while they complete their undergraduate degrees full-time and work part-time.

3.2 Assumptions

1. *We assume that each day and week will have the same probabilities.* Each day and week in the year will overall be extremely similar so we ignore any special circumstances.
2. *We observe that each day's usage (Sunday vs. Saturday vs. Monday etc.) does not differ substantially, which resultantly solves the issue of the grandma caring for her grandchildren.* Beyond the grandmother, the aforementioned assumption also resolves situational occurrences for each person, both now and in the future.

3. *The people in the given situations are in the US and not the UK.* The data provided is only relevant for the US.

4. *People will always use best quality streaming for Youtube (HD) and Netflix (Ultra HD), even if they only "need" less e.g. 1 mb/s for SD quality, they will never use it at that quality so we will assume that the bandwidth needed will be at the higher quality.* SD quality usually is unsatisfactory and people generally want to use the best quality for the best experience if it's available.

5. *Each pre-college school day is 5 hours, giving a total of 25 hrs/week. Every class will also be on a video conference format.* Since a general school day is around 7 hours, we made a general estimate that the amount of actual video-conferencing and internet usage would come to around 5 hours.

6. *College students also leverage online platforms for their classes in the wake of a pandemic. Their classes span an average of 16 hrs/week, via video conferencing softwares*[8]. In light of the additional time that college students have in a remote academic setting, it is safe to say that they will consciously look to take a more rigorous—albeit longer—college course load, leading us to take the higher bound of the class hrs/week (16 hrs).

7. *Teachers' work hours are the same as students (5 hours) and they do not work after school hours.* Justification: Due to the variability in each day for a teacher—inclusive of meetings and personal commitments—it's difficult to pinpoint their daily commitment to school-related curricula. As a result, we went ahead and assumed that they do not provide after school help, for the most consistency in results.

8. *People are awake to use the internet 16 hours a day.* 8 hours is a recommended amount of time for an average person to sleep[9].

9. *People only access the internet through one medium at a time and do not multitask.* Most people generally only consume the internet through one source at a time. Furthermore, this is a simplifying assumption that allows us to run the simulation and assign each person a specific task and internet consumption for a specific hour.

3.3 Model Development

3.3.1 Determining the Hours Each Person Spends on the Internet

Each person was first assigned a list of digital activities they would perform in a week. Each activity had attributes like time spent, lower bound of bandwidth usage, and upper bound of bandwidth usage. Age group, education level, and employment status were all factors used to determine this list of digital activities[6]. The age group was the first categorization to associate values with the ages of the people for the scenario. Utilizing the data provided[6], we were able to determine internet tendencies for entertainment purposes depending on the

age of the user. This included the amount of time they spent watching videos, listening to audio, etc. Next, in order to account for the global shift to remote learning and work, we added hours of schooling to those applicable, namely people in the age groups of 2-11, 12-17, and 18-34. Similarly, teachers must work through online, remote instruction and encounter similar internet usage as their students. As a result, school related activities were added on top of entertainment activities that were already present in the list of digital activities.

For employment, those in part-time jobs were considered to be in-person workers because remote part-time jobs are generally hard to obtain, especially for undergraduate students. As a result, jobs were not added onto the list of internet activities as they do not contribute to household bandwidth consumption.

3.3.2 Simulating Outcomes

Finally, we simulated the bandwidth demand for each scenario for every single hour for an entire week. Given that people are awake around 16 hours a day, we simulated $7 \cdot 16 = 112$ hours of internet usage. For each hour, the probability that a person would be performing a given internet activity is proportional to the percentage of time they spend on that activity over the entire week. For example, if a user spends 25 hours a week in school out of 112 hours, they would have a probability of 25/112 or 22.3% chance of attending school for any of the 112 hours. These random tasks would be selected for each person in the simulation and the individual bandwidth demand would be calculated. If the activity had a higher and lower bound to bandwidth usage, the actual bandwidth usage is determined by uniformly selecting a random floating point number between the bounds of the bandwidth usage. The individual bandwidth is then summed up in order to find the total household bandwidth demand for that given hour. This process is then repeated for each waking hour in the week.

We repeated the simulation 1000 times for every single scenario, representing 1000 weeks of internet usage for each scenario. In each week/trial, we recorded the maximum bandwidth demanded, since a household would require a bandwidth greater than the maximum bandwidth demanded at every single hour in order to successfully cover the internet needs for that week. The maximum values for all 1000 weeks were recorded and then sorted in ascending order. Finally, the 90th percentile and the 99th percentile values were calculated from the maximum bandwidth demand distribution. These represent the minimum bandwidth needed to cover 90% and 99% of the demand respectively as 90% of the bandwidth demands are less than the 90th percentile and 99% of the bandwidth demands are less than the 99th percentile.

3.4 Results

By running the program through each scenario 1000 times and plotting the data, we achieved 3 histograms of distributions of the maximum bandwidth needed for a given hour. In these histograms, shown in Figure 4, the vertical orange line represents the required bandwidth

to meet the total internet demand 90% of the time and the vertical red line represents the required bandwidth to meet the total internet demand 99% of the time.

