An Improved Regional Honey Production Model for the United States

Anthony B. Trimboli
Western Kentucky University, anthony.trimboli063@topper.wku.edu

Follow this and additional works at: http://digitalcommons.wku.edu/theses
Part of the Ecology and Evolutionary Biology Commons, and the Entomology Commons

Recommended Citation
http://digitalcommons.wku.edu/theses/1961

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact topscholar@wku.edu.
AN IMPROVED REGIONAL HONEY PRODUCTION MODEL
FOR THE UNITED STATES

A Thesis
Presented to
The Faculty of the Department of Agriculture
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Science in Agriculture

By
Anthony B. Trimboli

May 2017
AN IMPROVED REGIONAL HONEY PRODUCTION MODEL
FOR THE UNITED STATES

Date Recommended April 20, 2017

Stephen King
Stephen King, Director of Thesis

Dominique Gumirakiza

Linda Brown

Dean, Graduate School Date
Acknowledgements

This research could not have been completed without the assistance and support of many people. I would first like to thank my committee members, Drs. Stephen King, Dominique Gumirakiza, and Linda Brown for their help and guidance throughout this process. They each provided expertise and encouragement that was vital to this project. I would also like to thank Dr. Elmer Gray for encouraging me to expand my work done on honey bee statistics as a class project. Without his gentle, but persistent, suggestion I doubt this thesis would have become a reality.

I would like to thank the beekeepers of Kentucky for reigniting my interest in beekeeping. Many of them, especially Jon Benham, encouraged my research and took a true interest in it. May all of your hives flourish and produce hygienic, ankle biting, Varroa destroying bees!

I especially want to thank my wife Shannon for encouraging me to finish my graduate work, even when I was ready to throw in the towel. You have always been there for me through the good and the bad. I know I can be tough to deal with at times, can get sidetracked in an instant, and have a doctorate in procrastination, but you always find a way to encourage me to finish the task at hand. Your encouragement and support got me through the rough spots and kept me on track. Without you, I would have never gotten through this. Thank you.
Table of Contents

Acknowledgements............................................................................................................ iii
List of Illustrations ............................................................................................................. vi
  List of Figures ............................................................................................................. vi
  List of Tables ........................................................................................................... vi
Abstract ............................................................................................................................ viii
Introduction ..........................................................................................................................1
Review of Literature ..........................................................................................................3
  Importance of the Honey Bee to United States Agriculture ........................................3
  Factors Contributing to Colony Loss ..........................................................................7
  Foraging Behavior and How It Relates to Agriculture ............................................12
  Honey Production Records and Regions ..................................................................18
  Applied Uses and Value of Current Models ............................................................23
Data and Methodology .......................................................................................................25
  Background of the Problem .....................................................................................25
  Methodology ............................................................................................................25
  Data Description ......................................................................................................26
Empirical Analysis and Results ..........................................................................................28
  Geographically Based Models .................................................................................28
    Background Information ......................................................................................28
    Procedures for the Analysis of the Models ..........................................................30
    Specific Statistical Assumptions and Methods .....................................................31
  Ecoregion Based Models ............................................................................................32
Background Information .......................................................... 32
Procedures for the Analysis of the Models .................................. 33
Specific Statistical Assumptions and Methods .......................... 34
Cluster Based Models ............................................................... 39
Background Information .......................................................... 39
Procedures for the Analysis of the Models .................................. 40
Specific Statistical Assumptions and Methods .......................... 40
Results ...................................................................................... 46
Geographically Based Models .................................................... 46
Ecoregion Based Models ............................................................ 49
Cluster Based Models ............................................................... 53
Comparison to Determine an Improved Model for Regional Honey Production .... 57
Conclusions and Implications .................................................... 60
Literature Cited ........................................................................ 66
List of Illustrations

*Figures*

Figure 1: Map of the Nye model honey production regions ..............................................28
Figure 2: Map of the American Bee Journal honey production regions ............................29
Figure 3: Map of the Bee Culture honey production regions ............................................30
Figure 4: Map of the Ayers and Harman ecoregion based honey production regions .....33
Figure 5: Map showing level II ecoregions of the United States .................................35
Figure 6: Map of the ECO1 model honey production regions ..........................................36
Figure 7: Map of the ECO2 model honey production regions ..........................................37
Figure 8: Map of the ECO1A model honey production regions ......................................38
Figure 9: Map of the ECO2A model honey production regions ......................................39
Figure 10: Map of the COMP6 model honey production regions ...................................43
Figure 11: Map of the COMP7 model honey production regions ...................................43
Figure 12: Map of the COMP8 model honey production regions ...................................44
Figure 13: Map of the COMP9 model honey production regions ...................................44
Figure 14 Map of the COMP10 model honey production regions ...............................45
Figure 15: Map of the Page et al. (1987) model honey production regions ..................46

*Tables*

Table 1: Results of the Nye model statistical analysis ...............................................46
Table 2: Results of the American Bee Journal model statistical analysis ....................47
Table 3: Results of the Bee Culture model statistical analysis ..................................48
Table 4: Results of the ECO1 model statistical analysis ...........................................49
Table 5: Results of the ECO2 model statistical analysis ...........................................50
Table 6: Results of the ECO1A model statistical analysis ................................................51
Table 7: Results of the ECO2A model statistical analysis ................................................52
Table 8: Results of the COMP6 model statistical analysis ................................................53
Table 9: Results of the COMP7 model statistical analysis ................................................54
Table 10: Results of the COMP8 model statistical analysis ..............................................54
Table 11: Results of the COMP9 model statistical analysis ..............................................55
Table 12: Results of the COMP10 model statistical analysis ............................................56
Table 13: Results of the Page et al. (1987) model statistical analysis ...............................57
Table 14: Summary of all models analyzed.......................................................................58
Currently three systems are used to categorize honey production regions in the United States, one from the United States Department of Agriculture, one from the American Bee Journal used for its monthly U.S. Honey Crop and Markets report, and one from Bee Culture’s monthly regional honey price report. These systems follow political state boundaries and are based upon climate, bee forage, and regional beekeeping practices. While these systems are popular with the general beekeeping community, to our knowledge, their accuracy has not been studied.

Although differing geographic regions can vary in bee forage species availability, states with similar geography and flora should have similar honey production. This is not the case because states within the same honey production region vary in honey production, possibly due to smaller ecotype divisions within the larger honey production regions. Due to this ecotype gradient, some models divide the United States into far more regions based upon ecotypes and disregard political boundaries. While a model based on ecotypes that disregard state political boundaries may be more accurate, it is not currently possible to statistically evaluate them due to how honey production data are collected.

This study developed nine novel regional honey production models that regard political boundaries while attempting to satisfy ecotype similarity. The first four
alternative models are based solely on Level II ecoregions and were developed by a best-fit manual approach that minimized the number of ecoregions per honey production region. The five remaining models were created using statistical $k$-means partitioning cluster analysis and are purely data based. Also discussed is a linear regression model produced by Page et al. Differences within and between the models were analyzed using descriptive statistics and ANOVA in order to determine an improved model that describes regional honey production in the United States.

Many of the models, both preexisting and those developed for this study, had insignificant means and are not viable. Of those that had significant means, a $k$-means cluster based model was determined to be the statistically superior model and can be considered an improved regional honey production model for the United States.
Introduction

There are three systems used to categorize honey production regions in the United States, one developed by the United States Department of Agriculture (Nye 1980), one used by the American Bee Journal for its monthly U.S. Honey Crop and Markets report (Graham 2016), and one from Bee Culture’s monthly regional honey price report (Flottum 2017). These systems follow political state boundaries and are based upon climate, bee forage, and regional beekeeping practices. Two of these systems, the American Bee Journal and the Bee Culture models, are used to disseminate regional honey production data to beekeepers through the two most widely circulated industry journals in the United States.

Although it is clear that differing geographic regions vary greatly in bee forage species availability, states with similar geography and flora (ecoregions) should have similar honey production. This is not the case because states within the same honey production region can vary widely in honey production, possibly due to smaller ecotype divisions within the larger honey production regions. Due to this ecotype gradient, some models divide the United States into far more regions based upon ecotypes and disregard political boundaries (Ayers and Harman 1992). While a model based on ecotypes that does not follow state political boundaries may be more biologically accurate, currently it is not possible to statistically evaluate them due to how honey production data are collected by the USDA (2015).

This study developed nine novel regional honey production models that follow political boundaries while attempting to satisfy ecotype similarity. The first four alternative models are based solely on Level II ecoregions (EPA 2016) and were
developed by a best-fit manual and visual approach that minimized the number of
ecoregions per honey production region while maintaining contiguous regions and the
lowest number of regions possible. These models mimicked the style and function of the
popular American Bee Journal and Bee Culture models.

The five remaining models were created using statistical $k$-means partitioning
cluster analysis and are purely data based (Janowitz 2010). These models were computer
generated and no parameters for contiguous regions were set. The $k$-means models
differed markedly from the American Bee Journal and Bee Culture models. Also
discussed is a computer generated linear regression model produced by Page et al. (1987)
that was created with contiguous regions.

A set of parameters that defined an acceptable model was delineated for the study.
Differences within and between all models were analyzed using variance and ANOVA.
The set parameters were then used to accept or decline a model. All acceptable models
were then compared against each other in order to determine an improved model that
describes regional honey production in the United States.
Review of Literature

Importance of the Honey Bee to United States Agriculture

The European honey bee (*Apis mellifera mellifera*) is an important insect to United States agriculture for many reasons. Honey bees pollinate many crops and produce valuable hive products (honey, beeswax, and propolis). Honey bee pollinated produce had an estimated valued of $14.6 billion in the year 2000 (Morse and Calderone 2000). More recently, the White House (2014) reported that honey bees are responsible for more than $15 billion in agricultural and societal value. Honey bees are considered the third most economically important livestock species worldwide, trailing only cattle and pork (Traynor 2016).

The product that honey bees are best known for is the one for which they are named: honey. While honey is an important crop for the beekeeper, it does not have the economic impact of pollination services or the crops pollinated. For the year 2015, the USDA reported the total value of honey produced in the US was roughly $327 million. This value is acknowledged by the USDA as low because only producers with more than five colonies who can be counted as a farm are included in the survey. To be counted as a farm, an operation must sell $1,000 or more of agricultural products in the year (all products combined, not just honey or other apiary products) (USDA 2016a). Not included in the $327 million is the value of honey produced by farms with less than 5 colonies. The 2015 season is the first year for which this data has been collected. In 2015, farms with less than five colonies accounted for another 720,000 pounds of honey with an estimated value of $1.5 million. In addition, backyard beekeeping has become a popular hobby over the past ten years. These beekeepers generally have fewer than five hives, sell
less than $1,000 of farm products per year, and may give away a large proportion of their
honey harvest to family and friends. There are likely several million dollars of honey
produced by beekeepers who do not qualify as a farm under USDA guidelines.

