MathWorks Math Modeling Challenge 2023

Eltham College

Team #16474, London, England

Coach: Benjamin Eastley

Students: Oles Chaban, Atharv Gupta,Ben Robinson, Alice Sanderson, Ethan Southward

M3 Challenge TECHNICAL COMPUTING WINNER
$3,000 Team Prize

JUDGE COMMENTS

Specifically for Team # 16474 —Submitted at the Close of Triage Judging:

COMMENT 1: You presented good critical and analytical thinking,

COMMENT 2: I loved the idea of an SIR model, but I'm not sure about the population definitions, in particular the "Bike" population. However, I'm not sure what the third "recovered" population would be. Perhaps there would be a way to incorporate this idea further to include environmental factors.

COMMENT 3: Executive Summary: Good detail, model description, and result statements. Beware typos! (I didn't take off because of possible language barriers, but still…)

Q1: Good description of the SIR model, although the final use of the exponential model was simplistic. The integral was a good, unusual touch.

Q2: Interesting "impurity" model and improvements, but no other model explains ranking of factors aside from environmental awareness.

Q3: Good model and reasonable results but could have used more explanation.

+1 - cited sources and novel models

COMMENT 4: Great job with the summary!

Smart ideas, beautifully executed. Reasonable assumptions, strengths and weaknesses addressed. Learn about sensitivity analysis to improve your modeling.

Specifically for Team # 16474 —Submitted at the Close of Technical Computing Judging:

This team leveraged technical computing for all aspects of the challenge problem. For Part 1 they designed a simple but effective state transition model inspired by the classic SIR model. They used technical computing effectively, both to determine model parameters based on historic data, and to actually implement and evaluate the model's output. The team's MATLAB code for this part was clear and easy to follow. Judges were also impressed with the team's approach to Part 3. They designed a conceptual clean Monte Carlo model for traffic congestion. They implemented it effectively in MATLAB to analyze the impact of increased e-bike ridership on congestion. The results were shown clearly via plots that were all generated in code. One area for improvement was the approach to Part 2. While the team's Python code for this problem was well written and easy to follow, the use of a heavy-duty machine learning method like Random Forest Regression was not well justified, especially given the very small dataset being analyzed. The ML model also did not help the team to address the question of causation. Overall, this was a great paper that showed great creativity in the use of technical computing!

M3 test

#16474

March 4th 2023

Recent changes in modes of urban transportation is especially noticeable for the growth of e-bikes. Over the last 10 years, e-bikes have grown significantly in popularity [1] and are often touted as a solution to the challenges of traffic congestion, carbon dioxide emissions and obesity in 21st century cities. Our team aims to quantify the ways in which we expect the popularity of e-bikes to grow in the future; the factors fuelling this remarkable growth; and the effects this growth will have on the lives of citizens.

To forecast the sales of e-bikes, our team elected to use an SIR-inspired differential equation model, considering the spreading influence of e-bikes as they are normalised in society. We successfully model consumers as moving through three states: bike owners, e-bike owners and those without a bike. Leveraging an exponential fit on data from recent years, we integrate e-bike sales to estimate the initial state of our model. Solving the model numerically, we predict 1,329,887 e-bike sales in 2025 and 1,762,120 e-bike sales in 2028, both in the USA.

We evaluate a number of factors that might be fuelling the growth of e-bikes using a random forest regression. We input linear interpolation to produce a large dataset, from which we can train an accurate model to predict e-bike usage from five input factors. Using the mean decrease impurity as a measure for feature importance, we evaluate the importance of each factor relative to the others. Disposable income was found to be the most important factor, which is justified by the high cost of e-bikes [1]. The least important factor was found to be environmental awareness, likely due to the issues still surrounding e-waste when purchasing and disposing of e-bikes [8]. In order of decreasing importance, the full list of factors is: disposable income, urban population, battery price, gas price, and environmental awareness.

We quantify the effect of having more e-bikes in urban population, by accounting for the change in two important factors - congestion and carbon emissions. We first model the congestion of commuters with the current proportion of e-bikes across various different levels of maximum congestion (throughput), and then repeat the process after 10 years, with a new number of cars on the road. This gave us up to an 11% decrease in congestion after 10 years, and up to a 146.06g gram decrease in CO_2 emissions per commuter (who still uses a car). Which shows the positive impact of e-bikes on both the environment, and peoples' time.

Contents

1 Part I: The Road Ahead

1.1 Restatement of the Problem

We have been tasked with predicting how many e-bikes will be sold in the US in 2025 and in 2028.

1.2 Assumptions

1. *There will be no change in the rate of development of e-bike technology.* If there was any significant development this would be impossible for us to predict so is ignored in our model.
2. *People will not buy an e-bike unless they do not have a working bike.* Most people will not be persuaded to buy a new bike if they already have a functional bike. This is especially true given the significantly higher cost of an e-bike [1] relative to a conventional bike.[5]
3. *Any person will only buy at most 1 bike in any 10 year period.* For both e-bikes and standard bikes, most broken parts can be replaced, including the battery in e-bikes. The lifetime of a frame however, which cannot be replaced, is approximately 10 years [11]. We assumed there would be no bike sharing and ignored stolen and lost bikes to simplify the model.
4. *Only people between ages 15 and 64 buy e-bikes and the size of that population is constant.* Most children do not have e-bikes and the population over 64 are, on average, less able to use bikes and have less need of them as they retire and no longer have to commute. We have modelled this population as 217 million (65% [13] of 334 million [6]). This population will not change significantly in the next 5 years.
5. *The main reason for someone getting an e-bike is being influenced by meeting others with e-bikes.* Seeing others with e-bikes will help consumers understand the full benefits of e-bikes and including the impact of marketing and other factors is too complicated to be completed in the time frame.
6. *There is a homogeneous mixing of people with and without e-bikes.* As we are using an SIR model, we have to assume a homogeneous mixing of the 'infected' group (those with e-bikes) and the 'uninfected' group (those without bikes or e-bikes).

1.3 Variables

Parameter	Description	Value
M_0	Initial proportion of population with no bike	113,841,000 people
B_0	Initial proportion of population with a standard bike	100,000,000 people
E_0	Initial proportion of population with an e-bike	3,159,000 people
α	"Infection" rate, the probability of a non-bike owner buying an e-bike because of an interaction with an existing e-bike owner	2.580×10^{-9} bikes/yr $\times$ person^2
γ	The proportion of non-bike owners who will purchase a conventional bike in a year	0.1625 e-bikes/yr $\times$ person
z	(Lifespan of a standard bike in years)$^{-1}$	0.1 yr^{-1}
w	(Lifespan of an e-bike in years)$^{-1}$	0.1 yr^{-1}

Table 1: A summary of variables for our models. Values will be subsequently explained.

1.4 Model Development

1.4.1 Defining the System

We've used an altered SIR model as the spread of e-bikes in the U.S. is comparable to the spread of an epidemic where e-bike owners are 'infected' and may influence their friends to buy e-bikes. We define three states: e-bike owners (E), (conventional) bike owners (B) and people who do not own a bike (M).