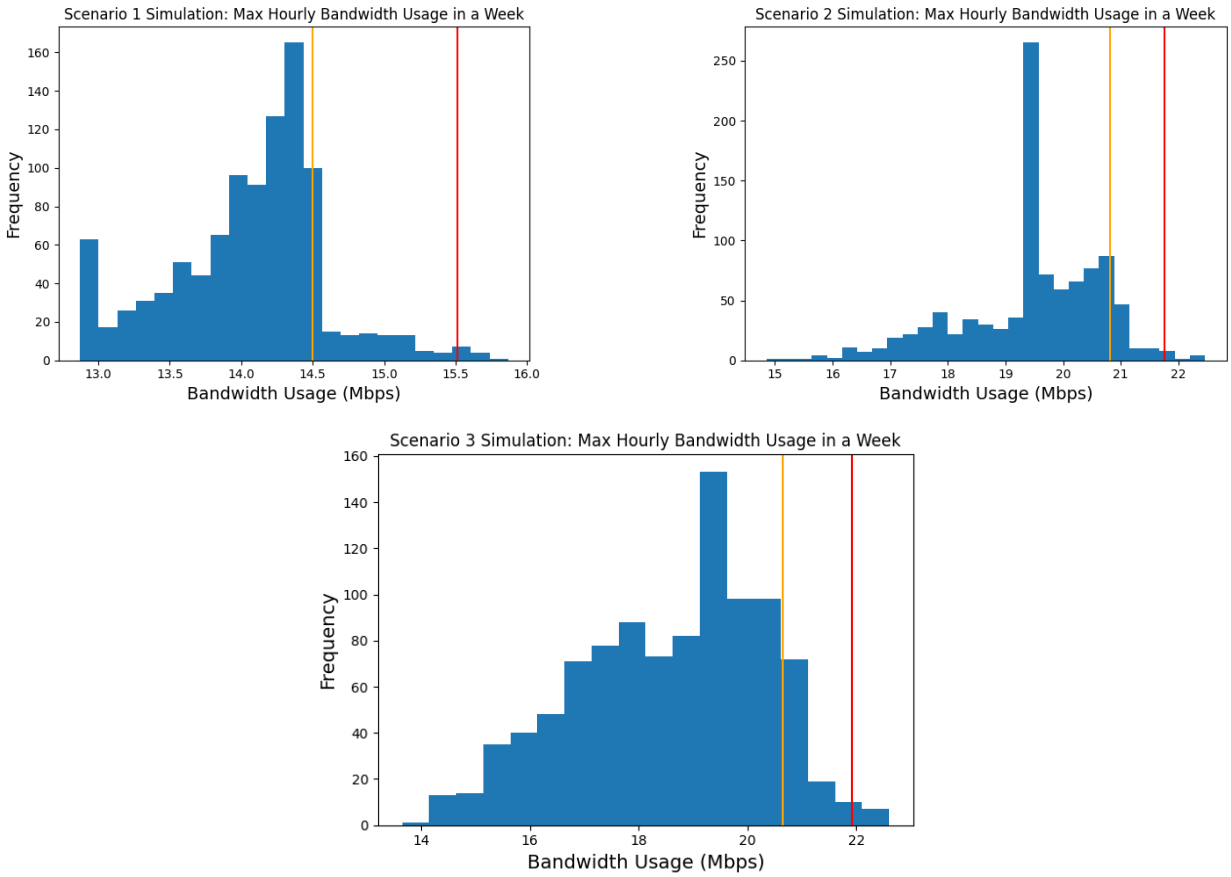

Figure 4: Frequency histograms of the distribution of maximum bandwidth usage for 1000 simulation trials.

Furthermore, as shown in the summarized table, in scenario 1, the required bandwidth to satisfy 90% of bandwidth demand is 14.5 Mbps and 99% of demand is 15.5 Mbps. In scenario 2, 20.8 Mbps is required for 90% of demand and 21.8 Mbps is required for 99% of demand. Finally, in scenario 3, 20.6 Mbps is required for 90% of demand and 21.9 Mbps is required for 99% of demand.

Scenario	90% Coverage	99% Coverage
1	14.5 Mbps	15.5 Mbps
2	20.8 Mbps	21.8 Mbps
3	20.6 Mbps	21.9 Mbps

Table 1: Summary of the minimum number of Megabits per Second of internet required to cover internet usage 90% and 99% of the time for each of the three scenarios

3.5 Strengths and Weaknesses

This model is accurate at predicting bandwidth usages and the amount of bandwidth required in order to satisfy those demands. All values predicted were between 14-22 Mbps. Our predictions are further supported by the FCC's household broadband guide[10], which recommends a medium service of 12 to 25 Mbps for moderate internet usage in households of 2-3 people. Moreover, this model is very flexible in considering multiple factors and internet usage caused by various internet activities. In the case that more data about internet usage for a new or existing activity comes up, our model and simulation can be easily adjusted to predict given the new factor.

