Defeating the Digital Divide: Internet Costs, Needs, and Optimal Planning

Team 15038

February 28, 2021

1 Executive Summary

Providing reliable and sufficient access to high-speed internet has become an important global issue, especially as the pandemic now forces many people to live, learn, and work online. This trend of virtual living is anticipated at least until the foreseeable future, and therefore, requires prompt implementation of efficient internet connection for all, particularly for those who are geographically isolated and economically vulnerable. To smoothly transition into this new technological era, many existing internet access issues need to be addressed, taking into consideration factors such as the cost of connectivity, individual's needs, and location of transmission nodes.

We predicted the cost of broadband access in dollars per Megabits per second over the course of the next ten years in the US and the UK with an exponential decay model. We used exponential regression to fit our model to past broadband pricing and speed data. We combined the Internet speed data from Ookla and Akamai by assuming the Internet provider market remained static and scaled Akamai's unweighted averages using the ratio of their data points for the overlap year of 2017. To obtain data for UK Internet prices, we applied logarithmic regression to data from the UK Office of Communications report and used that function to find the average monthly price for each year from 2010 to 2021. We obtained US Internet prices from various online sources. We then applied exponential regression to the US and UK data sets separately to obtain two models. Our models predict that broadband prices will continue to decrease over the next 10 years but at a diminishing rate, eventually reaching just 1 cent per Mbps in the UK and 2.5 cents per Mbps in the US.

Given the predicted cost reductions in Part 1, we also built a model to anticipate the Internet needs of a given household based upon the age groups of its members. We accomplished this using a Monte Carlo simulation to different probabilities a given person was doing one of several Internet activities at any given moment. The probabilities were determined using the hours per week each age group spent on each activity. This data was obtained from the Nielsen Corporation's 2020 report. For each scenario, we multiplied the number of people performing each activity by the broadband requirements for each activity, which we obtained from the 2017 Nerd Wallet Report and the 2019 CBTNuggets Report. We then tested our model with three scenarios. The Monte Carlo simulation found the broadband needed to meet each households requirement 90% of the time and 99% of the time.

Using our model from Part II, we treated each sub region as an individual, rather large household and simulated their maximum broadband demand in 99% of cases. We then assigned each sub region its numerical value for broadband demand and used simple human reasoning to determine tower placement and frequency band. Due to their relatively large radii of coverage and bandwidth capabilities, single towers can often fully service multiple sub regions simultaneously. As such, there is a functionally infinite amount of possible variations on tower placement and selection. We ran the data for regions A, B, and C through our model, and provided possible suggestions for tower placement in

region A.

2 The Cost of Connectivity

2.1 Problem Statement

Build a model to predict the cost per unit of bandwidth in dollars or pounds per Mbps over the next 10 years for consumers in the United States and the United Kingdom.

2.2 Assumptions

1. **There was no broadband access before 2000.**
    - **Justification:** Broadband Internet was invented in 2000 in the UK. [1].
2. **The US and UK broadband markets behave similarly.**
    - **Justification:** The US and the UK are both developed nations with strong economies, which means that they will have access to new Internet technologies at about the same time. In addition, they have similar cultures, which means that they will tend to have similar demand levels for Internet usage.
3. **The market shares of the cellular service providers across all data points is the same as in the year 2017 as found by the Ookla study.**
    - **Justification:** The main telecom giants that dominate the internet service landscape are difficult to challenge because of their ability to buy-out competing companies. This was in large part due to new legislation, such as the Telecommunications Act of 1996, enacted before the time period analysed. Therefore we can assume market shares will remain the same because of these companies and that their growth is reflective of the whole industry because of their merging power with budding companies [2] [3].
4. **The price for a 'Superfast Connection' plan will continue to decrease over time with the same behaviour as historical data for which a logarithm is an appropriate model for the next 10 years**
    - **Justification:** The model for predicting the price of this plan was built with data from the December 2014 to July 2019. This time period encapsulates the period of growth under 4G which is most recent example of uninterrupted growth under a broadband internet type. Therefore this model can be extrapolated for growth under 5G and later methods if devised. A logarithmic model is an appropriate model because predicted prices become infeasible (negative) only after 52 years from the 2010 baseline which is significantly beyond the next 10 years for which the behaviour is used.

2.3 Variables

Symbol	Definition	Units
P_{UK}	Monthly price of a "superfast" bandwith plan in the UK	Dollars
P_{US}	Monthly price of broadband internet plan in the US	Dollars
S_{UK}	Average wired internet speed within a given year in the UK	Megabits per second (Mbps)
S_{US}	Average wired internet speed within a given year in the US	Mbps
R_{UK}	Average pricing for wired internet within a given year in the UK	Dollars per Mbps
R_{US}	Average pricing for wired internet within a given year in the US	Dollars per Mbps
T	Years since 2000	Years
m	Months since January 2000	Months

2.4 Model Construction

When determining Internet speed for each year, we had to account for Ookla and Akamai's different methodologies. Having assumed that the market shares of companies do not vary, we scaled the Akamai data linearly using the ratio of Akamai's 2017 values to Ookla's 2017 values.

