ABSTRACT

GAYO, JAVIER. Species Authenticity and Detection of Economic Adulteration of Atlantic Blue Crab Meat Using VIS/NIR Spectroscopy. (Under the direction of Dr. S. A. Hale and Dr. S. M. Blanchard)

The application of Visible and Near-Infrared (VIS/NIR) spectroscopy to determine economic adulteration of crabmeat was determined. Crabmeat samples were adulterated in 10% increments according to weight. The adulterants chosen were surimi-based imitation crabmeat, due to its low cost and availability, and blue swimmer crabmeat, the most prevalent type of crabmeat imported into the United States. Several data pre-treatments and different chemometric analyses, Partial Least Squares (PLS), Principal Component Regression (PCR), and Multiple Linear Regression (MLR), were investigated to determine the predictive ability of VIS/NIR spectroscopy in detecting economic adulteration and species authenticity. In addition, wavelength variables selected by a genetic algorithm were evaluated to improve predictive ability of economic adulteration of crabmeat.

Absorption spectra of adulterated samples were dominated by water overtones. Absorption decreased for increasing level of adulteration. PLS was favored over PCR due to its predictive ability and lower number of latent variables used in model development. The first derivative data pre-treatment generated the best results, though similar results were gathered with the untreated data. First derivative data, using data from a correlogram, generated the best PLS model to detect economic adulteration of crabmeat adulterated with surimi-based imitation crabmeat. Economic adulteration

percentage was predicted within 2.5%. The second derivative data generated the highest errors and, thus, was not deemed an appropriate pre-treatment method.

For samples adulterated with blue swimmer, the first derivative data also generated the best results utilizing the full spectrum with a Standard Error of Calibration (SEC) and Standard Error of Prediction (SEP) of 5.64. Using a partitioned spectrum, however, the second derivative data performed better (SEC and SEP of 4.91 and 5.17, respectively). Regardless of the data pre-treatment, VIS/NIR spectroscopy was able to detect species authenticity and economic adulteration to less than 6% for samples adulterated with blue swimmer crabmeat. Variable selection via genetic algorithm enabled MLR to detect economic adulteration, making this the best overall method due to its low SEC and SEP (4.21 and 3.98, respectively). All of these findings provide a baseline for the design and development of a fast, reliable, and accurate technology for on-line detection of economic adulteration of crabmeat.

SPECIES AUTHENTICITY AND DETECTION OF ECONOMIC ADULTERATION OF ATLANTIC BLUE CRAB MEAT USING VIS/NIR SPECTROSCOPY

by

JAVIER GAYO

A dissertation submitted to the Graduate Faculty
of North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

BIOLOGICAL AND AGRICULTURAL ENGINEERING

Raleigh

2006

APPROVED BY:

Dr. S. Andrew Hale	Dr. Susan M. Blanchard
Co-Chair of Advisory Committee	**Co-Chair of Advisory Committee**
Dr. Frank Abrams	Dr. Donn Ward

DEDICATION

This work is dedicated to my parents. Thank you for your endless support and encouragement.

BIOGRAPHY

Javier Gayo was born in Bacelona, Spain in 1976. After a temporary move to the United States from 1988 to 1992, he returned home for a year and came back to finish high school and attend North Carolina State University. In 1999, he received his Bachelor of Science degree in Biological Engineering with a concentration in Biomedical Engineering. He continued his academic studies at NCSU and received his Master of Science degree in 2002 in Biological and Agricultural Engineering. His thesis was titled "Software Analysis Techniques for Odor Analysis and Classification Using the Electronic Nose."

ACKNOWLEDGEMENTS

I want to thank my co-advisors, Dr. Andy Hale and Dr. Susan Blanchard, for their support and guidance during my graduate studies. Thank you for your patience while I narrowed down the many directions initially discussed for my doctoral research. Thanks also to Dr. Donn Ward for his "blue crab" industry advice and to Dr. Frank Abrams for his help in emphasizing the engineering application(s) behind this research. And last, but not least, I would like to thank Dr. Riccardo Leardi for his quick replies to my many questions concerning spectral variable selection using genetic algorithms.