One significant weakness of this model is that the data utilized to calculate the probabilities was from the first quarter of 2020. While the data is not too far off, current internet usage tendencies will be different; currently, the model likely underpredicts the amount of needed bandwidth as people are using digital devices more due to being stuck at home in the midst of the pandemic. Moreover, people usually have activities they will perform at a certain time, like kids going to school from roughly 8 AM to 3 PM, and this model essentially has all activities being able to occur at any given hour interval.

4 Part III: Mobilizing Mobile

4.1 Restatement of the Problem

In this problem, we are tasked with developing a model for optimally placing cellular nodes in a given region using population, demographic, and bandwidth usage data.

4.2 Assumptions

1. *The price of 5G internet plans will remain constant across different regions.* Due to the small market share of 5G networks (namely under the domain of Verizon and AT&T), there's limited offerings for pricing, all of which hover around the $89 point according to a GSMA Intelligence report [11]. As a result, it made the most sense to assume that this price is homogeneous across all purchased 5G plans.

2. *The population densities remain constant across the hypothetical subregions created by M3. (eg: the Region A1, Region B3) provided by M3*[6]. In short, data set[6] provides information and shapes for the 3 hypothetical regions A, B, C and their respective subregions. While the M3 Data provided the population and area of the subregions in their hypothetical diagram, there is no way to account for fluctuations in population densities across the subregion.

3. *A majority of households that have annual incomes $\geq$ $103,689.32 and are located in a subregion with a minimum population density of 777 people/sq mile, will naturally*

choose the 5G plan (these values are derived later on in our approach). With 5G plans discretely offering faster speeds than their 4G opponents, a family in the position, both geographically and financially, would reasonably choose to do so.

4. *Walls and other natural barriers will not hinder 5G speeds inside the effective signal range.* It's nearly impossible to quantify the effects of such obstacles with the data provided[6]. Moreover, the impact will depend on the material composition of the barrier—another difficult metric to model.

4.3 Model Development

To optimize both internet connectivity and affordability for any given population, we sought to create an efficacious combination of 4G and 5G internet access points—capitalizing on their niche benefits, while mitigating any potential disadvantages. There are three relevant factors that need to be considered when differentiating between 4G and 5G networks: signal range, latency, internet speed. The following best characterize the distribution of such factors in the larger framework of the two networks: Simply put, 5G networks offer significantly faster internet speeds, which come at the expense of an abridged signal range. On the other hand, 4G networks have much more expansive coverage, but operate at higher latency rates, and thereby, slower speeds[12].

Given the more frequent use of 4G networks, largely due to their feasibility for populations anytime and anywhere, we decided to pinpoint the location of these nodes first in the hypothetical regions provided by the M3 Data[6]. Using principles from packing problem optimization, we created a mathematical model that can be used to place 4G towers at the points of maximum effectiveness.

Our mathematical model calculates the center of mass for each region, where mass refers to population, using the assumption that the population density is uniformly distributed throughout the region. By utilizing a small tiling of the region, we can approximate the moments of the x-axis and y-axis and obtain the coordinates of the center of mass. By placing the 4G towers at the centers of mass, we will guarantee that the maximum number of people can effectively utilize the bandwidth it provides. The number of internet nodes in a given region will depend on the area of the region and the effective signal range for the different internet towers.

First, set the origin at the bottom left of the region. While it does not necessarily have to be here, this keeps consistency between different regions. Additionally, let the side length of the square be Δx. We can approximate Δx using any subregion, whichever one is more convenient/has the least amount of area outside of the tiling. After we approximate the area of the subregion in terms of $(\Delta x)^2$, since we are given the areas of the subregions, we can approximate Δx.

As mentioned, we seek the centers of mass of each region. The center of mass, represented

as $(\overline{x}, \overline{y})$ has coordinates:

$$\overline{x} = \frac{M_y}{M}$$
$$\overline{y} = \frac{M_x}{M}$$

where M_x and M_y represent the moments about the x and y axis respectively, and M represents the area of the region (given in the data). Thus, in order to obtain the center of mass, we first find the moments. Since we are working under uniform mass density, we can simply let the density function $\delta(x, y) = 1$, thus making the moments equivalent to:

$$M_x = \iint_D y \, dA$$
$$M_y = \iint_D x \, dA$$

for domain D represents the domain of the moments and area of the region A. Since there is no easy way to bound domain D with functions, we instead use a Riemann sums to approximate these integrals. Using the original tiling, we will plot the centers of each square in region 2. The centers fully contained within the inner boundaries will be used to approximate the moments. For convenience, suppose P is the set of all centers contained within the region. If $P = (x_p, y_p)$, then we have that

$$M_x \approx \Delta A \sum_{(x,y)\in P} y$$
$$M_y \approx \Delta A \sum_{(x,y)\in P} x$$

Though we can't simplify the summation without knowing our actual region, we can calculate ΔA. By definition, we have that $\Delta A = \Delta x \cdot \Delta y$, but since we are tiling with squares, we have that $\Delta y = \Delta x$ so:

$$M_x \approx (\Delta x)^2 \sum_{(x,y)\in P} y$$
$$M_y \approx (\Delta x)^2 \sum_{(x,y)\in P} x$$

Calculating out these moments and then dividing by M will yield the coordinates of the center of mass.