2.4.1 United Kingdom

The UK Office of Communications (OfCom) 2020 report provided data on monthly prices between 2014 and 2019. We used that data to extrapolate the average monthly prices for broadband internet plans each year by performing a logarithmic regression using the form

$$P_{UK} = A \ln (m + e) + B \tag{1}$$

Where A and B are constants. We used months instead of years because the OfCom's data lists by month. We used the translation of e in order to prevent the model from having a vertical asymptote at $x = 0$. Using SciPy, we trained the logarithmic regression model and found A to be -16.892 and B to be 108.923. Using this model, we calculated the monthly price for each year by averaging the price for all 12 months of that year as determined by the model.

We decided to use an exponential decay model for the change in Bandwidth Price over time.

$$R_{UK} = Ae^{b}T \tag{2}$$

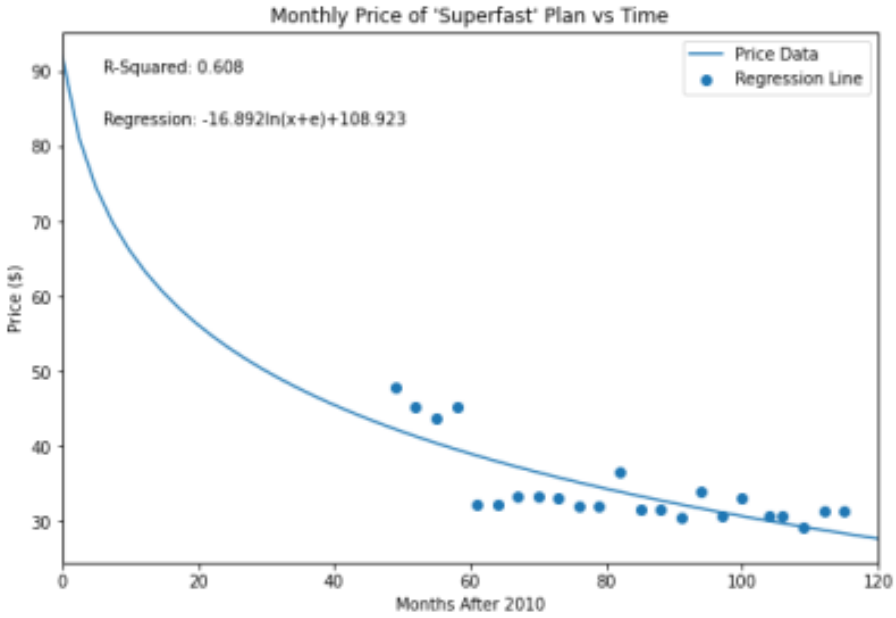


Figure 1: Logarithmic regression modelling monthly price of "superfast" broadband internet in the UK

Table 1: UK data table

Year	**Monthly Price ($)**	**Internet Speed (Mb/s)**	**Bandwidth Price($/Mb/s**
2010	39.805	7.952	5.006
2011	35.398	11.120	3.183
2012	33.183	15.322	2.165
2013	31.646	23.469	1.348
2014	30.490	27.283	1.118
2015	29.556	33.360	0.886
2016	28.772	39.438	0.730
2017	28.097	49.200	0.571
2018	27.505	81.100	0.339
2019	26.976	65.800	0.410
2020	26.499	55.200	0.480
2021	26.064	55.600	0.469

Where A and b are constants. This function fits the general shape of the data, which decreases at a diminishing rate. This function also makes logical sense because technological advancements eventually have diminishing returns

with regard to price. The cheaper broadband becomes, the more difficult it is to reduce the price further. Furthermore, an exponential decay model is good for extrapolation for two reasons. First, because it has a horizontal asymptote at 0, so it never predicts a negative price. Second, it does not have a vertical asymptote at x=0, so it does not predict extremely high values for small values of x.

Using the built in regression functions in Google Sheets, we found A to be 58.4 and b to be -0.264

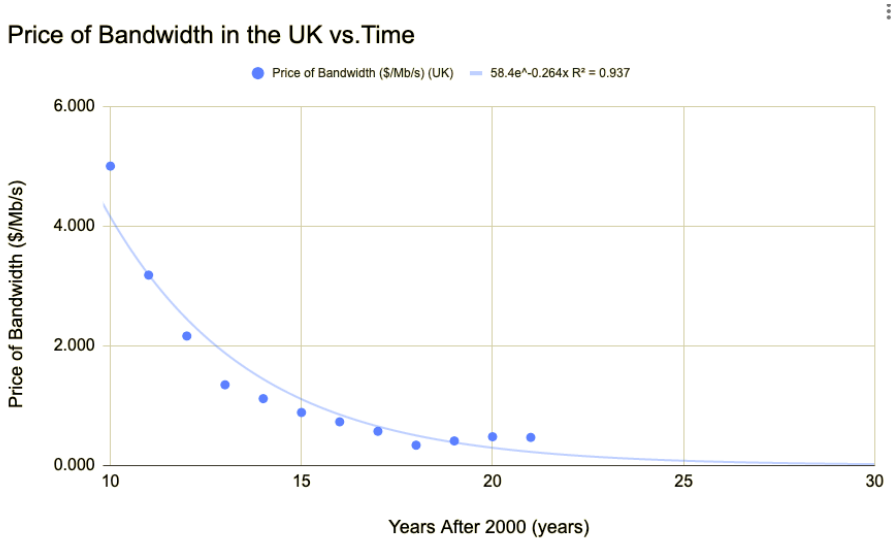


Figure 2: Exponential decay model of bandwidth prices in the UK

2.4.2 United States

Using various sources, we compiled data for the average cost of Internet in the United States from 2013 to 2020 [4–8]