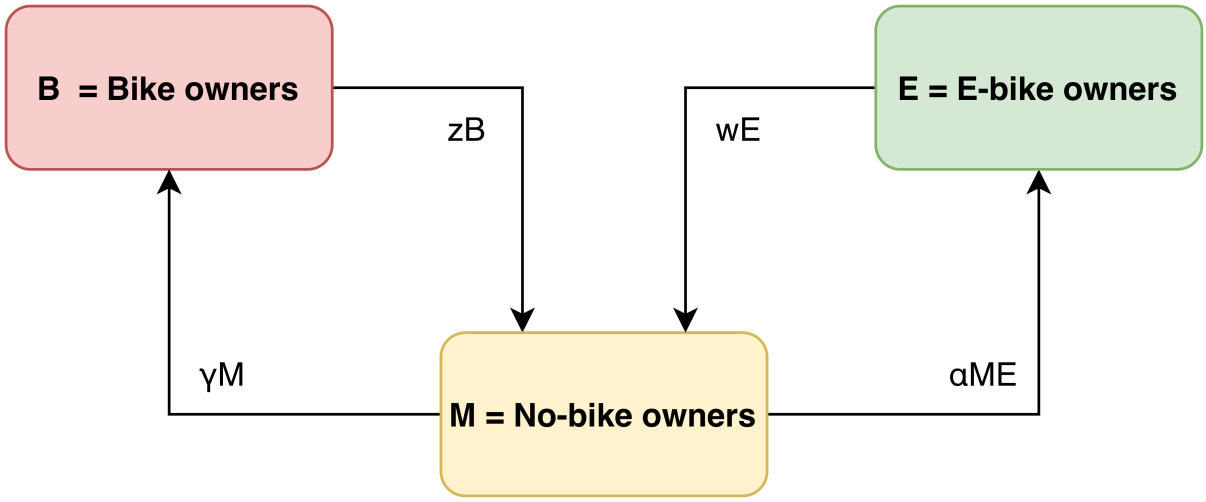


Figure 1: The states considered in our differential equations.

There are no direct transitions between states B and E. Limited data means modelling such transitions would not be feasible. Instead, we consider consumers moving first to the no-bike state before buying a new bike.

As per our assumptions, no one who currently owns a bike will buy a new bike. This means the number of e-bikes bought in a year can be modelled as the number of people who transfer from the 'no bike' to 'e-bike' state. Recall our assumption that e-bike purchases are mostly influenced by consumers seeing others with e-bikes and hence understanding the benefits of owning an e-bike; using this, we model the sales of e-bikes to be proportional to both the susceptible 'no-bike' population and the population who owns an e-bike. This is only valid for a well-mixed population, as in standard SIR models.

$$\frac{dE}{dt} = \alpha ME - wE \tag{1}$$

Where wE is the rate at which bikes fail, transferring owners to the no-bike state from which they can subsequently purchase either type of bike.

For the number of (conventional) bike owners, we assume bike sales are proportional to the size of the current population without a bike. Since conventional bikes are a stabilised market, we feel it should not be influenced by the same 'infection' dynamics as e-bikes. γ is the proportion of the no-bike population that will buy a bike in any given year.

$$\frac{dB}{dt} = \gamma M - zB \tag{2}$$

Where zB is an equivalent term for conventional bikes as to wE for e-bikes. It removes bike owners at the rate their bikes fail.

Given the differential equations 1 and 2, and the constraint that the rate of change of the total population $M + B + E$ is 0. We derive the final differential equation for our system:

$$\frac{dM}{dt} = wE + zB - \alpha ME - \gamma M \tag{3}$$

1.4.2 Determining Variable Values

To successfully model this system, we need to find the initial number of people in each state in 2022 (represented as M_0, E_0 and B_0) as well as the values of parameters α, γ, z and w.

Total eligible population ($M + E + B$): 217 million given by assumption 3.

Number of people with conventional bikes initially (B_0): This number is difficult to estimate given the length of time bikes have been sold for. Sources [15] appear to indicate approximately 100,000,000 bikes in the US in 2022.

Number of initial e-bike users (E_0): We used an exponential fit on to model bike purchases from 2012 until 2022 in the USA (see figure **??**). We found this to be the best fit short-term fit to 2022, with an R^2 value of 0.9182. We do not necessarily expect this trend to continue into the future, since there are a number of changing factors, and continued exponential growth is unlikely. However, over this shortened time frame, and taking into account the recent increase in "coolness factor", a brief exponential increase is plausible. We took 2012 as our starting year, as any e-bikes purchased beforehand would be no longer in circulation, as we found the maximum life of a bike to be 10 years, in accordance with source [11].

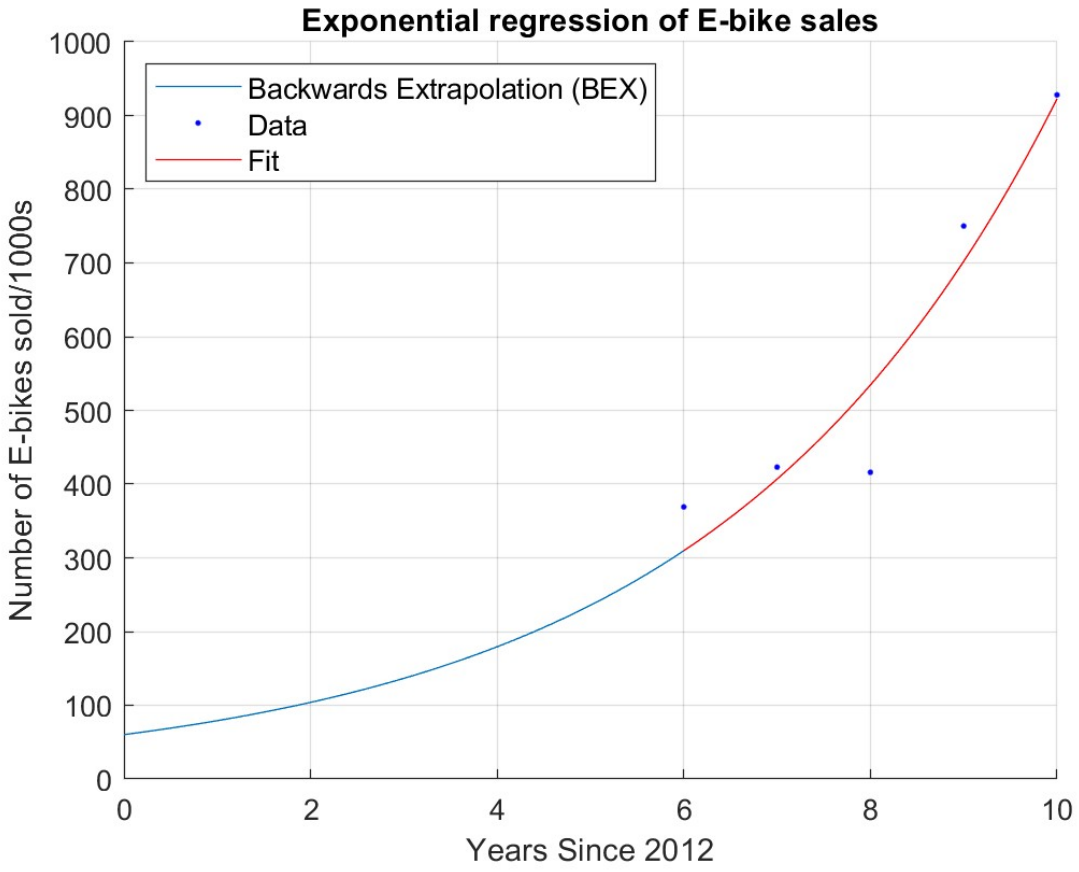


Figure 2: Regression of years since 2012 by number of e-bikes purchased.

