USING COMPUTER VISION TECHNIQUES TO BUILD A PREDICTIVE MODEL OF FRUIT SHELF-LIFE

A Thesis
presented to
the Faculty of California Polytechnic State University,
San Luis Obispo

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Industrial Engineering

by
Nandan G. Thor
June 2017

COMMITTEE MEMBERSHIP

TITLE: Using Computer Vision Techniques to Build a Predictive Model of Fruit Shelf-Life

AUTHOR: Nandan G. Thor

DATE SUBMITTED: June, 2017

COMMITTEE CHAIR: Jose Macedo, Ph.D.
Professor and Department Chair of Industrial and Manufacturing Engineering

COMMITTEE MEMBER: Reza Pouraghabagher, Ph.D.
Professor of Industrial and Manufacturing Engineering

COMMITTEE MEMBER: Daniel Waldorf, Ph.D.
Professor of Industrial and Manufacturing Engineering

ABSTRACT

Using Computer Vision Techniques to Build a Predictive Model of Fruit Shelf-Life

Nandan G. Thor

Computer vision is becoming a ubiquitous technology in many industries on account of its speed, accuracy, and long-term cost efficacy. The ability of a computer vision system to quickly and efficiently make quality decisions has made computer vision a popular technology on inspection lines. However, few companies in the agriculture industry use computer vision because of the non-uniformity of sellable produce. The small number of agriculture companies that do utilize computer vision use it to extract features for size sorting or for a binary grading system: if the piece of fruit has a certain color, certain shape, and certain size, then it passes and is sold. If any of the above criteria are not met, then the fruit is discarded. This is a highly wasteful and relatively subjective process.

This thesis proposes a process to undergo to use computer vision techniques to extract features of fruit and build a model to predict shelf-life based on the extracted features. Fundamentally, the existing agricultural processes that do use computer vision base their distribution decisions on current produce characteristics. The process proposed in this thesis uses current characteristics to predict future characteristics, which leads to more informed distribution decisions. By modeling future characteristics, the process proposed will allow fruit characterized as "unfit to sell" by existing standards to still be utilized (i.e. if the fruit is too ripe to ship across the country, it can still be sold locally) which decreases food waste and increases profit. The process described also removes the subjectivity present in current fruit grading systems. Further, better informed distribution decisions will save money in storage costs and excess inventory.

The proposed process consists of discrete steps to follow. The first step is to choose a fruit of interest to model. Then, the first of two experiments is performed. Sugar content of a large sample of fruit are destructively measured (using a refractometer) to correlate sugar content to a color range. This step is necessary to determine the end-point of data collection because stages of ripeness are fundamentally subjective. The literature is consulted to determine "ripe" sugar content of the fruit and the first experiment is undertaken to correlate a color range that corresponds to the "ripe" sugar content. This feature range serves as the end-point of the second experiment. The second experiment is large-scale data collection of the fruit of interest, with features being recorded every day, until the fruit reaches end-of-life as determined by the first experiment. Then, computer vision is used to perform feature extraction and features are recorded over each sample fruit's lifetime. The recorded data is then analyzed with regression and other techniques to build a model of the fruit's shelf-life. The model is finally validated. This thesis uses bananas as a proof of concept of the proposed process.

Keywords: agriculture, bananas, computer vision, feature extraction, food waste, LabVIEW, multiple regression, predictive modelling, produce grading

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

1. INTRODUCTION

Computer vision is one of the fastest growing and one of the most researched disciplines in this day and age. Computer vision is the hardware and software involved in capturing and analyzing an image with a computer. It has found a home on inspection lines because of its speed, accuracy, reliability, and objectivity. Computer vision excels on systems such as inspection lines at finding defects in well-defined objects. When a computer vision system has an ideal representation of what a, say, electronic component looks like, it is very easy to detect deviations from the norm. However, a field that is still in its infancy is applying computer vision to non-uniform objects. The purpose of this thesis is to present a process that applies existing computer vision techniques to build a predictive model of the shelf-life of fruit.

Increasingly, computer vision is being used in the agriculture industry. Agriculture is a good fit for computer vision because of the sheer volume of fruit that most factories process every day and the time, effort, and money that goes into quality control and distribution decisions. Today, some factories use computer vision to detect for bruises and other defects that would detract customers. This type of inspection falls under the broader category of fruit quality grading. In most agricultural businesses, this grading is done by humans. Essentially, the grading is a rating on a scale from one to five of the external quality of the fruit. Fruits with ratings of three and above are sold while those with ratings of two and below are discarded. This type of grading is a binary process that is beginning to be (somewhat reliably) performed by computers. Another, much more advanced, application is to use computer vision to classify distinct categories

of fruit with similar characteristics and place all of one group of fruit into one room, dose the entire room with ethylene, and then sell the entirety of that group when it is ripe. However, this takes up time, space, and money to store and wait for the fruit. It is also frequently wasteful and inaccurate.

The purpose of this thesis is to present a process that uses computer vision to extract color features of a large group of the fruit and then builds a predictive model of the shelf-life for the fruit (specifically bananas for the purposes of this thesis). While the thesis demonstrates the process on bananas, the process can theoretically be repeated for any climacteric fruit that changes colors distinctly as it ripens. Essentially, a camera would scan each piece of fruit, extract features, and then plug the values into the created model to get an accurate prediction of the shelf-life of that particular piece of fruit. This would allow fruit with a shorter shelf-life to be shipped across the state whereas fruit with a longer shelf-life would be shipped across the country. This would save money and space. It would also allow more fruit to be sold because fruit that may have been given an unacceptable rating would still be able to be sold (if the fruit is deemed to ripe to be shipped, it can still be sold locally). Fundamentally, instead of making distribution decisions based on current features (as is done in the industry today), this thesis proposes a model which allows current features to be used to predict future features, which are used to inform distribution decisions. Finally, the proposed technique would be advantageous because it allows for an objective, accurate, and concrete shelf-life prediction system as opposed to the current fruit grading procedure. This thesis develops an experimental process to undergo and the subsequent statistical analysis to build a predictive model of color features against shelf-life. The thesis uses bananas as a proof

of concept of the process. There are two experiments: the first one captures images of a few samples of fruit and then destructively measures sugar content. The purpose of the second experiment is to correlate color representations with sugar content to determine the end point of shelf-life. A sugar content of 23% indicates a ripe banana, which, for the purposes of this thesis will serve as the last day of data collection. The color of many samples of bananas on the first day that their sugar content reaches or exceeds 23% is recorded and the color range is then set as the end color of the second experiment. The second experiment involves capturing images of many samples of the fruit each day until it becomes spoiled. The color features are extracted every day using computer vision. For statistical analysis, each fruit is plotted to measure color as a function of time. Post-processing then occurs with different representations of color being calculated (using existing computer vision algorithms). Then, for each fruit, shelf-life is calculated at a given color value. This is repeated for each selected color and each fruit. Finally, shelf-life as a function of color can be modeled (with the corresponding standard deviation at each color) to ultimately give a model of shelf-life based on color.

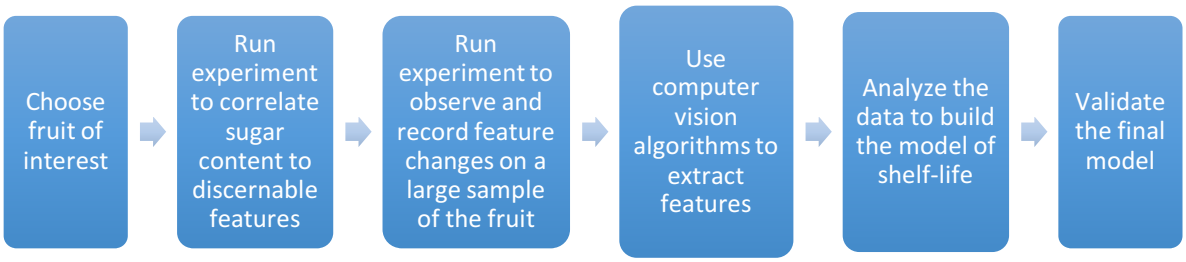


Figure 1: Outline of process.

2. LITERATURE REVIEW

2.1. Introduction:

The purpose of this literature review is to provide the readers with an understanding of how computer vision is applied in the agriculture industry today. This will identify current industry standards as well as active areas of research. Additionally, this literature review will serve to identify gaps in the current research.

Computer vision is the hardware and software involved in image processing of an object by a computer. Essentially, an object is analyzed by capturing an image of it and extracting features and characteristics for downstream processing (characterization, discrimination, model building, etc.). Over the past thirty years, computer vision has increasingly found a home in the agriculture industry. Common applications of computer vision systems in the agricultural industry involve automatic sorting, grading, and estimation of quality of fruits, vegetables, and meat. Traditionally, sorting and grading of fruits has been done by a panel of experts. However, in many ways, computer vision systems are better because they can be faster and more reliable than human graders. One of the advantages of computer vision systems are that computers can detect light in the ultraviolet and infrared spectrum, which humans cannot do (Cubero et al., 2011). Also, computers are generally much faster and more accurate (consistent) than humans. There is no bias in computer vision systems, unlike humans. That is, people are subjective and when they are only able to give a numerical value as a rating, there is not a detailed analysis of quality (Zhou et al., 2004). Finally, while there is a large up-front investment

in computer vision systems, in the long-run computer vision in the agriculture industry becomes much cheaper and more efficient than using human graders.

2.2. Current Issues:

The main challenge of using computer vision in the agriculture industry is that each piece of produce is unique. An apple from one orchard on one specific tree on one specific branch may look, feel, and taste completely different than a neighboring apple. Also, the produce naturally changes – it oxidizes, changes color, changes texture, changes quality etc. as the produce ripens. There is also the added complexity of the surface geometry and texture which misleads computer vision with shadows and raised surfaces. Computer vision is best when it has been trained on a set of images and has to categorize the exact same object. However, produce is imperfect and it is contentious as to what exactly the "perfect" fruit characteristics are. There is also a quantity versus quality argument: is it more important to get fruits analyzed correctly or quickly? Ideally both, but the answer will vary from industry to industry.

The next serious challenge has to do with controlling and replicating lighting effects. The intensity of light varies on different parts of produce, since fruit is not a uniform shape and since the surface of the fruit is often curved. So, using a computer vision system for grading produce must have consistent lighting that recognizes and addresses the problem of even lighting of uneven surfaces. Another hardware challenge is how to set up and integrate information from many cameras to obtain a complete representation of the fruit, while minimizing cost. This area involves systems integration:

how to get every component (camera, object, image, computer, encoder, etc.) to communicate with the other components to get a comprehensive computer vision result.

The majority of recent computer vision research involves creating faster, better algorithms in color analysis and shape analysis of images. There have been applications of genetic algorithms, neural networks, and other algorithms to computer vision use in the agriculture industry. These have come with various results and reproducibility. Importantly, the economic factors must be considered as well. One of the more pressing current issues in computer vision in the agriculture industry is trying to establish the best algorithms for computer vision of produce. Ideally, an algorithm that is fast, cheap, and accurate is desired. However, logistically, there usually has to be a trade-off of one of the aforementioned factors. Often, the best algorithms will vary from industry to industry and even between different pieces of produce.

2.3. Image Analysis:

2.3.1. Introduction:

"Appearance is a very important sensory quality attribute of fruits and vegetables, which can influence not only their market value, consumer's preferences and choice but also their internal quality to some extent. External quality of fruits and vegetables is generally evaluated by considering their color, texture, size, shape, as well as the visual defects. External quality inspection of fruits and vegetables manually is a time-consuming and labor intensive work" (Zhang et al., 2014). Indeed, when selecting

produce in a supermarket, appearance (color, shape, size) is the most important discriminating factor. It has been proven that the color of fruit can end up affecting the perception of the taste of the fruit (Abdullah et al., 2001). It is, then, very valuable to have a computer vision system that can analyze produce and make distribution decisions about the produce that will be appealing to customers. Further, computer vision systems that can grade fruit accurately will be able to predict which produce will sell in certain supermarkets (based on the features extracted using computer vision).

The motivating work that demonstrated that computer vision is an acceptable substitute for human grading comes from Nunes (2015). "Overall, there was a significant correlation between most of the [fresh fruits and vegetables] FFVs subjective quality attributes evaluated and the physicochemical analysis performed. Results from this study showed that subjective quality evaluations using rating scales can be a reliable and simple method to estimate changes in color, texture, water content, and ultimately changes in specific chemical components when FFVs are exposed to different environmental conditions. In the absence of a formal trained sensory panel this method can be easily used in research of industry settings (e.g., quality control at receiving)" (Nunes, 2015). Essentially, the work showed that computer vision extracted features do indeed mirror the observable physiological changes in fruit. Put another way, computer vision is an unbiased and accurate indicator of the physiological changes of ripening fruit. However, this is still an active and important area of research which is needed to validate the results from computer vision of produce.

2.3.2. Equipment:

Computer vision needs three things: a source of light, an object to be analyzed, and a receptor to detect the light (Vidal et al., 2013). Lighting is one of the most important aspects of computer vision. Illumination must be uniform, constant, and repeatable. "In the external quality assessment using computer vision, a good lighting system should provide uniform radiation throughout the scene, avoiding the presence of glare or shadows, and it must be spectrally uniform and stable over time" (Cubero et al., 2011). According to Zhang (2014), lighting needs to be controlled to reduce downstream filtering, to enhance distinction of the object of interest from the background, and to reduce reflection. There are two different types of lighting: front lighting is used to do analysis on the actual surface of the fruit (e.g. color analysis or defect determination). Back lighting is used to distinguish the shape of the object or the size of the object. Front lighting is for feature detection and back lighting is for general characteristics (Zhang et al., 2014). The use of either front or back lighting depends on the project and often both are used at different parts of the research. One unique way of controlling for uniform and replicable lighting conditions was done by Leemans (1998). The idea was to use a cylindrical tube with the inside painted white to reflect light. This allowed for even lighting to fill the entire tube instead of relying on point sources of light. Also, this technique has the advantage of not having a shadow (Leemans et al., 1998).

Cameras are the image capturing devices for computer vision. The most common cameras are charge-coupled device (CCD) cameras. Often, there are multiple cameras in order to capture different angles of the same object. "The most popular industrial

cameras are based on charge-coupled device (CCD), which consists of an array of sensors (pixels), each of which is a photocell and a capacitor" (Cubero et al., 2011). There, again, has to be a tradeoff between quality and quantity as well as price. There needs to be enough cameras to acquire a representative image of the object of interest. However, having a large number of cameras will be expensive and it will be very computationally taxing to coalesce many different images into one comprehensive image to analyze. Image acquisition is not limited to only cameras, there are also scanners, X-Rays, MRI, etc. (Cubero et al., 2011).

All of the equipment blends into real time vision systems. Generally, produce travels at a very high speed on inspection lines. In order to get clear images, a strobe light is generally used to capture one frame at a time. There should, also, be an encoder to adjust the speed of the conveyor belt. Indeed, there must also be a computer to analyze the images. Finally, there should be a primary and secondary output chamber to separate acceptable produce from unacceptable produce. (Cubero et al., 2011). An example system is shown in Figure 2.

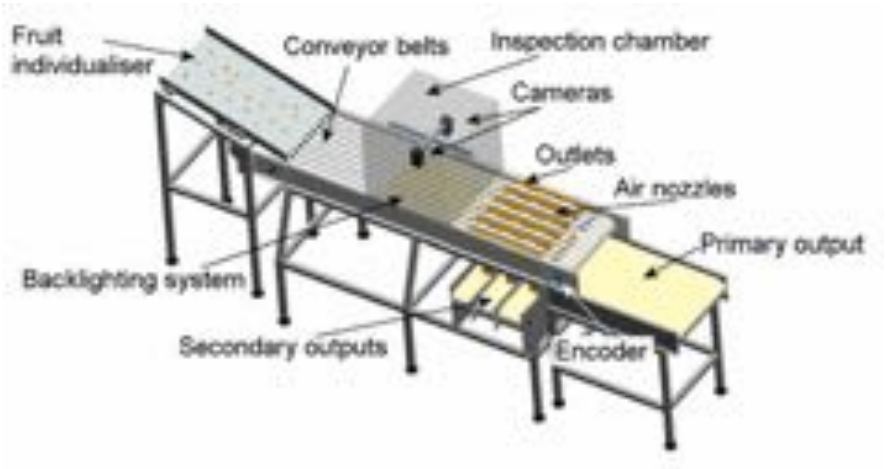


Figure 2: Computer vision system example. (Cubero, et al., 2011).

2.3.3. Color Analysis:

Color is the most important discriminating factor for consumers (Zhang et al., 2014). Therefore, most research has been devoted to creating computer vision systems that accurately represent colors. "Colour coordinates provided by these [CCD] devices are often referred to as the CIE 1931 colour space, in which they are denoted by X, Y and Z. Colorimeters are limited to the measurement of small regions or in applications where the integration of the colour all over the sample is of interest, which means that they are not well suited to measuring objects with a heterogeneous colour" (Cubero et al., 2011). The two main techniques for color feature extraction are Red, Green, Blue (RGB) (Figure 3) and Hue, Saturation, Intensity (HSI) (Figure 4). RGB gives a simple ratio of red, green, and blue color. RGB is generally used for discrimination, HSI is generally used for description. RGB is often seen as inferior for a few reasons. Firstly, each camera will provide a different RGB value for the same image. This is because the calculated RGB value is device dependent. This discrepancy can be overcome by transforming RGB values to standard RGB (sRGB) values – however this can take computational space and time. The other issue is that RGB color is not intuitive: it is not representative of how the human eye sees the world. For the above reasons, HSI is frequently preferred over RGB. It is important to note that, for the purposes of this thesis, the HSI values that are found throughout the thesis come from National Instruments LabVIEW (which was used for analysis) and its inherent 8-bit encoding of color. LabVIEW's 8-bit encoding produces scaled hue values between 0 and 255. Results may vary based on the software used and the encoding type (16-bit may produce different results than 8-bit). According to Vidal

(2013), RGB can be an issue because it is device dependent. To overcome this, RGB colors get converted to XYZ coordinates in order to standardize the RGB values. The XYZ coordinates are then converted to L*a*b* coordinates which are considered to be the most intuitive and the most accurate representation of reality as the human eye sees it (Figure 5). While this does transform RGB coordinates, the issue with the transformation is that the transformation is very computationally intensive and time-consuming (Vidal et al., 2013 and Pedreschi et al., 2006). Another way of using RGB color analysis was performed by Blasco (2003). Each image was analyzed for color using the RGB technique and then a Bayesian discriminant model was used to perform the RGB color analysis (Blasco et al., 2003). Shearer and Payne (1990) used RGB color analysis as a way to classify the quality of bell peppers. The analysis was done in a very computationally heavy way by classifying each pixel by RGB color analysis and then by further quantifying each pixel into one of eight possible color categories. This analysis was relatively accurate but too slow to be implemented on inspection lines (Shearer and Payne, 1990).

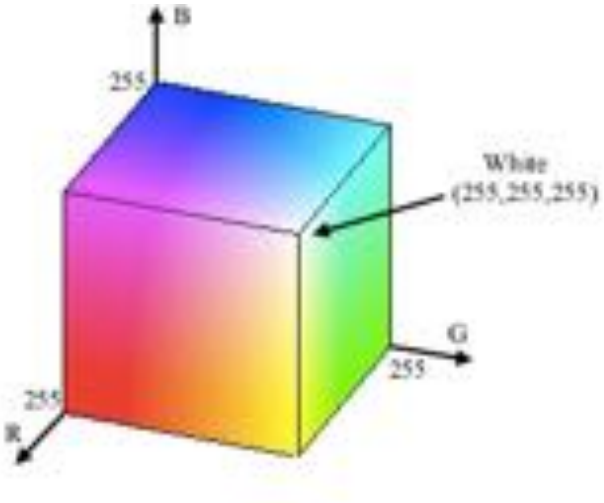


Figure 3: RGB color representation. (Pata, 2016).

Figure 4: HSI color representation. (Jewett, 2013).

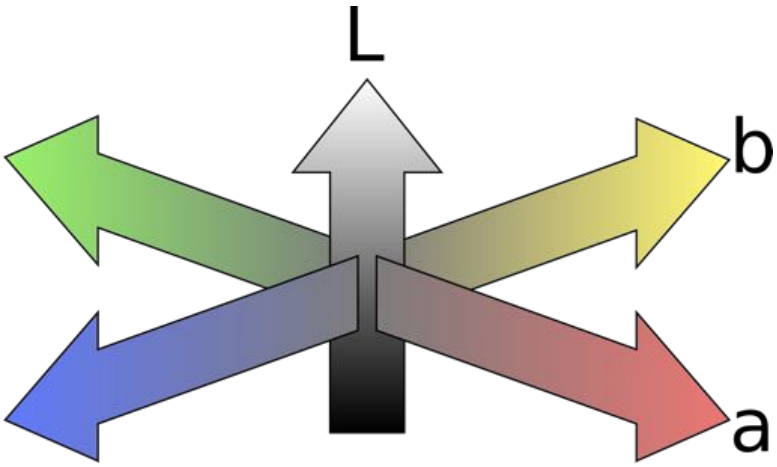


Figure 5: CIE L*a*b* color representation. (De, 2015).

Certain fruit can have different colors. Firstly, there is the main color that covers most of the produce's surface called the primary color. Secondly, there can be other colors called the fruit's secondary colors. A numerical estimate of the primary or secondary colors can be found by averaging the colors on the surface of the fruit (Blasco et al., 2003). If a fruit has a primary color – one color evenly distributed among the surface of the fruit, then averaging the primary color is an accurate way of determining quality. However, if the color of the fruit is non-uniform, the secondary color may be a more accurate indicator of quality (Jha et al., 2010). Clearly, color analysis will vary from between types of fruit.