We decided to use the same model type as we did for the UK because of our assumption that the markets behave similarly. This is supported by the similar shape of the scatter plot of bandwidth price levels vs. time. Using exponential regression in Google Sheets, we found A to be 89.6 and b to be -0.266.

2.5 Model Execution

From the exponential decay models created above, the cost per unit of bandwidth per second in dollars over the next 10 years for the United States and the United Kingdom could be predicted respectively by using the subsequent years as inputs and obtaining the output from each equation. In the year 2031, the

Table 2: US data table

Year	**Monthly Price ($)**	**Internet Speed (Mb/s)**	**Bandwidth Price($/Mb/s**
2013	90	29.96	3.00
2014	69.99	33.23	2.11
2015	62	43.63	1.42
2016	74.76	55.49	1.35
2017	66.17	70.8	0.93
2018	65.16	83.2	0.78
2019	72	111.7	0.64
2020	57	134.8	0.42

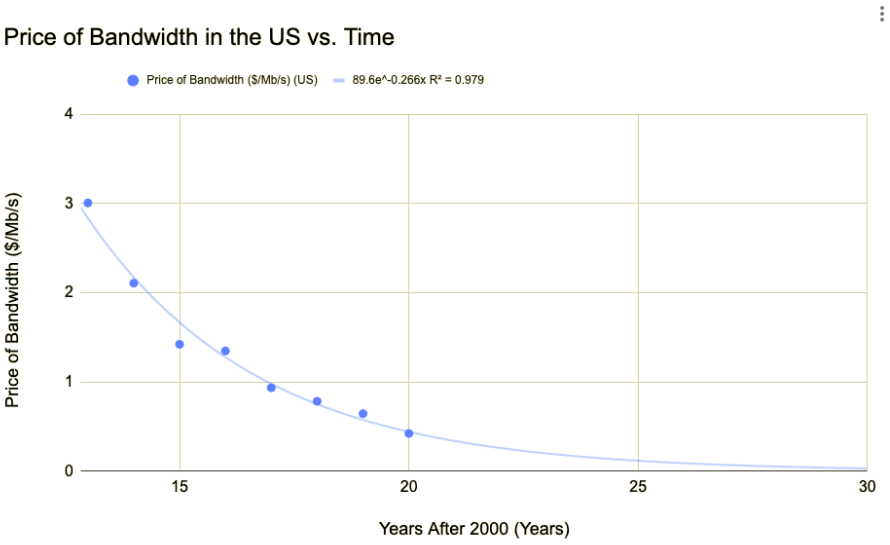


Figure 3: Exponential decay model of bandwidth prices in the US

cost for 1Mbps in the United Kingdom would end up as $ 0.010, while the cost for 1Mbps in the United States in the same year would be $ 0.024.

2.6 Results and Discussion

The models both predict slight reductions in price per Mbps in the US and UK through 2031. We believe these results to be logically sound. As efficiency-increasing technologies such as fiber optic lines are brought to market and implemented into the telecommunications network, bit rates will increase steadily without any major increases in cost. Additionally, competition between internet

Table 3: UK and US predictions

Year	**UK Bandwidth Price($/Mb/s)**	**US Bandwidth Price($/Mb/s)**
2022	0.175	0.258
2023	0.135	0.197
2024	0.103	0.151
2025	0.079	0.116
2026	0.061	0.089
2027	0.047	0.068
2028	0.036	0.052
2029	0.028	0.040
2030	0.021	0.031
2031	0.016	0.024

providers essentially ensures that unit pricing will continue to fall over time. The exponential decay factor b was similar for both countries, but with more data points, especially for the United States, it would be worth exploring whether the vast U.S. territory could result in a less steep decrease in unit bandwidth cost, since it is expected to take more time for the new technology to be implemented in a larger area.

2.6.1 Strengths and Weaknesses

The strength of our models is that the simplicity of their functions means they do not have erratic end behavior and thus can be used to extrapolate for years well outside the range of data used to train the model. This simplicity did not hinder the model either, as the exponential models for the UK and the US fit the data well, with R_2 values of .937 and .929 respectively. This indicates that the exponential models account for most of the variability in the price of bandwidth.

The main weakness of our models is that they do not include mobile internet data. Due to the relative complexity of cell service plans, which often include separate rates and limits for texting, phone calls, and internet access, consistent data on pricing and speed of mobile internet access is almost nonexistent. Some plans specify different limits for video streaming compared to internet browsing, another complicating factor. More importantly, wireless infrastructure is more liquid and turbulent than extensive wired networks. Specialized advancements such as LTE and 5G push the market forward in haphazard jolts, and also make it difficult to reduce bit rates to a representative average value. To avoid confounding the relatively consistent wired internet access data, we chose to exclude wireless internet from our model.