There are three other hive products that are not tracked by the USDA but are of
economic importance. These are beeswax, pollen and propolis. Beeswax is produced in
the greatest quantity and is used in a wide variety of products including candles,
processed foods, pharmaceuticals, cosmetics, textiles, polishes, and industrial materials
(Bogdanov 2016). Beeswax is preferred for many of these uses as because it is a natural,
renewable product that is produced by the wax glands on the ventral side of worker bees.
Because beeswax is made by the bees, it is perceived as organic and free from manmade
chemicals (Bogdanov 2016). However, this is hardly the case because many lipid soluble
agricultural chemicals have been isolated from beeswax, even from farms that do not
apply chemicals to their crops (Reed et al. 2010). This is due to the long foraging
distances of honey bees. Despite its chemical contaminants, wax still remains a valuable
commodity and the US demand is far greater than domestic production (Schmidt and

Pollen is collected by honey bees as their sole source of dietary protein. The bees
carry the pollen from the field to their hive as compacted pellets in hairy patches, called
pollen baskets, located on their rear legs (Caron and Conner 2013). The pollen pellets can
be collected with special pollen traps that are placed over the hive entrance. As the bees
enter the hive, the pollen trap brushes the pollen pellets from the pollen baskets and into a
collection device. Pollen is collected and sold as a nutritional supplement, natural allergy
aid, or as an ingredient for many holistic remedies. This is a specialty market and few
beekeepers go through the labor intensive process of collecting, cleaning, storing and marketing pollen. It is a high value product, however, and can add significant revenue to some apiaries (Schmidt and Buchman 1992).

Propolis is a sticky tree and plant resin that is collected by bees for use in sealing cracks and openings in their hives (Caron and Conner 2013). The substance has antimicrobial properties and is collected by some beekeepers who sell it for use in “natural” treatment preparations. The demand for propolis is high due to the low amount produced by beekeepers each year. The method of collection, cleaning and marketing is labor intensive which has kept production low. The high demand for and low supply of propolis has led to prices that can be quite high per unit volume (Schmidt and Buchmann 1992).

By far the greatest economic impact of honey bees comes from pollination and pollination contracts. Worldwide 52 of 115 leading global food commodities depend on honey bee pollination to set fruit or seed (vanEngelsdorp and Meixner 2010). In the United States, honey bees help support the cultivation of 90-130 crops grown through pollination (Bond et al. 2014). This pollination is most often provided by migratory commercial beekeepers through pollination contracts. Although honey bees are rarely the most efficient pollinator of many crops, they are the most important for large-scale crop monocultures due to the relative ease with which colonies can be moved, the ability to stimulate foraging behavior in a colony, and the large foraging population per colony (vanEngelsdorp and Meixner 2010). Even though honey bees are vitally important for pollination services, wild pollinators still play an important role and can increase fruit set even in crops utilizing honey bee pollination (Garibaldi et al. 2013).
Crops where honey bee pollination is employed range from nut and tree fruits to forage crops such as alfalfa and clover (Morse and Calderone 2000). In 2012, the ten most contracted crops for pollination in the U.S. were in order: almonds, sunflowers, canola, grapes, apples, sweet cherries, watermelons, plums, blueberries, and avocados (Bond et al. 2014). Almonds made up the majority of contracts and accounted for 44.6% of all pollination contract fees. The gross revenue from all pollination contracts for 2012 was estimated to be $655.6 million (Bond et al. 2014). This revenue will likely rise in the future because pollination fees have steadily increased over the past decade, in part, due to the increased demand for almond pollination, which pays a premium and to the increasing acreages of pollinator dependent crops (Bond et al. 2014, Aizen et al. 2008).

As the acreage of pollination service dependent crops continues to grow, the demand for managed bee colonies to carry out pollination services is also increasing. Although the current worldwide trend in pollinator populations, including native bee species, wasps, butterflies, beetles, and wild honey bees, has been a steady decline over the past decade, Aizen et al. (2008) found no current pollination shortage. The lack of a shortage is due to the current widespread availability of managed honey bee colonies for pollination services. They warn however, that this may not be the case in the future as the world continues to plant ever expanding acreages of pollinator dependent crops. Aizen and Harder (2009) found that managed honey bee colonies have, in fact, increased globally by ~45% over the last 50 years while at the same time pollination dependent crop acreage has increased by ~300%. The disparity in honey bee colony increase and pollinator dependent crop acreage increase is not expected to change in the foreseeable future (Aizen and Harder 2009).
Another aspect that must be addressed when discussing the lack of honey bee colonies for pollination services is the general decline in honey bee colony health and survivability over the last several decades. Even though total managed honey bee colony numbers have been slowly but steadily rising over the last 50 years, yearly managed colony losses have risen sharply over the last several decades (Neumann and Carreck 2010). Losses for the year spanning May 2015 to May 2016 have been preliminarily estimated to be 44% (Steinhauer et al. 2016). Many of these losses have been attributed to the highly publicized, but still little understood, phenomenon known as colony collapse disorder (CCD) (Oldroyd 2007, Ellis et al. 2010, Ratnieks and Carreck 2010).

Factors Contributing to Colony Loss

CCD was first documented in the United States in fall of 2006 when some beekeepers reported colony losses ranging from 30-90% (Ellis et al. 2010). Yearly colony loss above 30% is not unusual for scattered beekeepers, but the losses in 2006 were widespread and the symptoms did not correlate with any known honey bee diseases (Ellis et al. 2010). The phenomenon appeared to be a new disease and was named colony collapse disorder (Ellis et al. 2010).

CCD is characterized by a set of three specific symptoms: 1) the rapid loss of adult worker bees from the colony, 2) a noticeable lack of dead worker bees at or around the hive, and 3) a delayed invasion of hive pests (e.g. small hive beetles and wax moths) and kleptoparasitism from neighboring honey bee colonies (vanEngelsdorp et al. 2009). Since 1869 there have been at least 18 episodes of high colony mortality worldwide (vanEngelsdorp et al. 2009). Some of those episodes had similar symptomology to CCD indicating that it may not be an entirely new phenomenon (vanEngelsdorp et al. 2009).
Due to the severity of losses in 2006, a study was conducted to see if the trend continued into the 2007-08 apiary season. Findings showed that losses for that time period ranged from 31.2-36.0% for the apiaries that reported (vanEnglesdorp et al. 2008). CCD ranked as the 4th most common cause of colony loss and was reported by 9% of the responding apiaries (vanEnglesdorp et al. 2008). A similar study conducted on the 2008-09 season had similar results with 32% colony loss and ~10% of losses attributed to CCD (Williams et al. 2010). Beekeepers attributed CCD as only the 8th most important contributor to colony loss for the 2008-09 season (Williams et al. 2010).

Because CCD is hard to diagnose and it is self-reported the majority of the time (as opposed to being diagnosed by a trained state apiarist), these numbers may be low. Due to the low significance to colony loss attributed to CCD by beekeepers and the fact that it was never found to be the cause of a majority of colony losses, many people studying the phenomenon feel that other factors (Varroa mites (Varroa destructor), weather, etc.) may be of more importance to yearly colony loss (Williams et al. 2010).

The actual cause of CCD is still open for debate but it is becoming clear that it is not a simple disease caused by a single organism. Instead, it is a complex syndrome caused by the combination of many factors. Factors as diverse as climate change, genetically modified crops, a dwindling genetic pool, chemical residues, poor queen performance, diseases, and pests have all been studied as contributing factors in CCD (Oldroyd 2007, Conte 2008, vanEnglsdorp and Meixner 2010). Nothing definitive has been discovered for the cause of CCD in the U.S. Adding to the confusion is the fact that the colony loss problem is not restricted to the United States. Losses with similar
symptoms are occurring in Europe, the Middle East, and Japan (Neumann and Carreck 2010).

One thing that all of these locations have in common are the honey bees (Apis mellifera mellifera) found in those locations are susceptible to the Varroa mite (Varroa destructor) (Neumann and Carreck 2010). There are currently at least 20 recognized subspecies of the European honey bee (Dietz 1992). The Africanized bees of South America (Apis mellifera mellifera x Apis mellifera scutellata) and African honey bees (Apis mellifera scutellata) can survive Varroa mite and have some resistance to them (Neumann and Carreck 2010). Australia is currently Varroa destructor free (Neumann and Carreck 2010). However Varroa jacobsoni, a similar species that only infects the Asian honey bee (Apis cerana), was discovered in the port of Townsville, Queensland Australia in July of 2016 (Watson 2016).

The Varroa mite, like the honey bee, is not native to North America. Varroa mite was first discovered in North America in Wisconsin in September 1987 (Wenner and Bushing 1996). The point of origin for the introduced mites has never been conclusively proven. The Wisconsin bees originated from Florida hives and it is believed that Florida was the point of introduction for the United States (Wenner and Bushing 1996). The mites spread quickly and by 1995 could be found in all 48 contiguous states and Alaska (Wenner and Bushing1996).

While Varroa mites can be controlled in managed colonies through integrated pest management (IPM) practices and treatment with acaricides (Shimanuki et al 1992), there is always a reservoir of mites in wild or feral colonies from which re-infestation can occur. This reservoir mandates that beekeepers constantly monitor the mite populations in
their colonies and treat accordingly (Shimanuki et al. 1992). Acaricides have also been blamed as a possible contributor to CCD because levels of common acaricide chemicals (fluvalinate, amitraz, and coumaphos) are found in beeswax samples at significant levels (Johnson et al. 2010).

The Varroa mite is an ectoparasite and feeds on a honey bee’s hemolymph (Shimanuki et al. 1992). In doing so, the mite acts as a vector for many viruses including the Israeli acute paralysis virus (IAPV) (Watanabe 2008). IAPV has been linked to colonies diagnosed as dying from CCD (Watanabe 2008). This virus is also found in the Varroa free bees of Australia so it is not the sole cause of CCD, but may be a contributing factor (Watanabe 2008). IAPV causes honey bees to tremble and to be unable to fly. Bees stricken with this virus tend to walk out of the hive and a fan of dead and dying bees can be found in front of the hive entrance. This is not the case in colonies confirmed to have succumbed to CCD. In confirmed CCD colony losses, all or most of the adult bees are missing and no evidence of dead or dying bees can be observed. For this reason, some feel the paralysis viruses have little or no effect on CCD (Oldroyd 2007).