One of the most important papers in the field of color analysis for computer vision in the agriculture industry came from Pace (2014). The color of a certain region of lettuce was calculated by dividing that region's color by either white or brown. This analysis gave a ratio that allowed for comparison between different parts of the same object. The analysis was done on four separate images, each capturing a certain part of the produce in question (lettuce). The authors ultimately found that color was a significant indicator of quality – both green and brown and the respective ratio of each color (Pace et al., 2014). This paper was important because it showed that color is an accurate indicator of produce's quality and the paper introduced the use of a color ratio for more precise color analysis.

In 2008, Alfatni's experiment determined how ripe oil palm fruit was by calculating color intensity. They used RGB color intensity. The mean color intensity was simply the number of red pixels divided by the total number of pixels. They found that there was a correlation between the mean intensity color and ripeness (Alfatni et al.,

2008). Another, similar, experiment was done by Abdullah on the same oil palm fruit. In each image, RGB color analysis gives three separate values (R, G, and B) for each pixel in the image. A big problem with RGB color analysis is that it does not accurately map colors the way that humans do. That is, it is not an intuitive way of representing color. RGB should be converted to HSI for more understandable results. "The basic idea of machine vision discrimination analysis is to transform the multivariate hue distributions to univariate y such that y's derived from population for Gi for i = 1, 2, 3, . . . , g were separated as much as possible. The results can be interpreted in terms of the sample Mahalanobis' distance" (Abdullah et al., 2001). Mahalnobis' distances are essentially a multi-dimensional measure of standard deviation. It is a measure of how far away a point (P) is from a distribution (D) in all dimensions. In the case of pomegranates, the RGB color system was the most intuitive, with the most accurate indicator being the R/G ratio. There can be misclassification between the red and brown objects (using RGB analysis). The processing speed was very impressive: about 15 milliseconds per fruit. If speed is of high importance, the RGB model may be the correct approach (Blasco et al., 2009).

Computer vision systems can be trained to be even more accurate in differentiating colors than colorimeters, according to Diaz (2000). By being fully automated, computer vision systems can overcome the error that is inherent in human operation of colorimeters. The vision systems are also much more accurate at classifying olive quality states than olive experts because the experts tend to classify olives in the worse category whenever there is doubt (Diaz et al. 2000). In a similar example, when compared to the grading of a human panel, the computer vision approach was more accurate and much faster (Abdullah et al., 2001).

2.3.4. Size/Shape Estimation:

Shape estimation poses a serious challenge for computer vision systems. This is because of the non-uniformity of fruits. Each fruit will have its own unique shape and it is difficult to characterize an "ideal" fruit shape. In order to classify shape, generally, there needs to be a set of categories and criteria to distinguish which category the object falls in. That is, characteristics needs to be measurable (roundness, etc.) (Costa et al., 2011). Produce should fall within a specified shape tolerance otherwise people will not want to buy it. The producer should come up with an ideal shape and then allow for a certain variation tolerance, rejecting any fruit outside of that tolerance (Cubero et al., 2011). An excellent example of this was done by Leemans (1998). In his approach, each pixel of the acquired image was compared to an idealized model image. This was done using Mahalanobis distances. The main problem was that this was computationally intensive but it did give a very accurate model of how different the acquired image was from the idealized image. To speed up the process, the image was then segmented into subdivisions which were compared to the ideal image. This process is very successful at detecting bruises and discolorations (Leemans et al., 1998).

Size estimation is one of the main factors to distinguish commercial categories of produce. For produce producers, size is the most important characteristic (even more so than color). "Size, which is the first parameter identified with quality, has been estimated using machine vision by measuring either area ... perimeter … or diameter …" (Blasco et al., 2003). For size estimation, there are two general techniques. The first technique is pixel oriented: which consists of classifying each pixel as either belonging to the object

(foreground) or the background. The next technique is region oriented which entails taking previous knowledge to perform the segmentation (e.g. color changes, boundaries, changes in texture) (Blasco et al., 2007). This type of analysis was actually implemented earlier by Blasco: for size estimation, the image was treated as a binary product, with the foreground being the fruit and the rest being the background (Blasco et al., 2003). Size was then estimated by the number of pixels in the foreground. A unique segmentation technique was employed in an experiment by Blasco (2009). Instead of trying to segment the image, they used segmentation to try to distinguish the object from the background. This was helped by using a blue background (because they were interested in red and white pomegranate arils) so the foreground (object) and background can be easily distinguished. Size estimation can become more challenging with non-spherical produce (Cubero et al., 2011).

Perhaps the most complete analysis of image processing comes from Zhang (2014). "Image processing and analysis are performed in three levels. The low level processing, which is the basic processing of image, involves image acquisition and image preprocessing; the intermediate level processing, which is the make-or-break step in image processing and analysis, involves image segmentation, feature extraction, representation, and description; the high level processing, which is the key step of image analysis, involves recognition, interpretation and classification. The commonly used image processing and analysis techniques in the external quality inspection of fruits and vegetables …" (Zhang et al., 2014). The first step is to increase the contrast between the image and the background to allow for accurate processing. Then, according to Zhang's (2014) model, filters are applied to remove interferences and noise in the image and to

smooth out the image. Image segmentation is the next important step. Image segmentation allows the image to be broken up into areas of interest, with each area being analyzed independently and the result being a coalesced analysis of each subset. This can be done automatically in computer vision by comparing a pixel to its neighbor to determine edges or changes in brightness (Zhang et al., 2014). The end result of this preprocessing is to extract features. Then, these features can be measured. This allows a quantitative representation of qualitative features. Feature extraction is, often, the most important part of computer vision because the extracted features become the discriminating factors downstream. Both color and size analysis are based on the individual pixels in each subsection. Pixel-based analysis is computationally intensive. 2-D size can be estimated by edge detection. Texture analysis is also very important and can be more accurate at predicting ripeness than color (Zhang et al., 2014). Texture analysis is an adaptation of pattern recognition. First, the intensity of an area of pixels is measured and the values are stored in a matrix. Then, statistical methods are used to extract a statistical estimate of texture for a certain area (Zhang et al., 2014).

2.4. Comparing Computer Vision Results to Physiological Results

Because computer vision systems are a relatively new technology, there is an ever-growing field of "checking" computer vision results by measuring corresponding physiological parameters. The paper by Pace (2014) validated their computer vision system by checking computer vision results against physiological changes. In order to quantify whether lettuce was spoiled or not, ammonia level was found to be the best

indicator of freshness. Ammonia content was analyzed using a reagent (Pace et al., 2014). Color parameters were also measured by extracting chlorophyll. This determines how "green" the lettuce is. That is, results from the computer vision system were checked by extracting chlorophyll and correlating chlorophyll concentration to the "green" feature of the computer vision system. Another way of determining the color is by using a colorimeter (Pace et al., 2014). Similarly, Aimonino (2015) correlated the ratio of browned area on frozen fruits to their physiological changes. The physiological changes were measured using solutions with varying antioxidant concentrations to quantify the extent of browning. The results showed a strong correlation between the computer vision ratio of browning and the measured antioxidant content (Aimonino et al., 2015).

Another example of validation of computer vision results comes from Zhou (2004). "Regression analysis showed that the progression of lettuce browning corresponded well with days of storage, i.e. shelf life" (Zhou et al., 2004). What this shows is that color is an accurate indicator of shelf life (for lettuce). Color was represented by the mean color. Another experiment in the same vein came from Manninen (2015) and her work with bean color differences. Color analysis was performed by finding the mean color value, which was divided by the area to give different weights to different subsets. The mean color value was checked by extracting chlorophyll from each sample to quantify the green color. Results were proven to be effective because ANOVA showed no significant differences in color coordinates within the same type of bean but differences were significant between different types of beans. This technique was effective at finding minor color differences that human graders would

not be able to detect (Manninen et al., 2015). This concept was even further demonstrated by Jha (2010) who worked with mangoes. He found parameters for the qualities of mangoes: as the mango ripens, it weighs less, becomes shorter, and becomes softer. These differences were shown both in computer vision measurement and in physiological measurement of the mangoes as they ripened (Jha et al., 2010).

Another very important paper came from Zhang (2004), which showed definitively that computer vision is a correct substitute for human grading. The degree of "spoiling" that lettuce showed was determined by calculating the percentage of brown area as a ratio. At each sampling time, a panel of experts rated the lettuce on a scale of 1-5, with a score of 3 being the threshold for "unacceptable" quality. When using computer vision, they found that one piece of shredded lettuce may not be representative of the entire head of lettuce, so a petri dish was covered with different samples from the same head of lettuce. Most importantly, they found that the most significant color changes happened within 4-6 days of storage, as mirrored by the most significant physiological changes (Zhang et al., 2004). Velez-Rivera (2014) did a similar sort of analysis using mangoes. Classification was done using RGB color analysis which was then converted to HSB (hue, saturation, brightness) color coordinates. The color coordinates were then used to classify the mangoes into three phases of maturation (unripe, ripe, spoiled). Then, the results were compared to the ripening index classification using physiochemical properties (Velez-Rivera et al., 2014). The correlation was shown to be significant.

2.5. Economics of Food Waste

The loss and waste of food is not only a moral issue, but is is also an economic issue. "39 percent of total food loss, excluding loss at the farm level, was generated at the manufacturing stage" (Gunders, 2012). Certainly, an approach that targets in-factory distribution decisions will lower food waste at the manufacturing stage. Further, according to Gunders, 20% of produce loss occurs at the production level, and a further 19% in storage and distribution losses. "Estimated at retail market prices, $15.1 billion of fresh and processed fruit were lost from the US food supply in 2008. Of this amount, roughly $5.8 billion occurred at the retail level and $9.3 billion occurred at the consumer level. The amount for each individual fruit was a function of its price per pound and, more importantly, the quantity lost. Fresh apples, strawberries, peaches and grapes each had over one billion dollars' worth of losses at the retail and consumer levels—largely because these fruits are among the most commonly purchased and consumed" (Buzby 2011). So, the economic losses that come from spoiled fruit are monumental. Reducing fruit waste through better classification and distribution approaches by computer vision can reduce food waste, saving producers billions and reducing the price of fruit for consumers, making fruit more accessible to every person.

A paper was released by Gunders (2012) that investigated the current state of food waste in America. The paper identified the different sources of food waste and a brief description of each source. By identifying each possible source of waste in the retail fruit market, different areas of waste can be targeted and custom solutions can be tailored for each area.

> "Trimming. This includes removal of both edible portions (peels, skin, fat) and inedible portions (bones, pits, etc.). Processing efficiency. While most operations are quite efficient, some steps may lose more food than necessary. Improper handling. Various kinds of mishandling, such as deliveries needing refrigeration that sit too long on the loading dock, can damage products. Inconsistent refrigeration. Truck breakdowns and other mishaps can lead to spoilage due to lack of refrigeration. Rejected shipments. By the time a shipment is rejected, its contents have a shorter shelf life and may be difficult to sell before spoiling. Food displays. Excessive products may be displayed in order to create the effect of abundance, which is believed to increase sales. There can also be overstocking, over-trimming, and improper stock rotation. Ready-made food. Increases in this perishable category lead to greater discards at end of day. Label dates. Products that pass their "sell by" dates are removed from shelves. Pack size too large. Inflexible pack sizes lead to stores' ordering more than they expect to sell. Discarded product. The passing of holidays, promotion expiration, a high failure rate for new food products, and damaged packaging all lead to discarded product. Low staffing. With tight staffing, there is less labor to prepare food on-site and therefore less flexibility in repurposing minimally damaged products" (Gunders 2012).

One of the most important conclusion from Gunders (2012) is that the majority of produce is wasted between picking the fruits and vegetables and putting them on display in the supermarket. In this way, computer vision of produce can help reduce this waste by improving distribution decisions and having a less subjective fruit grading system. Further, "a packer of citrus, stone fruit, and grapes estimated that 20 to 50 percent of the produce he handles is unmarketable but perfectly edible" (Gunders, 2012). So, having a system that allows for fruit that will be unmarketable but is still edible to be sold, can be very valuable in reducing food waste.

Globally, one third of all food produced is thrown out. Significantly, fruits and vegetables are thrown out at a higher rate than any other food type (about 50%) (Iverson 2015). This is because they are quick to spoil, sensitive to their environment, and are difficult to transport. The total estimated value of global food waste is 1 trillion dollars per year. "In developed countries, food is mostly wasted by consumers – in supermarkets, restaurants, and cafeterias, and in our homes. Since people tend to pass by bruised apples, misshapen carrots, and other imperfect fruits and vegetables, they are in turn rejected by retailers and often discarded by farmers and suppliers" (Iverson 2015). Computer vision systems can be used to more accurately classify fruits and inform distribution decisions to save billions of dollars per year.

As demonstrated, food waste is both a moral and economic concern of significance. While computer vision is beginning to be introduced into the agriculture industry, there are still large gaps that need to be addressed. Ultimately, computer vision in the agriculture industry is still in its infancy, which affords many opportunities for research. The following section identifies a process to help fill in some of the gaps explained in this literature review.

3. PROPOSED PROCESS

3.1. Relevance of this Work:

The process proposed in this thesis is necessary for the growth of the application of computer vision to the agriculture industry. One area of research being done today uses computer vision to more accurately and efficiently grade produce. Essentially, research in the field justifies and expands on the use of computer vision in the agriculture industry. However, there is no work that explicitly uses computer vision to build a model of produce shelf-life. While fast and accurate grading of produce by computer vision systems is certainly valuable, it does not address the vast amounts of food waste that the current system produces. The other area of research in this field is using computer vision to predict quality features. That is, using computer vision to correlate color features to ammonia content or percentage of color that is brown. In all of these papers, the authors reflect on how computer vision can be used to correlate color to quality, and therefore, predict how the produce will fair in traditional grading systems but none go the extra step to model the quality features to shelf-life. This introduces uncertainty and ambiguity. So, no work exists that explicitly uses computer vision to predict shelf-life.

Fundamentally, existing common applications of computer vision for grading and quality estimation fail in that they use current features (color, size, etc.) to inform distribution decisions – there is no explicit prediction of shelf-life. This thesis proposes a process that uses current features to predict future features and correlate the future features to measures of ripeness and shelf-life. As shown in the literature review, no work has been done to use computer vision to explicitly predict shelf-life. The current

industry practices for fruit grading and supply chain distribution decisions vary wildly between companies and even within the same company between fruits. However, the most common industry process is to grade incoming fruit (using a very narrow range of acceptable color, size, shape, etc.). If the fruit does not fall in the strict feature range, it is discarded. If the fruit does fall in the range, it is stored with other fruit of similar characteristics until it is almost ripe, then shipped to where it is demanded. This current process generates high storage costs and usually leads to poor yield at the retail level (often, produce arrives too ripe or too unripe because of inaccurate quality predictions).

The process proposed in this thesis makes explicit the shelf-life calculations (through data calculations and regression), which ultimately leads to a more accurate model (as shown in the "Results and Discussion" section of this thesis). The more accurate model will decrease food waste two-fold. Firstly, if a piece of fruit is graded to be of sufficient quality to sell, this thesis model will be able to predict exact shelf-life more accurately than existing quality prediction techniques. If shelf-life can be predicted more accurately, supply chain distribution decisions can become more accurate, which leads to less food waste, lower storage costs, and fewer lost profits. Secondly, if a piece of fruit is graded to be of insufficient quality to sell (perhaps, it is already too ripe to ship to retail stores), the piece of fruit can still have its shelf-life predicted and possibly shipped locally where it can be sold before it becomes overripe (that is, widen the range of acceptable features for fruit grading). It is unknown what quality correlation models will recommend (because shelf-life is not explicitly predicted in quality correlation prediction systems). Fundamentally, because current quality correlation prediction

systems do not make explicit shelf-life predictions, they introduce uncertainty which generally leads to a worse model.

The goal of this thesis is to present a model that uses color to predict shelf-life. For bananas (which this thesis uses as a proof of concept), color changes over time, which will influence supply chain distribution decisions. This can be seen in Figure 6, which plots the average percent of each respective color that a banana can be over time. Supply chain distribution decisions when the color is primarily green will be starkly different than distribution decisions when the color of the banana is primarily yellow. However, in computer vision systems, color is most often expressed in RGB (red, green, blue) coordinates. This gives three coordinates for color (red, green, and blue). In order to simplify and improve calculations, color should be consolidated into one dimension. Therefore, RGB color is transformed to the hue value of the Hue, Saturation, Intensity (HSI) color representation and RGB color is transformed into the a* value of the L*a*b* color representation. By representing color in one dimension, building a model to predict shelf-life becomes both easier and more accurate. Also, explicit shelf-life calculations allow meaning (in the form of supply chain distribution decisions) to be attached to color values. For example, if a fruit has a certain color value, shelf-life can be used to determine where to ship the fruit, which allows for a concrete decision to be attached to color values.

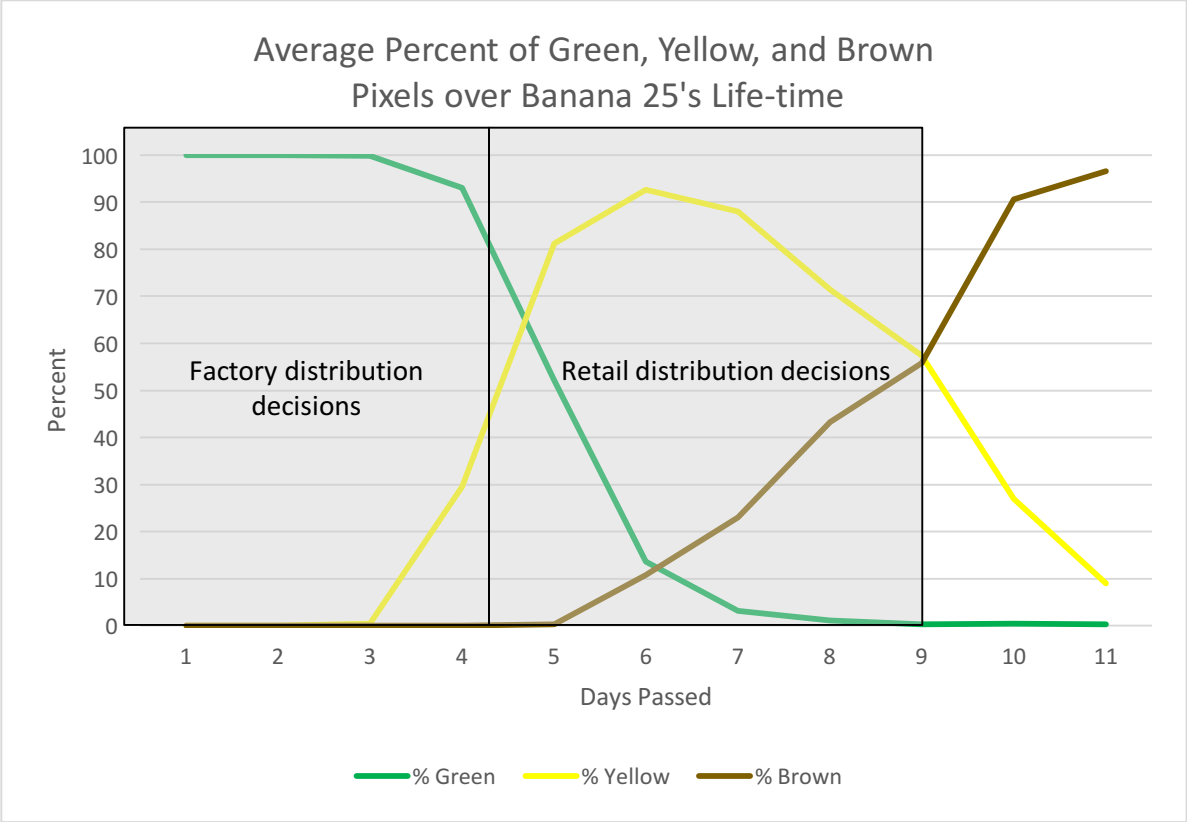


Figure 6: Average percent colors over life-time.

In order to build a model of shelf-life as a function of time, the RGB coordinates of a large sample of bananas are measured in an empirical observational study. Then, the RGB coordinates are transformed to hue and a* coordinates. An example of how the hue of a banana changes as it ripens is shown in Figure 7. Supply chain distribution decisions will vary greatly based on the color of the fruit. By building a model that explicitly predicts shelf-life, there is no ambiguity and decisions can be made based on the fruit's predicted color, instead of guessing what the future color of the fruit will be. By being able to predict exact shelf-life values, a degree of objectivity and accuracy is established that is unable to be found with traditional grading systems or current quality prediction systems.

Hue: 40.8 Hue = 28.6 Hue = 23.2

Figure 7: Hue change in a banana over its life-time.

3.2. Overview of Proposed Process:

The objective of this thesis is to develop a process that acts as the framework to build a predictive model of fruit shelf-life based on features extracted by computer vision. To that end, the following section outlines each step to take in the process and justifies each step.

The process proposed in this thesis develops certain discrete steps to undertake to build a model that predicts shelf-life based on empirically measured color features. The first step is to undertake an experiment that correlates color to sugar content (and, therefore, literature values of ripeness). The process uses color features to estimate sugar content to estimate ripeness. That is, sugar content is predicted by color, so color is the criterion for shelf-life. Using sugar content to estimate ripeness is different than using ethylene content to measure ripeness (most of the papers in the Literature Review use ethylene content measurements). The reason that this approach was chosen is simple: cost without loss of quality. A refractometer to measure sugar content can be purchased for less than $20, while a machine used to measure ethylene content costs tens of thousands of dollars. By using the cost effective method of subjectively measuring sugar content, the process allows for research to be done that may not have been economically viable using ethylene content. Research done by Tapre and Jain, 2012 and Soltani et al., 2010, have correlated sugar content to ripeness, validating it as an effective technique.