This regression had the equation $y = 60.21e^{0.2729x}$. We then estimated E_0 to be the integral

$$E_0 = \int_0^{10} (60.21e^{0.2729x})dx \approx 3,159,000$$

Number of people with no bikes initially (M_0): This value can be found by subtracting E_0 and B_0 from M_0, giving 113,841,000 people.

Proportion of regular bikes replaced each year (z): According to source [11], a bike frame lasts up to 10 years, thus 1/10 bikes break each year. $z = 0.1$

Proportion of e-bikes replaced each year (w): As per our assumptions, the proportion of e-bikes replaced each year will be taken as identical to z.

'Infection' rate of e-bike sales (α): Recall from equation 1 that the sales of bikes is modelled by the expression αME. Knowing there were 928,000 e-bikes sold in 2022 [1], we can rearrange to solve for α.

$$\alpha = \frac{\text{e-bike sales in 2022}}{ME} \approx 2.580 * 10^{-9} \text{e-bikes/yr} * \text{person}^2$$

Rate of conventional bike purchasing (γ): Online sources vary in their estimation of bike sales in the USA in 2022. Estimates around 18.5 million appear to be most plausible.[15] Rearranging the expression γM from equation 2 for the yearly sales of conventional bikes allows us to calculate γ:

$$\alpha = \frac{\text{conventional bike sales in 2022}}{M} \approx 0.1625 \text{bikes/yr} * \text{person}$$

This is how we obtained the values of the parameters in table 4.

1.5 Results

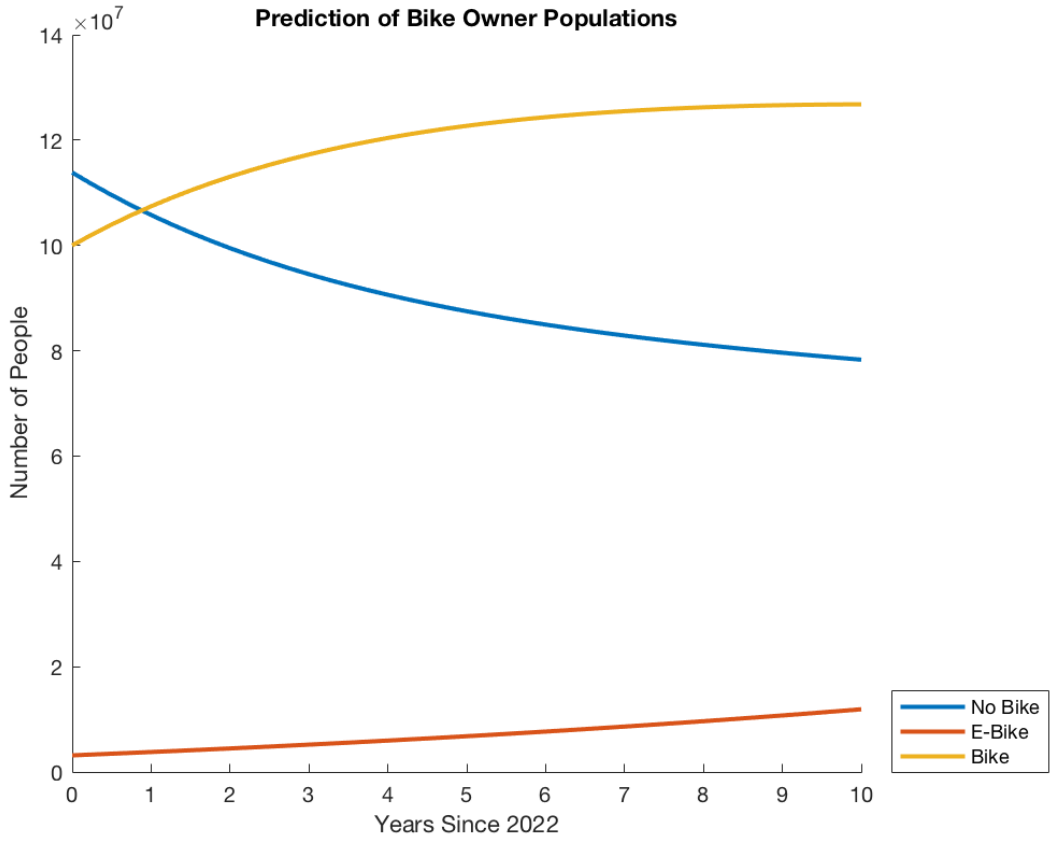


Figure 3: Predictions of our model over the next 10 years, evaluated numerically.

Evaluating the model numerically across the next 10 years predicts the number of e-bike owners to increase at an increasing rate. This is consistent with the data provided.

To calculate sales, we recall equation 1 where the sales of e-bikes was modelled as αME.

Year	Predicted Sales of E-bikes
2025	1,329,887
2028	1,762,120

Table 2: Sales of e-bikes in 2 and 5 years as predicted using our model

1.6 A Review of the Model

The model is an adapted SIR model with e-bike owners modelled as 'infected', standard bike owners as 'immune' and those without bikes as 'susceptible'. Our model does not include any transitions from bike to e-bike or vice versa.

Strengths

- By using an SIR-inspired model, we were able to incorporate the "word of mouth" spread of e-bikes, so we expect the model to accurately predict the way in which society comes to realise the advantages of e-bikes.
- Our model predicts reasonable sales for the next 5 years, particularly considering the artificially inflated sales of 2021 and 2022 due to a cycling "boom" in the aftermath of the pandemic [2].

Weaknesses

- Our model does not fully consider the impact e-bikes will have on the rate of people purchasing conventional bikes.
- Some people may be considered a non-bike owner but also be not susceptible to purchasing a bike (e.g. due to disability, location or other factors). Additionally, some people may be considered a non-bike owner but also not susceptible to purchasing a bike (e.g. due to disability, location or other factors). Adding this as an extra state may be an improvement to our model.
- Although the number of e-bikes sold does reduce the number of conventional bikes sold, the rate at which a non-bike owner buys a conventional bike is still constant in our model, which is unlikely to remain true as e-bike consumption increases.
- To simplify the model, we excluded those under 15 and over 64. To improve this model, we could include these age groups with slightly altered values of α and γ.

2 Part II: Shifting Gears

2.1 Restatement of the Problem

In this problem, we were tasked with finding the most significant factor in the increase of sales of e-bikes, out of environmental concern, gas prices, urban population, cost of batteries and personal finances. We aim to quantify the *relative* importance of each of these factors.

2.2 Assumptions

1. *There are no extreme data values not represented in our data set.* We assume data varies in an approximately linear way between recorded values. This allows us to perform piece-wise linear interpolation between data points to generate additional data for our model. We think this assumption is justified since the data provided [1] was recorded relatively frequently (each year).
2. *Feature importance in a random forest regression is a valid proxy for the real-world importance of factors in increasing the usage of e-bikes.* We believe it to be sensible to assume factors that have the greatest effect on predicted usage of e-bikes to be similarly important in causing people to start using an e-bike.

2.3 Model Development

2.3.1 Random Forest Regression

To understand the importance of factors in fueling growth in e-bike sales, we decided to use a random forest regression model. This allows us to consider many input variables and use these to predict the number of e-bike users. Through this process, we hope to understand which of the inputs has the greatest importance in forecasting the outcome, and can therefore be considered a significant factor in the growth of e-bike usage.