Another factor prevalent in the decision making process between 4G and 5G networks includes the price tag that each carries; at large, 4G systems are near ubiquitously cheaper than their 5G counterparts—making household income an important factor in the model. Hence, we decided that there had to be some minimum threshold for determining whether a city should implement 5G internet nodes.

To encompass the threshold, we opted for a criteria based model, in which a given region must meet both of the following criteria to be eligible for 5G network implementations.

The first is a population density $\geq$ 777 persons/miles^2. Since almost all 5G networks are located in urban centers because of the relatively small range of 5G signals, we want to ensure that planting a 5G tower will deliver benefits to the most consumers as possible. Urban clusters (regions that are slightly less packed than cities) are classified as having a population density of 777 people/square mile. Using this data point as a minimum will ensure that we maximize consumer utility. Furthermore, we found the top 10 leading cities with 5G networks have population densities greater than this value.

The second is an average annual household income $\geq \$103,689.32$. We gathered that the average American spends \$57.25 on internet service plans [13]. The average monthly income in the United States is \$5,555 [14]. Dividing these values yields that 1.03% of monthly income is allocated towards an internet plan. Moving onward, we used a constant of \$89 for the cost of a 5G plan. This value was derived from the GSMA Intelligence report that found the average 5G monthly cost in the United States [11]. In turn, this yielded the inequality:

$$\frac{(\text{Avg Annual Household Income})}{12} \cdot (0.0103) > 89.$$

When computed, this outputted a minimum household income = \$103,689.32 for a family to reasonably afford 5G service. If a given region meets both criteria, it'll be eligible for a 5G network in the context of our model. Ideally the web of 5G nodes should be placed in a central hotspot, the centroid of a population density curve. This will maximize the amount of people the network reaches as well as provide higher connection speeds to a larger number of people.

Region	Population Density (ppl per mi^2) [6]	Yearly Income (USD) [6]	5G Eligibility
A1	570	\$27,941.00	No
A2	1778	\$30,929.00	No
A3	1945	\$47,163.00	No
A4	168	\$34,273.00	No
A5	3452	\$30,425.00	No
A6	650	\$46,659.00	No
B1	1113	\$99,652.00	No
B2	486	\$134,375.00	No
B3	270	\$173,188.00	No
B4	487	\$112,306.00	No
B5	331	\$108,056.00	No
B6	646	\$147,500.00	No
B7	237	\$132,045.00	No
C1	3863	\$214,125.00	Yes
C2	11600	\$104,209.00	Yes
C3	10120	\$190,729.00	Yes
C4	5396	\$139,261.00	Yes
C5	10069	\$152,500.00	Yes
C6	3252	\$206,875.00	Yes

C7	5262	$151,731.00	Yes

Table 2: 5G Eligibility by Subregion

To calculate the coordinates of the 5G towers, we used the same mass-center approach, essentially applying the findings of the 4G model to a smaller land-mass. Similarly, the amount of 5G towers will depend strictly on the area of the given region and the effective signal range distance.

4.4 Results

Provided in Figure 5 are the results for 4G internet node placement for subregion B2 and 5G internet node placement for subregion C2.

Figure 5: Tiling of Subregion 2 with Origin Marked

As stated in the previous section, we can approximate Δx using any subregion within the same region. In this example, we utilize subregion 3, with the area being approximated relatively accurately to $22(\Delta x)^2$ as shown in Figure 6: The approximation is calculated by

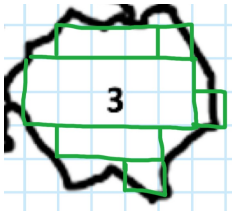

Figure 6: Approximate Tiling of Subregion 3

summing up the 19 outlined tiles and the remaining area in the region that is not contained in either of the outlined tiles, which is roughly equal to 3 full tiles. Since the area of subregion 3 is given to be 4.64 square miles, we can approximate Δx to be:

$$22(\Delta x)^2 \approx 4.64$$
$$\Delta x \approx 0.46$$

Now, we look for the center of mass.As aforementioned, in order to obtain the center of mass, we first find the moments. The approximations for the moments are

$$M_x \approx \Delta A \sum_{(x,y)\in P} y$$

$$M_y \approx \Delta A \sum_{(x,y)\in P} x$$

for similarly defined set P. Using the equivalence we obtained earlier, that is $\Delta A = (\Delta x)^2$, we have that

$$\Delta A = (\Delta x)^2 \approx .21$$

Next, we will plug in all P, which were marked in Figure 5. All dots have x or y values of form $(2n-1)\frac{\Delta x}{2}$ for integers $n \geq 1$, so we can multiply these values by the amount of times each value occurs. This gives us the following values for the moments:

$$M_x \approx (.21)\left(\frac{\Delta x}{2}\right)(1\cdot 1 + 3\cdot 2 + 5\cdot 4 + 7\cdot 4 + 9\cdot 5 + 11\cdot 3 + 13\cdot 3 + 15\cdot 2)$$
$$M_x \approx (.21)(101\Delta x)$$

$$M_y \approx (.21)\left(\frac{\Delta x}{2}\right)(5\cdot 3 + 7\cdot 4 + 9\cdot 8 + 11\cdot 6 + 13\cdot 3)$$
$$M_y \approx (.21)(110\Delta x)$$