3 Bit By Bit

3.1 Problem Statement

Create a flexible mathematical model to predict a given household's need for the internet over the course of a year.

3.2 Assumptions

1. **A given household is an American household.**
    - **Justification:** Detailed data could be found for the household in the United States, but sufficient data could not be obtained for the United Kingdom household.
2. **All individuals in ages 6-17 attend school daily.**
    - **Justification:** Due to Truancy Laws, the overwhelming majority (about 99% from calculation from the population data) of the U.S. school age individuals are enrolled in school [9] [10].
3. **Children aged 2-11 have the same internet usage habits as the subgroup aged 8-11**
    - **Justification:** Various privacy laws prevent the collection of data on children, making quantitative data on young children's internet usage hard to come by. Because our hard data cut off arbitrarily at 8 years for some of our usage categories, 8-11 data was assumed to be applicable to all children aged 2-11
4. **All employed adults of age 35 to 64 do not attend virtual school.**
    - **Justification:** Although some adults partake in education activities to complete their education or to learn new skills for their career, this is not the majority of adults [11]. Also, most activities do no take substantial time out of working adults' schedule to be considered for the virtual school category.
5. **Adults of age 65 and above do not have a virtual job.**
    - **Justification:** Many adults in the United States retire or give up their job as they become seniors. Although they continue to earn a living through various part-time jobs or self-employment, these jobs in sales or service sector tend to involve much more face-to-face interactions and work than online tasks.
6. **Current levels of virtual learning for k-12 and college students will remain constant for the foreseeable future**

- **Justification:** For obvious reasons, there is no long term data on the effects of the pandemic on schooling yet. To simplify our model, we decided to not attempt to predict how virtual schooling rates would evolve going into the future. Approximate rates of .53 remote, .28 in person, and .19 hybrid were treated as constants. [12]

7. **Each day any given member of a household will sleep 7.5 hours and be awake 16.5 hours**

   – **Justification:** Many general sources confirmed the average American sleeps approximately 7.5 hours each night. Since people do not use the internet when they are asleep, this 7.5 hour period was not included when calculating the probability that a household member was using the internet in a certain way.

3.3 Variable

Symbol	Definition	Units
$T_{i,j}$	Average number of hours spent by group i on activity j each week	Hours
$P_{i,j}$	Probability that a person in group i is doing activity j at any given waking moment	—

3.4 Model Construction

To calculate the probability of the online activities for each individual in the household, a full table of number of hours per week of each activity categorized by age group was needed. We produced this table by either using the given data [D4] Internet Media Consumption or calculating the value from the equation

$$h * p * t$$

where h is the number of hours per week of the activity, p is the percent of age group who participate in the activity, and t is the factor that accounts for the proportion of online time of the activity. The table is provided as Table IV below.

Using the values from the table, we ran a Monte Carlo Simulation. Each trial, an internet activity is randomly selected for each inputted household member, according to the probabilities listed in the table for each demographic group. The bandwidth required for each household member is then summed to determine the total bandwidth for that trial. The purpose of a Monte Carlo model is to glean a deterministic relationship through sampling. We chose this model for this question because of the multivariate complexity of the input in determining overall bandwidth usage as well as percentile values. The Monte Carlo

Table 4: Hours and Probability by Age Group and Activity

	Activity	Age Group						
		2-11	12-17	18-34 (student)	18-34 (worker)	35-49	50-64	65+
Hours	Video Games	2.72	4.18	3.63	3.63	1.73	0.47	0.17
	Virtual School	8.18	12.98	8.65	0	0	0	0
	Video Stream (Phone)	7.23	8.52	3.03	3.03	2.28	1.38	0.97
	Video Stream (Computer)	1.28	2.57	1.73	1.73	1.28	1.07	0.5
	Video Stream (TV)	7.72	4.52	6.95	6.95	6.87	4.95	3.2
	Virtual Work	0	0	3.524	7.048	5.36	5.36	0
	Email/Web Browsing/ Social Media	2.8	12.48	34.51	34.51	38.67	35.3	29.34
Probability	Video Games	0.024	0.036	0.031	0.031	0.015	0.004	0.001
	Virtual School	0.071	0.112	0.075	0.000	0.000	0.000	0.000
	Video Stream (Phone)	0.063	0.074	0.026	0.026	0.020	0.012	0.008
	Video Stream (Computer)	0.011	0.022	0.015	0.015	0.011	0.009	0.004
	Video Stream (TV)	0.067	0.039	0.060	0.060	0.059	0.043	0.028
	Virtual Work	0.000	0.000	0.031	0.061	0.046	0.046	0.000
	Email/Web Browsing/ Social Media	0.024	0.108	0.299	0.299	0.335	0.306	0.254

Table 5: Broadband Requirements for Activities

Activity	Broadband Requirement
Video Games	3
Virtual School	4
Video Streaming (Phone)	3
Video Streaming (Computer)	5
Video Streaming (TV)	25
Virtual Work	5
Email/Web Browsing/-Social Media	1

simulation model runs 10,000 trials and then samples the bandwidth usage to predict the minimum bandwidth to satisfy the desired percentage of time.