Random forest regression uses regression trees, which are a variant of their more well known cousins, decision trees. To form each tree, the data-set undergoes binary recursive partitioning:

1. The tree attempts a number of possible binary splits on the data (splitting into two partitions), each with a different condition.
2. From each possible split, the tree chooses the split that resulted in the smallest sum of deviations from each partition's mean.
3. This process continues recursively, splitting each branch again and again.

Random *forest* regression is a broader strategy which uses many of these possible trees to form a more robust model. The data-set is split into a number of random subsets, and each is used to produce a regression tree. To produce a prediction, the forest evaluates the output of each regression tree and averages the result. This helps minimise the error that might otherwise be introduced by an over-fitting tree.

When producing this forest, we must first partition our data into a training and testing data-set. This ensures that the testing data has not already been seen by the model and therefore can be used as a valid test of the forest's accuracy. We propose a partition of 20% of the data used for testing.

Input Factors	**Description**	**Source**
Gas Price	Higher gas price would encourage more people to travel using any type of bikes - a proportion of which will be e-bikes	Data-set given [1]
Disposable Income	Families and individuals with a larger sum of funds left after essential spending will lead to greater spending of non-essential commodities (like e-bikes)	Data-set given [1]
Environmental Perceptions of US citizens - split between great deal, fair amount, only a little and not at all	Greater concern for the environment will lead to more frequent bike usage - a proportion of which will include e-bikes	Data-set given [1]
Urban Population	A larger urban population leads to a greater number of commuters, of which a certain proportion will use e-bikes	Source [10]
Battery Charging Price	Consumers are more likely to buy e-bikes if they are less expensive to use. We were unable to find figures for the cost of the whole e-bike over time thus we looked at the price to charge it	Source [3]

Table 3: Input factors used in our random forest model

Entering certain numbers for the above input factors will allow the model to estimate a value for the number of e-bike users.

2.3.2 Calculating Relative Importance

To calculate the relative importance of each factor that we included, we decided to use the feature importance values, produced by the forest. If we achieve a high R^2 value, we can be confident that the relative importance of the factors (as determined by the forest) is also accurate.

"Mean decrease impurity" is an index often used for evaluating feature importance. As per [4], is often described as "the total decrease in node impurity...averaged over all trees of the ensemble." This relies on the concept of impurity, which for a regression tree can simply be defined as the variance of the partition (relative to the partition's mean).[14] Other definitions of impurity do exit, however they serve the same purpose of measuring the homogeneity of the data with more diverse partitions receiving a larger value.

$$Impurity = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{4}$$

Where N is defined as the total number of values in the partition, y_i refers to the y'th output value in this partition and μ is the mean of the partition.

Immediately inputting our data into the random forest regression tools provided by the SciKit Learn library[12] allowed us to quickly evaluate the effectiveness of this model. We found the model was extremely volatile, with each different train/test set selection producing vastly different outcomes. Furthermore, our correlation of determination value was rarely high enough to give us sufficient confidence in our model. To address this, we describe two improvements to the model below.

2.3.3 Interpolating Yearly Data

When iterating on our random forest regression model, we found 11 data points was insufficient to produce an acceptable R^2 model accuracy and reasonable outcomes for factor importance. Additionally, a number of factors had missing values, further reducing the data's usability.

To combat this, we performed linear interpolation at smaller steps between the 11 existing (yearly) data points. Missing values at the edge of the yearly data were identified and extrapolated using the linear interpolation model between the preceding two data points. We believe this to be a valid method given our assumption that the data we have accessed to encapsulates all major trends in the data.

2.3.4 Consolidating the environmental data

In order to use the environmental data given in source [1], we first had to combine the four different survey responses for the question "How much do you personally worry about the quality of the environment?" into a final figure. This allowed us to quantitatively compare how much the US population as a whole was concerned about the environment over the course of our time period. In order to do this, we assigned a value to each of the possible responses, with "A great deal" being 1, "A fair amount" being 0.67, "Very little" being 0.33 and "Not at all" being 0. We multiplied the proportion of people who responded with each category by that category's rating, then added these products together to obtain a final figure for each year. The graph of this figure over time is shown in 4, compared with a graph of proportion of people with each response to the survey over time.

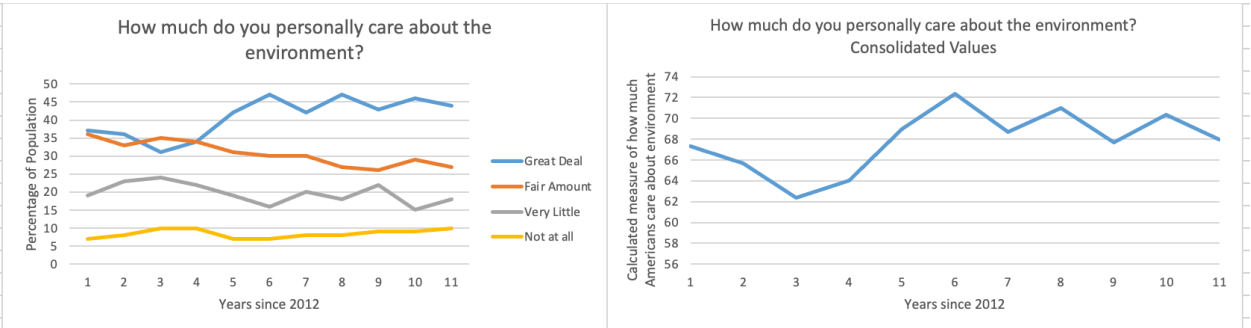


Figure 4: (Left) % of different responses against years since 2012 (Right) our calculated value for average interest in the environment against years since 2012

These graphs appear to follow similar trends, suggesting this is a good method for quantifying the survey responses. We experimented with applying different weightings to the different responses in order to perform a basic sensitivity analysis and found that the graph produced did not change significantly.

2.4 Results

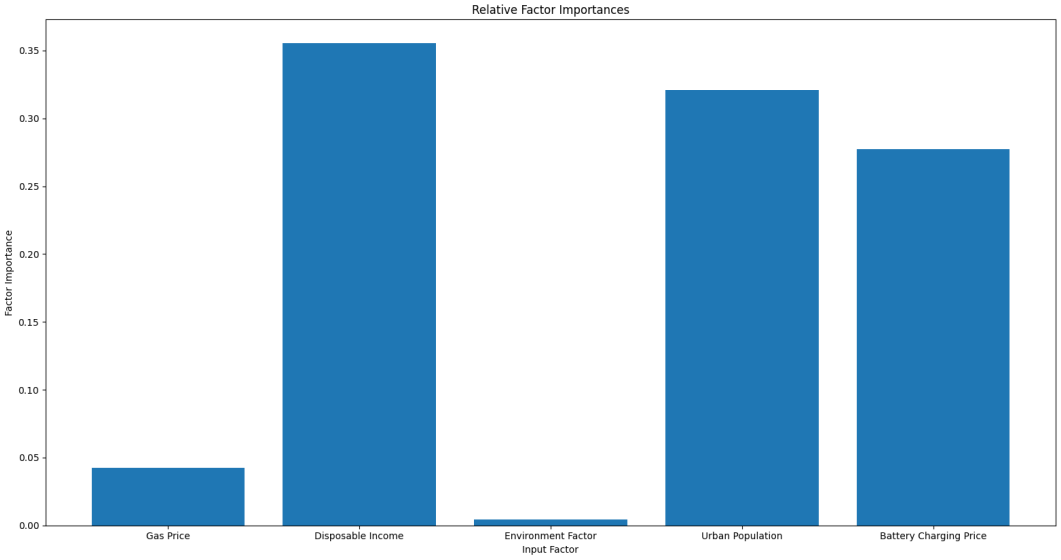


Figure 5: Factors that contributed to the growth of e-bike usage, measured using mean decrease impurity.