TABLE OF CONTENTS

LIST OF FIGURES

2.0 REVIEW OF LITERATURE

3.0 MANUSCRIPT #1

4.0 MANUSCRIPT #2

5.0 MANUSCRIPT #3

APPENDIX

LIST OF TABLES

APPENDIX

1.0 INTRODUCTION

The blue crab industry, which mainly processes the Atlantic blue crab (*Callinectes sapidus*), represents an economically important segment of the United States seafood industry. During the last decade, however, domestic landings have decreased significantly due to declining crab populations, poor water quality, negative impacts of hurricane seasons, and exploitation of available resources. Over the years, the accumulation of these factors has led to an increase in industry production costs. In addition, the U. S. blue crab industry has been negatively affected by an increase in crabmeat imports of blue swimmer crabs (*Portunus pelagicus*) from Asia and the Indo-Pacific.

In comparison with the domestic crab, imported crabs are cheaper economically and of poorer quality. The increase in blue crab industry production costs and availability of an imported economically cheaper type of crab have tempted some seafood processors to mix imported crab meat with domestic crab meat while selling it as "Atlantic blue" crabmeat. The aim of this illegal practice, i.e. economic adulteration by seafood processors, is financial gain through consumer fraud.

Species authenticity, the substitution of one species for another, is the most common type of economic adulteration in the seafood processing industry. Quality monitoring of seafood products, usually done by sensory methods via trained and consumer panels, is not an adequate methodology to detect adulteration. Seafood processing often requires removal of morphological properties between similar species; therefore, visual detection of species authenticity and economic adulteration in seafood products is more challenging.

Traditionally, this problem has been addressed by wet chemistry methods, such as gas chromatography and mass spectrometry. These expensive methods, however, are of limited value because sample preparation, extraction, and analysis can take a long time to conduct, making them not suitable for on-line control monitoring. In addition, trained personnel are needed to operate machinery and interpret results. For these reasons, an alternative technology is needed to detect economic adulteration and species authenticity in the blue crab industry.

Visible and Near-Infrared (VIS/NIR) spectroscopy is an affordable, powerful, and objective tool that usually requires minimal sample preparation and destruction and, if used properly, can generate results in less time. In addition, it does not require a highly trained individual to operate the machinery and interpret the results. VIS/NIR spectroscopy has been used to detect adulteration in many high-value foods, such as fruit juices, wine, grain, honey, and some meat products. Research, however, has not focused on economic adulteration of seafood products, such as crabmeat. Based on a review of the literature, VIS/NIR spectroscopy, in combination with chemometrics, has great potential for on-line monitoring, making this an excellent technology for early detection of economic adulteration of crabmeat.

The aim of this research was to detect economic adulteration and species authenticity of the Atlantic blue crab (*Callinectes sapidus*) through the use of VIS/NIR spectroscopy. The specific objectives of this research were to:

- Investigate the feasibility of VIS/NIR spectroscopy in determining economic adulteration of the Atlantic blue crab using surimi-based imitation crabmeat;
- Detect species authenticity of the Atlantic blue crab using blue swimmer crabmeat as the adulterant;

- Evaluate the performance of current chemometric techniques in detecting economic adulteration;

- Perform wavelength selection on the VIS/NIR spectral data to detect the most important wavelengths for economic adulteration detection.

2.0 LITERATURE REVIEW

2.1 *Callinectes sapidus:* The United States Blue Crab Industry

Callinectes sapidus, often referred to as "blue crab," is naturally distributed from Nova Scotia to northern Argentina (Rugolo *et al*., 1998). The Latin scientific name describes three attributes of the blue crab: *Calli-*, *-nectes*, and *sapidus*, which mean beautiful, swimmer, and savory, respectively (Ward, 2000). Therefore, the blue crab is a beautiful swimmer that tastes good.

The United States blue crab industry, an important segment of the domestic seafood industry, is distributed along the Southeast region, along the Gulf of Mexico to Texas and from New Jersey to Florida (Rugolo *et al*., 1998), though it is prevalently located in the Chesapeake Bay, an area that "has provided more crabs for human consumption than any body of water in the world" (Johnson *et al*., 1998). Between 1935 and 1945, the species was artificially introduced into the Northern Aegean Sea (Enzerross *et al.*, 1997). Due to gradual migration of the species, distribution also includes France, Denmark and along the Western Black, Northern Adriatic, and the Mediterranean seas.

The origins of the soft and hard crab industries date back to 1873 (Crisfield, Maryland) and 1878 (Hampton, Virginia), respectively. During the middle to late twentieth century, however, the blue crab industry experienced an enormous expansion, mainly due to the development of tamper-evident packaging and the availability of pasteurization and rapid freezing technologies (Johnson *et al*., 1998).