In this thesis, twenty bananas are chosen as proof of concept of the process. Every day of experimentation, two bananas are chosen, their colors are measured by taking pictures of the banana and then using computer vision to find the colors, and

finally the refractometer is used to destructively measure sugar content. The result of this experiment is a color range where the sugar content of the banana first becomes "ripe" (Figure 13). Ultimately, this step is necessary because measures of "ripeness" are fundamentally subjective, so this first experiment determines a color range that serves as the end-point of the second experiment (large scale data collection that the model is actually built off of). The second experiment uses thirty-six bananas and takes a picture of each banana on every day of the experiment. Each day, after the picture is taken of each banana, computer vision is used to extract features and the color of every banana is recorded every day. The first day that the experimental banana falls into the "ripe" color range (determined in experiment one) indicates the end of life for that specific banana. Because we are interested in distribution decisions, the first day of ripeness serves as the subjective end of life for the thesis. The justification of this is that retail stores will not accept bananas after they are ripe because they will not be able to sell them. After large scale data collection, RGB color coordinates (the default color representation) are transformed into HSI color coordinates and L*a*b* color coordinates.

While RGB (red, green, blue) coordinates are valuable and ubiquitous, they are very sensitive to lighting conditions and vary from camera to camera. It is, therefore, necessary to convert the extracted RGB coordinates to HSI (Hue, Saturation, Intensity) coordinates. This conversion deemphasizes the issue of lighting sensitivity because the image is deconstructed into color (hue) and intensity (brightness) instead of being deconstructed into red, green, and blue pixels. This simplifies and standardizes color processing by separating the color dimension from the other dimensions of an image. Also, HSI is standardized to not vary between devices as RGB does. That is, because of

differences in the amount and quality of RGB filters between imaging devices, RGB coordinates of the same image can vary between different devices. However, HSI values will be the same across devices because RGB coordinates are translated to sRGB (standard red, green, blue) before being converted to HSI, which is calculated to not take RGB filter differences between devices into account. Standard RGB accomplishes this by finding the largest coordinate of the red, green, and blue values and uses that as a non-zero reference baseline, while the other two colors become zero. As can be seen by the RGB and HSI visual representations (Figures 3 and 4), RGB is an additive representation whereas HSI is subtractive. This allows for a richer and more accurate description of color. The algorithms to transform RGB to HSI are presented in the "Analysis" section of this thesis.

The CIE L*a*b* coordinate system is an intuitive way of representing color. In this model, color is deconstructed into three dimensions (similar to how color is broken down in the human eye). In the human eye, one type of cone photoreceptor detects red and green wavelengths while another detects blue and yellow wavelengths of light. Rods detect black and white wavelengths. The CIE L*a*b* representation of color mirrors the human eye by deconstructing color into three dimensions: a* (red to green), b* (blue to yellow), and L* (black to white) but the representation range goes beyond the human vision spectrum (Figure 5). Ultimately, this produces an intuitive way of representing color that is independent of lighting conditions (unlike RGB). Also, similar to HSI, the conversion removes the device dependency of RGB by standardizing the conversion between devices because RGB is converted to sRGB (standardized RGB) before being

converted to L*a*b*. The algorithms to transform RGB to L*a*b* are presented in the "Analysis" section of this thesis.

Another representation of color is to group the colors a banana can exhibit (yellow, green, and brown) into distinct hue bins. The captured region of interest of the image of the banana can be represented as a certain percentage of each color. This allows for the use of existing conventions of color (the hue color wheel) to inform image representation. Essentially, every pixel in the analysis region of interest is transformed to its respective hue coordinate, and then the percent of pixels in the region of interest that fall into the "green", "yellow", and "brown" hue bins are calculated. This allows for a representation of the color of each banana as a percentage of the three colors that the banana can exhibit. This is done because it gives a more precise representation of color (percentage of pixels that fall into three hue bins) as opposed to an average hue value.

After the transformation of color coordinates, analysis occurs. Because there are discrete values for "experimental day", color coordinates must be standardized in order to build the model. That is, a range of color coordinates are chosen, in this case hue values of {40, 38, 36, 34, 32, 30} and a* values of {-10, -8, -6, -4, -2, 0}, and exact shelf-life values are calculated. However, it is unlikely that hue values exactly the same as the range of interest are observed for each banana. So, interpolation is used to estimate shelf-life of existing data and extrapolation is used to estimate shelf-life of data outside the range of empirically observed data. It is very important to note that no assumptions of linearity are made in this process. This is one of the reasons why, ultimately, this process produces a more accurate model of shelf-life.

The first step of the interpolation and extrapolation technique (using data to calculate remaining shelf-life) is to plot every banana's hue as a function of time. This is done to get an idea of the overall shape of the line that is created by plotting the hue of a ripening banana. Then, a hue range of interest is created. In the case of this thesis, the hue range {40, 38, 36, 34, 32, 30} is chosen because (1) the data is most linear there and (2) early-life data is generally more important for distribution decisions. However, because each banana's hue was measured empirically on every discrete day, it is rare that an experimental banana exhibited a hue of exactly 40. Therefore, interpolation must be done. For the hue value of 40, the closest point with a hue value larger than 40 and the closest point with a hue value smaller than 40 are chosen and interpolated between. The interpolation creates an equation, which can be evaluated to find the exact time measurement of the specific hue value of 40. This allows discrete, exact time measurements to be taken from a standardized hue range. Once the hue value of interest is found by interpolation, the number of days that specific banana lasted is found and the calculated time value is subtracted from the number of days lasted. This standardizes each banana (because they all lasted a different amount of days) and produces exact shelf-life estimates for a discrete range of hues ({40, 38, 36, 34, 32, 30}).

If the data point of interest lies outside of the empirically observed range (i.e. if the banana started at a hue value of 39), then extrapolation occurs. The second order polynomial with the smallest R-Squared value is created based on existing data. Then, the hue point of interest can be extrapolated. It is subtracted from the total number of days lasted to find the explicit shelf-life. A second order polynomial is used because it produces the largest R-Squared value without overfitting. All of the second-order

polynomials had R-Squared values of 0.99 and there were generally five data points to extrapolate from. This interpolation and extrapolation process occurs for every banana and every point in the hue range of interest. After all of the data points for every banana have been calculated, the shelf-life values for each hue point in the range of interest are averaged and then plotted. The best-fit line is, finally, calculated to create a model of shelf-life based on color.

The same process (interpolation, extrapolation, and averaging) occurs with the a* of the L*a*b* representation. Percent color is not used to build a model because it is the least significant (the process to determine significance follows this section). The exact calculations, examples, and graphs are shown in the "Analysis" section of this thesis.

For the purposes of comparison between the interpolation and extrapolation technique and regression, regression is also used to build models of shelf-life using color representations. To build the regression model, initially, best subsets regression is performed using each color feature (hue average, hue minimum, hue maximum, percent green, percent yellow, percent brown, a* average, a* minimum, and a* maximum) as the independent variables and shelf-life as the dependent variable. This regression is done in the software Minitab. However, because of the correlation among color features, the autocorrelation factor is well above the commonly accepted literature value of 10. Even after removing the correlated variables, the VIF is too large (very close to 10). So, linear regression is performed instead of multiple regression, using hue average in one linear regression and a* average in the other linear regression (with shelf-life as the response in each performed regression). This makes sense considering the interpolation and extrapolation technique uses the same predictors.

This process produced a coefficient P-Value and R-Squared value for each feature which is used to explain how much variation in shelf-life the feature accounts for. The P-Value found was used to inform which extracted features had the most significant effect on shelf-life determination. This information is then used to determine which features are kept and modeled upon and which features are disregarded. Essentially, this regression step shows the significance of each factor in a larger model and objectively determines which features will provide the most accurate model.

The mean absolute deviation (MAD), is calculated to evaluate each model's prediction power. Similar to the first validation step, each data point is entered into the model and the predicted shelf-life value is calculated. Then, the difference between the predicted value and the actual value is calculated. The absolute value of the difference is taken. Finally, the mean of all of the calculated absolute deviations of each data point is found. In theory, the better the predictive power of the model, the smaller the MAD. To this end, both MAD and "percentage of data points correctly predicted by the model" will be used to objectively determine the best model. This is because they are the most direct objective measure of accuracy for each model.

A final experiment is undertaken to validate each model. This is done in order to objectively test each model and to determine the distribution of prediction errors. This is the final step to determine which model is best and to speculate on error distribution. The final validation model is also advantageous because it allows another testing of sugar content to validate the "ripe" hue range established in the first experiment.

While there are no actual numbers as to how much waste the agriculture industry exactly produces by not having explicit shelf-life models, the Literature Review shows

that about 40% of produce is wasted in the manufacturing stage of the produce distribution process. The manufacturing stage is defined as between receiving the fruit at the factory and the produce leaving the factory. This stage includes grading, sorting, and storage. Therefore, the benchmark of current food waste is 40% at the level that this thesis investigates (and because shelf-life accuracy is not explicitly calculated), 60% accuracy will be the benchmark point for the models proposed in this thesis.

This section has outlined the proposed process and justified each step. The next section will describe exactly the process to undergo for each of the two experiments introduced above.

4. EXPERIMENTAL PROCEDURES AND METHODOLOGIES

This section will introduce each of the two experiments introduced above. The criteria for choosing fruit as well as the exact steps that must be taken are elaborated upon. The required software and hardware as well as the procedure for each experiment are discussed.

There are two fundamental features that the fruit of interest must possess in order to work in the proposed model building process. The first feature is that the fruit must be capable of ripening after it is picked. The post-harvest ripening is a characteristic of climacteric fruit. If the fruit does not ripen after harvesting, there would be no shelf-life model to build. The second feature is that the fruit must exhibit noticeable changes in color (or some other feature such as size or shape) as it ripens. That is, computer vision must be able to detect the color (or other feature) changes as the fruit ripens. The fruit used in this thesis is bananas because they have very distinct unripe (green), ripe (yellow), and spoiled (brown) states. Bananas are also climacteric, so they will ripen after they are harvested. Bananas were also chosen as the fruit for this thesis because they are one of the most important agricultural commodities (in terms of units sold and widespread distribution). Other fruits that the process presented in this thesis could work with are: apples, melons, tomatoes, avocado, blackberries, etc. It is important to note that fruit grown in different conditions (soil composition, temperature, humidity, sun exposure, etc.) may exhibit different color features. The results that this thesis presents are valid only for the specific bananas that were tested. The following section outlines the experimental processes that were undertaken.

4.1. Experiment 1:

Experiment 1 Procedure:

Objective:

Determine the hue value that corresponds to a ripe banana (23% sugar content) (Tapre and Jain, 2012 and Soltani et al., 2010). This hue value will be used as the end-point of the data collection for the second experiment.

Materials:

EARTH brand Cavendish Bananas (*Musa acuminata*, Costa Rica) (n = 20)

Canon EOS 20-D camera (in raw and .jpeg mode) (Figure 10)

Sony Computer Vision camera (XCD-X710CR #100230) (Figure 9)

Camera mounting device

Blank sheet of paper: this will be a blank sheet of paper with colored boxes printed on it for reference

Software: National Instruments LabView Vision Builder

Macro Ring Lite MR-14 EX lighting source

Refractometer (Soyan ATC: 600316291330)

Thermometer

Pre-Experiment Check List – Repeat Everyday:

Lighting: Lighting should be as consistent among measurements as possible. Use the same light source each time and try to block out any external light. In the thesis experiment, all external light was blocked and the only source of illumination was the Macro Ring lighting source for the Canon camera. The Sony computer vision camera used ambient lighting.

Temperature: Temperature should be kept as uniform as possible to store the fruit for the extent of the experiment (about two weeks). The fruit was stored in a room that had a constant temperature of 63 degrees Fahrenheit. If this is not possible, take and record the temperature at measurement time.

Background: The background of the image should be the same across samples and measurements – in this case, use a blank piece of paper with colored boxes printed on it.

Camera: Ensure that the same cameras are used throughout the experiment.

Camera Distance: It is vital that the camera is always the same distance away from the object of interest. Take measurements and verify the distance before each measurement.

Time of day of measurement: Try to keep the time of day the same as often as possible. During this thesis, measurements were taken at 11:00 AM.

The experimental setup for this thesis is shown in Figure 8.

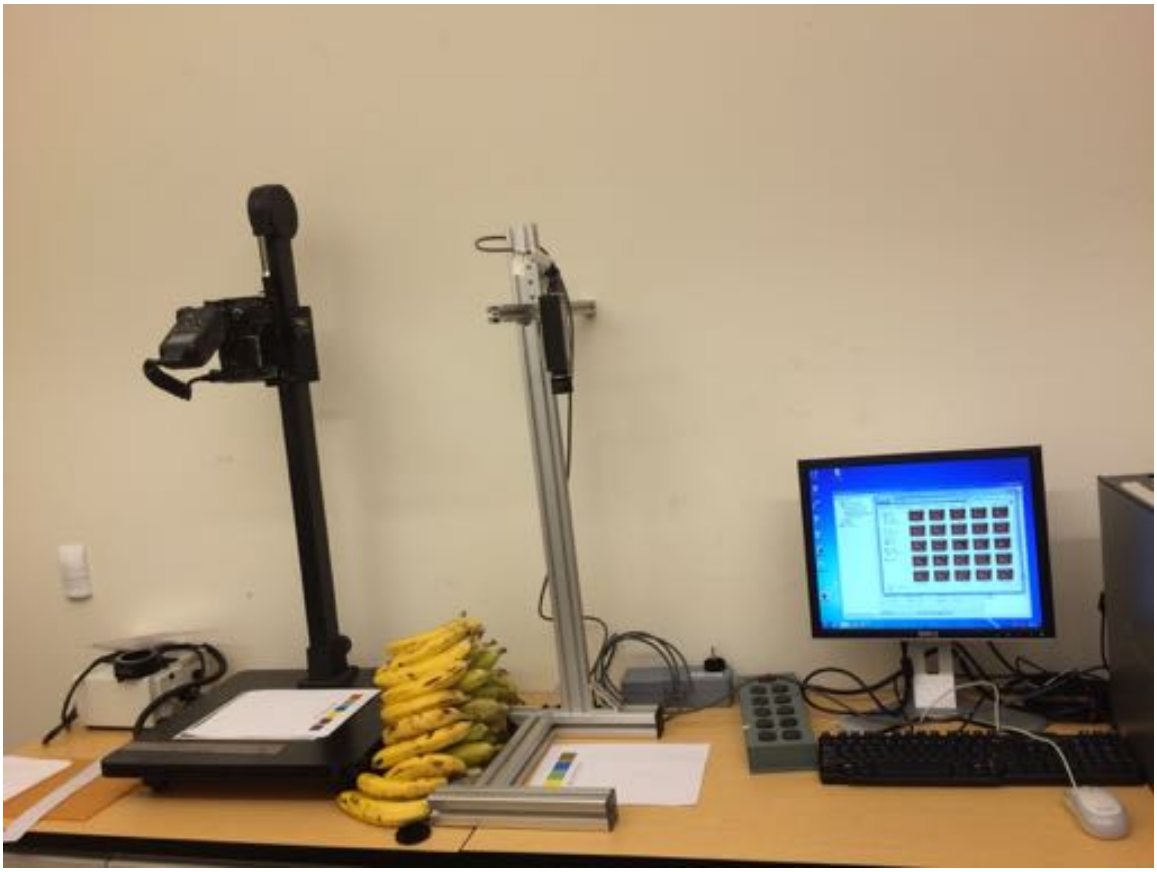


Figure 8: Experimental procedure set-up.

Figure 9: Sony computer vision camera set-up.

Figure 10: Canon Eos 40-D camera set-up.

Method:

1. Starting with 20 bananas as mentioned above, label each banana so that it can be distinguished at a later time.
2. Set up the experiment by putting the first banana on the blank sheet of paper. Place the piece of fruit directly under the camera.
3. Configure the camera using the viewfinder so that the image acquired captures only the entire fruit.
4. Measure and record the distance of the camera from the banana. Ensure that this distance is consistent for every image acquisition.
5. Set up a repository for the images on the computer that the camera is attached to. The computer should have National Instruments LabVIEW already installed on it.
6. Capture the image.
7. Save the image to the computer.
8. If two cameras are used, repeat steps 1-7 with the other camera.

9. Calibrate the refractometer by dropping a small amount of water onto the device.
10. Cut out a small segment of the selected banana. Bananas are selected numerically (1 through 20), with two bananas tested per day. For example, on experimental day three, bananas # 5 and # 6 are tested.
11. Extract the juice from the segment of banana.
12. Drop the extracted juice onto the sacchorometer.
13. Measure and record the sugar content of the juice in percent brix.

14. Remove the destructively measured banana from experimental procedure.
15. Repeat procedure for each banana selected to be tested on the specific day.
16. Repeat steps 1-15 for each day until all of the remaining tested bananas have all had their sugar content destructively measured.

Computer Vision Feature Extraction:

1. Load the images into LabVIEW Vision Builder by pressing "Simulate Acquisition" and navigate the file path to the folder with the acquired images.
2. Choose a consistent 150 pixel by 150 pixel square region of interest to perform processing. An example region of interest can be found in Figure 11.
3. Use the "Measure Colors" feature to extract and record color features. This should be done in RGB, as that is the inherent color representation in LabVIEW and transforming color representations happens subsequently.
4. As a validation step, ensure that the background paper's squares has the same color measurements among fruits, days, and devices. Any samples with deviations in color measurements should be removed from the experiment.
5. In LabVIEW, convert RGB to HSI, CIE L*a*b*, and "percent brown, yellow, and green" color coordinates using the appropriate formulas (found in the Analysis section of this thesis).
6. Compute Hue average, Hue minimum, Hue, max, percent yellow, percent green, percent brown, a*, a* minimum, and a* maximum (the computation instructions can be found in the Analysis section of this thesis).

7. Repeat analysis for each acquired image.
8. Analyze the refractometer data by comparing it to the colors extracted using computer vision. Note the color of the sample bananas in which the brix content equals or exceeds 23%. Use this color (range) to set as the endpoint of shelf-life. Shelf-life for each day is the end day point minus the experimental day.
9. We now have the values for the dependent variable (shelf-life) for each sample banana (the experimental day subtracted from the day that the banana's color falls into the range calculated in step 2). Record the shelf-life values for every sample banana. The end point for data collection was an average hue of 25 and an average a* value of 4 (Figure 13). The data and results are shown in the "Results and Discussion" section of this thesis.

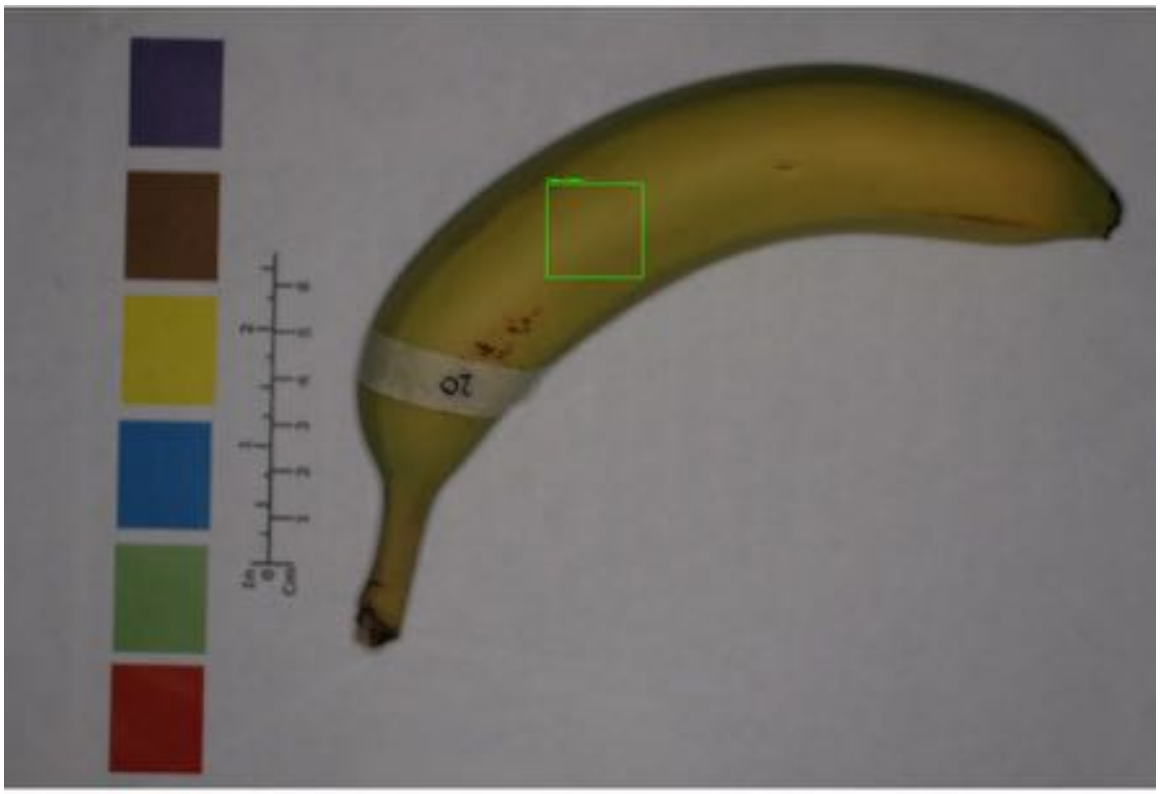


Figure 11: Example region of interest.

4.2. Experiment 2:

Experiment 2 Procedure:

Objective:

Now that we have an end-point for data collection (experiment 1), large-scale data collection for extraction of features of sample images can be done to build a predictive model of shelf-life.