Our model considered environmental concern, gas prices, urban population, cost of batteries and personal finances, and was successfully able to determine the relative importance's of the factors it was given, producing a set of reasonable results.

Overall, the economic factors were the most influential on the sales of e-bikes out of the ones we considered, which is plausible, as e-bikes are considered a luxury, rather than a necessity. This is consistent with research rehashing the much greater price of e-bikes relative to normal bikes.[1][5]

The least important factor for predicting the use of e-bikes was environmental viewpoint. This could be justified by the most environmentally aware consumers opting to take even greener modes of transport. It is common knowledge that rechargeable batteries pose a significant risk to the environment,[8] possibly deterring some consumers.

In order of decreasing importance: disposable income, urban population, battery price, gas price, environmental awareness.

2.5 A Review of the Model

Strengths

- As discussed, the output of the model seems entirely reasonable. It is easy to justify the relative importance it predicts.
- Using linear interpolation and by transforming the environmental viewpoint data, we were able to achieve a high R^2 value (> 0.9999). Whilst this high correlation of determination would be expected with the linear extrapolation process we used, it is a good indication that the different input factors are being used in the correct way by the random forest.
- Random forest modelling is robust against any values that do not contribute much change to the number of e-bike users.
- Regression trees tend to over-fit the data. Random forest overcomes this by generating many of these trees with randomly selected data, thereby improving the accuracy.

Weaknesses

- We were not able to find the change of e-bike cost over time; this may have been a significant factor, especially considering the significance of personal finances. We hope using battery price as a proxy for this value managed to encapsulate the trend.

- In the future, we would like to consider a broader range of factors. It remains to be seen how this model would scale if the number of parameters was increased drastically.

3 Part III: Off the Chain

3.1 Restatement of the Problem

In this problem, we were tasked with finding the impacts on carbon emissions and traffic congestion as more people choose to travel by e-bikes instead of using gas powered vehicles.

3.2 Assumptions

1. *Once someone has purchased an e-bike, they will use this bike for every commute.* We have also modelled that everyone who buys an e-bike is commuting, since this is a big motivation for buying e-bikes and in our first model, everyone buying an e-bike is between 15 and 64, so is of commuting age.
2. *The people who buy e-bikes are proportionally distributed between the different commuter groups, excluding cyclists.* We assumed that people driving to work were equally likely to buy e-bikes as those taking the train, bus or other modes of public transport. It is difficult to find data for how the mode of transport you already use affects your likelihood of buying an e-bike and it is likely that this the impact of this factor would be small so we ignored it for this model.
3. *Bikes and e-bikes do not affect road congestion.* Bikes can avoid most traffic and modelling the congestion caused by bikes was too complicated to include in the given time frame.
4. *Congestion is measured as the total time a population spends commuting.* From our model, the most intuitive way of seeing the congestion, is by taking the mean travel time, and then multiplying by the total population. By including the whole population it then allows us to look at carbon emissions as a whole more easily.
5. *The difference in congestion only significantly effects commutes in urban areas, and all e-bikes are used to commute in urban areas.* From our second model, we can see that urban population has a very significant effect on the increase in e-bikes, and therefore assume that the vast majority of them are used in urban areas. This also allows us to use our first model to predict the increase in bike users years in the future, and therefore the change in number of cars on the road.
6. *Carbon emissions are proportional to time spent commuting.* This allows us to not have take into account differences in emissions of idle and moving vehicles, which we could not find suitable data for.
7. *All commuters begin their commute between 07:15 and 8:45 and start times within this window are normally distributed.* This allows us to set more discrete start times for the commutes, and model the congestion before the work day more accurately.
8. *Commute times, independent of congestion, are normally distributed.* To generate random numbers for commute times that our model starts with, we need to normally distribute the data that we found for commute times. This allows us to get a larger, still accurate, range of values for the commute time. We included 'independent of congestion' in the assumption, as in our model we introduce a threshold through-put that the roads can handle, which then changes the time that commutes spend in the system.
9. *The data we used for commute time [9] does not account for congestion that may occur during the journey.* In order for us to accurately generate results for the change in commute time due to the increase in e-bikes, it is necessary for us to know the congestion, and by assuming there is none in the data, it allows us to implement it into our model.
10. *Average carbon emissions is same for all cars*, and will stay constant for the following 10 years.

3.3 Variables

Parameter	Description	Value
T_0	Initial commute time	T_0 N(21.71,13.33^2)
Ω	Maximum road throughput	Varies in model
P_0	Population of commuters in our model	1,000 commuters
C_r	Number of cars removed after 10 years	6,242,000 cars
CO_2	Amount of CO_2 emitted per minute by car	30.24gmin^{-1} [7]
P_1	Proportion of people who drive alone to work	67.8 [1]
P_2	Proportion of people who drive in a carpool of 2 to work	5.9[1]
P_3	Proportion of people who drive in a carpool of 4 to work	1.2[1]
P_4	Proportion of people who drive in a carpool of 4 to work	0.8[1]

Table 4: A summary of variables for our models. Values will be subsequently explained.

3.4 Model Development

The distribution of commute times was assumed to be normally distributed. We used the data from source [9] which gave us the proportion of the population within different brackets of commute times. We wanted to ignore those with no commute, which made up 26% of people from this data set. In order to find the proportion of people with different commute times, we divided the proportions from each remaining bracket of commute time by 0.74. This gave us the proportion of people for each class of commute length shown in table 6.

Time taken (minutes)	Adjusted proportion of commuters (%)
<15	32
15-29	40
30-59	22
60-120	5
>120	1

Table 5: Proportion of commuters who fall into each category of commute length to the nearest minute

As the final two class widths represented very small percentages of the population, they were skewing results for mean and standard deviation, and thus we decided to ignore them. The third equation's upper limit contained up to 95% of the data, again skewing the data, thus was ignored.

$$15.5 = -0.4677\sigma + \mu$$
$$29.5 = 0.5828\sigma + \mu$$

Computing the simultaneous equations gives $\mu = 21.73$ mins and $\sigma = 13.33$ mins.

To model the journey times of commuters, we generated a population of 1000 car commuters. Using a normal distribution, we spread commuters over a 90 minute period over which they start their commute. Using the above normal distribution with the calculated parameters, we also gave each commuter a base commute time. It was necessary to correct both these random variables to be within the acceptable bounds of the considered 90 minute period.