Now that we have the values of our moments, we can plug these values back in to our original equations for $\overline{x}$ and $\overline{y}$. Given that the mass M, which represents the area of the region, 4.35, we obtain the following values:

$$\overline{x} = \frac{M_y}{M} \approx 2.443$$

$$\overline{y} = \frac{M_x}{M} \approx 2.243$$

Thus, for the given tiling, a 4G tower will be placed at $(2.443, 2.243)$.

We will show the similar steps for another example, specifically subregion 2 of Region C shown in Figure 7, in which a 5G tower would be placed.

Once again, let the side length of the square be Δx. Although not shown, we can approximate Δx using subregion 3 of Region C, which is approximately the size of a rectangle with side lengths approximately $2.5\Delta x$ and $8\Delta x$. Since the area of subregion 3 is given to be 0.1, we can approximate Δx to be:

$$(2.5\Delta x)(8\Delta x) \approx 0.1$$
$$20\Delta x \approx 0.1$$
$$\Delta x \approx \sqrt{\frac{1}{200}}.$$

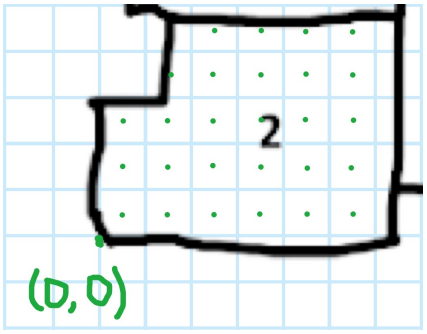

Figure 7: Tiling of Subregion 2

We have the same approximations for the moments:

$$M_x \approx \Delta A \sum_{(x,y)\in P} y$$

$$M_y \approx \Delta A \sum_{(x,y)\in P} x$$

Next, plugging in all of the P and substituting in $\Delta A = (\Delta x)^2 = \frac{1}{200}$ gives us the following values for the moments:

$$M_x \approx \frac{1}{200}(89\Delta x)$$

$$M_y \approx \frac{1}{200}(67\Delta x)$$

Now that we have the values of our moments, we can plug these values back in to our original equations for $\overline{x}$ and $\overline{y}$. Given that the mass M, which represents the area of the region, 0.14, we obtain the following values:

$$\overline{x} = \frac{M_y}{M} \approx 0.225$$

$$\overline{y} = \frac{M_x}{M} \approx 0.169$$

Thus, for the given tiling, a 5G tower will be placed at $(0.225, 0.169)$.

Overall, we can then reproduce the above steps for all of the subsections and regions. For other subsections in the same region, the same tiling dimensions can be used, but for other regions the tiling dimensions will be different. However, the steps can be followed in the exact same pattern just with different values of Δx. Note that if one wanted to have all of the regions around locations with the same tiling, the coordinates can be converted into square miles and then translated accordingly.

4.5 Strengths and Weaknesses

Our model is particularly accurate at determining the optimal point for placing 4G and 5G nodes due to our thorough mathematical approach. Moreover, given the rather tedious

process of establishing 5G networks, our model is able to filter through cities based on indicators that will determine if the region is a suitable area for 5G internet.

However, our model does not deem 4G and 5G networks in a given region as mutually exclusive; therefore, there is a chance for overlap between the two networks. Families who decide to enroll in the 4G plan, in spite of qualifying for the 5G plan, would cause an operational inefficiency in our model.

5 Conclusion

5.1 Further Studies

Our first model is formed on the basis of a layered decision making tree-encompassing correlations, proportional relationships, and a random forest regression. As a result, it becomes increasingly difficult to identify the exact process that our algorithm followed to eventually arrive at its output. Our second model relies heavily on data procured from the first quarter of 2020—when COVID-19 was a paramount confounding variable in the internet usage for any given household. With the pandemic creating a new sense of normality, we can expect to see definitive changes in our use of web-based services, and in turn, access to the internet. Our third model doesn't distinguish particular subregions in terms of the type of connectivity—4G or 5G—provided. As a result, households that, under geographic and fiscal circumstances, qualify for 5G may choose to willingly implement 4G in their homes, leading to inaccuracies in our initial predictions. Recreating our models to account for and track each factor with specific weighting would greatly improve the accuracy of our results.

5.2 Summary

We provided an estimation of the projected internet costs in the US and the UK over the next decade. By first training a machine learning model on factors such as population densities, cost of living, current average download speeds, and current average price per Mbps on data from the continental United States, we were then able to apply the model to the US and UK as a whole. The algorithm ultimately predicted that US prices would drop by $0.23 per Mbps and UK prices would drop by $0.57 over the next decade.