Listing 1: Monte Carlo model

```
def Monte(household, precentile, trials=10_000):
  sums=[]
  # enumerate through household people categories
  for i, num in enumerate(household):
    # run trials
    for _ in range(trials):
      # skip if none
      if num!=0:
        # randomly draw weighed sample
        draw = choice(ds.index, num, p=ds.iloc[:, i])
        sum=0
        # sum sample bandwiths
        for activity in draw:
          sum+=bandwidth[activity]
        # append sample to sums
        sums.append(sum)
  # make array
```

```
sums=np.asarray(sums)
# graph and output in chart title
pd.DataFrame(sums,
   columns=["Trials"]).plot(kind='hist',
      subplots=True,sharex=True,sharey=True,
      title=f'input: {household}, precentile: {precentile
          }, model: {int(np.percentile(sums, precentile))}
          Mbps',
      grid=True, bins=100, xlabel="Mbps")
```

3.5 Model Execution

The input is an integer array with each element representing the amount of members in the household-to-be-estimated that fall into our defined age categories and a integer from 0 to 100 representing the percent of time we want the internet bandwidth to satisfy. There is another parameter trials, which is optional and set by default to 10,000. After running and sampling our trials, we find the minimum bandwidth to satisfy 90% and 99% of time by finding the 90^{th} and 99^{th} percentile respectively from the sampling data as the model output.

Listing 2: Monte Carlo use statements

```
# test cases for bandwidth
# case 1
Monte([1, 0, 2, 0, 0, 0], 90)
Monte([1, 0, 2, 0, 0, 0], 99)
```

3.6 Results and Discussion

Our model found the third household of three undergraduate students to require the most broadband, the first household of an unemployed and employed couple with a 3-year-old to require the next most, and the second household of a retired woman with her 2 grandchildren to require the least. The second household and first household had the same requirement of 28 Mbps to cover 99% of the time but had a slightly slower requirement for 90% time, with 5 Mbps compared to the first household's 7 Mbps. This makes sense because the college students require Internet both for online learning and for their part-time jobs and college students use a significant amount of Internet for other leisure activities. The first and second households have similar requirements because they both have two people who use Internet a significant amount (the couple and the two school-aged children) and one person who does not use the Internet a lot (the 3 y/o and the grandmother).

Figure 4: Monte Carlo Simulation Graphs of Broadband Requirements

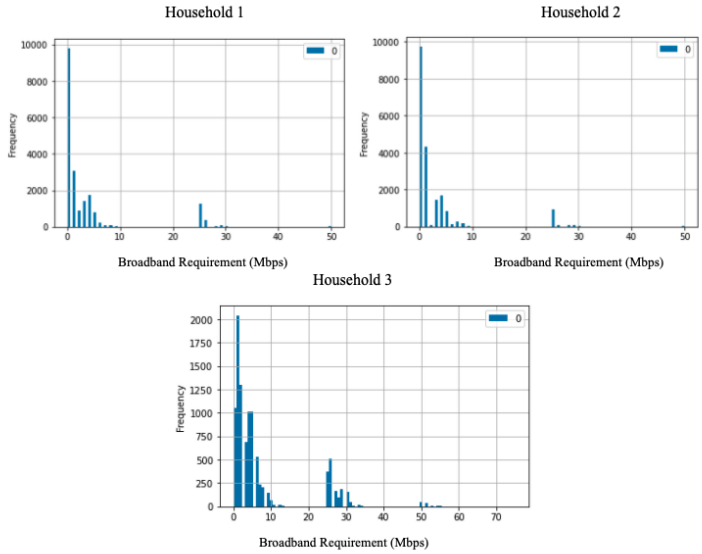


Table 6: Predictions for Minimum Broadband Requirements

	Minimum Broadband Required (Mb/s	
Household	**90% of the time**	**99% of the time**
Unemployed and employed 30 y/o couple w/ 3 y/o	7	28
Retired 70 y/o with 2 school-aged grandchildren	5	28
Three undergraduate students w/ part time jobs	26	50

3.6.1 Strengths and Weaknesses

The strength of this simulation model is that it can account for every possible Internet use scenario based on the household. This makes it flexible and

applicable to any combination of the predetermined age groups. Because the Monte Carlo simulation was run 10,000 times, we are confident that the output distribution is representative of the household's Internet use over the course of a long period of time.

The weakness of this model is that it does not account for different levels of virtual work and/or education, instead using the average number of hours spent by the entirety of each group. This makes the model less flexible. In the future, we would address this issue by adding parameters between 0 and 1 to the Monte Carlo simulation that would define the weight of education and work to account for different levels of participation. They would be multiplied by a constant probability, thus modifying the probability the person is doing virtual school or work based on their level of participation.

4 Mobilizing Mobile

4.1 Problem Statement

Develop a model that produces an optimal plan for distributing cellular nodes in a region.