Materials:

EARTH brand Cavendish Bananas (*Musa acuminata*, Costa Rica) (n = 36)

Canon EOS 20-D camera (in raw and .jpeg mode)

Sony Computer Vision camera (XCD-X710CR #100230)

Camera mounting device

Blank sheet of paper: this will be a blank sheet of paper with colored boxes printed on it for reference

Software: National Instruments LabView Vision Builder

Macro Ring Lite MR-14 EX lighting source

Thermometer

Pre-Experiment Check List – Repeat Everyday:

Lighting: Lighting should be as consistent among measurements as possible. Use the same light source each time and try to block out any external light. In the thesis

experiment, all external light was blocked and the only source of illumination was the Macro Ring lighting source for the Canon camera. The Sony computer vision camera used ambient lighting.

Temperature: Temperature should be kept as uniform as possible to store the fruit for the extent of the experiment (about two weeks). If this is not possible, take the temperature at measurement time and block if necessary. For this thesis, the fruit was stored in a room that had a constant temperature of 63 degrees Fahrenheit.

Background: The background of the image should be the same across samples and measurements – in this case, use a blank piece of paper with colored boxes printed on it.

Camera: Ensure that the same cameras are used throughout the experiment.

Camera Distance: It is vital that the camera is always the same distance away from the object of interest. Take measurements and verify the distance before each measurement.

Time of day of measurement: Try to keep the time of day the same as often as possible. During this thesis, measurements were taken at 11:00 AM.

Method:

1. Label each banana so that it can be distinguished at a later time.
2. Set up the experiment by putting the first banana on the blank sheet of paper. Place the piece of fruit directly under the camera.
3. Configure the camera using the viewfinder so that the image acquired only captures the entire fruit.
4. Measure and record the distance of the camera from the banana. Ensure that this distance is consistent for every image acquisition.

5. Set up a repository for the images on the computer that the camera is attached to. The computer should have National Instruments LabVIEW already installed on it.
6. Capture the image.
7. Save the image to the computer.
8. If two cameras are used, repeat steps 1-7 with the other camera.
9. Repeat steps 1-8 for each day until the hue reaches the end-point hue established in Experiment 1.

Computer Vision Feature Extraction:

1. Load the image into LabVIEW Vision Builder by pressing "Simulate Acquisition" and navigate the file path to the folder with the acquired images.
2. Choose a consistent 150 pixel by 150 pixel square region of interest to perform processing.
3. Use the "Measure Colors" feature to extract and record color features. This should be done in RGB.
4. As a validation step, ensure that the background paper's squares has the same color measurements among fruits, days, and devices. Any samples with deviations in color measurements should be removed from the experiment.
5. Convert RGB to HSI, CIE L*a*b*, and "percent brown, yellow, and green" color coordinates using the appropriate formulas (found in the Analysis section of this thesis). This should be done for every banana in the sample.

6. Compute Hue average, Hue minimum, Hue, max, percent yellow, percent green, percent brown, a*, a* minimum, and a* maximum (the computation instructions can be found in the Analysis section of this thesis). This should be done for every banana in the sample.
7. Repeat analysis for each acquired image.

Figure 12 shows an example of how the data may look after analysis:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Fruit	Day	Hue Avg	Hue Min	Hue Max	%G	%Y	%B	a	a Min	a Max	Shelf-Life
2	1	1	27.3	17	31	0.1	99.82	0.63	1	-1.6	8	4
3		2	27.1	12	30	0.21	98.44	4.47	1.2	-1.2	16.6	3
4		3	26.2	16	30	0.04	86.64	20.56	2.1	-1.8	13.5	2
5		4	24.5	10	29	0	58.54	54	3.7	-1.7	15.12	1
6		5	23.2	8	29	0	45.6	73.93	4.7	-5.67	15.32	0
7	2	1	28.2	21	32	11.2	99.6	0.09	0.2	-3.02	8.8	4
8		2	27.5	20	31	2.22	99.92	0.65	0.9	-2.37	3.86	3
9		3	27	22	31	0.32	99.37	4.62	1.3	-2.05	7.23	2
10		4	25.3	12	30	0.1	70.81	39.82	3	-1.56	16	1
11		5	22.7	0	50	0.36	37.85	74.74	4.7	-5.34	14.92	0
12	3	1	27.7	20	31	1.43	99.79	0.36	0.5	-2.11	8.68	6
13		2	27.1	10	30	0.1	99.66	1.7	1.3	-1.2	22.86	5
14		3	27	20	30	0.04	99.62	2.48	1.2	-1.7	5.95	4
15		4	26.9	18	30	0.04	97.65	7.82	1.4	-1.7	11.81	3
16		5	25.5	12	30	0	76.89	39.16	2.8	-1.05	12.77	2
17		6	24.5	13	28	0	52.95	65.1	3.7	0.15	12.95	1
18		7	23.6	11	22	0.05	43.12	70.27	4.6	-2.79	12.896	0
19	4	1	27.9	24	32	3.65	99.53	0.12	0.4	-3.17	5.34	4
20		2	26.9	18	30	0.22	98.33	5.68	1.4	-1.34	8.13	3
21		3	26.1	16	30	0.01	91.04	25.06	2.4	-0.98	8.101	2
22		4	25.2	12	29	0	70.82	49.61	3.2	-0.13	13.29	1
23		5	23.3	10	9	0	38.77	75.58	5	-0.38	16.04	0

Figure 12: Example data spreadsheet.

Notes:

While it is possible to run each experiment separately, both experiments were run concurrently for the purposes of this thesis. This was done to ensure that the conditions among both experiments were consistent. However, in order to run both experiments at the same time, the fruit must be clearly and carefully labeled and it should be noted that the data collection experiment often runs longer than it must. It is also important to note

that a dry run of data collection (practice using the refractometer and imaging devices) may be helpful and is encouraged.

There were two cameras used in both experiments. The first, Canon EOS 20-D, is a CMOS (complementary metal-oxide semiconductor) sensor camera and the other camera, Sony XCD-X710CR, is a CCD (charged couple device) sensor camera. The Canon camera carries a few advantages over the Sony camera. Firstly, the Canon camera has more RGB filters which allows for a more precise representation of color. Secondly, the Canon camera allows for manual overrides of certain features (ISO speed, color balance, storage mode etc.). Thirdly, the Canon camera has a higher resolution which allows for better image quality. The advantage of the Sony camera is that it has a shorter exposure time and better shutter mechanics, so images can be captured faster, with less distortion. However, there is no "correct" camera for this process because each situation will be unique. For example, if a company highly priorities speed of capture for high speed grading, a CCD camera (the Sony camera) may be the correct camera. If a company emphasizes accuracy over speed, a CMOS camera (the Canon camera) may be recommended. In the thesis, accuracy was prioritized over speed because there was no time pressure (and a consistent theme in this thesis is accuracy over speed), the Canon camera was the most suitable.

The only signal that is extracted from the images, using LabVIEW, is the RGB coordinates of every pixel in the image. Because of the scope of this thesis, color was determined as the only feature to be extracted, as it has been shown to be the most accurate indicator of quality (see the Literature Review). So, the only signal of interest was color. Because the Canon camera captured images in "Raw" mode (without any

compression of the images), there was no post-processing done by the camera. The unprocessed image was loaded into LabVIEW and the RGB color coordinates are extracted. Then, the RGB color coordinates are transformed (as illustrated above) to HSI and L*a*b* coordinates. So, the only signal collected for analysis in LabVIEW is the unprocessed RGB color coordinates.

This section of the thesis has developed the exact steps to be taken in order to perform the two experiments for data collection. The next section will delve into direct results from the experiment, before any analysis occurs.

5. DIRECT EXPERIMENTAL RESULTS

5.1. Sugar Content:

The purpose of this section is to introduce results that come directly from the experiments (i.e. without analyzing the data). First, sugar content as a function of color is correlated, and then replicability between devices is described.

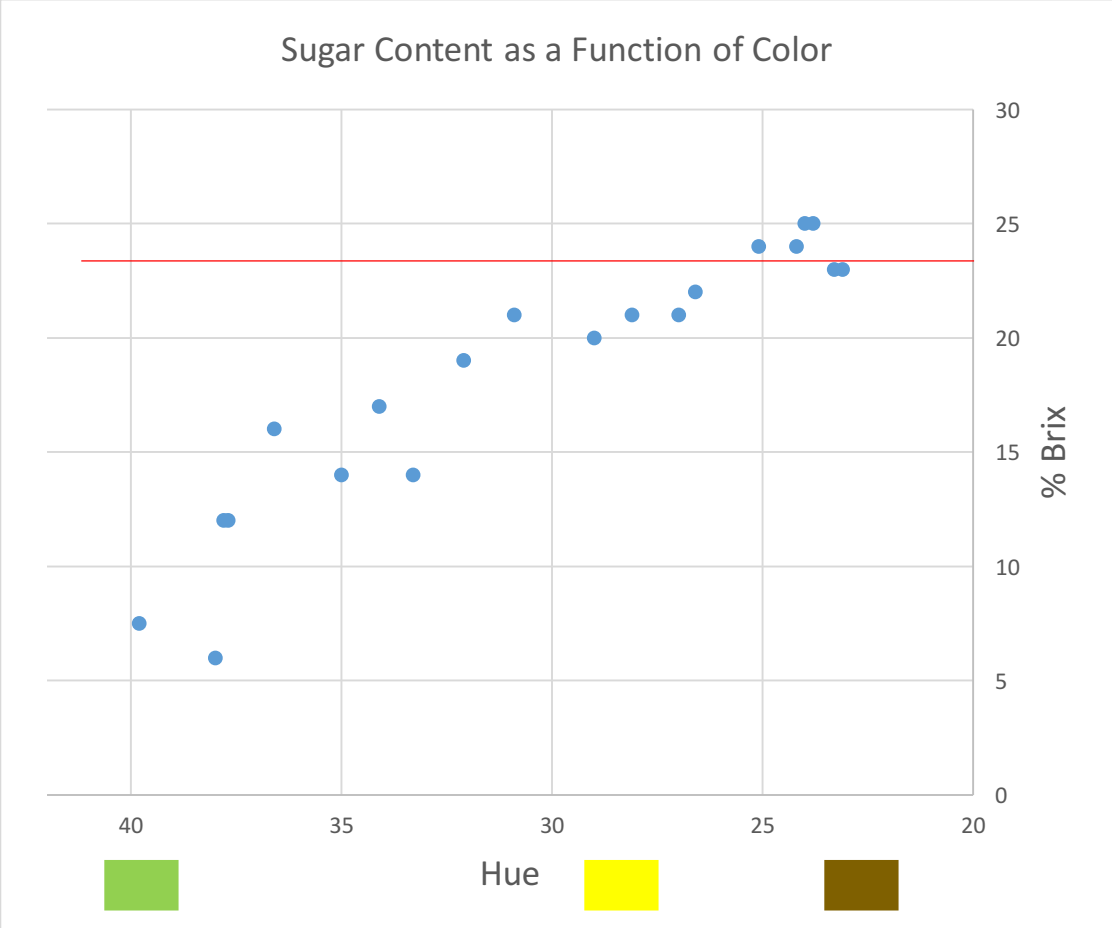


Figure 13: Sugar content over time.

The sugar content of each tested banana was calculated and plotted as a function of color (average hue). This allows a range of "ripe" colors to be calculated. According to the literature (Tapre and Jain, 2012 and Soltani et al., 2010), a brix content of 23% indicates the point at which a banana is ripe. Based on the collected data, a ripe banana has a hue value between 25 and 23 (Figure 13). Please note that the hue value axis is reversed in order to represent a more natural way of showing sugar content over time. Thus, the first day that the hue value of the banana of interest falls into (or goes past) the hue range (25), the lifetime of that banana is ended and the banana is removed from the experiment. This process allows for a subjective way to determine the end-point for each banana in the experiment. There are many other factors that can be used to find an objective end-point such as: ethylene content, firmness, curvature, length, width, etc. Hue was used to predict sugar content and sugar content was used to measure ripeness. Sugar content was chosen as the measure of ripeness for this experiment because it is the cheapest and simplest to measure, there already is literature correlating sugar content to ripeness, and there is a well-defined process used to measure sugar content in fruits with a refractometer. In order to determine whether the correlation between sugar content and hue is significant, the values of both hue and brix are loaded into Minitab. The Pearson correlation coefficient is -0.937 with a P-Value of less than 0.0001. Thus, the correlation is significant.

It would be possible to fit a model to the data shown in Figure 13, in order to use hue to predict % Brix. However, because the scope of the thesis is to find a model that predicts shelf-life using color, fitting a model to this data would be unnecessary. It is important to keep in mind that hue predicts sugar content which measures stages of

ripeness. So, the relevant part of this data is the range of the hue where the sugar content first exceeds 23% because we are only interested in the last day that the fruit can be sold to retail stores (the first day of ripeness).

5.2. Replicability Between Devices

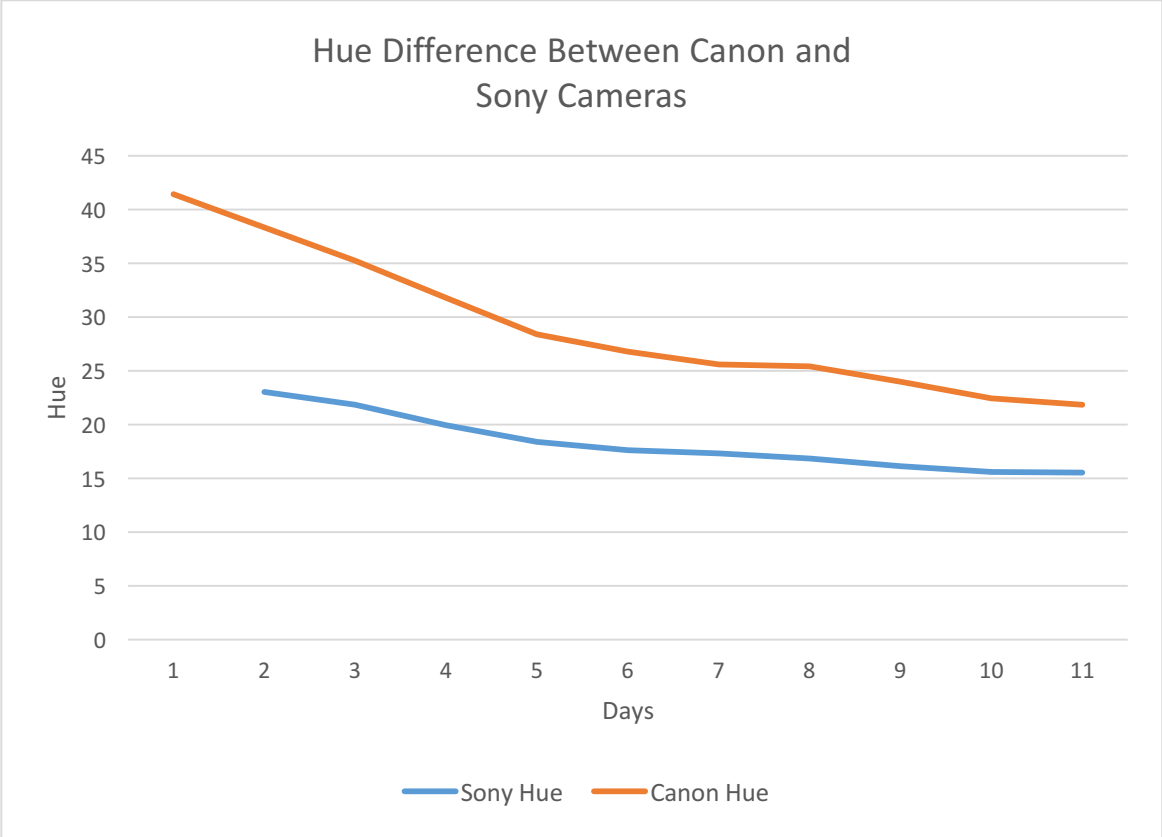


Figure 14: Replicability of devices.

Every day, the bananas were photographed by the Canon camera and by the Sony computer vision camera. Computer vision feature extraction (as above) is performed and the results can be compared between devices (Figure 14). For every banana, hue features were extracted and plotted for both the Canon and Sony cameras but the hue values were

not the same. This can be due to many factors. Firstly, the Canon camera was used in isolation of external light and used a very controlled light source, whereas the Sony camera used ambient fluorescent light. Next, the distance of the camera from the banana was held constant among days and fruit but varied between the Canon device and the Sony device. The variation in distance was present because the camera's viewfinders had different ranges (if the same distance was used between devices, the Sony camera would only capture half of the banana). The Sony camera had a much smaller viewfinder than the Canon camera. It is also true that the RGB filters in each of the cameras were different. That is, both cameras use different sensor arrays. It seems that the Canon camera was more accurate (more RGB filters) in terms of color vision in comparison to the Sony camera, based on the difference in range of both cameras exhibited in Figure 14. The light source for the Canon camera was consistent and controllable, whereas the ambient light for the Sony camera varied day-to-day, which could also be the source of difference between images. These differences were vast enough to not be reconcilable by RGB transformation to HSI or L*a*b*. Most importantly, this analysis shows that the model is only valid for bananas in the very specific thesis conditions (they must have been captured with the Canon EOS 20D camera, illuminated by the Macro Ring, stored at 63 degrees Fahrenheit, etc.).

The purpose of this section was to introduce two important results that came directly from the experiments. The next section explores how to analyze data in order to ultimately build the final models.

6. ANALYSIS

After data has been collected from the experiments, the data must be analyzed in order to build the models. First, RGB coordinates (from computer vision feature extraction) are transformed to L*a*b* and HSI color coordinates. Then the method to actually build the models using interpolation and extrapolation is explained. Finally, an overview of how to measure accuracy and how to perform regression for model comparison is developed.

6.1. Converting RGB coordinates to HSI coordinates:

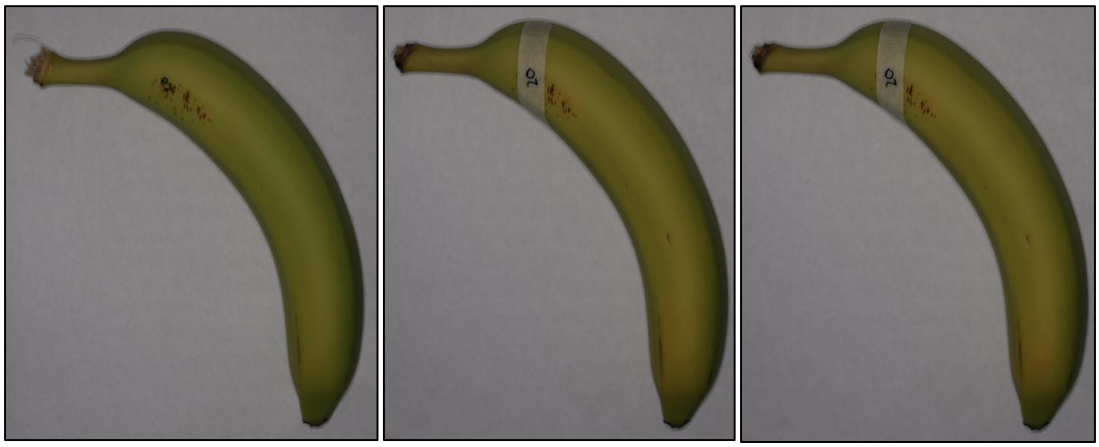
As explained above, RGB color coordinates must be transformed to HSI color coordinates in order to build the models. Each pixel in a given (and consistent between individual bananas) region of interest (Figure 11) was classified using RGB coordinates and then converted to HSI. The consistent region of interest was 150 pixels by 150 pixels. A script was written to automate the conversion between RGB and HSI is shown in Figure 15. Please note that the calculations were performed in LabVIEW but the script in this figure is written in Python to show the mathematical logic behind the conversion.

```
r = R/(R+G+B)
g = G/(R+G+B)
b = b/(R+G+B)

if b <= g
    hue = acos((0.5*(r-g)+(r-b))/(sqrt((r-g^2)+(r-b)*(g-b))))

if b > g
    hue = (2*pi)- acos((0.5*(r-g)+(r-b))/(sqrt((r-g^2)+(r-b)*(g-b))))

saturation = min(r,g,b)

intensity = (R+G+B)/(3*255)
```

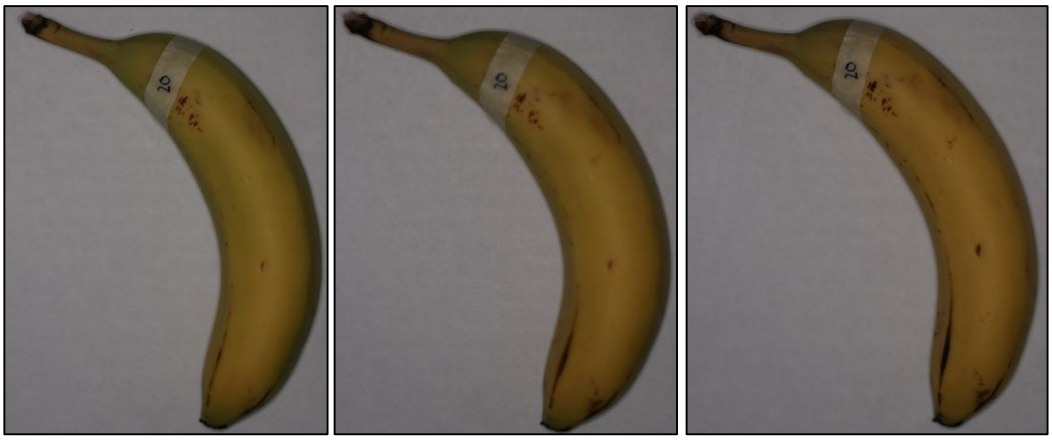
Figure 15: The script to convert between RGB and HSI.