Next, we utilized the computational capabilities of code to run a simulation with a thousand trials to accurately determine the required bandwidth to cover 90% and 99% of internet usage for three scenario households. Using factors like age, education level, and employment status to determine the internet usage tendencies of individuals, we were able to obtain a distribution of maximum bandwidth usage for each household in a week. Looking at the 90th and 99th percentiles allowed us to figure out the minimum bandwidth needed to cover the projected internet usage.

Finally, we developed a method to supply a given area with internet capabilities using 4G

and 5G towers depending on characteristics of a region. Variables like annual household income and population density allowed us to determine whether a region is suitable for use of 5G towers, which provides better connection at a greater cost and lower signal radius. In order to place cell nodes in the most optimal locations, the center of mass of each region was found, allowing us to place the fewest 4G and 5G towers while maximizing area coverage.

Ultimately, ensuring access to both stable and high-speed internet sources curates a paradigm shift for future generations—in which technical or fiscal disparities don't govern end outcomes for any individual. Investing in the infrastructure to efficiently set up such network capabilities is fundamental to establishing a better tomorrow. With the continued understanding and modeling of disruptive technologies including, but not limited to 5G towers and bandwidth allocation, we can foster a resilient working class with the tools at their disposal to stand out, conquer obstacles, and reach excellence.

6 References

[1] - California's Future: Population. Jan. 2018, www.ppic.org/wp-content/uploads/r-118hj2r.pdf.

[2] - Pettinger, Tejvan, et al. "Why Does the Cost of Living Keep Rising?" Economics Help, 20 Oct. 2016, www.economicshelp.org/blog/11826/inflation/why-does-the-cost-of-living-keep-rising.

[3] - "List of States by Population Density." List of States By Population Density, state.1keydata.com/state-population-density.php.

[4] - Cost Of Living Index by State 2021, worldpopulationreview.com/state-rankings/cost-of-living-index-by-state.

[5] - Average Internet Speed by State 2021, worldpopulationreview.com/state-rankings/average-internet-speed-by-state.

[6] - Defeating the Digital Divide Data, MathWorks Math Modeling Challenge 2021, https://m3challenge.siam.org/node/523.

[7] - Shibu, Sherin. "Here's Where People Shell Out the Most and the Least for Internet." PCMAG, PCMag, 16 Nov. 2020, www.pcmag.com/news/heres-where-people-shell-out-the-most-and-the-least-for-internet.

[8] - "How College Differs from High School." Academic Support Programs — Baylor University, https://www.baylor.edu/support_programs/index.php?id=88158.

[9] - "How Much Sleep Do We Really Need?" Sleep Foundation, 31 July 2020, https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need.

[10] - "Household Broadband Guide." Federal Communications Commission, 11 Mar. 2020, www.fcc.gov/consumers/guides/household-broadband-guide.

[11] - "Intelligence Brief: How Much Will We Pay for 5G?" Mobile World Live, 13 June 2019, www.mobileworldlive.com/blog/intelligence-brief-how-much-will-we-pay-for-5g.

[12] - Huang, Michelle Yan. "Why You Shouldn't Get Too Excited about 5G Yet." Business Insider, Business Insider, 16 Oct. 2020, www.businessinsider.com/5g-high-speed-internet-cellular-network-issues-switch-2019-4.

[13] - Catherine McNally Sr. Writer, et al. "How Much Is the Average Internet Bill?" Reviews.org, 8 Dec. 2020, www.reviews.org/internet-service/how-much-is-internet.

[14] - "Q: How Much Do Average Jobs Pay per Month in 2021?" ZipRecruiter, www.ziprecruiter.com/Salarie Salary-per-Month.

7 Appendix

7.1 Part 1: The Cost of Connectivity

q1.py

```
# import necessary libraries
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

# read the data from file
df = pd.read_csv("q1.csv")

# drop the state column as it is not needed for prediction
df.drop("State", axis=1, inplace=True)
# convert values from string to float
df = df.astype(float)
# drop rows with any missing values
df.dropna(axis=0, inplace=True)

# create feature and target dataframes
X = df.drop("Cost per Mbps", axis=1)
y = df["Cost per Mbps"]

# initialize random forest regression model
model = RandomForestRegressor(n_estimators=20, max_depth=5)
# training
model.fit(X, y)
# prediction
preds = model.predict(X)
# RMSE calculation
error = mean_squared_error(y, preds, squared=False)

print(error)

# regression functions for download speeds
def download_speed_US(x):
  a = 1.4994*10**(-168)
  b = 1.21291
  year = 2020 + x
  return a*b**year

def download_speed_UK(x):
  a = 2.3341*10**(-183)
  b = 1.23357
  year = 2020 + x
  return a*b**year

# calculating US internet cost for the next 10 years starting from current values
for i in range(10):
  print(model.predict(np.array([[33.86*(1+i*0.013099), 101.90425531914893*(1+i*0.022), download_speed_func(i)]]))[0])

# calculating UK internet cost for the next 10 years starting from current values
for i in range(10):
  print(model.predict(np.array([[725*(1+i*0.013099), 102*(1+i*0.022), download_speed_UK(i)]]))[0])
```