4.2 Assumptions

1. **Demographic proportions of each region are constant across sub regions**

    - **Justification:** In order to use our model from Part II, we needed demographic data for each sub region, so we modeled each sub region to have the same demographic breakdown as the totals given.

2. **Households only use 4G or 5G wireless internet for smartphones**

    - **Justification:** Any TV or computer usage would run through wired internet.

3. **Households with only a smartphone would use it 2x as much as households with a smartphone as well as other devices**

    - **Justification:** The factor of 2 is an arbitrary multiplier, but basic reasoning suggests that a lack of other devices to use would make households with only a smartphone use it significantly more often. It would be impossible to calculate the true increase in smartphone usage without more comprehensive data, which either does not exist or was not accessible to us.

4.3 Model Construction

Rather than create a complex two-dimensional geographical model, we decided to utilize our model from part II to determine a single figure for broadband need to satisfy 99% of cases within each sub region. Because our model in part II calculates total internet usage for a single household, with no distinction between broadband and wired connection, slight modifications were needed. Because of our assumption that only smartphone usage would factor into broadband need, many internet usage cases were simply disregarded. In households where the only internet access was through a smartphone and therefore broadband, we made the further assumption that the smartphone would be used twice as much as it would have otherwise to compensate for the lack of alternative internet access. Our final alteration to the model in Part II was to expand it to model each sub region as an individual household. This required no change to the model itself, but the demographic information that we calculated for each sub region had to be translated into numerical values for the amount of people within each demographic group we had outlined for Part II. Some minor assumptions

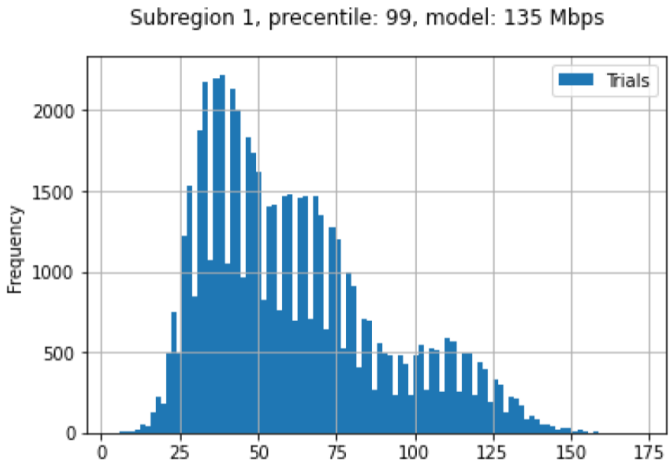
were made to fill in mild discrepancies between the group cutoffs in D8 and in our own table for Part II.

Figure 5: Activity Probability Available to Cellular Users

Activity	Probability (2-11)	Probability (12-17)	Probability (18-34 students)	Probability (18-34) worker	Probability (35-49)	Probability (50-64)	Probability (65+)	Bandwidth
Video Games	0.0255	0.0392	0.0341	0.0341	0.0162	0.0044	0.0016	3
Virtual School	0.0768	0.1209	0.0812	0	0	0	0	4
Video Streaming (Phone)	0.0679	0.08	0.0285	0.0285	0.0214	0.013	0.0091	3
Virtual Work	0	0	0.0177	0.0352	0.0268	0.0268	0	5
Email/Web Browsing/Social Media	0.0263	0.1	0.321	0.324	0.3631	0.3315	0.2755	1
None	0.8035	0.6599	0.5175	0.5782	0.5725	0.6243	0.7138	0

4.4 Model Execution

We modeled each sub region in regions A as an individual household, producing the following frequency distributions with the disclosed maximum broadband needed to meet demand 99% of the time. Region A:



1: 135 Mbps 2: 263 Mbps 3: 242 Mbps 4: 60 Mbps 5: 234 Mbps 6: 258 Mbps
Region B:
1: 584 Mbps 2: 329 Mbps 3: 201 Mbps 4: 183 Mbps 5: 383 Mbps 6: 370 Mbps 7: 281 Mbps
Region C:
1: 452 Mbps 2: 497 Mbps 3: 318 Mbps 4: 402 Mbps 5: 405 Mbps 6: 317 Mbps 7: 546 Mbps

From here, we can simply hand pick locations for towers to meet the demands of each sub region, using the data from D9- Mobile Broadband Frequency Band

Characteristics. A simple area calculation using the provided radius for each tower was utilized. In region A, for example, 2 mid-range towers placed literally anywhere within the region would exceed all demand.

4.5 Results and Discussion

Our model simplifies the complex demographic information provided for each region - information that could be reasonably obtained for any geographic region in the world - into single numerical values that express the maximum possible bandwidth demand in 99% of cases. This makes it easy to prioritize tower selections and placements that can conform to real world conditions that are impossible to model simply, like real estate availability and local building regulations.

4.5.1 Strengths and Weaknesses

The model from Monte Carlo Simulation allows us to express total maximum demand for each sub region and region. It assigns a simple scalar quantity to each region, which can easily be managed by hand.

However, this model does not actually provide specific instructions for placing towers or selecting frequency bands. Its simplicity is a double edged sword of sorts; it allows for flexibility when applied to the real world, and can adapt to outside factors; but it does not even consider the outside factors.