Average hue is calculated by first finding the RGB coordinates of every pixel in the 150 pixel by 150 pixel region of interest (which is the same region of interest between and among bananas). This is done using LabVIEW's "Region of Interest" command and finding summary statistics. Then, each pixel in the region of interest is converted from RGB to HSI (using the script above). Finally, the average hue value is calculated by dividing the sum of the hue values of the pixels in the region of interest by the total number of pixels (for this thesis, 22500 pixels) in the region of interest. Hue maximum is the largest hue value in the region of interest and hue minimum is the smallest hue value in the region of interest. All of the calculations are done in LabVIEW.

The average hue values over the lifetime of a selected banana are shown in Figure 16.



Day 1: Hue = 40.8 Day 2: Hue = 37.8 Day 3: Hue = 35.1



Day 4: Hue = 31.9 Day 5: Hue = 28.6 Day 6: Hue = 26.6

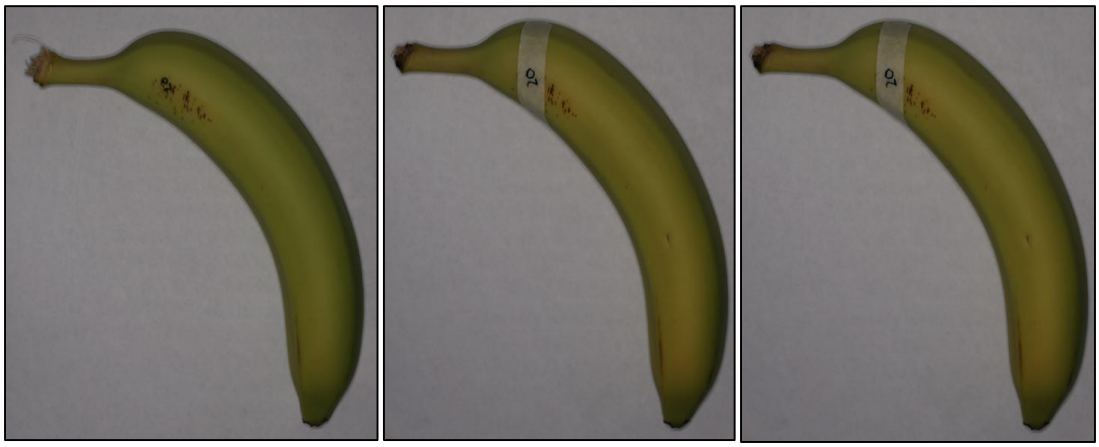
Day 7: Hue = 25.6 Day 8: Hue = 25.4 Day 9: Hue = 23.7

Figure 16: Example hue change over life-time of banana 20.

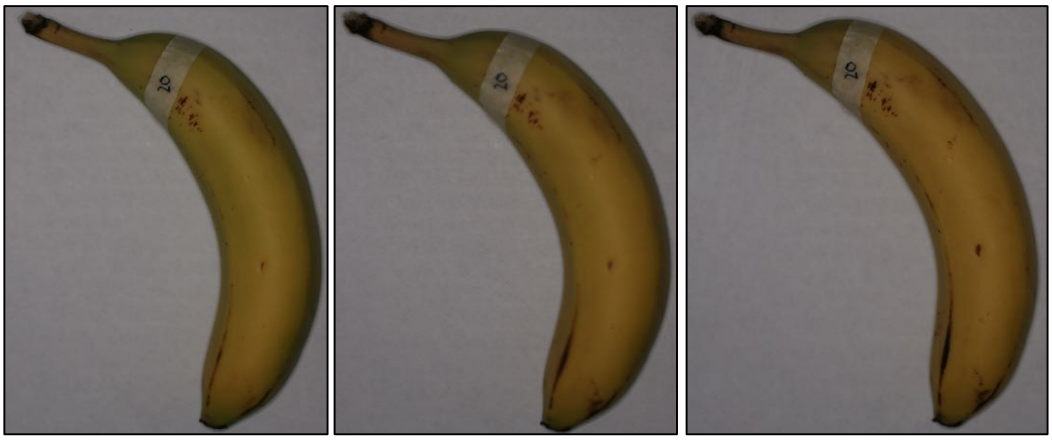
6.2. Converting RGB Coordinates to L*a*b* coordinates:

For the reasons articulated above, RGB color coordinates must be converted to L*a*b* color coordinates. The mathematical conversion of RGB to L*a*b* is done in LabVIEW using the command "rgbtocolor2". First, the RGB coordinates are converted to sRGB to remove the device dependency of RGB. Then, the sRGB coordinates are transformed into CIE XYZ coordinates which is a representation of the color space with the same limits as the human eye. Finally, the CIE XYZ coordinates can be transformed into the CIE L*a*b* representation, which extends the color ranges beyond human vision. This is all done inherently in LabVIEW's "rgbtocolor2" function.

To find the average a* value, each pixel in a given region of interest is represented by its RGB coordinates. This is done using LabVIEW's "Region of Interest" command and finding summary statistics in the same way demonstrated in the "Converting RGB coordinates to HSI coordinates". Then, the RGB coordinates of each pixel are converted to CIE L*a*b* coordinates (using the "rgbtocolor2" function, as above). The average a* value is calculated by finding the sum of the a* values of every pixel in the region of interest and dividing by the total number of pixels. The a* maximum value was found by taking the largest a* value in the region of interest and the a* minimum value was found by taking the smallest a* value in the region of interest. An example of the a* change over the lifetime of a banana is demonstrated in Figure 17, which shows the a* lifetime change of banana 20.



Day 1: a* = -8.4 Day 2: a* = -6.8 Day 3: a* = -5.2



Day 4: a* = -2.7 Day 5: a* = -0.1 Day 6: a* = 1.6

Day 7: a* = 2.5 Day 8: a* = 2.6 Day 9: a* = 3.8

Figure 17: Example a* change over life-time of banana 20.

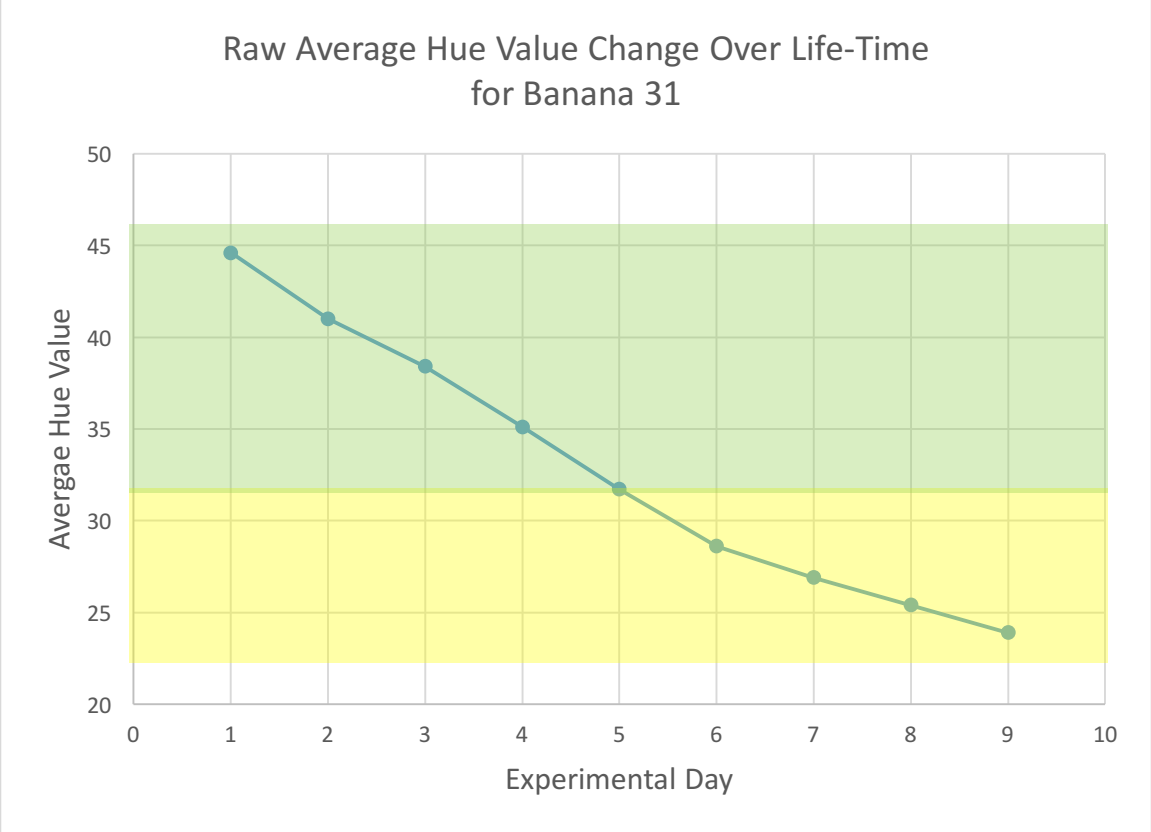
6.3. Converting RGB to Percent Yellow, Green, and Brown:

RGB coordinates are also converted to percentages of the three colors exhibited by a ripening banana: green, yellow, and brown. This is done by calculating the RGB coordinates of each pixel in a given region of interest (that is consistent among and between individual bananas). Each pixel is then converted from RGB coordinates to the HSI representation (as shown in the "Converting RGB to HSI section" of this thesis) (Figure 4). Then, the total number of pixels is found using the LabVIEW "Region of Interest" command and calculating summary statistics. A color wheel is then consulted (Figure 18) to find the literature hue values for green, yellow and brown. The green color hue is between 60 and 32, the yellow color hue is between 32 and 23, and brown color hue is between 23 and 10. Then, using LabVIEW, the number of pixels classified as green (the pixel's hue values fall between 60 and 32) is calculated. Finally, the number of green pixels is divided by the total number of pixels to get the percentage of green pixels. This procedure (finding the number of pixels in a given range and then dividing by the total number of pixels) is then done for the pixels classified as yellow and the pixels classified as brown.

Figure 18: Hue color wheel (Work with Color, 2016).

6.4. Data Calculation of Remaining Shelf-Life:

Every day of data collection, the hue value of every sample banana was found using computer vision (as described above in the experimental procedures). At the end of data collection, the hue value over every banana's lifetime is recorded. Then, the hue values of every banana in the sample can be plotted as a function of days passed (time). An example of this is seen in Figure 19. This plotting gives a visual representation of how the hue values of the bananas in the sample change over the respective banana's lifetime. The changing hue values over time for every sample banana can be seen in Figure 20.



Figures 19: Raw hue value change over lifetime for banana 31.

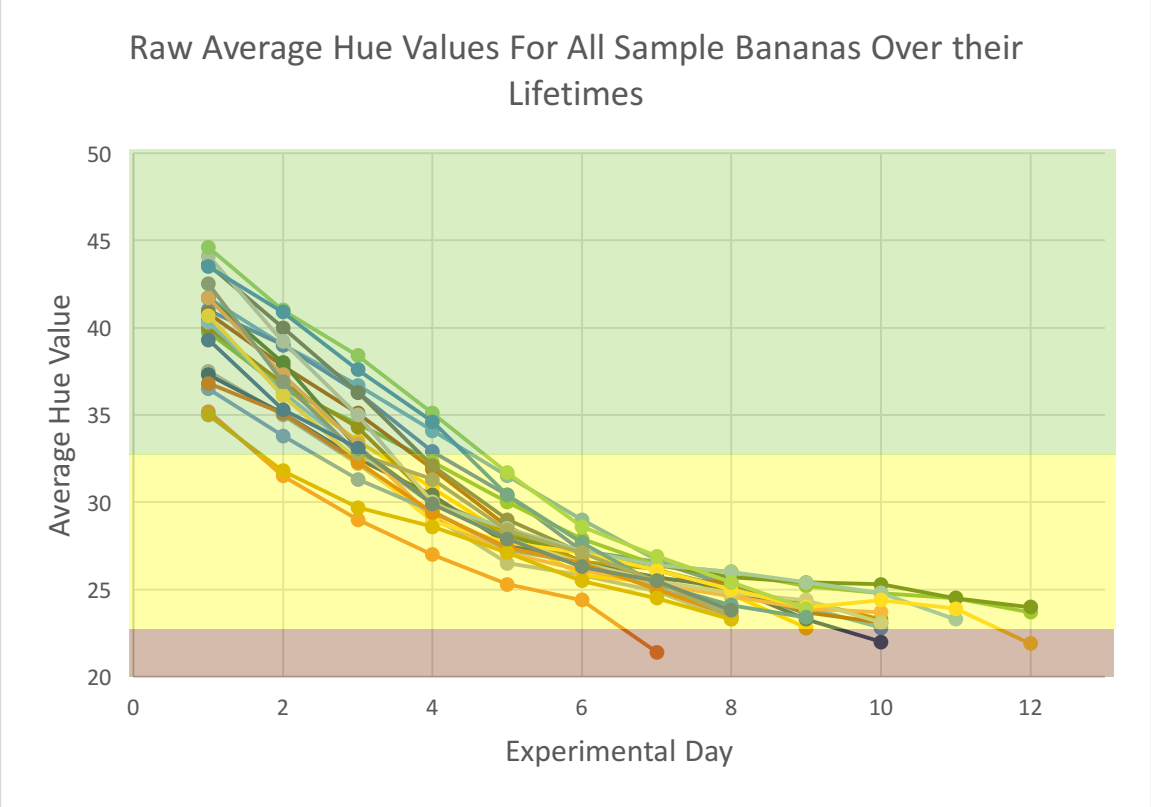


Figure 20: Raw hue value for all samples of bananas as a function of time.

The same analysis as above is performed on the a* data (as opposed to the hue data). The results of plotting the raw a* value over a sample banana's lifetime is shown in Figure 21. The results of plotting all of the a* value changes over every sample banana's lifetime is shown in Figure 22.

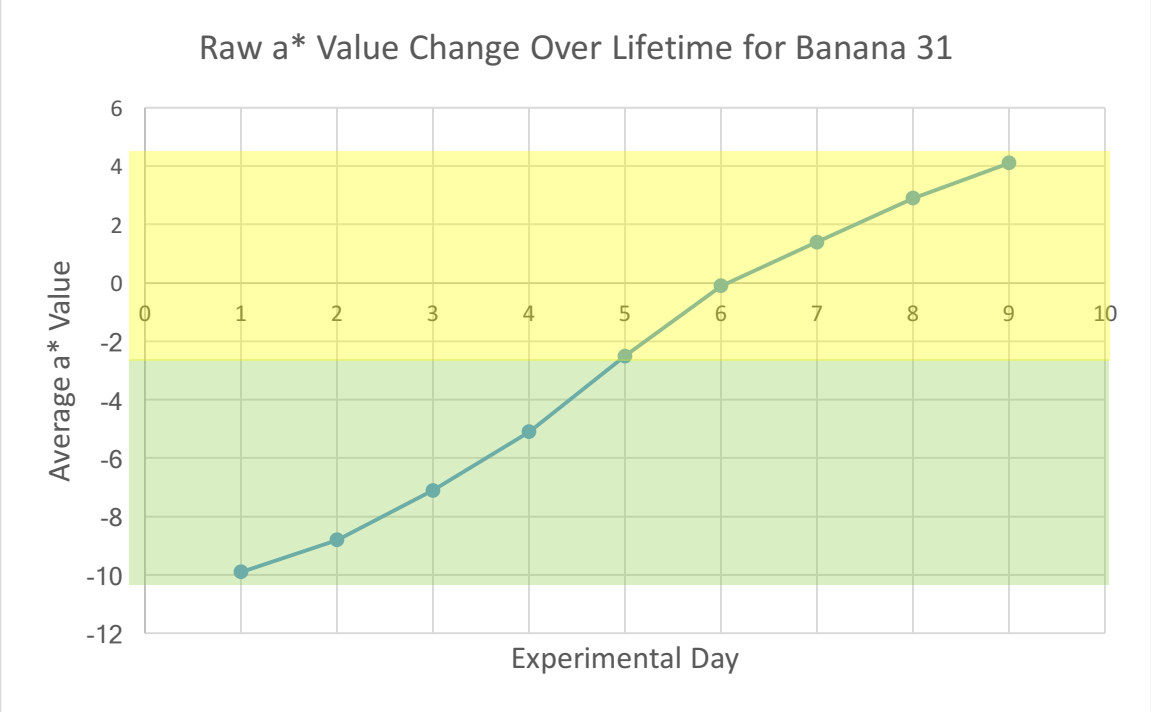


Figure 21: Raw a* value change over life-time for banana 31.

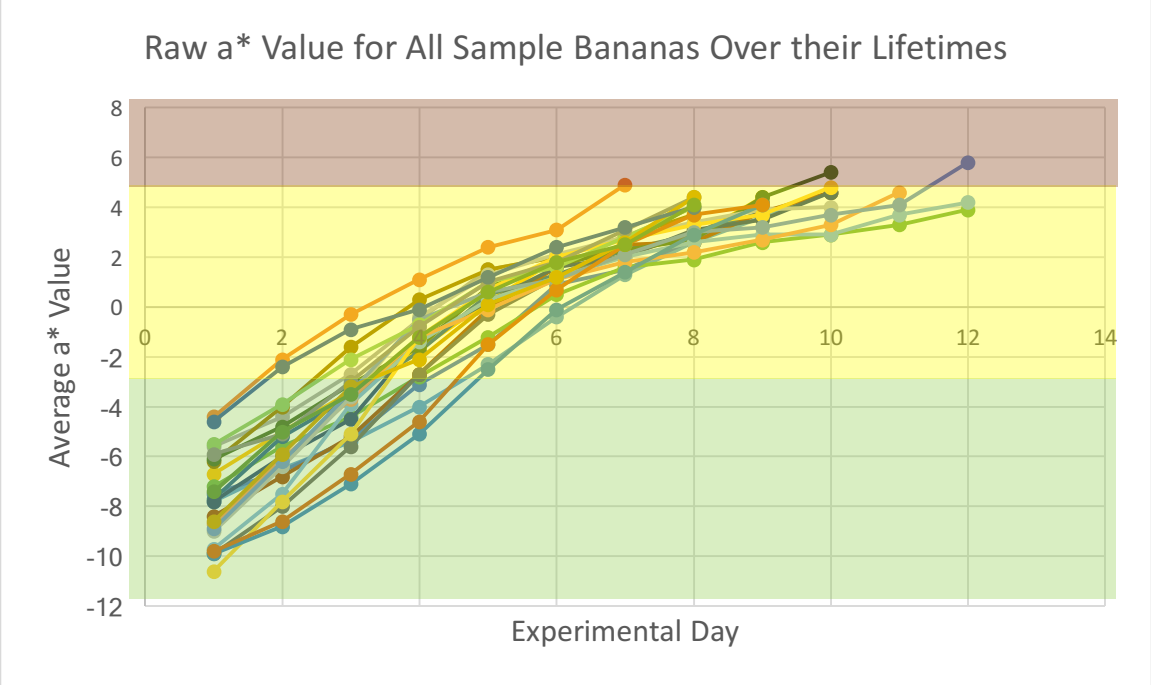


Figure 22: Raw a* value changes for all sample bananas as a function of time.

Keeping in mind that this model is supposed to be deployed to make supply chain distribution decisions, and the increase in variation of 'experimental days' as the fruit nears the end of its lifetime, a subjective decision to only use the data with a hue value above 28 was made. The most relevant and useful data (with respect to supply chain distribution decisions) is data early in the fruit's life. In general, prediction accuracy increases as the fruit gets closer to its shelf-life. For example, it is easier (but, in this case, less useful) to determine a banana's shelf-life one day before it spoils as opposed to fourteen days before it spoils. So, the important data (for the purpose of this thesis) is the data early in the fruit's lifetime. Thus, any data with a hue value less than 28 was cut out of the analysis (except to record the last day which informed shelf-life of each individual banana). The same sort of analysis was done on the a* data. For the same reasons as enumerated above, data points with an a* value larger than 0 were discarded.

In order to build the model, hue has to be standardized among each banana in the sample because each banana starts at a unique hue and has a unique shelf-life. That is, it is necessary to calculate every sample fruit's shelf-life at a specific given hue. However, it is common that the exact hue value of interest was not recorded (if the hue value of interest is 40, the fruit may have gone from a hue of 41.2 on day one to a hue of 39.9 on day two), so it is necessary to interpolate between existing data points. For each hue value: {40, 38, 36, 34, 32, 30}, the line between the two points closest to the point of interest was interpolated. That is, the closest point greater than the point of interest and the closest point less than the point of interest are selected and interpolated between. For example, if the point of interest is 38 and the two closest data points are (1, 39) and (2,

37) (where the y-coordinate is the hue value, and the x-coordinate is the experimental day of measurement), the slope and intercept are calculated using the two existing points to produce the line ‘hue’ = -0.5 * ’days’ + 20.5. This is shown in Figure 23.