7.2 Part 2: Bit by Bit

main.py

```
from Simulation import Simulation
from Person import Person
from Activity import *
import matplotlib.pyplot as plt

# Simulation 1: Couple in early 30s

# Create a two new person objects, teacher and lookingForWork, and initialize it with
# all the internet related activities he/she will do during a week
teacher = Person([TraditionalTV(23.8), TVGameConsole(1.73), TVInternetDevice(6.87), InternetComputer(4.77),
                  VideoComputer(1.28), TotalSmartphone(26.98), VideoSmartphone(2.28), AudioSmartphone(0.85),
                  TotalTablet(5.85), VideoTablet(1.07), AudioTablet(0.22), School(25)])

lookingForWork = Person([TraditionalTV(23.8), TVGameConsole(1.73), TVInternetDevice(6.87), InternetComputer(4.77),
                         VideoComputer(1.28), TotalSmartphone(26.98), VideoSmartphone(2.28), AudioSmartphone(0.85),
                         TotalTablet(5.85), VideoTablet(1.07), AudioTablet(0.22)])
# Create a new simulation with two people, teacher and lookingForWork, and simulate internet usage per hour for a week
# The simulation is repeated 1000 times
simulation1 = Simulation([teacher, lookingForWork], 1000)
simulation1.run()

# Output results
print("Simulation 1: ")
print("90% Threshold " + str(simulation1.maxBandwidthWeek[round(simulation1.trials * 0.9) - 1]))
print("99% Threshold " + str(simulation1.maxBandwidthWeek[round(simulation1.trials * 0.99) - 1]))

# Simulation 2: Retired Woman with Grandchildren

# Create three Person objects with related activities
retired = Person([TraditionalTV(50.6), TVGameConsole(0.17), TVInternetDevice(3.20), InternetComputer(3.17),
                  VideoComputer(0.5), TotalSmartphone(17.73), VideoSmartphone(0.97), AudioSmartphone(0.42),
                  TotalTablet(7.22), VideoTablet(0.65), AudioTablet(0.10)])

child1 = Person([TraditionalTV(12), TVGameConsole(2.72), TVInternetDevice(7.72), School(25)])
child2 = Person([TraditionalTV(12), TVGameConsole(2.72), TVInternetDevice(7.72), School(25)])

# Create new simulation with all 3 objects and run 1000 trials
simulation2 = Simulation([retired, child1, child2], 1000)
simulation2.run()

print("Simulation 2: ")
print("90% Threshold " + str(simulation2.maxBandwidthWeek[round(simulation2.trials * 0.9) - 1]))
print("99% Threshold " + str(simulation2.maxBandwidthWeek[round(simulation2.trials * 0.99) - 1]))

# Simulation 3: Three undergraduates

# Create three Person objects
student1 = Person([TraditionalTV(11.35), TVGameConsole(3.63), TVInternetDevice(6.95), InternetComputer(3.97),
                   VideoComputer(1.73), TotalSmartphone(24.83), VideoSmartphone(3.03), AudioSmartphone(1.62),
                   TotalTablet(3.88), VideoTablet(0.98), AudioTablet(0.20), School(16)])

student2 = Person([TraditionalTV(11.35), TVGameConsole(3.63), TVInternetDevice(6.95), InternetComputer(3.97),
                   VideoComputer(1.73), TotalSmartphone(24.83), VideoSmartphone(3.03), AudioSmartphone(1.62),
                   TotalTablet(3.88), VideoTablet(0.98), AudioTablet(0.20), School(16)])

student3 = Person([TraditionalTV(11.35), TVGameConsole(3.63), TVInternetDevice(6.95), InternetComputer(3.97),
                   VideoComputer(1.73), TotalSmartphone(24.83), VideoSmartphone(3.03), AudioSmartphone(1.62),
                   TotalTablet(3.88), VideoTablet(0.98), AudioTablet(0.20), School(16)])

# Create and run simulation
simulation3 = Simulation([student1, student2, student3], 1000)
simulation3.run()

print("Simulation 3: ")
print("90% Threshold " + str(simulation3.maxBandwidthWeek[round(simulation3.trials * 0.9) - 1]))
print("99% Threshold " + str(simulation3.maxBandwidthWeek[round(simulation3.trials * 0.99) - 1]))

# Plots
# Plots the histograms for all three scenarios

plt.hist(simulation1.maxBandwidthWeek, bins='auto')  # arguments are passed to np.histogram
plt.title("Scenario 1 Simulation: Max Hourly Bandwidth Usage in a Week")
plt.xlabel("Bandwidth Usage (Mbps)", fontsize = 14)
plt.ylabel("Frequency", fontsize = 14)
plt.axvline(x = simulation1.maxBandwidthWeek[round(simulation1.trials * 0.9) - 1], color = "orange")
plt.axvline(x = simulation1.maxBandwidthWeek[round(simulation1.trials * 0.99) - 1], color = "red")
plt.show()

plt.hist(simulation2.maxBandwidthWeek, bins='auto')  # arguments are passed to np.histogram
plt.title("Scenario 2 Simulation: Max Hourly Bandwidth Usage in a Week")
plt.xlabel("Bandwidth Usage (Mbps)", fontsize = 14)
plt.ylabel("Frequency", fontsize = 14)
plt.axvline(x = simulation2.maxBandwidthWeek[round(simulation2.trials * 0.9) - 1], color = "orange")
plt.axvline(x = simulation2.maxBandwidthWeek[round(simulation2.trials * 0.99) - 1], color = "red")
plt.show()