References

[1] https://www.computerworld.com/article/3412338/a-history-of-uk-broadband-roll-out–bt–openreach-and-other-major-milestones.html

[2] Fu, Hanlong Mou, Yi Atkin, David. (2015). The Impact of the Telecommunications Act of 1996 in the Broadband Age.

[3] The Cable Communications Policy Act of 1984: A Balancing Act on the Coaxial Wires, 19 Ga. L. Rev. 543 (1985)

[4] https://www.bbc.com/news/magazine-24528383

[5] https://www.ncta.com/broadband-facts

[6] https://www.analysisgroup.com/uploadedFiles/Content/Insights/Publishing/Broadband_Competition_Report_November_2016.pdf

[7] https://www.statista.com/chart/11963/the-most-and-least-expensive-countries-for-broadband/

[8] https://www.reviews.org/internet-service/how-much-is-internet/: :text=On

[9] https://www.census.gov/newsroom/press-releases/2019/school-enrollment.html

[10] https://www.kidsdata.org/topic/34/child-population-age-gender/tablefmt=141loc=1tf=108ch=1433,926,927,1434,1435,372,78,77,79 sortColumnId=0sortType=asc

[11] https://www.bls.gov/careeroutlook/2017/article/older-workers.htm

[12] https://www.educationnext.org/pandemic-parent-survey-finds-perverse-pattern-students-more-likely-to-be-attending-school-in-person-where-covid-is-spreading-more-rapidly/

5 Code Used

5.1 Part I

Listing 3: Modelling from provided data [1]

```
# import statements
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy as sp

# import data from xlsx of Defeating the Digital Divide
    Data,
# MathWorks Math Modeling Challenge 2021,
# https://m3challenge.siam.org/node/523.
df = pd.read_excel('/content/M3_Challenge.xlsx',
                   sheet_name='Sheet1',
                   skiprows=[0], header=None,
                   names=("Date",
                          'M_After_2014',
                          'M_after_2010',
                          'Price_GBP',
                          'Price_USD'));

# format data
df.set_index("Date",
             inplace=True)

# define function to model plan cost
def func(x, a, c):
  return a*np.log(x+np.e)+c

# fit data
cof, cov = sp.optimize.curve_fit(
    xdata=df.M_after_2010,
    ydata=df.Price_USD,
     f=func)
cof

# calculate r^2

df['residuals'] = df.Price_USD - func(df.M_after_2010,
    cof[0], cof[1])
df['residuals_sq'] = df.residuals**2
```

```
df['y_dif_sq'] = (df.Price_USD-df.Price_USD.mean())**2
r_squared = 1-df.residuals_sq.sum()/df.y_dif_sq.sum()
r_squared = np.round(r_squared, 3)
r_squared

# create plot
fig, ax = plt.subplots(figsize=(9,6))
# make scatter plot
ax.scatter(df.M_after_2010,
           df.Price_USD);
# plot log regression line
plt.plot(np.linspace(0, 120),
         func(np.linspace(0, 120),
              cof[0],
              cof[1]))
# add labels to chart
plt.xlabel("Months After 2010")
plt.ylabel("Price ($)")
plt.legend(("Price Data",
            "Regression Line"),
            loc=0)
plt.title("Monthly Price of 'Superfast' Plan vs Time")
plt.xlim([0, 120])
# annotate r^2 and regression
ax.annotate(f'R-Squared: {r_squared}',
            xy=(60, 45),  xycoords='data',
            xytext=(0.05, 0.95),
            textcoords='axes fraction',
            horizontalalignment='left',
            verticalalignment='top',
            )
cof[0] = np.round(cof[0], 3)
cof[1] = np.round(cof[1], 3)
ax.annotate(f'Regression: {cof[0]}ln(x+e)+{cof[1]}',
            xy=(60, 45),  xycoords='data',
            xytext=(0.05, 0.85),
            textcoords='axes fraction',
            horizontalalignment='left',
            verticalalignment='top',
            )
plt.show();

# import data from xlsx of Defeating the Digital Divide
    Data,
# MathWorks Math Modeling Challenge 2021,
# https://m3challenge.siam.org/node/523.
```

```
ds = pd.read_excel(io='/content/M3_Challenge_(1).xls',
                   sheet_name="Sheet3",
                   skiprows=[0, 12],
                   header=None,
                   names=("Years", "Price"));

# format data
ds.set_index(ds["Years"]-2000,
             inplace=True)
ds.dropna()

# define function to fit bandwith price
def funcB(x, a, c):
  return a/(x)+c

# fit data
cof, cov = sp.optimize.curve_fit(xdata=ds.index,
                                 ydata=ds.Price,
                                 f=funcB)
cof

# calculate r^2

ds["residuals"] = ds.Price - funcB(ds.index, cof[0], cof
    [1])
ds["residuals_sq"] = ds.residuals**2
ds["y_dif_sq"] = (ds.Price-ds.Price.mean())**2
r_squared = 1-ds.residuals_sq.sum()/ds.y_dif_sq.sum()
r_squared = np.round(r_squared, 3)
r_squared

# create plot
fig, ax = plt.subplots(figsize=(9,6))
# make scatter plot
ax.scatter(ds.index, ds.Price);
# plot log regression line
plt.plot(np.linspace(0, 30),
         funcB(np.linspace(0, 30),
               cof[0],
               cof[1]))
plt.plot(np.linspace(0, 30),
         np.zeros(50))
# add labels to chart
plt.xlabel("Years after 2000")
plt.ylabel("Price of Bandwidth ($/Mb/s)")
plt.legend(("Regression Line",
```

```
                "y=0",
                "Price Data"),
                loc=0)
plt.title('Price of Bandwidth ($/Mb/s) vs. Year')
plt.xlim([0, 30])
plt.ylim([-5, 50])
# annotate r^2 and regression
ax.annotate(f'R-Squared: {r_squared}',
                xy=(0, 25), xycoords='data',
                xytext=(0.15, 0.95),
                textcoords='axes fraction',
                horizontalalignment='left',
                verticalalignment='top',
                )
cof[0] = np.round(cof[0], 3)
cof[1] = np.round(cof[1], 3)
ax.annotate(
    f'Regression: ({cof[0]}/x){cof[1]}',
                xy=(0, 25), xycoords='data',
                xytext=(0.15, 0.85),
                textcoords='axes fraction',
                horizontalalignment='left',
                verticalalignment='top',
                )
plt.show();
```