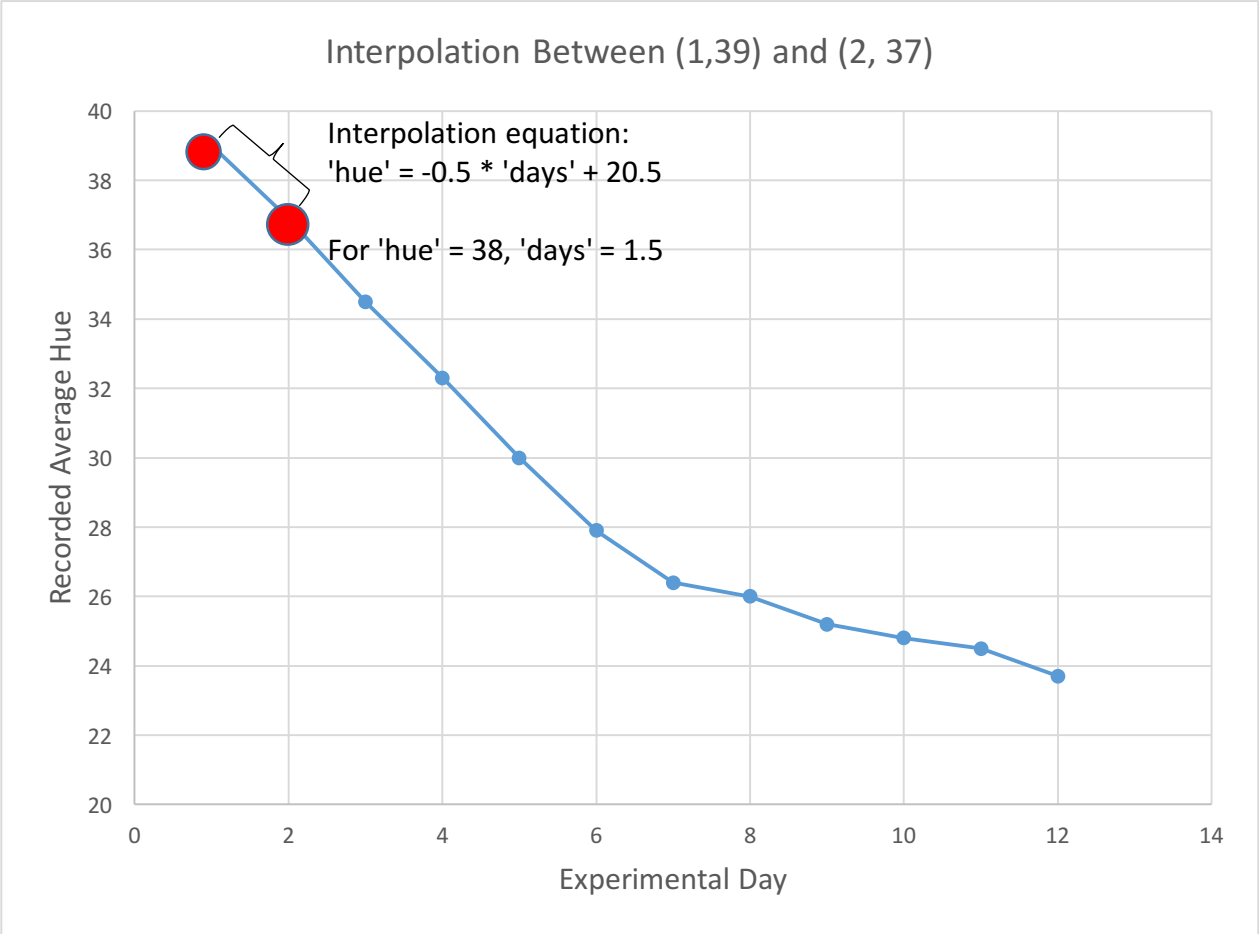


Figure 23: Example interpolation.

The hue value of interest (38) is then plugged into this equation to get an exact value for the ‘days’ x-coordinate corresponding to a 38 ‘hue’ y-coordinate. If the exact data point of interest exists in the dataset (i.e. if (1, 38) already exists without interpolating), the existing ‘days’ x-value is taken.

Finally, the 'days' value is subtracted from the number of total days that the specific fruit lasted to compute the shelf-life value at each point of interest. Because each banana has its own unique shelf-life that is used in the calculation, the last subtraction step also standardizes all of the data, which allows for comparison between bananas. A visual representation of this shelf-life calculation is shown in Figure 24. This process (interpolating between the two nearest points to calculate the 'days' value for the point of interest and then subtracting that value from the total days lasted) is repeated for each data point collected from each fruit.

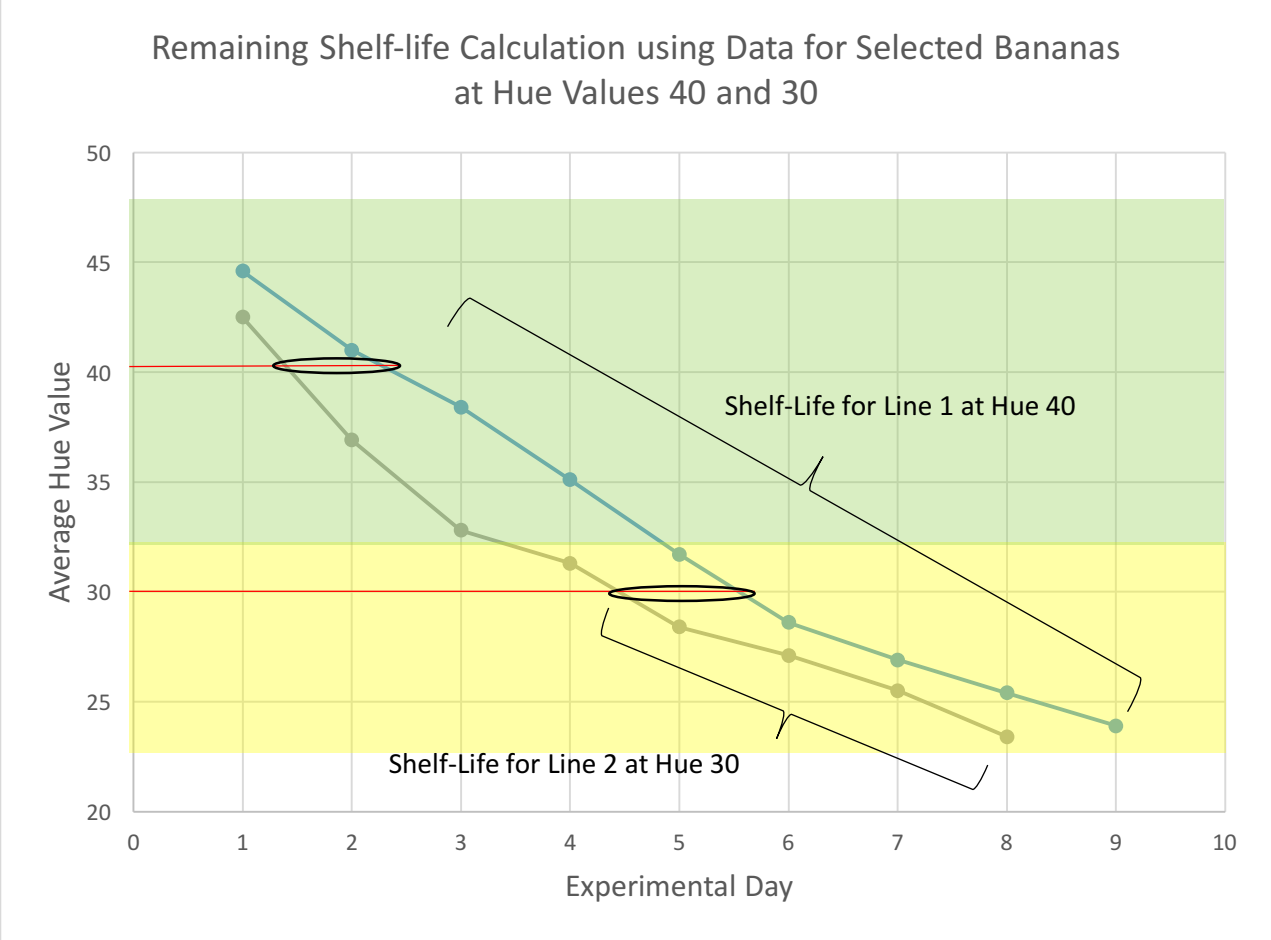


Figure 24: Example interpolation of hue values 40 and 30.

Some of the data points that must be calculated lie outside of the observed range. For example, if the largest hue value collected was 38, the predicted shelf-life value for 40 must be extrapolated (there are no two nearest points greater than and less than the point of interest). This is done by finding the second-order polynomial that best fits the collected data and then extrapolating the point out of the data range. For example, if the best-fit polynomial of a fruit in the hue range of 30 to 40 was given by the polynomial 'days' = 0.001 * 'hue'^{2} – 0.5 * 'hue' + 20 and the hue value of interest of 40 was out of range, 40 is entered into the polynomial to calculate the value for 'days'. This predicted 'days' value is then subtracted from the total number of days that the fruit lasted. The reason that a second-order polynomial is fit is because it provides the highest R-Squared value without over-fitting, based on the number of data points found for each banana.

A second-order polynomial was used when there were at least five data points in the hue range of interest (30 to 40). However, in one instance, there was only three data points in the hue range of interest, so a linear model was used to describe the data. This was done to prevent over-fitting and thereby losing prediction accuracy. All of the best-fit lines for extrapolation have R-squared values over 99.5%.

In order to ensure that extrapolation yields reasonable results, the slopes of all of the bananas that do have discrete hue values larger than 40 (that is, the bananas that did not need extrapolation) are calculated between the hue points of 36 and 40 (the range of extrapolation for some of the bananas). The average and range of these slopes is found. This is done in order to ensure that none of the extrapolation points are unreasonably out of the range of the empirical data. The average slope (for interpolated banana's data points between hue values of 36 and 40) is 0.3207, with a maximum of 0.5936 and a

minimum of 0.1933. None of the bananas extrapolation slopes fell outside of this range, so the extrapolations are reasonable. That is, the slope of the extrapolation lines fell in the range of the existing interpolation slopes.

After all of the interpolation and extrapolation, every sample banana has a discrete shelf-life value at every point in the range {30, 32, 34, 36, 38, 40}. The results of plotting the newly-calculated shelf-life values for every sample banana is shown in Figure 25. Please note that the X and Y axes are flipped as the data changes from empirically observed 'days' to calculated 'shelf-life'.

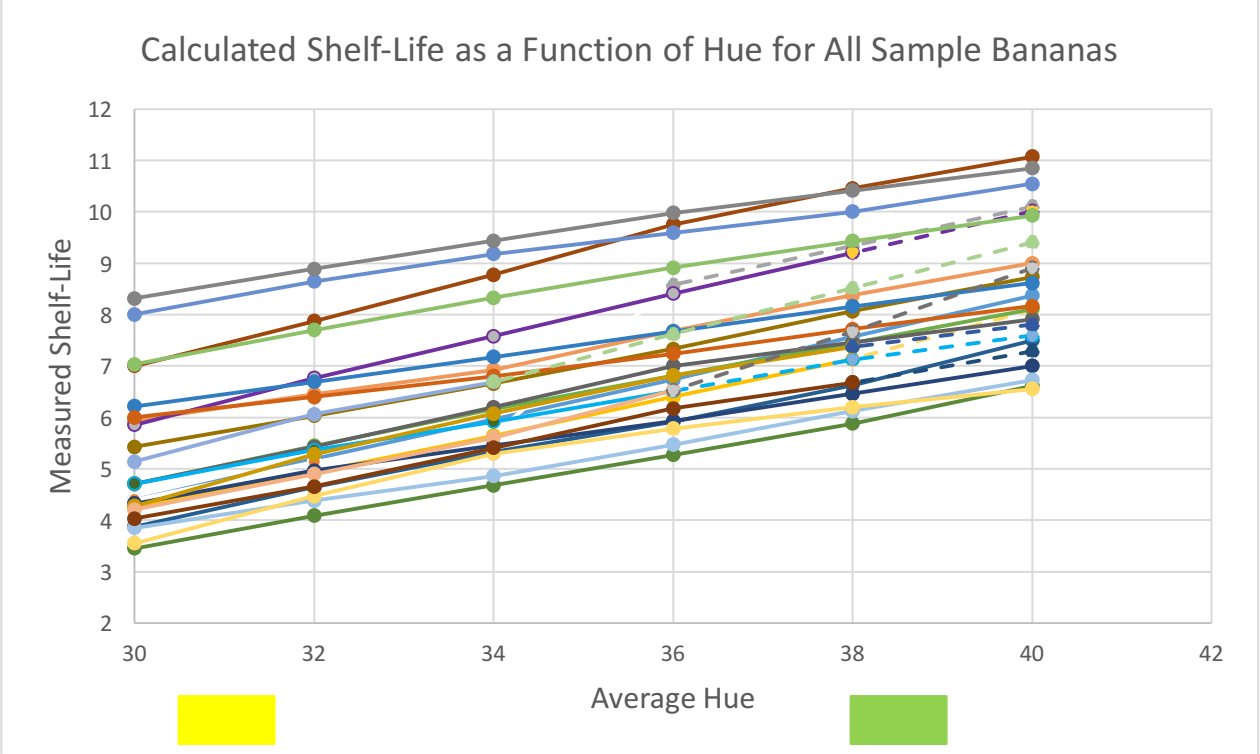


Figure 25: Data calculation hue shelf-life calculation for all bananas.

The same type of analysis was done on the a* data. The a* range of interest is {-10, -8, -6, -4, -2, 0}. The nearest neighbors above and below the a* of interest were found and interpolated between to find the exact 'days passed' value. This value is then subtracted from the total days the individual banana lasted to finally calculate the shelf-life. Similar to the hue analysis above, a third-order polynomial was fit to the data to extrapolate a* values outside of the empirically recorded data. The R-squared value of every best-fit polynomial was over 99.5%. The results of the calculation of remaining shelf-life using the data for a* is shown in Figure 26.

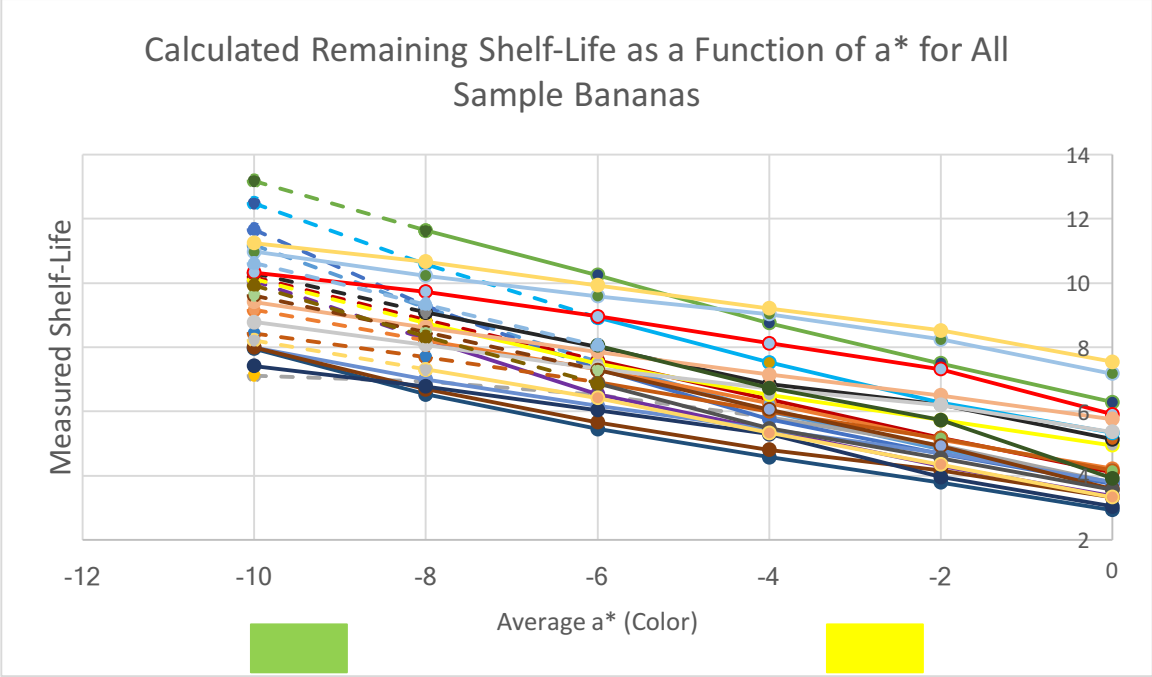


Figure 26: Data calculation a* shelf-life calculation for all bananas.

After all of the shelf-life data has been calculated for each sample, using all of the data, descriptive statistics need to be calculated in order to create the final shelf-life model. First, the values at each data point of interest {40, 38, 36, 34, 32, 30} are

averaged. Then, the standard deviation of each data point of interest is calculated. At this point, there is a corresponding average shelf-life value (and standard deviation) to every hue value in the range of interest (Table 1).

The hue values and average shelf-life days are plotted and then run through Minitab's regression function to find the best line that describes the relationship between hue value and shelf-life. The averages are plotted with the standard deviation error bars to ultimately produce the final model of shelf-life as a function of color.

The same analysis is then performed on the a* data. Every data point in the a* range of interest {-10, -8, -6, -4, -2, 0} has an average shelf-life value and its corresponding standard deviation. This is shown in Table 2. Finally, the best fit line that describes the relationship between a* value and shelf-life is calculated.

6.5. Testing for Significance:

The models created above can be validated against the existing data. Each data point can be treated, in isolation, as a validation point because there is an empirically found shelf-life value for each point. That is, by isolating each data point of each fruit, using the model to predict shelf-life, and then comparing the model's predicted result to the empirically found shelf-life value, the data can be validated upon itself. First, each data point is entered into the model and the shelf-life for each individual data point is calculated. Then, the predicted shelf-life is subtracted from the actual shelf-life. The absolute value of the difference is calculated. Because the shelf-life value has a resolution of one day, if the absolute difference is less than one day, that data point is correctly predicted. If the absolute difference is larger than one day, that data point is incorrectly predicted.

Also, the mean absolute deviation (MAD), is calculated to evaluate each model's prediction power. Similar to the first validation step, each data point is entered into the model and the predicted shelf-life value is calculated. Then, the difference between the predicted value and the actual value is calculated. The absolute value of the difference is calculated. Finally, the mean of all of the calculated absolute deviations of each data point is found. In theory, the better the predictive power of the model, the smaller the MAD.

In order to measure the accuracy of the interpolation and extrapolation technique, regression (using Minitab) must be performed to allow for comparison. First, in order to build the best model, best subsets regression is performed using the software Minitab.

Initially, every color feature is used in the regression. That is, hue average, hue minimum, hue maximum, percent green, percent yellow, percent brown, a* average, a* minimum, and a* maximum are used as the independent variables and shelf-life is used as the dependent variable. However, the correlation factor (the Variance Inflation Factor (VIF)) was shown to be too high: over 200 when the literature values for the accepted maximum VIF is between 5 and 10. Even after removing the highly correlated hue average and a* average features, the models still have VIFs that are too large (close to or above 10). Thus, instead of using multiple regression, linear regression is performed twice. In the first linear regression, hue average is used as the independent variable in the linear regression with shelf-life being the dependent variable. Similarly, in the second linear regression, a* average is used as the independent variable and shelf-life is the dependent variable. This also makes sense because regression is used as a comparison tool for the interpolation and extrapolation technique, which only use hue average and a*, respectively to predict shelf-life.

Note that no interaction terms are included in any of the regression models. This is because adding interaction terms did not increase the adjusted R-Squared of any of the models. Also, keeping in mind that regression is used as a point of comparison to the interpolation and extrapolation technique, because the interpolation and extrapolation technique does not take into account interaction effects, the regression technique does not consider interaction either.

7. MODEL RESULTS AND DISCUSSION

7.1. Final Models:

This section of the thesis presents the final calculated models of shelf-life, discusses regression models as a point of comparison, and introduces a validation experiment in order to test the models.

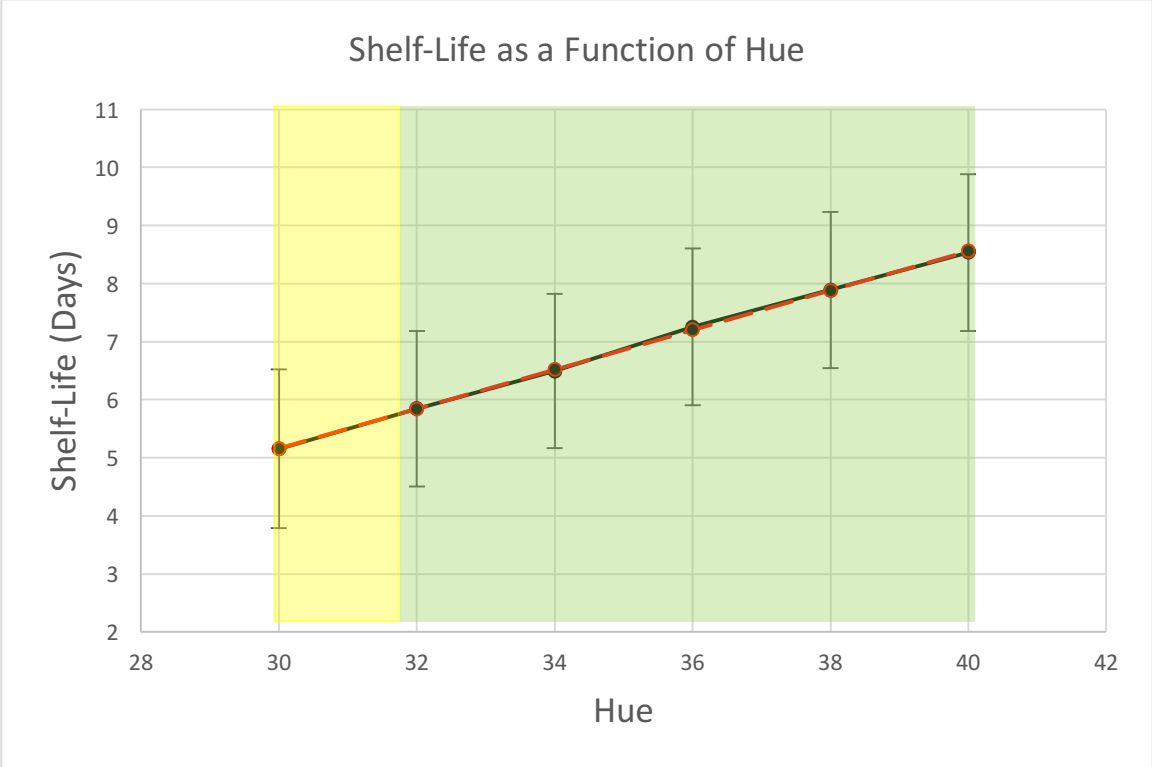


Figure 27: Shelf-life model as a function of average hue.