```

```
plt.hist(simulation3.maxBandwidthWeek, bins='auto')  # arguments are passed to np.histogram
plt.title("Scenario 3 Simulation: Max Hourly Bandwidth Usage in a Week")
plt.xlabel("Bandwidth Usage (Mbps)", fontsize = 14)
plt.ylabel("Frequency", fontsize = 14)
plt.axvline(x = simulation3.maxBandwidthWeek[round(simulation3.trials * 0.9) - 1], color = "orange")
plt.axvline(x = simulation3.maxBandwidthWeek[round(simulation3.trials * 0.99) - 1], color = "red")
plt.show()
```

Simulation.py

```
from Person import Person
from Activity import *

# Class to run simulations
class Simulation:

    # Store number of trials and the various Person objects
    # Create variables to store bandwidthUsed, which stores all the raw generated bandwidth data
    # and maxBandwidthWeek which stores the max amount of bandwidth used on any single hour  for all trial weeks
    # in the situation
    def __init__(self, personList, trials):
        self.trials = trials
        self.personList = personList
        self.bandwidthUsed = []
        self.maxBandwidthWeek = []

    # Run the simulation
    def run(self):
        # Run a specified number of trials
        for trial in range(self.trials):
            # Store the bandwidth used every hour for this trial week
            bandwidthPerWeek = []

            # Each trial consists of simulating one week and calculating bandwidth used at every single hour during
            # that week
            for hour in range(112):
                # Stores the total amount of bandwidth used by all people during the hour
                totalHourBandwidth = 0

                # Calculate the amount of bandwidth used by each individual and sum it up to get total bandwidth used
                # in that hour
                for person in self.personList:
                    totalHourBandwidth += person.getRandomActivityBandwidth()

                bandwidthPerWeek.append(totalHourBandwidth)

            self.bandwidthUsed.append(bandwidthPerWeek)
            self.maxBandwidthWeek.append(max(bandwidthPerWeek))

        # Sort the maximum bandwidth usage for each week in ascending order
        self.maxBandwidthWeek.sort()
```

Person.py

```
import random

# Person class to store internet browsing tendencies of a person
class Person:
    def __init__(self, activityList):
        # Store a list of activities that the person would perform in a week that requires internet
        self.activityList = activityList

        # Calculate endpoints of intervals for each subsequent activity in order to randomly select activities to
        # perform in a given hour proportional to how long is spent on that activity in a week.
        self.probabilityThreshold = []
        tempSum = 0
        for activity in activityList:
            tempSum += activity.hours
            self.probabilityThreshold.append(tempSum)

    # Randomly select an activity using the calculated interval endpoint
    def getRandomActivityBandwidth(self):
        randNum = random.uniform(0, 112)
        for i in range(len(self.probabilityThreshold)):
            if randNum < self.probabilityThreshold[i]:
                return self.activityList[i].getBandwidthUse()
        return 0
```

Activity.py

```python
import random

# Parent Class Activity contains the general attributes and methods for an activity requiring internet
class Activity:
    # Store values about hours spent, lower bound of bandwidth use for a task, and upper bound
    def __init__(self, hours, minBandwidth, maxBandwidth):
        self.hours = hours
        self.minBandwidth = minBandwidth
        self.maxBandwidth = maxBandwidth

    # Returns a random amount of bandwidth used between the lower and upper bounds
    # If both bounds are equal to each other (data about variation in bandwidth use wasn't available),
    # return the constant bandwidth value
    def getBandwidthUse(self):
        if self.minBandwidth == self.maxBandwidth:
            return self.minBandwidth
        else:
            return random.uniform(self.minBandwidth, self.maxBandwidth)

# Child classes that have predefined lower and upper bounds for bandwidth in Mbps.
class School(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 2, 5)


class TraditionalTV(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 6.5, 6.5)


class TVGameConsole(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1, 3)


class TVInternetDevice(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 5, 8)


class InternetComputer(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1.5, 1.5)


class VideoComputer(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 2, 4)


class TotalSmartphone(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1, 1)


class VideoSmartphone(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 5, 8)


class AudioSmartphone(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1, 1)


class TotalTablet(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1, 1)


class VideoTablet(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 5, 8)


class AudioTablet(Activity):
    def __init__(self, hours):
        Activity.__init__(self, hours, 1, 1)
```