5.2 Part II

Listing 4: Modelling from provided data [1]

```
# import statements
import numpy as np
import pandas as pd
from numpy.random import choice

# load data
ds = pd.read_excel(io="/content/M3TableQ2.xls",
    sheet_name="Sheet8")

# format relative frequency data
ds.dropna()
ds=ds.drop(columns="Activity")
# format bandwith usage data
bandwidth=ds.iloc[:, 7]
```

```
ds=ds.iloc[:, 0:7]

# input has format: number of people in household of ages
    (2-11) (12-17) (18-34) (35-49) (50-64) (65+)

def Monte(household, precentile, trials=10_000):
  sums=[]
  # enumerate through household people categories
  for i, num in enumerate(household):
    # run trials
    for _ in range(trials):
      # skip if none
      if num!=0:
        # randomly draw weighed sample
        draw = choice(ds.index, num, p=ds.iloc[:, i])
        sum=0
        # sum sample bandwiths
        for activity in draw:
          sum+=bandwidth[activity]
        # append sample to sums
        sums.append(sum)
  # make array
  sums=np.asarray(sums)
  # graph and output in chart title
  pd.DataFrame(sums, columns=["Trials"]).plot(kind='hist'
      ,subplots=True,sharex=True,sharey=True,
                            title=f'input: {household},
                                precentile: {precentile},
                                model: {int(np.percentile(
                                sums, precentile))} Mbps',
                            grid=True, bins=100, xlabel="
                                Mbps")

# test cases for bandwidth
# case 1
Monte([1, 0, 2, 0, 0, 0], 90)
Monte([1, 0, 2, 0, 0, 0], 99)
# case 2
Monte([0, 2, 0, 0, 0, 1], 90)
Monte([0, 2, 0, 0, 0, 1], 99)
# case 3
Monte([0, 0, 3, 0, 0, 0], 90)
Monte([0, 0, 3, 0, 0, 0], 99);
```

5.3 Part III

Listing 5: Modelling Towers from Part II Model, Region A [1]

```
# import statements
import numpy as np
import pandas as pd
from numpy.random import choice

# load data
data = pd.read_excel(io="/content/M3_Challenge_regionA.
    xls", sheet_name="Sheet9")
ds = pd.read_excel(io="/content/M3TableQ2.xls",
    sheet_name="Sheet8")


# format relative frequency data
ds.dropna()
ds=ds.drop(columns="Activity")
# format bandwith usage data
bandwidth=ds.iloc[:, 7]
ds=ds.iloc[:, 0:7]
ds

# input has format: number of people in household of ages
    (2-11) (12-17) (18-34) (35-49) (50-64) (65+)

def Monte(household, precentile, subregion, trials=10_000
    ):
  sums=[]
  # enumerate through household people categories
  for i, num in enumerate(household):
    # run trials
    for _ in range(trials):
      # skip if none
      if num!=0:
        # randomly draw weighed sample
        draw = choice(ds.index, num, p=ds.iloc[:, i])
        sum=0
        # sum sample bandwiths
        for activity in draw:
          sum+=bandwidth[activity]
        # append sample to sums
        sums.append(sum)
  # make array
  sums=np.asarray(sums)
```

```
  # graph and output in chart title
  pd.DataFrame(sums, columns=["Trials"]).plot(kind='hist'
      ,subplots=True,sharex=True,sharey=True,
                              title=f'{subregion},␣precentile
                                  :␣{precentile},␣model:␣{int(
                                  np.percentile(sums,␣
                                  precentile))}␣Mbps',
                              grid=True, bins=100, xlabel="
                                  Mbps")

# test cases for bandwidth
for col in data:
  if "Category" not in col:
    Monte(data[col].values, 99, col)
```

*Only data import statements differed to get data for different regions