Many other possible contributing causes of CCD have been hypothesized including already-known and studied diseases and parasites such as European and American foulbrood, tracheal mites (which vectors the CCD similar Isle of Wright Disease that struck apiaries on a small English island in 1906), the protozoan Nosema apis (a common honey bee endoparasite), and any of a number of viruses that are mainly vectored by the Varroa mite. Of these, viruses are considered the most likely candidates for contributing to CCD (Oldroyd 2007).
Another possible contributor to CCD are agricultural chemicals (Yang et al. 2008, Johnson et al. 2010, Mullin et al. 2010). Nearly every chemical used inside the hive to control Varroa and tracheal mites, small hive beetles, and wax moths can be found in beeswax, honey, pollen and the bees themselves (Mullin et al. 2010). A study by (Mullin et al. 2010) detected traces of the acaricides fluvalinate and coumaphos in 98% of the beeswax sampled.

Also found in nearly every hive are common agricultural chemicals applied to crops, lawns, golf courses and landscape plants. Mullin et al. (2010) found residues of over 121 agricultural chemicals (pesticides, herbicides and fungicides) in samples of wax and honey. Even trace amounts of long-banned chemicals such as DDT, DDD, and chlordane were found in another study (Johnson et al. 2010). The levels of agricultural chemicals found in bee colonies are not high enough to cause morbidity in the bees, but there may be long term effects to the immune system, foraging behavior, or reproductive ability (Yang et al. 2008).

An agricultural chemical that has been researched for a possible connection to CCD is imidacloprid (Yang et al. 2008, Johnson et al. 2010). Imidacloprid is a neonicitinoid systemic pesticide that has gained popularity over the past decade. Being systemic, the pesticide is translocated throughout the plant and can be detected in pollen and nectar (Johnson et al. 2010). Imidacloprid has been shown to adversely affect foraging behavior in honey bees at minute, sub-lethal doses and may contribute to CCD (Yang et al. 2008).
Honey bees forage on a wide variety of flowering plants in order to gather the sugar rich nectar they require as an energy source, the protein rich pollen they need to raise brood, and the resinous propolis they use for sealing the hive (Caron and Conner 2013). Honey bees also forage for water which is needed for the honey ripening process and for evaporative cooling of the hive. While it may seem as if the landscape offers an endless array of plants from which a honey bee may forage, an individual bee tends to have a limited foraging area at any given point in time (Caron and Conner 2013). This foraging area may be as small as an individual tree or bush or a small patch of flowers. If the quality of the utilized forage decreases, the bee will switch to an alternate source. The bee will then concentrate on a limited area of the new forage (Caron and Conner 2013).

In all colonies a portion of the foraging force will act as scout bees which go afield and actively search out new forage sources. Scent and sight both play a role in attracting honey bees to forage sources (Caron and Conner 2013). In monocrop fields with few landmarks for the bees to use for orientating, some bees continue to forage small patches of the crop while others in the colony tend to forage over the entire field (Gary 1992).

The time spent per foraging trip depends on many variables. A major factor in the amount of time spent afield is whether the foraging trip was spent gathering nectar or pollen (Caron and Conner 2013). Trips spent gathering pollen are considerably shorter than those spent gathering nectar. Nectar gathering trips vary considerably in length depending on the amount of nectar available (Caron and Conner 2013). Nectar abundance
is dependent on the species of flower, developmental stage, climatic conditions, and competition from other pollinators (Fowler et al. 2016).

Temperature also has an effect on how much honey bees will forage. The minimum temperature that bees will forage at is 13°C (Caron and Conner 2013). At temperatures above 43°C, nectar and pollen foraging activity may cease, but water foraging activity will continue because water is needed to cool the hive (Gary 1992). The optimum air temperature for honey bee flight is between 19°C and 30°C (Caron and Conner 2013).

The age of the forager also has a direct correlation to the distance and time traveled nearby (Caron and Conner 2013). Young bees that are new to the foraging force travel shorter distances from the hive than older, more experienced bees. This is believed to be due to older bees having had more time to learn the foraging area nearby (Caron and Conner 2013). Honey bees typically forage close to the hive (within a 3.2 km radius) if enough satisfactory nectar producing forage is available, but are capable of flying up to 6.5 km from the hive if forage is scarce nearby (Caron and Conner 2013).

Studies by Beekman and Ratnieks (2000) in Sheffield, England found differences in foraging distance between hives in the same location at different times of the year. In May, the average foraging distance was 1 km while at the same location in August the average foraging distance had increased to 5.5 km with a maximum distance of 9.5 km. Beekman and Ratnieks (2000) hypothesized nectar foraging distance may be greater in patchy environments where forage patch size and quality is not consistent. Another study did not find any significant difference in average nectar foraging distance between simple and complex landscapes, but did find a significant difference in pollen foraging distance.
with bees foraging further for pollen in simple landscapes (Steffan-Dewenter and Kuhn 2003). Steffan-Dewenter and Kuhn (2003) also found that time of year appears to have a greater impact on foraging distance than the complexity of the landscape.

Pollen foraging trips are significantly shorter than nectar forging trips even though the bees may be traveling the same distance (Caron and Conner 2013). The shorter foraging times are attributed to fewer flowers needing to be visited to gather a full load of pollen and to bees being less selective on pollen quality (Caron and Conner 2013). Colony size does not seem to have an effect on foraging distance when nectar and pollen resources are abundant, but as resources diminish larger colonies (20,000+ bees) will forage significantly further than small colonies (~6,000 bees) (Caron and Conner 2013). Small colonies also gather significantly more pollen (Beekman et al. 2004). The average honey bee forager will fly approximately 800 km in her lifetime (Caron and Conner 2013).

The number of flowers visited per nectar foraging trip varies greatly depending on the flower species and weather conditions (Fowler et al. 2016). Large flowered species with ample nectaries, such as tulip poplar (*Liriodendron tulipfera*) may only require one flower visit per trip (Caron and Conner 2013). Plants with small flowers and less ample nectaries such as sweet clover (*Melilotis* sp.) can require as many as 1,000 flower visits per trip (Caron and Conner 2013). Weather events, such as rain, can also affect the number of flower visits required because heavy rainfall can wash the nectar from erect blossoms such as those found on tulip poplar (Caron and Conner 2013). Drought can have an adverse impact on nectar production and increase the number of flower visits per trip.
The amount of nectar collected per trip varies but averages 40 mg and can be as high as 70 mg – nearly the body weight of the foraging bee (Caron and Conner 2013). Nectar can be up to 40% sugar and can include sucrose glucose, maltose and fructose (Caron and Conner 2013). Honey bees show a strong preference for sucrose and nectars with a high percentage of sucrose are highly attractive to them (Caron and Conner 2013). When nectar is scarce, honey bees will forage on other sources of sugar such as sugar water solutions provided by beekeepers, hummingbird feeders, and even trash cans with cans and cups of sugary drinks (personal observation).

The location and distance of a forage source is communicated by a returning scout or forager to the other foragers in the hive in the form of a “dance (Gary 1992).” There are two main dances used by honey bees to communicate resource locations: the round dance, which is used to communicate a resource < 80 m from the hive, and the wagtail dance (sometimes called the waggle dance) used to communicate a resource found at >80 m from the hive (Gary 1992). There are variations, known as dialects, in the dances performed by the various races of bees. The Italian race of honey bee has a third dance, known as the sickle dance, that is used to convey the location of a resource found between 10 and 30 m from the hive (Gary 1992).

Although all of the dances are slightly different, they convey the same information – the angle of travel the resource can be found in relation to the sun, and the distance from the hive (Caron and Conner 2013). How bees determine the distance to a forage source was thought to be related to how much energy the bee expends during the flight (Esch and Burns 1996). Studies by Esch and Burns (1996) found evidence that this might not be the case. They proposed an optical flow hypothesis where bees use the
speed of retinal imaging motion perceived from the ground to estimate the distance flown. The scientist who decoded the wagtail dance, von Frisch, had considered flight time to be a possibility for honey bee distance estimation. There is evidence that bees can determine time and combined with flight speed from optical flow could determine the distance to a forage source (Esch and Burns 1996). Although the dances of honey bees have been decoded, exactly how the bees determine the information they convey is still poorly understood (Esch and Burns 1996).

The relatively predictable foraging behavior of honey bees, their large colony size, and the ease by which their colonies can be moved by truck, are what make them important as pollinators in agriculture. The often recommended hive density for crop pollination is one hive per acre, but may be higher if the crop is difficult to pollinate or has a high density (Caron and Conner 2013). Even though honey bees are a predictable, easily managed and moved pollination force for a wide variety of crops, there are limitations to their pollination capabilities. Honey bees are rarely the most efficient pollinator for many crops. Some crops are more efficiently pollinated by native pollinator species such as orchard mason bees, bumblebees, solitary bees, and squash bees (Winfree et al. 2007). What honey bees lack in pollination efficiency they make up for with sheer numbers of foragers.

The presence of wild pollinators seems to enhance fruit set even when honey bees are present. In some instances honey bees are actually supplemental to wild pollinators (Garibaldi et al. 2013). Studies on watermelon pollination showed that native pollinators could provide sufficient pollination on >90% of the farms studied and the native pollinators deposited significantly more pollen per flower visit (Winfree et al. 2007).
These studies tend to indicate that in some areas and with specific crops, pollination provided by native pollinators may be adequate and no honey bee supplementation is needed. This could prove crucial because the number of acres of pollinator dependent crops planted worldwide are increasing at a faster rate than the number of managed honey bee colonies is increasing (Aizen and Harder 2009). Complete reliance on native pollinators may not be possible in all areas as native pollinator numbers are also declining worldwide (Potts et al. 2010).

The changing agricultural landscape also impacts the efficiency of honey bee pollination services. The trend towards vast acreages of monoculture crops and clean field edges in developed countries seems to have a negative impact on pollination rates (Garibaldi et al. 2011). Garibaldi et al. (2011) found that fruit set decreased as the distance from natural areas increased. This was attributed to the lack of native pollinators present in the fields. Honey bee visits to the target crop remained high as alternate forage was limited or nonexistent, but fruit set remained low due to pollination inefficiency of the honey bee (Garibaldi et al. 2011). To help combat the lack of pollinator habitat in agricultural areas, several United States government set aside programs now include pollinator initiatives (Decourtye et al. 2010). These programs pay for fallow fields, field edges, conservation buffers, and roadsides to be seeded with pollinator flower mixes (Decourtye et al. 2010).