The ultimate goal of the thesis is to present a process to predict shelf-life of fruit based on distinct features extracted using computer vision. Ultimately, the desired end-

product to show proof of concept of the process is a model that predicts shelf-life based on color features. After data collection, processing, remaining shelf-life calculation, and modeling, a function of shelf-life based on hue is created. It is described by the equation: 'Shelf-life' = 0.3401 * 'Hue' - 5.04 (shown as the dotted red line in Figure 27). The average shelf-life (and corresponding standard deviation) for every data point in the hue range of interest is shown in Table 1. So, if in an agricultural factory setting, a banana comes off an inspection line with a hue value of 37, the predicted shelf-life is 0.3401 * 37 - 5.04, which evaluates to 7.54 days of shelf-life. If a banana much closer to ripeness is measured to have a hue of 30, the shelf life is 0.3401 * 30 - 5.04, which is 5.16 days. These different predictions of shelf-life can then be used to inform supply chain distribution decisions. The predictions can also be used as an alternative to traditional, subjective fruit quality grading systems (the banana with the hue value of 30 can be shipped locally instead of being discarded for being "too ripe" in a traditional fruit quality grading system).

Table 1: Average shelf-life values for Hue.

Hue	Average Shelf-Life (days)	Standard Deviation
40	8.536	1.347
38	7.890	1.342
36	7.253	1.350
34	6.491	1.328
32	5.848	1.339
30	5.153	1.367

The model for shelf-life as a function of a* is given by 'Shelf-life' = 4.455 – 0.5213 * 'a*' (shown as the dotted red line in Figure 28). The results can be interpreted similarly as in the model for hue. The average shelf-life (and corresponding standard deviation) for each a* hue point of interest is shown in Table 2.

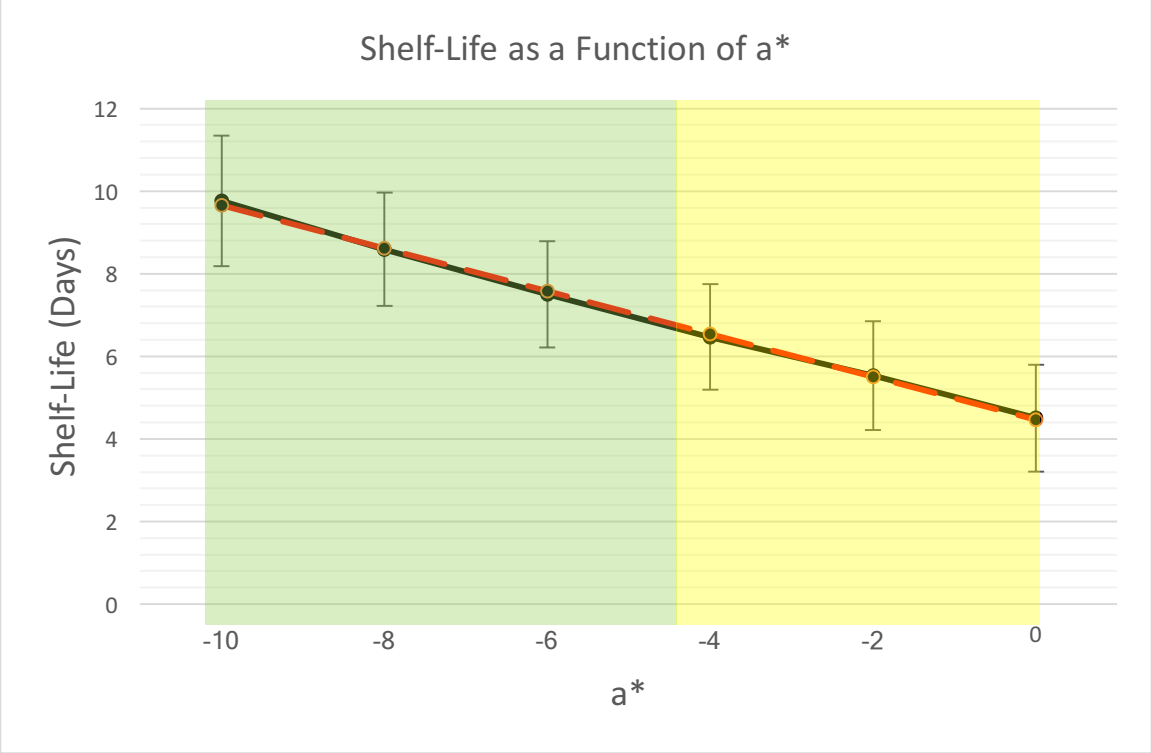


Figure 28: Shelf-life model as a function of average a*.

According to the standard hue color wheel (as shown above in Figure 18), a hue value between 40 and 32 corresponds to the color green, whereas a hue value between 32 and 30 corresponds to the color yellow. This color representation is indicated on the graph of the model. Depending on the expected application, it may be possible to repeat the process that was undergone in this thesis to find shelf-life predictions of hue values

outside of the range of 30 to 40. So, with respect to being able to model shelf-life using color features, the thesis is successful.

Table 2: Average shelf-life values for a*.

a* Value	Average Shelf-life (days)	Standard Deviation
-10	9.763	1.583
-8	8.592	1.364
-6	7.500	1.286
-4	6.467	1.278
-2	5.536	1.320
0	4.505	1.292

7.2. Percent Yellow, Green, and Brown

The percentage of green, percentage of yellow, and percentage of brown are calculated each day of every tested banana's lifetime and then plotted. The primary color of the banana can be determined as the color with the highest relative percentage in the given day. The shape of each color's line makes sense in terms of how a banana ripens: the banana starts green, then turns yellow, then turns from yellow to brown. The graph above is an average of the color percentages for every banana tested. The difference among bananas is the amount of time each color spends at a certain percentage (for example, some bananas had 100% green for one day, while some had 100% green for four days) and the relative shape of the 'percent yellow' curve. The banana's hue enters the ripe range (hue value of 25 to 23), when the percentage of brown reaches about 70% and the percentage of yellow comes down to about 50%. This can be seen in Figure 6. Usually, this is one day after the percent yellow and percent brown lines intersect. So, while hue was used in this thesis to determine the end of life, it may be valid to use factors such as percent yellow, percent brown, and percent green. Hue and a* were used in this thesis because they were the most significant in the regression analysis and for simplicity: hue and a* involves minimal post-processing (when looking realistically at models being implemented in computer vision systems, simple is better). The T-Value of a* is -8.19, the T-Value of hue is 8.19 and the T-Value of percent brown is -7.71. So, a* and hue are the more significant factors. Thus, no model was built using only the percent green, percent yellow, and percent brown data.

7.3. Model Accuracy:

Using the models created above and the validation process outlined in the "Analysis" section of this thesis, results for model validation were computed. The hue model correctly characterized (with a one-day measure of error) 64.89% of the bananas in the experiment. The correct prediction of 64.89% of bananas is an improvement on the current banana grading system which is estimated to lose 40% of produce (Geiling, 2015 and Gunders 2012). For validation of the a* model, the a* model correctly predicted (within one day) 64.84% of sample fruit. However, there are many reasons that the models do not correctly characterize every fruit. Firstly, every banana ripens at its own rate and the model is a representation of the average shelf-life of bananas. Each banana had different starting features: hue value, size, shape, curvature, etc. Of the features, hue value had the biggest effect on shelf-life but the other features had some effect as well. So, by discounting the other features, accuracy of prediction is lost. However, accuracy must be lost in order to create a simple, reasonable model that could be implemented in the industry (realistically, a model that only takes into account one factor). Finally, a fruit becomes ripe by producing the chemical ethylene, and the rate of ethylene production is determined by factors undetectable by computer vision. The MAD of the hue model is 0.9645 and the MAD of the a* model is 0.9819, both of which can be considered successful because the resolution of the shelf-life response is one day.

7.4. Regression and Comparison of Models:

Another method of building a model (without the calculation of remaining shelf-life using data, as above) is to use regression to find the P-Values of the coefficients to determine the most significant features. This analysis is done by loading the raw data into the software Minitab and using the inherent regression function. This is done in order to compare the accuracy of the interpolation and extrapolation model-building technique to the more traditional regression model-building technique. The results of using the data to calculate the remaining shelf-life to build the model and regressing the raw data to build the model are compared at the end of this section. As demonstrated in the "Analysis" section above, because of autocorrelation of features and because regression is used as a comparison tool for the interpolation and extrapolation technique, linear regression is used instead of multiple regression. That is, two linear regressions are performed: the first uses hue as the predictor and shelf-life as the response. The second linear regression uses a* as the predictor and shelf-life as the response.

Best Subsets Minitab output:

In Appendix B, we can see the results of the best subsets regression, using Minitab.

Multiple Regression Minitab output:

The results of one of the multiple regressions is shown in Appendix C. The VIFs are well over the maximum accepted value of 10, so hue average is removed.

Multiple Regression without Hue Average Minitab output:

The results of the multiple regression without Hue Average are shown Appendix D. Even after removing the highly correlated Hue Average feature, autocorrelation is too large.

Linear Regressions Minitab output:

The results of the linear regressions for average hue and average a* are shown in Appendix E and F (respectively). The adjusted R-Squared values for the linear regression models are shown in Table 3. The table demonstrates that the linear regression model of a* average ends up as the best (largest adjusted R-Squared value) model.

Table 3: Regression results.

Model	Adjusted R-Squared
Linear Regression: Hue	77.35%
Linear Regression: a*	80.66%

For the regression model of only hue ('Shelf-Life' = -8.984 + 0.448 * 'hue'), 52.21% of data is correctly predicted within one day and the MAD is 1.340. For the regression model of only a* ('Shelf-Life' = 3.838 – 0.624 * 'a*'), 55.45% of data is predicted correctly within one day and the MAD is 1.529. Comparing these results to the

model built by calculating remaining shelf-life based on the raw data, the hue model using data calculations is slightly better (64.89%) and the a* model (64.84%) is slightly better as well at characterizing existing data. Further, the hue model has a slightly smaller MAD (0.965) and the a* model has a slightly smaller MAD (0.9819). Ideally, we want the MAD to be as close to zero as possible but any value less than one should be considered successful (because the resolution of the shelf-life response is one day). The results of all of the models can be seen in Table 3. A 2-sample proportion test performed in Minitab with $H_0 = P_1 = P_2$ and $H_A = P_1 \neq P_2$ to determine whether the difference between the data calculation method of a* average and the linear regression model of a* average is statistically significant. The null hypothesis could be rejected.

The reason that the shelf-life calculation using raw data method outlined above is preferred to regression lies in the assumptions of regression. Among the assumptions in linear regression, the assumption that there is a linear relationship between the color features and shelf-life is the most likely assumption to be violated in the regression analysis for this thesis. In terms of the linearity assumption, because there are thirty-six bananas, a linear model for each fruit resulted in differing R-squared values. For three of the thirty-six total bananas, the R-Squared value of a linear best-fit line was 1. For the rest (thirty-three) of the bananas, the R-Squared values for a linear best-fit line ranged from 0.85 to 0.99. This is the reason why extrapolation in the data calculation section used third-order polynomials instead of linear best-fit lines. It also explains why the regression model was the worst of the validated models. Another reason that the data calculation method is preferred is because most of the error of the regression model came early in the banana's life-times (as shown in Figure 29). The data calculation model has

error spread evenly throughout the banana's life-times (as shown in Figure 30). When considering distribution decisions, early life data is more important (as outlined in the "Data Calculation of Remaining Shelf-life" section of this thesis), so the data calculation method is preferred. It is also important to keep in mind that HSI and L*a*b* may be correlated because they come from the same RGB coordinates. For all of the above reasons, the data calculation method is superior.

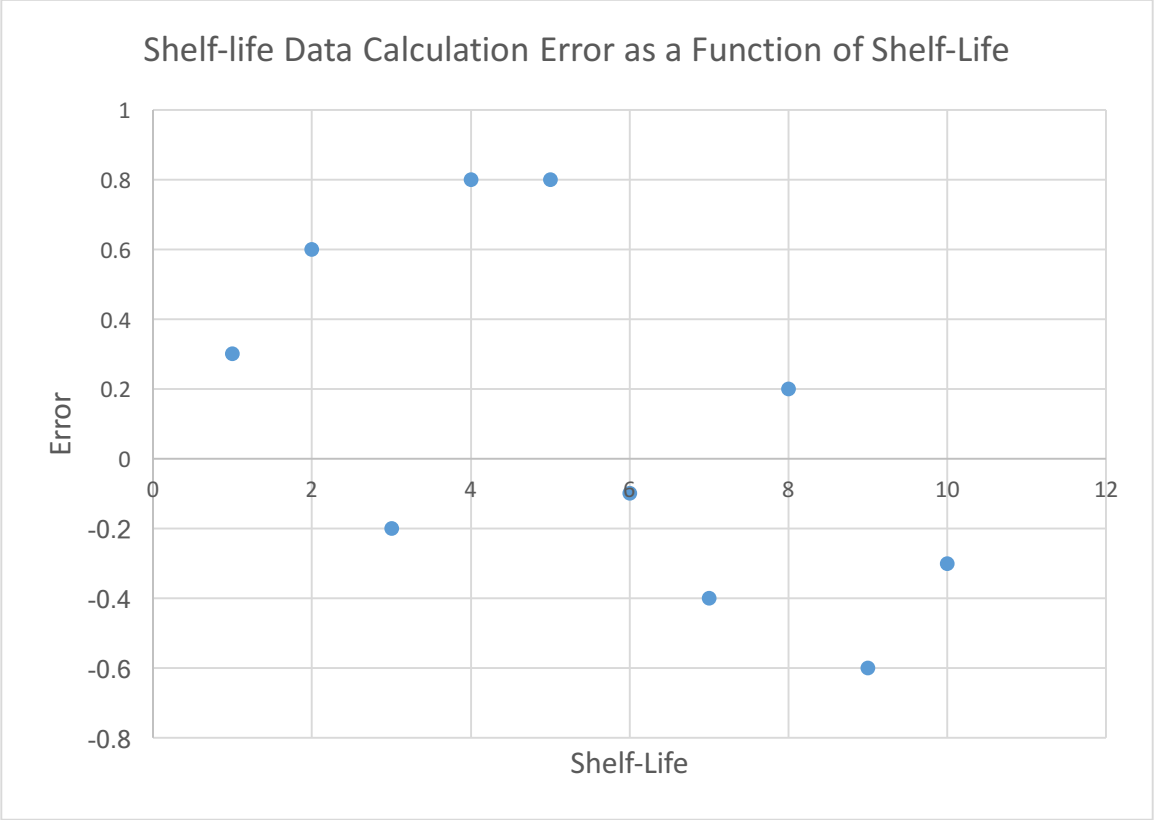


Figure 29: Random error of data calculation model.

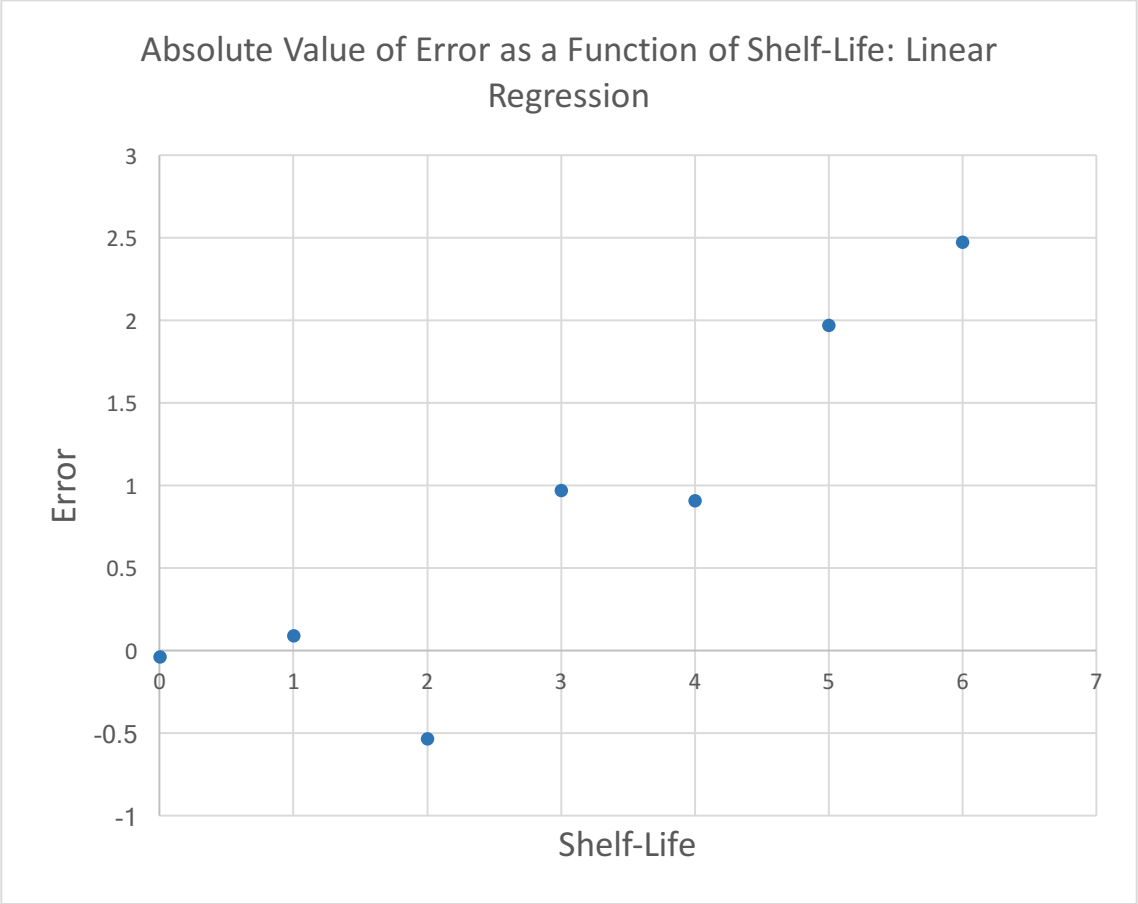


Figure 30: Increasing error of regression model.

Table 4: Summary accuracy data for every model.

Model	Correct Prediction	MAD
Data Calculation: Hue	64.89%	0.965
Data Calculation: a*	64.84%	0.982
Linear Regression: Hue	52.21%	1.34
Linear Regression: a*	55.45%	1.529

7.5. Model Validation:

In order to formally validate the models, a final experiment is undertaken. The conditions of the first two experiments are replicated (lighting, camera distance, temperature, background, camera, and time of measurement) for the validation experiment. 12 of the same brand of bananas are purchased. The color values (average hue, average a*, and percent brown) of each banana are extracted (in the same process outlined in the Analysis section) and the values are entered into each respective model. The models validated are the three models that have the best accuracy (Hue Data Calculation, a* Data Calculation, and Linear Regression: a*). The models produce exact shelf-life values so the predictions are rounded to the nearest integers larger than the shelf-life prediction and smaller than the shelf-life prediction. For example, if the model outputs a shelf-life of 4.612 days, the banana's color is measured on the fourth day and on the fifth day. If the banana has not fallen into the "ripe" hue range by the fifth day, the measurements continue each subsequent day until the banana's hue is determined to be "ripe". The day that the banana's measured average hue falls into the "ripe" hue range (less than a hue value of 25) is the day that the experiment ends for that banana.

Because the different models produce different estimates, the banana's hue is measured at least twice for each model. For example, if a banana's shelf-life is estimated by the three models to be 4.612 days, 4.813 days, and 5.441 days, the banana will be measured on days four, five, and six (unless the hue falls below 25 before the sixth day). Alternatively, if the model is very close to an integer, the banana will be checked on the day before, the day on, and the day after. For example, if the shelf-life is predicted to be

4.991, the banana is tested on days four, five, and six. The model correctly identifies the shelf-life if the observed shelf-life has less than one-day measure of error from the model prediction. The results of the validation are shown in Table 4.