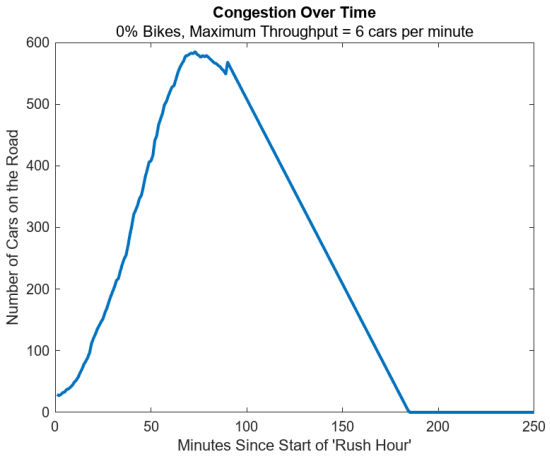
As a simplifying assumption, road networks (for our purposes) have a maximum hourly throughput. We define ω to be the maximum throughput of cars per minute. Whilst it is likely impossible to calculate this for a given real-world read network, we will consider varying values of ω to understand how bike usage affects congestion on a range of possible road networks.

In code, we generate each commuter agent and add them to the road when their commute starts. When the length of their commute comes to pass, the commuter is then removed from the road. However, as a model for congestion across the network, commuters may only be removed at the maximum rate of ω. This may result in some commuters remaining on the road for longer, hence increasing their commute length.

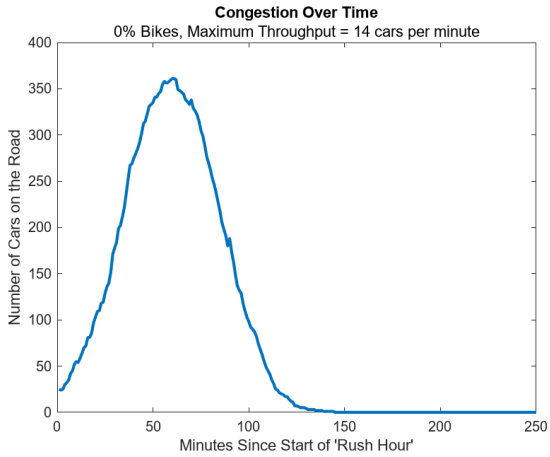
Now, we compare this situation with the one where a number of people who where using cars are now using bikes. We calculated this using our model from part 1. This predicted that the total number of e-bike users in 10 years time would be 11908314.61. Since our model for congestion already includes e-bike users in 2022, we subtract this initial value of 3159000 from total users in 10 years time to give 11908314.61 − 3159000 8749314.607. We then used this and the figures given in [1] to calculate how many cars these new e-bike users would remove from the road. We did this by performing the following calculation:
% of car decrease = number of new e bikes $* \frac{P_1 + 0.5 * P_2 + 0.33 * P_3 + 0.25 * P_4}{100}$

We conclude that 4.95% of the population will own an e-bike in 10 years time. Considering the optimal scenario where all of these e-bike owners no-longer drive to work, we model the car commuter population to have decreased by 4.95%. Comparing these models provides a maximum reduction in commute length by e-bikes in 10 years time.

3.5 Results

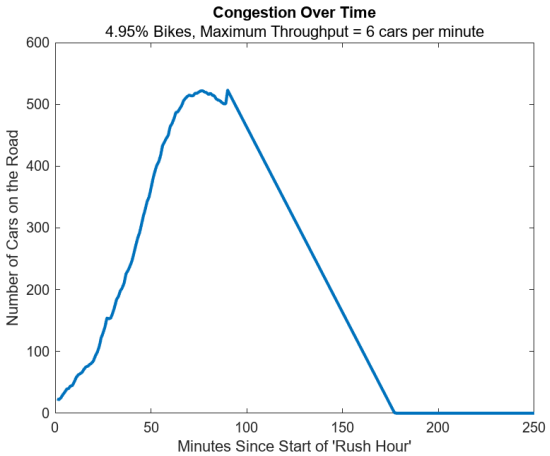


$\Omega = 6$

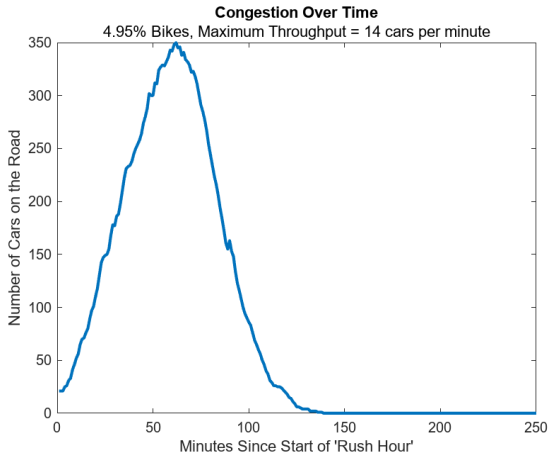


$\Omega = 14$

Figure 6: Graphs of congestion with number of cars on the road in 2022



$\Omega = 6$



$\Omega = 14$

Figure 7: Graphs of congestion with number of cars on the road in 2032

Where the decrease in congestion appears linear, the maximum throughput is being completely used at each time step. We might consider the road network to be at peak capacity.

Plotting the mean congestion time against the max throughput of the road system yields the following plot:

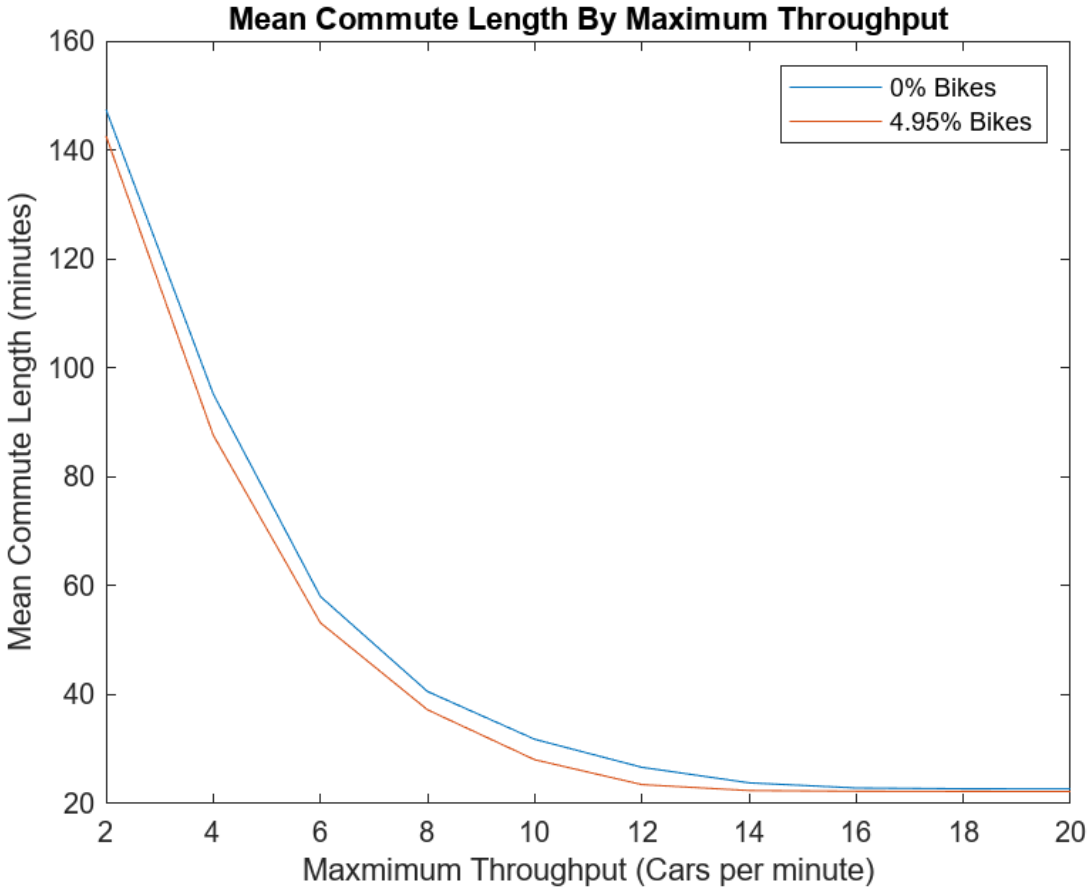


Figure 8: Plot of the mean commute length for a commuter population of 0% bike users and 4.95% bike users.