Any increase in forage diversity appears to increase native pollinator success (Williams and Kremen 2007). Williams and Kremen (2007) found solitary bees, *Osmia lignaria*, nesting in areas of low forage diversity had significantly lower reproductive success. They also found that solitary bees living in an area of high habitat connectivity
fared better than those in highly fragmented, though diverse, habitats. A study from the tropics shows an increase in honey yield for honey bee colonies in close proximity to forest lands (Sande et al. 2009). It is clear that a diverse habitat is not only better for all pollinator species, but also plays a direct role in the efficiency of crop pollination.

One of the greatest impacts on honey bees foraging agricultural fields come from the agricultural practices used to maintain the crops. In developed countries the use of agricultural chemicals – pesticides, herbicides, and fungicides – is standard operating procedure. Time of application is vitally important, especially for pesticides, in order to limit pollinator kill (both native bees and honey bees) (Xerces Society 2017). Pesticides should only be applied at night when pollinators are not actively foraging (Xerces Society 2017). Some systemic pesticides in the neonicotinoid family have been found in nectar and pollen from treated plants and have also been shown to cause abnormal foraging behavior in bees (Yang et al. 2008). These effects occurred at far below lethal dosages and included delayed foraging, failure to return to the foraging site, and bees that just disappeared (Yang et al. 2008).

Honey Production Records and Regions

The production of honey in the United States has been tracked by the USDA’s National Agricultural Statistics Service (NASS) since 1939, with a suspension of data collection between 1982 and 1985 (Page et al. 1987). Until 2015, honey production data only included apiaries with five or more hives that also qualify as a farm under USDA guidelines (farms producing at least $1,000 in gross revenue from farm raised products). In 2015, the USDA began collecting data from apiaries with fewer than five hives that
still qualify as a farm. There is no national data recorded by the USDA for apiaries that do not qualify as a farm, regardless of how many hives are involved.

Data are collected by the USDA in two ways. The first is a voluntary yearly survey in which all bee keepers can participate. Only data from those apiaries qualifying as farms under USDA guidelines are used for statistical purposes. Although the survey is voluntary, not submitting production records and hive number records can adversely impact a producer’s eligibility in some Farm Service Agency programs such as the livestock indemnity program (USDA 2015). A sample of producers that participated in the survey and who qualify as an apiary under the USDA guidelines are chosen for participation in the NASS census which is conducted every five years (USDA 2016b).

Separate from NASS, many states have mandatory reporting of numbers of hives and of honey production (Potter 2016). These state records provide a way to cross check NASS records. However, several states do not require any reporting of apiaries or their products (Potter 2016).

Data prior to 1939 is scarce, although an attempt has been made to reconstruct honey yields for specific geographical locations for specific tree species (Maxwell 2012). Reconstruction of honey yield data for some tree species has been done using tree ring data and is fairly accurate (Maxwell 2012). Unfortunately, the technique is limited to long-lived tree species and cannot be used to reconstruct historical yields from areas where the major flows are from herbaceous or shrubby species (Maxwell 2012).

Because flora, weather patterns, and even beekeeping methods vary across the United States, honey production is normally discussed in regard to honey production regions. Currently there are three primary regional honey production mapping systems.
The first is the system described by Nye (1980) in which the United States is divided into seven geographical regions. In this system some states are split and fall into two or more honey production regions. The second system, used by the American Bee Journal (Graham 2016), divides the United States into eight honey production regions with all states belonging to only one region (no split states). The last system, used by Bee Culture magazine (Flottum 2017), divides the United States into seven regions with all states belonging to only one region. All of these systems are based on geographical and climatic similarities between states because the geography has a direct (but not complete) influence on the flora of the region. All models take regional beekeeping methods into consideration.

In order to fully understand how honey production units are created, it is important to understand how the United States is divided geographically and ecologically. Jefferson (1917) wrote about the importance of understanding how the United States is divided geographically. Fenneman (1928) reported a need to divide the earth into smaller sections in order to better describe each of them. He contended that no continent was simple enough to be fully described by a single physiogeographic unit (a unit of land with the same basic geographic features) and that many divisions would be necessary to fully describe the United States.

Fenneman designed his divisions to be as geographically homogenous as possible while covering the greatest amount of land possible to limit the total number of regions. Prior to Fenneman’s work, several attempts to divide the US physiogeographically were made, however, these attempts did not have strongly defined boundaries and units overlapped or left spaces unaccounted (Fenneman 1928). The new system corrected the
problems of overlap and omission and divided the US into major divisions, provinces and sections. The US is divided into eight major divisions, the eight major divisions are subdivided into 25 provinces, and the provinces are divided into 78 sections. These regions are static because the geology of a region does not change in any measurable way during the scope of human time.

While geology is important to soil type, climate, and to some extent the flora of a region, it does not give a complete overview of a region’s ecology. Over the last fifty years, more emphasis has been placed on creating divisions based on the total ecological profile for a region (CEC 1997). These divisions are known as ecoregions. Ecoregions have changed many times over the past half century because the methods used to delineate them have evolved.

Many of the early attempts at classifying ecoregions were based on forest and climate classification systems and were mainly climate-based (Baily 1976). In the mid-1980s, a more holistic approach to ecoregion classification was undertaken. In these approaches a full range of physical and biotic factors were considered while at the same time realizing the ecoregions are not always dominated by a single factor (Omernik 1987). The current ecoregion divisions were created by the Commission for Environmental Cooperation (CEC) in 1997 using similar holistic methods and by gathering the opinions of scientists and ecologists in order to create a data matrix for building each ecological level.

In the current CEC system of ecoregions, North America is divided into four levels: Level I, Level II, Level III and Level IV ecoregions – a system similar to the one used for physiogeographic regions. Level I ecoregions are the largest division and the
continental US is divided into 12 Level I ecoregions. The Level I ecoregions are subdivided into Level II ecoregions of which there are 25 in the continental US. The Level II ecoregions are further subdivide into Level III ecoregions which number 105 in the continental US. The smallest divisions are the Level IV ecoregions numbering 967 in the continental US (EPA 2016). As these are ecoregions there is the possibility of them being changed in the future due to a change in ecoregion assignment protocol.

Ayers and Harman (1992) created a model for describing US honey production regions based on ecoregions. This model divides the US into eleven honey production regions, totally ignores political state boundaries and is based solely on ecoregion delineations. Although this is a viable modeling method, and of value to beekeepers as an aid in determining apiary management decisions, it is of little value for calculating regional statistics due to the current USDA data collection and reporting protocol which collects data by state. Data collection by county would be required for a meaningful statistical analysis to be conducted on a solely ecoregion based honey production model since states may contain several ecoregions.

At least one attempt has been made to create a computer generated, statistics based regional honey production model. Page et al. (1987) used USDA data for the 42 year period of 1939 to 1981 to create a linear regression model based on average honey yield per state for the entire 42 year period. The linear regression was calculated using a computer-assisted algorithm that was repeated until all states met two basic criteria: 1) states within a region were correlated and 2) states within a region must be contiguous. The linear regression yielded three models. The first model was created at $r \geq 0.000$ and resulted in 13 regions, with three of those regions being single states. The second model
was created at \( r \geq 0.295 \) and resulted in 17 regions, with only one single state region. The final model was created at \( r \geq 0.100 \) and resulted in 13 regions with 2 single state regions. Page et al. (1987) considered the model created at \( r \geq 0.100 \) to be the best model even though it had two single state regions. To deal with the single state regions, Page et al. proposed combining the single states with an adjacent region showing a positive correlation.

*Applied Value and Uses of Current Models*

Regional honey production models are important for several reasons. Models such as those used by the American Bee Journal and Bee Culture are used by the two most widely circulated beekeeping industry journals to disseminate information on regional honey harvest, market honey prices, and news about regional weather events or disease outbreaks (Graham 2016, Flottum 2017). Although these models are accepted by beekeepers, and have a logical, easy to understand format, a thorough review of the literature did not find any scientific study as to the validity of them.

Attempts have been made to create scientifically sound models such as the ecoregion based model of Ayers and Harman (1992) and the regression analysis based model by Page et al. (1987). Although both models were created using scientific process neither has become popular or widely used. Perhaps this is due to the Ayers and Harman (1992) model having regions which do not follow political boundaries. The original model by Page et al. (1987) had a large number of regions (13), and two of those were single state regions. These are possible reasons for the lack of adoption of the Page et al. (1987) model.
Another useful aspect of a statistically sound regional honey production model is that it may be used for detecting pest and disease outbreaks before large-scale colony loss is experienced. This was the impetus that led Page et al. (1987) to develop their regression based model. They felt that any change in regional state makeup from year to year could indicate a pest or disease problem and warrant further in the field investigation. With the increase in honey bee colony loss due to *Varroa* mite vectored viruses and CCD this may be the most important aspect of a well-designed and statistically accurate regional honey production model.
Data and Methodology

Background of the Problem

Nye (1980), Graham (2016), Flottum (2017), Ayers and Harman (1992) and Page et al. (1987) have provided several honey production regional models. Each system has its pros and cons. However, a thorough literature review discovered no research that directly assesses the accuracy and scientific correctness of the four models. It may also be possible to develop new models that combine the best principles from each of the four established models.

This study seeks to answer the question:

Is there a regional honey production model that demonstrates the true state-to-state honey production for the United States? This study sought to determine an improved model for honey production regions in the United States from several alternatives.

Methodology

In this study, nine novel regional honey production models were developed. Four of these models follow political boundaries while attempting to satisfy ecotype similarity. These models are based solely on Level II ecoregions and were developed by a best fit manual approach that minimized the number of ecoregions per honey production region. The ecoregion models were compared statistically to each other.

The remaining five models are purely data based and were created using computer generated statistical $k$ - means partitioning cluster analysis (Jain and Dubes 1988). These computer generated models were compared statistically to each other and to the Page et al. (1987) computer generated linear regression model.
Once calculations were finalized, all models were compared statistically to each other in order to determine the regional honey production model represented an improved regional honey production model for the United States based on a comparison of the overall variance of each model. Models with a lower overall variance should have more homogenous data sets. This would indicate that those models with low variance are closer to demonstrating the true relationships of honey production between states.

An acceptable model was defined as having no split-states because there is no way to conduct statistical tests on them due to the USDA data collection protocol (USDA 2015). The model should have a low number of regions ($\leq 10$) for ease of information dissemination and user acceptance. This is a subjective criteria and was chosen because models with more than 10 regions looked too busy and were confusing (personal observation). Acceptable models will have low average variance between regions. Low average variance is a criteria as it is an indicator of data homogeneity. The average p-value for the model, between regions, must be statistically significant ($p \leq 0.05$). A significant p-value indicates that the model is valid.