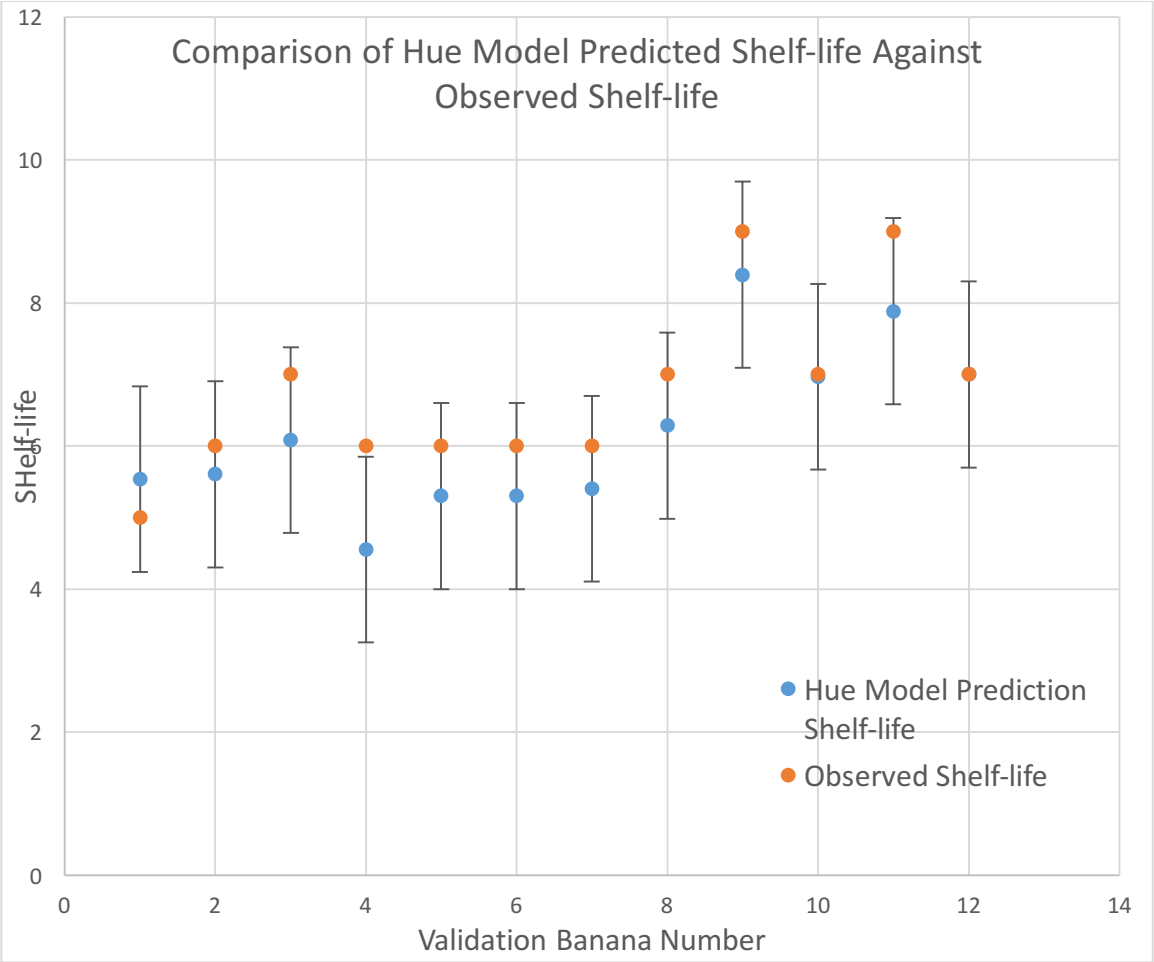


Figure 31: Validation experimental results.

Table 5: Validation results.

Model	Number of Bananas Correctly Predicted	Percent of Bananas Correctly Predicted
Data Calculation: Hue	9	75%
Data Calculation: a*	9	75%
Linear Regression: a*	7	58.3%

Then, because the hue data calculation method has performed the best thus far in the thesis, the results of the hue data calculation shelf-life prediction (with the standard deviation calculated in the "Analysis" section of this thesis) are plotted against the empirically observed shelf-life. This is shown in Figure 31. It is important to note that all but one of the empirical shelf-life observations fall within the standard deviation of the prediction. It is also important to note that almost all of the predictions are less than the empirically observed shelf-life values. It is not possible to know exactly what caused the discrepancy but it is possible that the difference lies in the fact that the validation bananas were measured (touched and moved) fewer times than the original experimental bananas. Another possible explanation could be a change in farming conditions or other difference at the farming and harvesting process (although the same brand of banana was bought at the same store).

As a final validation step, the sugar content of ten of the twelve banana was tested (using the method outlined in Experimental Procedure 1) on the first day that the hue value dropped below 25. All of the bananas that were tested during the validation step had a sugar content above 23%.

This section of the thesis has presented the shelf-life models, regression, and a validation experiment. A summary of the thesis and suggestions of future work follows.

8. SUMMARY AND FUTURE WORK

8.1. Summary:

Ultimately, this thesis has demonstrated proof of concept of a process to use computer vision to build a predictive model of shelf-life of fruit. Computer vision is becoming widespread in many industries and is, therefore, a burgeoning field of research. It is most widely used on inspection lines of electronic components. Computer vision holds many advantages to its human counterpart: it is faster, more reliable, cheaper in the long-term, and more accurate. Naturally, computer vision is starting to become implemented in other industries. One of the most important fields that computer vision is starting to move into is the agriculture industry. Computer vision excels at doing a well-defined process many times. This lends well to industries such as electronics production because every piece of one electronic component must look the same. This means that a computer can rapidly capture an image of an electronic component coming onto an inspection line, compare certain regions to an ideal (correctly produced) image of the component, and then make a decision to sell or discard the component. In the agriculture industry, every piece of produce is unique, so the process described above for electronic components is no longer valid.

Currently, most companies in the agriculture industry either do not use computer vision or use it in a binary grading system. That is, computer vision is used to extract features and then make a decision to either sell or discard the produce. However, this process leads to excessive food waste and an overall inefficient process. Another application of computer vision (specifically for bananas) is to receive all of the fruit

before they ripen and then dose them with ethylene before being shipped to retail locations. However, this process does not account for variation of beginning characteristics between fruits, it costs money to store and wait for the fruit, and frequently leads to fruit and profit loss. Fundamentally, the existing processes that do use computer vision base their distribution decisions on current characteristics. The process proposed in this thesis uses current characteristics to predict future characteristics, which leads to more informed distribution decisions. By modeling future characteristics, the process proposed will allow fruit characterized as unfit to sell by existing processes to still be utilized (i.e. if the fruit is too ripe to ship across the country, it can still be sold locally) which decreases food waste and increases profit. Further, by changing shipping locations instead of date shipped, money is saved in storage costs and holding excess inventory.

The process put forth by this thesis involves a distinct methodology that can be replicated for a large variety of fruit. First, the fruit of interest must be selected. This methodology is valid for fruit that fulfills two criteria: (1) the fruit must continue to ripen after it has been harvested (a characteristic of climacteric fruit) and (2) the fruit must exhibit a distinct change in color as it ripens. After the fruit has been selected, two experiments must be undertaken. The first experiment measures sugar content through the lifetime of a group of the fruit of interest. The purpose of this experiment is to determine the characteristic color of the fruit when it reaches a certain sugar content (the sugar content that corresponds to "ripeness" for that fruit). Essentially, this experiment is used to determine the end-point for the second experiment. In the second experiment, a large set of fruit is observed over each fruit's respective life-time. The end-point of the

fruit's life is determined by the color corresponding to "ripe" found in the first experiment. The second experiment is large-scale data collection.

After the data has been collected, computer vision techniques are used to extract features. Then, computer vision algorithms are applied to convert color representations into useful data. Data late in the fruits' lives are cut out of analysis and the remaining shelf-life data is calculated. Finally, a model is built to predict the shelf-life of fruit based on color characteristics. The model is then validated. While the thesis presents the process to undergo, the proof of concept of the process is done with bananas.

8.2. Future Work

This thesis lays the groundwork for a myriad of further experiments as well as a multitude of alternations of the process. Naturally, the process may be able to be applied to any fruit that adheres to the two criteria enumerated above. Future work can possibly apply the process to many fruits. Among the best candidates for future fruit are mangoes, apples, avocados, blackberries, melons, and tomatoes. This is because they are climacteric and have the most distinct color changes in their life-time of ripeness.

While this thesis has focused on the distribution aspect of fruit decision making, there is possibly an application of the thesis process earlier in the fruit's lifetime. On the farm, the process presented in this thesis can be used to determine time to ripeness, time to harvest, etc. (perhaps with the help of virtual reality, the thesis process can be integrated with current virtual reality glasses). Another related application of the process proposed in this thesis would be in the growing area of robot farming and harvesting. Instead of measuring shelf-life, the robot harvesters could measure days until harvest using a similar methodology but different inputs.

The process proposed in this thesis also has the potential to use machine learning techniques, which is the topic of many current papers in the field. For classification of "unripe", "ripe", and "spoiled", instead of using sugar content measurements, machine learning techniques such as k-means clustering and neural networks can be used. Further, in the model building step of the process, neural networks can be used for regression to build the model. Other machine learning techniques may be used to mine the big data that this thesis process produces, including identifying trends and model building.

In terms of how the process itself can be changed, the possibilities are almost endless. Different features (other than color), may be used to build the model. These features can be size, shape, firmness, etc. Also, different representations of color (CIE XYZ, HSL, etc.) may be used. Different measures for "ripeness" may be used – ethylene content, firmness, smell, etc. The analysis step may also be altered to produce different models, depending on what the object of experimentation is. Different validation steps may be taken.

The model may also be able to be useful in consumer applications. The development of a mobile application that incorporates data from the thesis process can be very valuable. The idea would be to capture an image of a fruit and run it through the model (previously created by the process developed in this thesis) in order to produce a shelf-life estimate. This would allow consumers to plan their fruit consumption while shopping and would have the overall benefit of reducing food waste. There is another application of the information from this thesis to the field of dynamic pricing. Similar to how the price of hotel rooms or airplane tickets increases as space decreases, a model of the price of the fruit as the color changes may be created. As the fruit ripens, its hue value decreases, so the price decreases by a certain amount for each hue value decrease.

Stepping away from the realm of agriculture, this process can theoretically be applied to anything that breaks. Treating the day that the object breaks as the "shelf-life", the process could be applied to other industries. For example, there is a possible application to predicting meantime to failure of certain electronics, given that there are discernable, changing features over the electronics' life-time.

BIBLIOGRAPHY

Abdullah, M. Z., Guan, L. C., & Azemi, B. M. (2001). Stepwise discriminant analysis for colour grading of oil palm using machine vision system. *Food and bioproducts processing*, *79*(4), 223-231.

Aimonino, D., Barge, P., Comba, L., Gay, P., Occelli, A., Tortia, C., (2015). Computer vision for laboratory quality control on frozen fruit. *Chemical Engineering Transactions, 44,* 175-180

Alfatni, M. S. M., Shariff, A. R. M., Shafri, H. Z. M., Saaed, O. M. B., & Eshanta, O. M. (2008). Oil palm fruit bunch grading system using red, green and blue digital number. *Journal of Applied Sciences*, *8*(8), 1444-1452.

Blasco, J., Aleixos, N., & Moltó, E. (2003). Machine vision system for automatic quality grading of fruit. *Biosystems Engineering*, *85*(4), 415-423.

Blasco, J., Aleixos, N., & Moltó, E. (2007). Computer vision detection of peel defects in citrus by means of a region oriented segmentation algorithm. *Journal of Food Engineering*, *81*(3), 535-543.

Blasco, J., Cubero, S., Gómez-Sanchís, J., Mira, P., & Moltó, E. (2009). Development of a machine for the automatic sorting of pomegranate (Punica granatum) arils based on computer vision. *Journal of Food Engineering*, *90*(1), 27-34.

Buzby, J. C., Hyman, J., Stewart, H., & Wells, H. F. (2011). The Value of Retail-and Consumer-Level Fruit and Vegetable Losses in the United States. *Journal of Consumer Affairs*, *45*(3), 492-515.

Costa, C., Antonucci, F., Pallottino, F., Aguzzi, J., Sun, D. W., & Menesatti, P. (2011). Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. *Food and Bioprocess Technology*, *4*(5), 673-692.

Cubero, S., Aleixos, N., Moltó, E., Gómez-Sanchis, J., & Blasco, J. (2011). Advances in machine vision applications for automatic inspection and quality evaluation of fruits and vegetables. *Food and Bioprocess Technology*, *4*(4), 487-504.

De, J. (2015, May 13). CIE Lab colorspace as coordinate system (A1). Retrieved from https://openclipart.org/detail/218703/cie-labs-colorspace-as-coordinate-system-a1

Diaz, R., Faus, G., Blasco, M., Blasco, J., & Moltó, E. (2000). The application of a fast algorithm for the classification of olives by machine vision. *Food Research International*, *33*(3), 305-309.

Geiling, Natasha, (2015, August 8). Selling Ugly Fruits and Vegetables Could be Key to Solving America's Food Waste Problem, Retrieved April 6, 2016, from http://thinkprogress.org/climate/2015/08/19/3692594/ugly-fruit-and-vegetable-food-waste-campaign/

Gunders, D. (2012). Wasted: How America Is Losing Up to 40 Percent of Its Food from Farm to Fork to Landfill. NRDC, 12(06), 26-26. Retrieved November 19, 2015, from https://www.nrdc.org/food/files/wasted-food-ip.pdf

Iverson, L. (2014, February 25). Healthy Food for a Healthy World: Wasted Food, Wasted Nutrients. Retrieved November 19, 2015, from http://www.thechicagocouncil.org/blog-entry/healthy-food-healthy-world-wasted-food-wasted-nutrients-4

Jewett, T. (2013). HSB: Hue, saturation, and brightness. Retrieved April 12, 2016, from http://www.tomjewett.com/colors/hsb.html

Jha, S. N., Narsaiah, K., Sharma, A. D., Singh, M., Bansal, S., & Kumar, R. (2010). Quality parameters of mango and potential of non-destructive techniques for their measurement—a review. *Journal of food science and technology*, *47*(1), 1-14.

Leemans, V., Magein, H., & Destain, M. F. (1998). Defects segmentation on 'Golden Delicious' apples by using colour machine vision. *Computers and Electronics in Agriculture*, *20*(2), 117-130.

Manninen, H., Paakki, M., Hopia, A., & Franzén, R. (2015). Measuring the green color of vegetables from digital images using image analysis. *LWT-Food Science and Technology*.

do Nascimento Nunes, M. C. (2015). Correlations between subjective quality and physicochemical attributes of fresh fruits and vegetables. *Postharvest Biology and Technology*, *107*, 43-54.

Pace, B., Cefola, M., Da Pelo, P., Renna, F., & Attolico, G. (2014). Non-destructive evaluation of quality and ammonia content in whole and fresh-cut lettuce by computer vision system. *Food Research International*, *64*, 647-655.

Pata, P. (2016, April 02). Image Processing Toolbox. Retrieved from http://radio.feld.cvut.cz/matlab/toolbox/images/color4.html

Pedreschi, F., Leon, J., Mery, D., & Moyano, P. (2006). Development of a computer vision system to measure the color of potato chips. *Food Research International*, *39*(10), 1092-1098.

Shearer, S. A., & Payne, F. A. (1990). Color and defect sorting of bell peppers using machine vision. *Transactions of the ASAE*, *33*(6), 2045-2050.

Soltani, M., Alimardani, R., & Omid, M. (2010). Prediction of banana quality during ripening stage using capacitance sensing system. *Australian Journal of Crop Science*, *4*(6), 443.

Tapre, A. R., & Jain, R. K. (2012). Study of advanced maturity stages of banana. *Int. J. Adv. Eng. Res. Stud. I (III)*, 272-274.

Vélez-Rivera, N., Blasco, J., Chanona-Pérez, J., Calderón-Domínguez, G., de Jesús Perea-Flores, M., Arzate-Vázquez, I., ... & Farrera-Rebollo, R. (2014). Computer Vision System Applied to Classification of "Manila" Mangoes During Ripening Process. *Food and bioprocess technology*, *7*(4), 1183-1194.

Vidal, A., Talens, P., Prats-Montalbán, J. M., Cubero, S., Albert, F., & Blasco, J. (2013). In-line estimation of the standard colour index of citrus fruits using a computer vision system developed for a mobile platform. *Food and Bioprocess Technology*, *6*(12), 3412-3419.

Work With Color. (n.d.). Retrieved April 14, 2016, from http://www.workwithcolor.com/cona-hue-ranges-map-02.png

Zhang, B., Huang, W., Li, J., Zhao, C., Fan, S., Wu, J., & Liu, C. (2014). Principles, developments and applications of computer vision for external quality inspection of fruits and vegetables: A review. *Food Research International*, *62*, 326-343.

Zhou, T., Harrison, A. D., McKellar, R., Young, J. C., Odumeru, J., Piyasena, P., ... & Karr, S. (2004). Determination of acceptability and shelf life of ready-to-use lettuce by digital image analysis. *Food research international*, *37*(9), 875-881.

APPENDIX

(a)

Area	Author (Main)	Year	Significance
Advantages of Computer Vision	Cubero	2011	Machines can see spectrums of visions that humans cannot.
	Zhou	2004	Humans cannot give a detailed level of analysis for quality because they only assign a number.
Importance of Appearance	Zhang	2014	Appearance is important because it determines market value and it can indicate the internal quality.
	Abdullah	2001	The perception of the appearance of the fruit can affect taste.
	Nunes	2015	Numerous computer vision features have been validated by measuring the physiological changes of fruit.
Equipment	Vidal	2013	Showed the hardware requirements for computer vision.
	Cubero	2011	Demonstrated the importance of lighting for repeatable results.
	Zhang	2014	Emphasized the difference between front lighting and back lighting.
	Leemans	1998	Unique way of controlling lighting.
	Cubero	2011	Showed the different kinds of image acquisition devices and the necessary components of a computer vision system.
Color Analysis	Zhang	2014	Reiterated the importance of color for computer vision.
	Cubero	2011	Showed the difference between RGB and HSI color analysis.
	Vidal	2013	Put forth the reasons that HSI is preferred over RGB.
	Blasco	2003	Used RGB color analysis with a Bayesian discrimination model.
	Shearer	1990	Implemented a unique way to classify pixels using RGB analysis.
	Jha	2010	Distinguished the importance of primary and secondary colors.
	Pace	2014	Implemented the idea of a color ratio and showed that color can be an accurate indicator of quality.
	Alfatni	2008	Did color analysis by using RGB color intensity.

	Abdullah	2001	Showed mathematically why HSI is superior to RGB.
	Blasco	2009	Showed that RGB color analysis can be done very quickly.
	Diaz	2000	Computer vision systems can be significantly better than human panels.
Size/Shape Estimation	Costa	2011	Introduced the idea of comparing parameters to an "ideal" fruit.
	Leemans	1998	Introduced the idea of segmentation.
	Blasco	2003, 2007, & 2009	Developed the idea of pixel-oriented and region-oriented approaches to estimate size.
	Zhang	2014	Introduced a complete model for image processing.
Comparing Computer Vision Results to Physiological Results	Pace	2014	Showed that computer vision results were valid by measuring ammonia content.
	Aimonino	2015	Validated computer vision results by measuring antioxidant content.
	Zhou	2004	Demonstrated that color was an accurate indicator of quality.
	Manninen	2015	Demonstrated that computer vision systems are more accurate than humans and used chlorophyll content to validate results.
	Jha	2010	Showed the physiological changes that happen when fruit ripens.
	Zhang	2004	Found that most physiological changes mirror computer vision results that changes happen within 4-6 days.
	Velez-Rivera	2014	Showed that there was a correlation between computer vision classification and physiochemical changes
Economics of Food Waste	Buzby	2011	Estimated the total cost of fruit waste.
	Gunders	2012	Identified different sources of food waste.
	Iverson	2015	Quantified how much food and fruit is lost per year.

(b)

Best Subsets Regression: Shelf-Life versus Hue Avg, Hue Min, ...

Response is Shelf-Life

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	Hue Avg	Hue Min	Hue Max	%B	a	a Min	a Max
1	80.7	80.7	358.5	80.4	105.2	1.2590					X		
1	77.6	77.5	415.5	77.3	158.0	1.3572							X
2	85.5	85.3	273.6	85.0	27.8	1.0970				X	X		
2	84.7	84.6	287.8	84.3	40.3	1.1246	X			X			
3	86.3	86.1	261.2	85.7	16.1	1.0684				X	X	X	
3	86.1	86.0	264.7	85.5	18.0	1.0728		X		X	X		
4	86.7	86.4	255.7	86.0	11.1	1.0545		X		X	X	X	
4	86.7	86.4	257.0	85.9	11.5	1.0556	X			X	X	X	
5	87.0	86.7	251.5	86.2	6.9	1.0424	X	X		X	X	X	
5	86.9	86.6	253.6	86.1	8.6	1.0463	X			X	X	X	X
6	87.3	86.9	249.0	86.4	4.8	1.0349	X	X		X	X	X	X
6	87.1	86.7	253.7	86.1	8.7	1.0442	X	X	X	X	X	X	
7	87.3	86.9	251.0	86.3	6.4	1.0363	X	X	X		X	X	X
7	87.3	86.9	251.1	86.3	6.5	1.0365	X	X	X	X	X	X	X
8	87.3	86.9	253.2	86.2	8.2	1.0382	X	X	X	X	X	X	X
8	87.3	86.8	254.3	86.1	8.2	1.0383	X	X	X	X	X	X	X
9	87.3	86.8	256.5	86.0	10.0	1.0402	X	X	X	X	X	X	X

(c)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.04237	87.04%	86.74%	86.24%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	17.09	4.25	4.02	0.000	
Hue Avg	-0.364	0.147	-2.48	0.014	142.85
Hue Max	-0.0296	0.0115	-2.57	0.011	2.20
%B	-0.02464	0.00370	-6.66	0.000	3.12
a	-1.169	0.225	-5.20	0.000	176.42
a Min	0.1513	0.0451	3.35	0.001	8.37

Regression Equation

Shelf-Life = 17.09 - 0.364 Hue Avg - 0.0296 Hue Max - 0.02464 %B - 1.169 a + 0.1513 a Min

(d)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.05450	86.68%	86.43%	86.01%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6.645	0.480	13.86	0.000	
Hue Max	-0.0306	0.0117	-2.62	0.009	2.19
%B	-0.02899	0.00329	-8.80	0.000	2.42
a Min	0.1329	0.0450	2.95	0.004	8.14
a	-0.6270	0.0513	-12.22	0.000	8.98

Regression Equation

Shelf-Life = 6.645 - 0.0306 Hue Max - 0.02899 %B + 0.1329 a Min - 0.6270 a

(e)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.36259	77.45%	77.35%	76.99%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-8.9836	0.4884	-18.39	<0.0001	
Hue (avg)	0.44483	0.01611	27.61	<0.0001	1.00

Regression Equation

Shelf-Life = -8.9836 + 0.44483 Hue (avg)

(f)

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.25900	80.75%	80.66%	80.39%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3.83798	0.08529	45.00	<0.0001	
a*	-0.62366	0.02044	-30.52	<0.0001	1.00

Regression Equation

Shelf-Life = 3.83798 – 0.62366 a*