We summarise the percentage changes in mean journey times below:

Max Throughput (Cars per minute)	Mean Commute Length (minutes) (0% Bikes)	Mean Commute Length (minutes) (4.95% Bikes	Percentage Decrease (%)
6	58.02	53.19	8.33
10	31.79	28.04	11.80
14	23.79	22.37	5.96

Table 6: Proportion of commuters who fall into each category of commute length to the nearest minute

Using the data for $\omega = 6$ and the average car emitting $30.24gmin^{-}1$, we model reductions in CO_2 to be up to $146.06g$ per car commuter per rush hour due to reduced traffic. Car commuters swapping to e-bikes may reduce emissions by up to $657.12g$ per person per rush hour.

3.6 A Review of the Model

Strengths

- The outputs of the model seem reasonable given the percentage change in cars.

Weaknesses

- The distribution of commute times was difficult to model using the data we found because it only included large intervals of time. We were unable to find data for commute times that was not already contained into classes, and thus had to approximate the mean and standard deviation by standardising. The data given did however result in 5 equations, that could be paired into 10 simultaneous equations;solving each gave very obscure data for the mean and standard deviation - for example, with standard deviation larger than mean. Thus, we only used the first two equations to find μ and σ, resulting in an inaccurate representation of the distribution of commute times. An improvement would be obtaining actual commute time data for US working population.
- The data we used for commute time without congestion already factored in did in fact have congestion already factored in. This is because we could not find data for commute times without congestion but to improve the model we might model this by finding the average commute distance and average speed without traffic to give average commute time without traffic.
- To improve the results, we would like to run the simulation many more times and average of the results. This form of Monte Carlo simulation would provide more consistent results. However, considering the large population size, the results are still very similar between runs at the moment.

4 Conclusion

4.1 Further Studies

For our model for part I, we suggest considering the change in population. Introducing a state in which people who are "not susceptible" to buying an e-bike would also be a valid improvement. As with all our models, additional data and time would allow us to better fit existing trends and fine tune the output. For example, how should the extremely large growth in e-bike sales of late (fuelled by an external factor, the pandemic) be considered by the model.

In part II, we would like to have been able to consider a greater range of factors and see how the model performs when the number of factors considered is scaled. One factor we would like to consider is seasonal temperature and people's attitudes to cycling as the weather changes.

For part III, the model would benefit from additional Monte Carlo simulations. Furthermore, the model of traffic congestion could be improved by modelling roads as a graph or otherwise.

4.2 Summary

With our first model, we provided an estimate for the number of sales of e-bikes in the years 2025 and 2028 (1.33 million and 1.76 million respectively) after looking at data for the proportion of people with no bikes, standard bikes and electric bikes. We then also calculated the change of the number of people in each state, using differential equations from an adaptation of an SIR model. With our second model, we used the random forest regression algorithm to find which variables had greatest effect on the number of e-bikes sold. Disposable income, urban population and battery prices were found as some of the most important factors with an R^2 value of >0.9999. For the third model, we simulated multiple runs of traffic congestion over time with 6 cars per minute and 14 cars per minute. We ran traffic with 0% bikes and 4.95% bikes, and found e-bikes are beneficial for the environment and for congestion times.

5 References

[1] MathWorks Math Modeling Challenge 2023. *Ride Like the Wind*. URL: https://m3challenge.siam.org/node/596.

[2] Adrienne Bernhard. "The great bicycle boom of 2020". In: *BBC* (). URL: https://www.bbc.com/future/bespoke/made-on-earth/the-great-bicycle-boom-of-2020.html.

[3] BloombergNEF. *Price to charge lithium ion cells in USD per kWh*. URL: https://www.sustainable-bus.com/news/battery-packs-prices-2022-bloomberg-nef/.

[4] Friedman Breiman. *Classification and regression trees*. 1984.

[5] BSXInsight. *How Much Does a Bike Cost? Top Full Guide 2023*. Accessed on 04/03/2023. URL: https://www.bsxinsight.com/how-much-does-a-bike-cost/.

[6] United States Census Bureau. *U.S. and World Population Clock*. Accessed on 04/03/2023. URL: https://www.census.gov/popclock/.

[7] Ravalli County. *How much CO_2 is produced by a car per hour*. Accessed on 04/03/2023. URL: https://ravalli.us/DocumentCenter/View/229/Vehicle-Idling#:~:text=An%5C%20hour%5C%20of%5C%20automobile%5C%20idling,can%5C%20contribute%5C%20to%5C%20global%5C%20warming..

[8] Charlotte Duck. "Batteries are charging our planet, but what's the cost?" In: *National Geographic* (). URL: https://www.nationalgeographic.com/science/article/partner-content-audi-nl-the-cost-of-batteries.

[9] Jack Flynn. *15+ AVERAGE COMMUTE TIME STATISTICS [2023]: HOW LONG IS THE AVERAGE AMERICAN COMMUTE?* URL: https://www.zippia.com/advice/average-commute-time-statistics/#:~:text=As%5C%20of%5C%202021%5C%2C%5C%2026%5C%25%5C%20of,commute%5C%20both%5C%20ways%5C%20each%5C%20day..

[10] macrotrends. *U.S. Urban Population 1960-2023*. Accessed on 04/03/2023. URL: https://www.macrotrends.net/countries/USA/united-states/urban-population.

[11] Aled Nelmes. *How Long Will an Aluminium Bike Frame Last?* Accessed on 04/03/2023. URL: https://southerndistributors.co.uk/2021/07/22/how-long-will-an-aluminium-bike-frame-last/.

[12] Open source. *Sci-kit Learn*. URL: https://scikit-learn.org/.

[13] Statista. *Age distribution in the United States from 2011 to 2021*. Accessed on 04/03/2023. URL: https://www.statista.com/statistics/270000/age-distribution-in-the-united-states/.

[14] MLlib - Decision Trees. *Apache Spark*. URL: https://spark.apache.org/docs/1.3.0/mllib-decision-tree.html.

[15] Zippa. *US Bicycle Industry Statistics By Consumers*. Accessed on 04/03/2023. URL: https://www.zippia.com/advice/bicycle-industry-statistics/.