Data Description

All data were sourced from archived US Honey Production Reports located on the USDA NASS website (USDA 2015). The data analyzed for this study covers the 29-year period from 1986 to 2014 and includes values for honey production in pounds per hive for each year. The states of Connecticut, Delaware, Maryland, Massachusetts, Nevada, New Hampshire, New Mexico, Oklahoma, Rhode Island and South Carolina do not have individual published data. Data collected by the USDA-NASS are through a voluntary survey and an every five years sampling census. Only those apiaries with five or more
hives that also qualify as a farm by the USDA (> $1,000 gross farm product sales/year) are included in the dataset. Many beekeepers did not qualify for the survey or census under the USDA guidelines that were in place during the time period studied and the data collection protocol has since been updated to correct for this (O’Rear 2017). For these reasons, these data represent a sample of honey production numbers for the US, not a complete dataset. Statistical calculations were conducted accordingly.

There are several limiting assumptions associated with these data:

1. It is assumed that all qualifying apiaries have been included in the survey. Qualifying apiaries are defined as having five or more colonies producing honey and are considered a farm by the USDA (> $1,000 gross farm product sales/year).

2. It is assumed that honey production for apiaries having fewer than five colonies is inconsequential.

3. It is assumed that honey was counted for the state in which it was produced.

4. Due to privacy issues (USDA 2015), this study is limited to 40 states because 10 states do not have data available to the public.
Empirical Analysis and Results

Geographically Based Models

Background of the Problem

The three currently accepted and most widely used regional honey production models are from Nye (1980) (Figure 1), the US Honey Crops Markets report of the American Bee Journal (ABJ) (Graham 2016) (Figure 2), and the Regional Honey Price Report of Bee Culture (BC) magazine (Flottum 2017) (Figure 3). Nye divides the US into seven honey production regions based on geography, flora and beekeeping methods. In this model, the states of New York, Pennsylvania and California are split between two regions. The state of Texas is divided between three regions.

Figure 1: Map of the Nye model honey production regions.
The ABJ model divides the US into eight honey production regions. There are no split states with this model.

Figure 2: Map of the American Bee Journal model honey production regions.
The BC model divides the US into seven regions with no split states.

![Map of the Bee Culture model honey production regions.](image)

**Figure 3:** Map of the Bee Culture model honey production regions.

Procedures for the Analysis of the Models

Do any of the existing geographically based regional honey production models (Nye, American Bee Journal, and Bee Culture) demonstrate a true regional honey production model for the United States? To obtain an answer for this question a research protocol was developed to evaluate existing geographical models based on the average variance for each model.

The first step in this process was to determine descriptive statistics for each region in each model and preform an ANOVA for each model. Once the statistical calculations were completed, the models were compared using the following criteria:

- no split states
- a low number of regions (≤ 10)
- low average variance for the model
- statistical significance (p ≤ 0.05) for the model

Those models with significant means that also met all other criteria were deemed acceptable models and considered in the determination for an improved regional honey production model for the United States.

Specific Statistical Assumptions and Methods

There is one additional limiting assumption associated with this phase of the study: States that contain two or more honey production regions have been included in the region that covers the most area. Data for state colony numbers are collected in the same manner as honey production data (USDA 2015). Therefore it is not possible to ascertain how hives are distributed throughout a state, and which region has the majority of production. It is assumed that all colonies were in the larger region although they were not.

Statistics were calculated for all models using the following protocol:

Step one: Calculate descriptive statistics (Ẋ, s²) for each region of the three geographical regional honey production models.

Step two: Test for significant differences in means using ANOVA for all regions in the three geographical models. ANOVA conducted at a 0.05 level of confidence.

Hypothesis for all ANOVA calculations:

\[ H_0 = \text{There are no significant differences in the means.} \]

\[ H_A = \text{There are significant differences in the means.} \]

Step three: Determine an average of the variances for the regions of the three geographical models.
All descriptive statistics equations and ANOVA equations used are from Kachigan, (1986). Any additional statistical techniques are discussed under the methods for that particular regional model.

For the Nye model, split states were placed in the region that covered the majority of the state. Texas was omitted from the study because it falls into three regions with no region having a clear majority of coverage. No ANOVA values could be determined for the Southwest region of the Nye model because it only contained one state (Arizona) for which data were available.

Ecoregion Based Models

Background Information

An ecoregion-based approach to regionalizing honey production in the US is perhaps the most appropriate. It has long been known that each honey producing region of the country has specific plants that are important for nectar flow (Lovell, 1926). These major nectar plants tend to follow specific ecoregions. Ayers and Harman (1992) attempted to divide the US into honey production regions based on ecoregions (Figure 4).
This model did not follow political state boundaries and, therefore, was not conducive to calculating statistical data for the regions due to the data collection methods of the USDA. In order for a model to be useful for statistical analysis and for ease of use by the public, it is important to adhere to political state boundaries.

Procedures for the Analysis of the Models

Do any of the ecoregion based regional honey production models (ECO1, ECO2, ECO1A, and ECO2A) demonstrate a true regional honey production model for the United States? To obtain an answer for this question a research protocol was developed to evaluate existing geographical models based on the average variance for each model.
The first step in this process was to determine descriptive statistics for each region in each model and perform an ANOVA for each model. Once the statistical calculations were completed, the models were compared using the following criteria:

- no split states
- a low number of regions (≤ 10)
- low average variance for the model
- statistical significance (p ≤ 0.05) for the model

Those models with significant means that also met all other criteria were deemed acceptable models and considered in the determination for an improved regional honey production model for the United States.

Specific Statistical Assumptions and Methods

Four models were developed for this phase of the project, (ECO1, ECO2, ECO1A, and ECO2A) using Level II ecoregions (Figure 5).
Figure 5: Map showing Level II ecoregions of the United States (EPA 2016).

The continental United States is divided into 25 Level II ecoregions. Most states consists of two or more of these ecoregions. The first model developed (ECO1, Figure 6) was a “best fit” by visualization of all the states into honey production regions following ecoregion borders.
Because the majority of states have several ecoregions within their borders, it was often a judgement call where to place a state. California and Florida became independent regions because both states have unique ecoregion composition.

The best-fit approach resulted in the formation of 10 honey production regions. Once the regions were set, statistics were calculated following the protocol outlined in the statistical techniques section. ANOVA was impossible for one multi-state region, because Arizona was the only state in the region with available data. ANOVA was also impossible for the single state regions of California and Florida. Descriptive statistics ($\bar{X}$, $s^2$) were calculated for these single state regions.

Statistics were calculated for all models using the following protocol:

Step one: Calculate descriptive statistics ($\bar{X}$, $s^2$) for each region of the four ecoregion honey production models.
Step two: Test for significant differences in means using ANOVA for all regions in the four ecoregion models. ANOVA conducted at a 0.05 level of confidence.

Hypothesis for all ANOVA calculations:

- $H_0 = $ There are no significant differences in the means.
- $H_A = $ There are significant differences in the means.

Step three: Determine an average of the variances for the regions of each ecoregion based model.

All descriptive statistics equations and ANOVA equations used are from Kachigan (1986). Any additional statistical techniques are discussed under the methods for that particular regional model.

Once the statistics were calculated on ECO1, a second model (ECO2, Figure 7) was developed using the findings from the first statistical test as a guide.

**Figure 7:** Map of the ECO2 model honey production regions.
States that were a judgement call between two regions were moved to the alternate region and statistical analysis repeated. This was done to find the combination that yielded the lowest average variance indicating a dataset with more homozygosity. The number of regions was kept constant at 10, and all but two of the multi-state regions were adjusted. California and Florida remained independent regions.

Two additional models based on ECO1 and ECO2 were created that combined the single state regions of Florida and California with the regions that most closely resembled them in ecoregion makeup. This reduced the total number of regions to eight and eliminated the single state regions. These adjusted models were named ECO1A (Figure 8) and ECO2A (Figure 9). Statistical analysis was conducted on the adjusted models.

![Map of ECO1A model honey production regions](image)

**Figure 8:** Map of the ECO1A model honey production regions.
Cluster Based Models

Background Information

Creating honey production units using geographical and ecoregion data by a best-fit approach has limitations and drawbacks. One may form regions with an unknown bias due to a model they have seen in the past, personal bias, or misinterpretation of the available data or misinterpretation of the protocol (Hoyt and Kearns 1999). A common factor in all human created models is that all of the states in a region are contiguous. This makes sense to the human mind and creates easy to follow and understand regional maps.

The natural world is hardly ever that simple, however, and noncontiguous states may have similar honey production. Although the states are in different ecoregions, with differing flora and climate, there is no rule that dictates that dissimilar ecoregions cannot have similar honey production. This would create scattered, non-contiguous honey production.
production regions - something a human is not likely to create. In order to discover if any such non-contiguous regions exist and whether they may describe a more statistically accurate model, computer-generated modeling was employed.

Procedures for the Analysis of the Models

Do any of the cluster based regional honey production models (COMP6, COMP7, COMP8, COMP9, COMP10) or the Page et al. (1987) model demonstrate a true regional honey production model for the United States? To obtain an answer for this question a research protocol was developed to evaluate existing geographical models based on the average variance for each model.

The first step in this process was to determine descriptive statistics for each region in each model and perform an ANOVA for each model. Once the statistical calculations were completed, the models were compared using the following criteria:

- no split states
- a low number of regions ($\leq 10$)
- low average variance for the model
- statistical significance ($p \leq 0.05$) for the model

Those models with significant means that also met all other criteria were deemed acceptable models and considered in the determination for an improved regional honey production model for the United States.

Specific Statistical Assumptions and Methods

There are two additional limiting assumptions associated with this phase of the study: 1. It is assumed that the states for which there is no data would not have an effect on the regions created.
2. It is assumed that the \( k \)-means models and the Page et al. (1987) model are in parity even though they were created using different data sets.

Generating a mathematical model is computationally complex and requires the use of specialized computer software. For this study the \( k \)-means clustering application of XLStat was employed. Cluster analysis renders a dataset into similar clusters according to the algorithm utilized. There are many forms of cluster analysis. For this study, the type of cluster analysis algorithm used is \( k \)-means clustering (Fraley and Raftery 1998).

In \( k \)-means clustering a dataset is partitioned so that \( n \) observations are partitioned into \( k \) clusters (Jain and Dubes 1988). The clusters are formed around similar means, known as centroids, determined by the clustering algorithm. The value of \( k \) can be any number of clusters the researcher chooses as long as the number of clusters is \( \leq n \) (Jain and Dubes 1988). For each \( k \)-value there will be several centroid arrangements possible so it is important to run multiple repetitions and iterations for each \( k \) (Kaufman and Rousseeuw 1990, Fraley and Raftery 1998, Janowitz 2010). All \( k \)-means calculations in this study were conducted with 50 repetitions and 500 iterations.

Formula for \( k \)-means clustering:

\[
\arg\min_S \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2
\]

Where for a given set of yearly state honey production averages \((x_1, x_2, ..., x_n)\), where each yearly state honey production average is a \( d \)-dimensional real vector, \( k \)-means clustering aims to partition the \( n \) states into \( k \) sets \((k \leq n) S = \{S_1, S_2, ..., S_k\}\) so as to minimize the within-cluster sum of squares, and where \( \mu_i \) is the mean of points in \( S_i \) (Jain and Dubes 1988).
Five cluster models were developed with 6, 7, 8, 9, and 10 clusters (COMP6, COMP7, COMP8, COMP9, COMP10, Figures 10-14). These cluster numbers give parity with the other models in the study. The $k$-means clustering algorithm used for this study was from the XLStat package. Descriptive statistics and ANOVA were calculated for all models and comparisons made.

Statistics were calculated for all models using the following protocol:

Step one: Calculate descriptive statistics ($\bar{X}, s^2$) for each region of the five computer generated regional honey production models.

Step two: Test for significant differences in means using ANOVA for all regions in the five computer generated models. ANOVA conducted at a 0.05 level of confidence.

Hypothesis for all ANOVA calculations:

$H_0$ = There are no significant differences in the means.

$H_A$ = There are significant differences in the means.

Step three: Determine an average of the variances for the regions of each computer generated model.

All descriptive statistics equations and ANOVA equations used are from Kachigan, 1986.

The maps for the COMP regions show the states for which no data are available in dark grey because the computer program could not assign them to a region. For this reason, the maps for the COMP models appear incomplete compared to other models in this study.
Figure 10: Map of COMP6 model honey production regions.

Figure 11: Map of the COMP7 model honey production regions.
Figure 12: Map of the COMP8 model honey production regions.

Figure 13: Map of the COMP9 model honey production regions.
In addition to the five cluster models, an existing linear regression model from Page et al. (1987) was examined. This model consist of 13 regions including two single state regions (Figure 15) and utilized a dataset spanning the years 1939-81. Data were available for all states. For this study, the two single state regions were combined with adjacent regions with which the states had a positive correlation. This technique was described in Page et al. (1987) as a way to remove the single state regions. The result of this combination was an 11 region model with no single state regions and all regions contiguous.

**Figure 14:** Map of the COMP10 model honey production regions.
Results

Geographically Based Models

The results of the statistical analysis for the Nye model are found in Table 1.

Table 1: Results of the Nye model statistical analysis.

<table>
<thead>
<tr>
<th>Region</th>
<th>N</th>
<th>$\bar{X}$</th>
<th>$S^2$</th>
<th>F</th>
<th>p</th>
<th>F Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. East</td>
<td>116</td>
<td>48.8</td>
<td>384.550</td>
<td>61.690</td>
<td>3.00E-04</td>
<td>2.686</td>
</tr>
<tr>
<td>N. Central</td>
<td>261</td>
<td>63.8</td>
<td>285.588</td>
<td>15.542</td>
<td>1.50E-18</td>
<td>1.975</td>
</tr>
<tr>
<td>S. East</td>
<td>319</td>
<td>60.9</td>
<td>477.726</td>
<td>47.602</td>
<td>1.12E-56</td>
<td>1.861</td>
</tr>
<tr>
<td>Plains</td>
<td>116</td>
<td>74.7</td>
<td>431.106</td>
<td>19.696</td>
<td>2.52E-10</td>
<td>2.686</td>
</tr>
<tr>
<td>Mountain</td>
<td>145</td>
<td>60.5</td>
<td>362.541</td>
<td>32.671</td>
<td>3.15E-19</td>
<td>2.436</td>
</tr>
<tr>
<td>S. West</td>
<td>29</td>
<td>57.7</td>
<td>133.993</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>West</td>
<td>87</td>
<td>50.8</td>
<td>156.788</td>
<td>4.924</td>
<td>0.001</td>
<td>3.105</td>
</tr>
</tbody>
</table>

**Average Variance:** 318.899

**Average p-Value:** 2.17E-04
The Nye model has the highest average variance (318.899) of the geographical models. Due to the lack of data for states in the Southwest region, a complete statistical analysis could not be conducted. Another flaw with this model is that it does not follow political state boundaries and splits several states into one or more honey production regions. The current system for collecting and reporting honey production data are on a state basis rather than a county basis. Because the Nye model splits states, it creates a situation where calculating meaningful statistics is not possible for six of the seven regions. The average p-value was significant \((p = 2.17E-04 \leq 0.05)\) which indicates that the model is valid, but for these reasons given above this model cannot be considered a true representation of regional United States honey production.

The results for the statistical analysis for the American Bee Journal model are found in Table 2.

**Table 2:** Results of the American Bee Journal model statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n)</td>
<td>(\bar{X})</td>
</tr>
<tr>
<td>N. East</td>
<td>145</td>
<td>43.3</td>
</tr>
<tr>
<td>Mideast</td>
<td>146</td>
<td>46.6</td>
</tr>
<tr>
<td>S. East</td>
<td>116</td>
<td>67.7</td>
</tr>
<tr>
<td>E. Central</td>
<td>145</td>
<td>64.4</td>
</tr>
<tr>
<td>W. Central</td>
<td>203</td>
<td>72.1</td>
</tr>
<tr>
<td>S. West</td>
<td>116</td>
<td>74.4</td>
</tr>
<tr>
<td>Mountain</td>
<td>145</td>
<td>60.5</td>
</tr>
<tr>
<td>West</td>
<td>87</td>
<td>50.8</td>
</tr>
<tr>
<td><strong>Average Variance</strong></td>
<td><strong>298.670</strong></td>
<td><strong>Average p-Value</strong></td>
</tr>
</tbody>
</table>

The American Bee Journal model had the second highest average variance (298.670) of the geographical models. The average p-value was significant \((p = 1.88E-03 \leq 0.05)\) indicating that this model is valid. This is an acceptable model as it contains no split-states. There are also no regions in this model for which statistical calculations
cannot be conducted. This model represents a viable regional view of United States honey production.

The results for the statistical analysis for the Bee Culture model are found in Table 3.

Table 3: Results of the Bee Culture model statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Ẋ</th>
<th>S²</th>
<th>F</th>
<th>p</th>
<th>F Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>145</td>
<td>48.3</td>
<td>322.657</td>
<td>51.869</td>
<td>1.14E-26</td>
<td>2.436</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
<td>46.6</td>
<td>116.575</td>
<td>3.814</td>
<td>0.006</td>
<td>2.436</td>
</tr>
<tr>
<td>3</td>
<td>174</td>
<td>72.8</td>
<td>467.789</td>
<td>27.418</td>
<td>3.12E-20</td>
<td>2.268</td>
</tr>
<tr>
<td>4</td>
<td>232</td>
<td>66</td>
<td>268.905</td>
<td>9.728</td>
<td>1.44E-10</td>
<td>2.051</td>
</tr>
<tr>
<td>5</td>
<td>174</td>
<td>74</td>
<td>408.831</td>
<td>15.227</td>
<td>2.48E-12</td>
<td>2.268</td>
</tr>
<tr>
<td>6</td>
<td>116</td>
<td>60.5</td>
<td>245.087</td>
<td>29.047</td>
<td>5.73E-14</td>
<td>2.686</td>
</tr>
<tr>
<td>7</td>
<td>116</td>
<td>49.9</td>
<td>146.531</td>
<td>4.302</td>
<td>0.007</td>
<td>2.686</td>
</tr>
</tbody>
</table>

**Average Variance:** 282.339  **Average p-Value:** 1.86E-03

The Bee Culture model had the lowest average variance (282.339) of the geographical models. This indicates homogeneity of data between states from the same region, which implies a similarity in honey production. The p-value was also significant (p = 1.86E-03 ≤ 0.05) indicating that the model is valid. These statistical findings coupled with this model’s low number of regions, lack of split states, shows that the Bee Culture model is an acceptable geographically based regional representation of United States honey production.
Ecoregion Based Models

The results of the statistical analysis for the ECO1 model are shown in Table 4.

**Table 4:** Results of the ECO1 model statistical analysis.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>2</td>
<td>232</td>
</tr>
<tr>
<td>3</td>
<td>116</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
</tr>
<tr>
<td>5</td>
<td>116</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
</tr>
<tr>
<td>7</td>
<td>174</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
</tr>
</tbody>
</table>

**Average Variance:** 299.63  **Average p-Value** 0.123

Note: Highlighted values are not significant at $p \leq 0.05$.

The ECO1 model had an average variance of 299.630. In region 4, there is no evidence to reject the null hypothesis because $p = 0.861 > 0.05$. In other words, there are nonsignificant means across the regions. This makes the overall p-value ($p = 0.861 > 0.05$) and a nonsignificant overall p-value ($p = 0.123 > 0.05$); suggesting that the overall ECO1 model is insignificant. This may be linked to a flaw in the dataset. In this model, 30% of the regions cannot have complete statistical analysis conducted due to them containing a single state, or only having one state with available data. This lack of important statistical information renders it impossible to properly evaluate this model.
The results of the statistical analysis for the ECO2 model are shown in Table 5.

**Table 5**: Results of the ECO2 model statistical analysis.

<table>
<thead>
<tr>
<th>N</th>
<th>X</th>
<th>S^2</th>
<th>F</th>
<th>p</th>
<th>F Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>174</td>
<td>64.3</td>
<td>404.378</td>
<td>47.614</td>
<td>1.71E-30</td>
</tr>
<tr>
<td>2</td>
<td>203</td>
<td>44.8</td>
<td>121.741</td>
<td>8.726</td>
<td>1.98E-08</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>70.6</td>
<td>480.815</td>
<td>30.680</td>
<td>2.47E-18</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>58.4</td>
<td>225.150</td>
<td>0.150</td>
<td>0.861</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>84.4</td>
<td>428.744</td>
<td>1.703</td>
<td>0.188</td>
</tr>
<tr>
<td>6</td>
<td>203</td>
<td>65.2</td>
<td>205.577</td>
<td>4.711</td>
<td>2.00E-04</td>
</tr>
<tr>
<td>7</td>
<td>116</td>
<td>47.2</td>
<td>83.326</td>
<td>0.760</td>
<td>0.519</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>83.5</td>
<td>273.116</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>56.4</td>
<td>308.887</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>57.7</td>
<td>133.993</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Average Variance:** 266.572  **Average p-Value:** 0.224

Note: Highlighted values are not significant at \( p \leq 0.05 \).