6 Appendix

6.1 Code for Part I: The Road Ahead

6.1.1 Exponential Regression of Bike Sales

```
% Reading data table
dataDir = 'C:\MATLABdata\';
dataFile = 'M3Downloads.csv';
Ebikes = readtable([dataDir, dataFile]);

% Defining x and y (years after 2012, and number of E-bikes sold)
x = Ebikes.Year - 2012;
y = Ebikes.SoldBikes1000s;
% Defining backwards extrapolation, to go to 2012
extrapolation_x = 0:0.1:10;

% Fitting exponential curve to dataset
[mdl, gof] = fit(x, y, 'exp1')

% Plotting the output curve + backwards extrapolation
figure;
hold on
plot(extrapolation_x, mdl(extrapolation_x));
plot(mdl, x, y);
legend(["Backwards Extrapolation (BEX)", "Data", "Fit"], "Location","northwest", "FontSize",5);
xlim([0, max(extrapolation_x)]);
xlabel("Years Since 2012");
ylabel("Number of E-bikes sold/1000s");
title("Exponential regression of E-bike sales");
grid on
hold off
```

6.1.2 Evaluating Differential Equations Numerically

```
% Define our variables
N = 217000000;
alpha = 0.000000002580;
gamma = 0.1625;
w = 1 / 10;
z = 1 / 10;
pars = [alpha gamma w z];

% Define our initial state
y0 = [113841000; 3159000; 100000000];

% Numerically solve for the next 10 years
tSpan = [0 10];
[t, y] = ode45(@diffEq, tSpan, y0, [], pars);

% Plot the output
hold on
plot(t, y(:, 1), "LineWidth", 2);
plot(t, y(:, 2), "LineWidth", 2);
plot(t, y(:, 3), "LineWidth", 2);
legend("No Bike", "E-Bike", "Bike", "Location", "southeastoutside");
title("Prediction of Bike Owner Populations");
xlabel("Years Since 2022");
ylabel("Number of People");
```

```

% Compute sales of e-bikes each year
% sales = alpha * M * E
sales = alpha .* y(:, 1) .* y(:, 2);
writematrix([t y sales], "bike-predictions.csv");


function f = diffEq(~, y, pars)
    f = zeros(3, 1);

    % dM/dt = wE + zB - alpha*ME - gamma*M
    f(1) = pars(3) * y(2) + pars(4) * y(3) - pars(1) * y(1) * y(2) - pars(2) * y(1);

    % dE/dt = alpha*ME - wE
    f(2) = pars(1) * y(1) * y(2) - pars(3) * y(2);

    % dB/dt = gamma*M - zB
    f(3) = pars(2) * y(1) - pars(4) * y(3);
end
```

6.2 Code for Part II: Shifting Gears

6.2.1 Piecewise Linear Interpolation

```

```

6.2.2 Random Forest Regression (Python)

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
from matplotlib import pyplot

# Perform random forest regression and return the r^2 and feature importances
def fitModel(features, values):
    # Split the data
    featuresTrain, featuresTest, valuesTrain, valuesTest = train_test_split(features, values,
    test_size=0.2)

    # Fit the model
    model = RandomForestRegressor(n_estimators=150)
    model.fit(featuresTrain,valuesTrain)

    # Evaluate the model
    prediction = model.predict(featuresTest)
    r2 = metrics.r2_score(valuesTest, prediction)

    # Return model statistics
    return (r2, model.feature_importances_)

# Read in the data
data = pd.read_csv("C:/Users/rawat/OneDrive - Eltham College/Desktop/interpolated_factors.csv")

# Separate the features and outputs
readableFeatureCols = ["Gas Price", "Disposable Income", "Environment Factor", "Urban Population",
"Battery Price"]
featureCols = ["GasPrice", "DI", "Environment", "UrbanPopulation", "Battery"]
```

```
features = data[featureCols]
values = data["EBikeUsers"]

# Remove feature names from the csv
features = features.drop(0, axis=0)
values = values.drop(0, axis=0)

# Record statistics across all runs
numRuns = 100
sumR2 = 0
sumImportances = np.zeros(len(featureCols))

# Run the model repeatedly
for i in range(numRuns):
    modelStats = fitModel(features, values)
    sumR2 += modelStats[0]
    sumImportances = np.add(modelStats[1], sumImportances)
    if modelStats[0] < 0:
        print("Unexpected negative R^2 value")

# Calculate averages
avgR2 = sumR2 / numRuns
avgImportances = np.divide(sumImportances, numRuns)

# Plot the data in a bar chart
pyplot.bar(readableFeatureCols, avgImportances)
pyplot.ylabel("Factor Importance")
pyplot.xlabel("Input Factor")
pyplot.title("Relative Factor Importances")
pyplot.show()

# Print the data for inspection
print(avgR2)
print(avgImportances)
```

6.3 Code for Part III: Off the Chain

```
% Constants...
bikeProportions = [0 0.0495];
mean = 21.73;
sd = 13.33;
totalPopulation = 1000;
avgLengths = zeros(10, 2);

% For all modelled proportions with bikes
for i = 1:length(bikeProportions)

    % Adjust the population who no longer drive and
    % hence no longer commute to congestion
    population = round(totalPopulation * (1 - bikeProportions(i)));

    % Sensibly distribute start times
    startRange = 90;
    startSd = startRange / 4;
    startMean = startRange / 2;

    % Normally distribute (and correct) start times.
    startTimes = round(commutestartrange / 2 +
```

```
        commutestartsd * randn(population, 1));
    startTimes = max(startTimes, 1);

    % Normally distribute (and correct) predicted lengths
    predLengths = round(mean + sd * randn(population, 1));
    predLengths = max(predLengths, 1);

    % Model this situation for all possible throughputs
    for maxThroughput = 2:2:20
        l = modelCongestion(population, startRange, startTimes, predLengths, maxThroughput, bikeProp
        avgLengths(maxThroughput / 2, i) = l;
    end
end

% Plot our final graph of mean commute length decrease with bikes
figure;
plot(2:2:20, avgLengths);
legend("0% Bikes", "4.95% Bikes");
title("Mean Commute Length By Maximum Throughput");
xlabel("Maxmimum Throughput (Cars per minute)");
ylabel("Mean Commute Length (minutes)");

function avgLen = modelCongestion(population,
    startRange, startTimes, predLengths, maxThroughput, bikeProportion)

    extrapolationTime = 160;
    endTimes = zeros(population, 1);
    congestion = zeros(startRange + extrapolationTime, 1);

    % For every time step, model change in congestion
    for t = 1:(startRange + extrapolationTime)
        entered = 0;
        left = 0;

        % Update the state of the population
        for k = 1:population
            if startTimes(k) == t
                entered = entered + 1;
            end

            % For people who haven't finished their commute and
            % who can be removed from the road
            if (endTimes(k) == 0 &&
                (startTimes(k) + predLengths(k)) <= t && left < maxThroughput)

                left = left + 1;
                endTimes(k) = t;
            end
        end

        % Update the new congestion value
        if t > 1
            congestion(t) = congestion(t - 1) + entered - left;
        else
            congestion(t) = entered - left;
        end
    end

```

```
80      % Plot this congestion-time graph
81      figure;
82      plot(1:(startRange + extrapolationTime), congestion, "LineWidth", 2);
83      title("Congestion Over Time");
84      subtitle((bikeProportion * 100) + "% Bikes, Maximum Throughput = " + maxThroughput + " cars per
85      xlabel("Minutes Since Start of 'Rush Hour'");
86      ylabel("Number of Cars on the Road");
87
88      % Return the mean commute length
89      endTimes(endTimes == 0) = startRange + extrapolationTime;
90      avgLen = sum(endTimes - startTimes) / population;
91  end
```