The average variance of the ECO2 model was 266.573. Regions with nonsignificant means increased from one to three and the ECO2 model was also insignificant because the average p-value is greater than the threshold (\( p = 0.224 > 0.05 \)). This model is not viable. It also suffers from having 30% of its regions incapable of having complete statistical analysis conducted. Therefore, this model is not an acceptable regional honey production model for the United States.
The results of the statistical analysis for the ECO1A model are shown in Table 6.

**Table 6:** Results of the ECO1A model statistical analysis.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>2</td>
<td>232</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
</tr>
<tr>
<td>5</td>
<td>116</td>
</tr>
<tr>
<td>6</td>
<td>87</td>
</tr>
<tr>
<td>7</td>
<td>203</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
</tr>
</tbody>
</table>

**Average Variance** 300.0285  **Average p-Value** 0.123

Note: Highlighted values are not significant at $p \leq 0.05$.

This model had an average variance of 300.029, similar to the ECO1 model from which it was created. Combining the single state regions of California and Florida with regions that bordered them did not have any great effect on the statistical outcome. The number of nonsignificant regional means remained unchanged as indicated by the higher p-values. The average p-value also remained high ($p = 0.123 > 0.05$) indicating a nonviable model. This model must be rejected as an acceptable representation of United States regional honey production.
The results of the statistical analysis for the ECO2A model are shown in Table 7.

**Table 7:** Results of the ECO2A model statistical analysis.

<table>
<thead>
<tr>
<th>N</th>
<th>$\bar{X}$</th>
<th>$S^2$</th>
<th>F</th>
<th>p</th>
<th>F Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>174</td>
<td>64.3</td>
<td>404.378</td>
<td>47.614</td>
<td>1.71E-30</td>
</tr>
<tr>
<td>2</td>
<td>203</td>
<td>44.8</td>
<td>121.741</td>
<td>8.726</td>
<td>1.98E-08</td>
</tr>
<tr>
<td>3</td>
<td>174</td>
<td>72.8</td>
<td>467.579</td>
<td>27.418</td>
<td>3.12E-20</td>
</tr>
<tr>
<td>4</td>
<td>87</td>
<td>58.4</td>
<td>225.150</td>
<td>0.150</td>
<td>0.861</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>84.4</td>
<td>428.744</td>
<td>1.703</td>
<td>0.188</td>
</tr>
<tr>
<td>6</td>
<td>203</td>
<td>65.2</td>
<td>205.577</td>
<td>4.711</td>
<td>2.00E-04</td>
</tr>
<tr>
<td>7</td>
<td>145</td>
<td>49.0</td>
<td>140.186</td>
<td>4.165</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>29</td>
<td>57.7</td>
<td>133.993</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

**Average Variance:** 265.9185  **Average p-Value** 0.150

Note: Highlighted values are not significant at $p \leq 0.05$.

This model was nearly identical statistically to the ECO2 model from which it was developed, and had a nearly identical average variance (265.9185). There were two regions with nonsignificant means and the overall model is nonviable. This model is also not an acceptable regional honey production model for the United States.
Cluster Based Models

The results of the statistical analysis for the COMP6 model are shown in Table 8.

**Table 8:** Results of the COMP6 model statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
<td>54.5</td>
</tr>
<tr>
<td>2</td>
<td>174</td>
<td>63.9</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
<td>54.6</td>
</tr>
<tr>
<td>4</td>
<td>348</td>
<td>77.6</td>
</tr>
<tr>
<td>5</td>
<td>261</td>
<td>43.0</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>78.1</td>
</tr>
</tbody>
</table>

**Average Variance:** 257.228  **Average p-Value:** 0.008

This model had an average variance of 257.228. The p-values for the five regions for which ANOVA could be conducted and all indicate significant means (p < 0.05) and an overall significant model (p = 0.008 < 0.05). This indicates that this is a viable model. There was one single state region and all multi-state regions were noncontiguous. The computer generated clusters showed an improvement in average variance even with a low number of regions. The low average variance indicates that this cluster based model has good data homogeneity within the regions. This model is an acceptable representation of regional United States honey production.
The results of the statistical analysis for the COMP7 model are shown in Table 9.

**Table 9:** Results of the COMP7 model statistical analysis.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>2</td>
<td>174</td>
</tr>
<tr>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td>4</td>
<td>116</td>
</tr>
<tr>
<td>5</td>
<td>377</td>
</tr>
<tr>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>7</td>
<td>116</td>
</tr>
</tbody>
</table>

**Average Variance:** 245.482  
**Average p-Value:** 0.078

Note: Highlighted values are not significant at  \( p \leq 0.05 \).

This model had an average variance of 245.482. COMP7 had four regions with nonsignificant means since their corresponding p-values are high (\( p > 0.05 \)) and the overall average p-value indicates that the model is barely insignificant (\( p = 0.078 > 0.05 \)). This is a nonviable model at p-value of 0.05, but viable at p-value of 0.10.

Results of the statistical analysis for the COMP8 model are shown in Table 10.

**Table 10:** Results of the COMP8 model statistical analysis.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td><strong>X</strong></td>
</tr>
<tr>
<td>1</td>
<td>203</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>174</td>
</tr>
<tr>
<td>5</td>
<td>174</td>
</tr>
<tr>
<td>6</td>
<td>174</td>
</tr>
<tr>
<td>7</td>
<td>87</td>
</tr>
<tr>
<td>8</td>
<td>116</td>
</tr>
</tbody>
</table>

**Average Variance:** 235.910  
**Average p-Value:** 0.002
This model had an average variance of 235.910. This model also included a single state region for which ANOVA could not be conducted. The p-values are small and therefore indicate significant means for all regions which ANOVA calculations were possible. The overall average p-value also indicates that the overall model is significant \( (p = 0.002 < 0.05) \). This model is an acceptable representation of regional honey production in the United States.

Results of the statistical analysis for the COMP9 model are shown in Table 11.

**Table 11:** Results of the COMP9 model statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th><strong>Descriptive Statistics</strong></th>
<th><strong>ANOVA</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N )</td>
<td>( \bar{X} )</td>
</tr>
<tr>
<td>1</td>
<td>145</td>
<td>53.6</td>
</tr>
<tr>
<td>2</td>
<td>87</td>
<td>57.8</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>64.6</td>
</tr>
<tr>
<td>4</td>
<td>145</td>
<td>71.2</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
<td>78.6</td>
</tr>
<tr>
<td>6</td>
<td>290</td>
<td>43.4</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>58.6</td>
</tr>
<tr>
<td>8</td>
<td>116</td>
<td>84.8</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>78.1</td>
</tr>
</tbody>
</table>

**Average Variance:** 258.025  **Average p-Value:** 0.159

Note: Highlighted values are not significant at \( p \leq 0.05 \).

This model had an average variance of 258.025. There was one single state region for which ANOVA could not be calculated. Two of the regions had nonsignificant means; leading to a high average p-value \( (p = 0.159 > 0.05) \). This model is not viable and is not an acceptable model of regional honey production in the United States.
Results of the statistical analysis for the COMP10 model are shown in Table 12.

**Table 12:** Results of the COMP10 model statistical analysis.

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>( \bar{X} )</td>
</tr>
<tr>
<td>1</td>
<td>145</td>
<td>59.2</td>
</tr>
<tr>
<td>2</td>
<td>58</td>
<td>57.1</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>72.4</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>66.1</td>
</tr>
<tr>
<td>5</td>
<td>174</td>
<td>84.8</td>
</tr>
<tr>
<td>6</td>
<td>232</td>
<td>48.3</td>
</tr>
<tr>
<td>7</td>
<td>58</td>
<td>60.7</td>
</tr>
<tr>
<td>8</td>
<td>145</td>
<td>40.2</td>
</tr>
<tr>
<td>9</td>
<td>58</td>
<td>71.2</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>59.9</td>
</tr>
</tbody>
</table>

**Average Variance:** 215.96270  **Average p Value** 0.228

Note: Highlighted values are not significant at \( p \leq 0.05 \).

This model had an average variance of 215.962 and had five regions with insignificant means since their p-values are greater than the threshold. There was one single state region for which ANOVA could not be calculated. The overall average p-value indicates nonsignificant means because \( p = 0.228 > 0.05 \). This model must be rejected as an acceptable model for regional honey production in the United States.
Results of the statistical analysis for the model from Page et al (1987) are shown in Table 13.

**Table 13:** Results of the Page et al. (1987) model statistical analysis

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Statistics</th>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Ẋ</td>
</tr>
<tr>
<td>1</td>
<td>58</td>
<td>48.1</td>
</tr>
<tr>
<td>2</td>
<td>116</td>
<td>56.1</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>72.7</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>59.0</td>
</tr>
<tr>
<td>5</td>
<td>29</td>
<td>73.9</td>
</tr>
<tr>
<td>6</td>
<td>261</td>
<td>64.9</td>
</tr>
<tr>
<td>7</td>
<td>87</td>
<td>51.9</td>
</tr>
<tr>
<td>8</td>
<td>174</td>
<td>72.7</td>
</tr>
<tr>
<td>9</td>
<td>58</td>
<td>48.2</td>
</tr>
<tr>
<td>10</td>
<td>29</td>
<td>34.6</td>
</tr>
<tr>
<td>11</td>
<td>87</td>
<td>47.2</td>
</tr>
</tbody>
</table>

**Average Variance:** 244.792 **Average p-Value** 0.090

Note: Highlighted values are not significant at  \( p \leq 0.05 \).

This model had the most regions of any model in the study. The average variance was 244.792. There were two regions with nonsignificant means. The overall model is barely insignificant at 0.05, but significant at 0.10 (\( p = 0.090 > 0.05 \)). Although the model did not contain any single state regions, there were two regions for which ANOVA could not be calculated due to a lack of state data. This model is not viable and must be rejected as an acceptable representation of regional honey production in the United States.

**Comparison to Determine an Improved Model for Regional Honey Production**

Table 14 summarizes the models analyzed in this study. The preferred model for this study was defined as having no split states, a low number of regions...
(≤ 10), low average variance for the model, and the data for the model would show statistically significant differences at p-value ≤ 0.05.

**Table 14:** Summary of all models analyzed.

<table>
<thead>
<tr>
<th>Model</th>
<th># Regions</th>
<th>Contiguous Regions</th>
<th>Split States</th>
<th>Average S²</th>
<th>Average p-Value</th>
<th># of Single State Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABJ</td>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>298.670</td>
<td>0.002</td>
<td>0</td>
</tr>
<tr>
<td>BC</td>
<td>7</td>
<td>Yes</td>
<td>No</td>
<td>282.339</td>
<td>1.86E-03</td>
<td>0</td>
</tr>
<tr>
<td>NYE</td>
<td>7</td>
<td>Yes</td>
<td>Yes</td>
<td>318.899</td>
<td>2.17E-04</td>
<td>0</td>
</tr>
<tr>
<td>ECO 1</td>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>299.63</td>
<td>0.123</td>
<td>2</td>
</tr>
<tr>
<td>ECO 2</td>
<td>10</td>
<td>Yes</td>
<td>No</td>
<td>266.573</td>
<td>0.224</td>
<td>2</td>
</tr>
<tr>
<td>ECO 1A</td>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>300.029</td>
<td>0.123</td>
<td>0</td>
</tr>
<tr>
<td>ECO 2A</td>
<td>8</td>
<td>Yes</td>
<td>No</td>
<td>265.919</td>
<td>0.15</td>
<td>0</td>
</tr>
<tr>
<td>COMP 6</td>
<td>6</td>
<td>No</td>
<td>No</td>
<td>257.228</td>
<td>0.008</td>
<td>1</td>
</tr>
<tr>
<td>COMP 7</td>
<td>7</td>
<td>No</td>
<td>No</td>
<td>245.482</td>
<td>0.078</td>
<td>0</td>
</tr>
<tr>
<td>COMP 8</td>
<td>8</td>
<td>No</td>
<td>No</td>
<td>235.91</td>
<td>0.002</td>
<td>1</td>
</tr>
<tr>
<td>COMP 9</td>
<td>9</td>
<td>No</td>
<td>No</td>
<td>258.026</td>
<td>0.159</td>
<td>1</td>
</tr>
<tr>
<td>COMP 10</td>
<td>10</td>
<td>No</td>
<td>No</td>
<td>215.963</td>
<td>0.228</td>
<td>1</td>
</tr>
<tr>
<td>PAGE</td>
<td>11</td>
<td>Yes</td>
<td>No</td>
<td>244.792</td>
<td>0.09</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: Highlighted values are not significant at p ≤ 0.05.

Starting with the requirement that the model must have a statistically significant average p-value (p ≤ 0.05) the following models are declared nonviable:

ECO1, ECO2, ECO1A, ECO2A, COMP7, COMP9, COMP 10,and the model from Page et al. (1987)

After eliminating all of the statistically nonviable models only five models remain:

The Nye Model, the American Bee Journal model, the Bee Culture model, COMP6 and COMP 8

The Nye model can be eliminated for having several split states. This leaves the American Bee Journal and Bee Culture models from the geographical models category
and two cluster-based models: COMP 6 and COMP 8. Looking solely at the last requirement (low average variance), both geographical models and COMP6 can be eliminated as they all have a higher average variance than COMP8. As determined by the criteria set forth in this study COMP8 is the statistically acceptable improved model of honey production in the United States.
Conclusions and Implications

According to the assumptions set for this study, COMP8 is the improved model for regionalizing honey production in the United States. However, the limiting assumptions must be reviewed before COMP8 can be proclaimed an improved honey production model. There were six limiting assumptions made for this study. In reality, all six limiting assumptions for this study are false, but had to be considered true in order to calculate statistics. The assumptions and how each could affect the outcome of the study are discussed below.

The first limiting assumption stated all qualifying apiaries were included in the dataset. This is assumption is false for several reasons. First, the USDA collects only a sample of the data. The data are extrapolated, and a honey yield is calculated for each state, but this number is an estimate based on a sample. Statistical techniques are adjusted to work with this sample data, but it is never as accurate as working with complete population data. However, the USDA data collection technique is consistent across all states, which should mitigate errors.

Another aspect of data collection are the regional differences in how beekeepers react to volunteering data. In states where there is mandatory beekeeper reporting, beekeepers may be more likely to partake in NASS surveys. Some states do not require beekeepers to report hive numbers and honey sales on the state level. It has been my personal experience that beekeepers in these states are often unwilling to release hive numbers or honey production data. There may be several reasons for this including distrust of the government, insurance liability, not wanting to report taxable income, etc. Even if there were mandatory national reporting, some beekeepers would still evade it
just as some people evade filing yearly income taxes. Because the same dataset is common to all models in this study, any anomalies should affect all models equally.

The second and the sixth limiting assumptions both dealt with split states. The Nye model is the only model in the study to which these assumptions apply. Because the USDA data collection protocol collects data on a state basis and not on a county basis, there is no way to deal with the ramification of split states. Split states can be placed into a region as was done in this study with California, New York and Pennsylvania; or they can be dropped from the study as was done with Texas. Neither of these options is statistically valid. Models with split states cannot be considered for regional honey production under the current data collection scheme. Therefore, the Nye model is eliminated from further discussion.

Assumption number three addressed the production of apiaries with fewer than five hives. It was assumed their production is inconsequential to the total yearly yield for a state. Because there are so many hobby and backyard beekeepers this is likely not the case. There is no way to collect or estimate this yield due to lack of mandatory reporting in all states. The USDA, in 2015, started to collect data from apiaries with fewer than five hives that qualified as a farm. It may be possible in the future to include this data into honey production modeling projects. However, the new data still does not take into account honey production from apiaries that do not qualify as a farm. Examples of such apiaries include beekeepers who do not sell more than $1,000 in farm products either because they do not produce enough or because they give most or all of their honey away.
For this study, the lack of data from small apiaries should not make a difference because the dataset was consistent in collection protocol and parameters across all states. However, it would be more correct to call the models created and tested “regional honey production models for apiaries with more than five hives that qualified as a farm.” In the future, the models can be adjusted as data are made available from the smaller apiaries. There will likely always be a lack of reliable data or honey production estimates for backyard and hobby beekeepers.

The fourth limiting assumption deals with honey produced in one state, but reported as produced in another state. The problem of misreported honey is mainly due to large commercial migratory beekeepers who move their hives interstate as they follow the nectar flow north, or are moving colonies for pollination contracts. Commercial beekeepers can have several hundred to several thousand hives. The amount of honey reported as produced in the wrong state may be substantial, but we could find no research estimating the amount. Although there may be an effect on honey yield data there is no way to determine how much of an effect. For this study, the assumption of there being no improperly reported honey yield must be followed and is consistent across all models. It is unlikely that a solution to this data collection problem will occur anytime in the future.

The final limiting assumption deals with the missing data from several states due to privacy issues. Data are unavailable for the states of Connecticut, Delaware, Maryland, Massachusetts, Nevada, New Hampshire, New Mexico, Oklahoma, Rhode Island and South Carolina. There is a major lack of data for the northeast and the southwest making meaningful statistical analysis impossible for honey production regions located in those areas. The lack of data are consistent over all models so the analysis of this study is valid,
although incomplete. However, the assumption that the missing state data has no effect on regional makeup, or the statistical calculations for a region, is false.

The model by Page et al. (1987) is the only model in this study that was developed using a dataset from all 48 contiguous states. The data used to develop this model were collected prior to 1986. After 1986, states were allowed to request data not be made public. This model may be the most accurate model because it was created using a complete dataset for all contiguous states.

Unfortunately, it is impossible to determine what impact the missing state data would have had on the regional statistics of each model. Although the regions created in the ECO models would have remained the same, the statistical data calculated would have changed. The COMP models would been impacted to the greatest extent. With a complete dataset, all states would have been assigned to a region and complete statistical calculations would have been possible. Models that had one or more single state regions may have had no single state regions, and continuity of regions may have increased. Although the Nye, American Bee Journal and Bee Culture models had preset regions, a complete state dataset would have allowed for a more accurate statistical analysis.

After evaluating the effect of the limiting assumptions on the outcome of the study COMP8 still meets the parameters for an improved honey production model. This holds true only for this study, the dataset utilized, the limiting assumptions made, and the parameters followed to define the best-fit model. Other models analyzed were also statistically suitable models, but had higher average variance.

The COMP8 model was statistically strong. This model may show a true image of honey regions in the United States, but lack simplicity for dissemination of information to
the beekeeping public. The lack of contiguous regions and a single state region make this model more cumbersome and complicated for public use. Though not as useful for relaying honey production information, the COMP8 model may be more useful for scientific research such as economic studies of honey production, studies of how honey production is similar across ecoregions, or tracking changes in honey production due to disease or pests.

The COMP8 model is an improved model for regional honey production in the United States. It is far different in regional layout and makeup than the popular American Bee Journal and Bee Culture models. This will require education and training of the beekeeping public in order for this model to gain wide acceptance. Model COMP8 may also require an adjustment in the way regional honey production, and other pertinent regional information, is disseminated by popular beekeeping journals to the beekeeping public.

By determining regional state composition by $k$-means clustering, the statistical strength of the COMP8 model improved over both the popular American Bee Journal and Bee Culture models. This increase in statistical validity indicates that this model may be used to detect shifts in honey production that could signal a pest or disease outbreak (Page et al. 1987). Although COMP8 is a new way of looking at regional honey production in the United States, it is a statistically sound and improved model.

A complete state dataset is required in order to prove or disprove the validity of any of the models created or tested in this study. It may be possible to obtain the missing state data from the USDA under an agreement of confidentiality. This will allow the creation of complete COMP models and for complete statistical calculations for
all regions in all models. The ECO models and the ABJ model would not change regionally as the regions were set independent of honey production records. The outcome of the statistical analysis would differ, however, and may indicate the models are better or worse than originally calculated. Of special interest is how the Page et al. (1987) model, created with complete state data from 1939-81, would change when tested with recent data. Any changes would indicate shifts in honey production over time that may be correlated with Varroa mite spread, CCD, or changes in honey production due to climate change (Page et al. 1987). In the future, a solution that allows dissemination of the state data to the public while still maintaining beekeeper privacy may be developed. Until then, regional honey production models will be based on incomplete data. The lack of a complete dataset is the greatest limiting factor of this study.

The ecoregion models could also be enhanced by the use of GIS technology to aid in the creation of regions. By using GIS, the percentage of area covered by an ecoregion in a state could be determined. This data could then be used to better group states into regions. This may result in noncontiguous regions (much like k-means clustering), but may give a more accurate ecoregion model.

The statistical study of honey bees and their products is just now gaining ground. With the worldwide decline in pollinator species, research opportunities in the field are sure to increase. This study has laid basic groundwork that can be built upon to help researchers learn more about the fascinating creature known as the honey bee.