2.1.1 *Marketing of the Blue Crab*

The marketing of the Atlantic blue crab is sorted into four groups: *blue*, *peeler*, *soft*, and *soft and peeler*. Crabs in the *blue* category are hard shelled and often sold whole in the live-trade market. Aside from the live-trade market, *blue* crabmeat is also sold according to where the meat is obtained from the shell (Johnson *et al*., 1998). Though not described here, the categories of meat include lump, jumbo lump, backfin, special, and claw. In a process called molting, crabs shed their exoskeleton to replace it with a new one (Chaves and Eggleston, 2003). *Peelers* are hard crabs with signs of molting and *soft* describes crabs that have recently molted and, thus, contain a carapace that has not yet hardened (Rugolo *et al*., 1998). Crabs in the *soft* and *peeler* categories are often sold alive, though they can also be sold frozen, then cooked and eaten as a whole. Besides this difference in delicacy tastes, soft and peeler crabs are also more expensive, averaging about $5.40 per pound compared to the $1.60 average per pound of *blue* (hard) crabs (Sutherland, 2006).

Figures 1 and 2 show the average percentage of total landings and associated economic value, respectively, for each type of crab grouping for the years 2000 to 2004 (NMFS, 2006). Crabs marketed as *blue* represent the biggest sector in both total landings (75.50 thousand metric tons [M.T.]) and economic value ($131.72 million [M.]), followed by *peeler* (1.63 thousand M.T. and $8.71 M.), *soft* (0.36 thousand M.T. and $3.99 M.), and *soft and peeler* (0.74 thousand M.T. and $8.84 M.). Due to the prevalence of the *blue* category over the rest, further discussion of the blue crab industry and the associated "blue crab" import industry, will focus just on crabs marketed as *blue*.

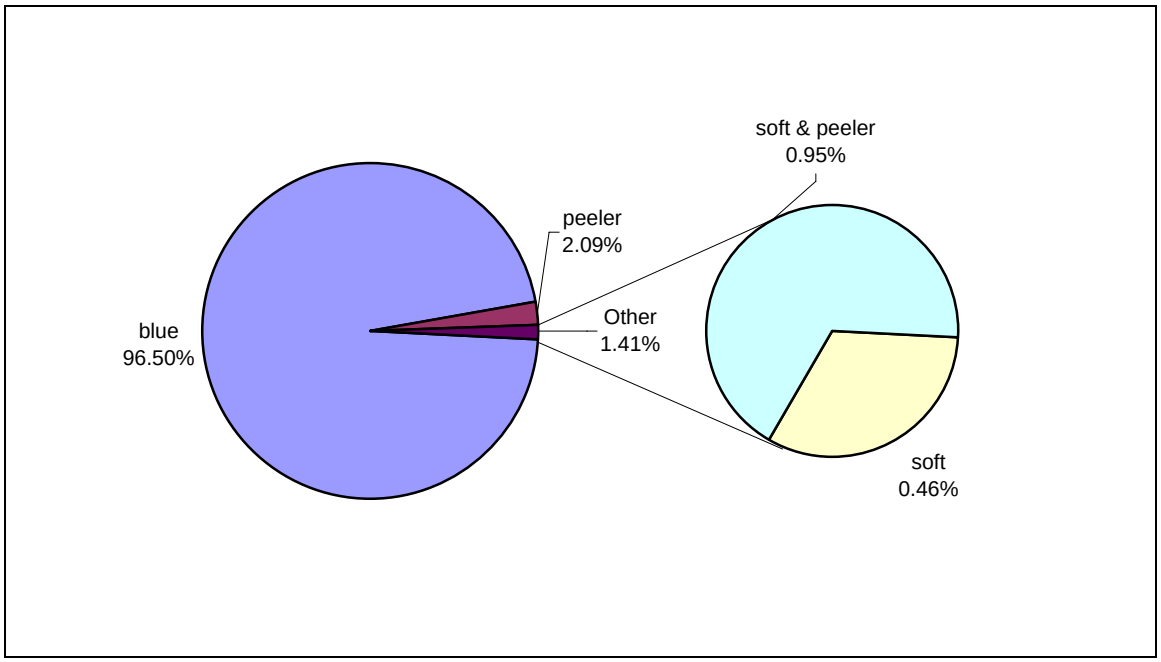

Figure 1: Percentage representing the average metric tons of Atlantic blue crab marketed for sale during 2000 to 2004.

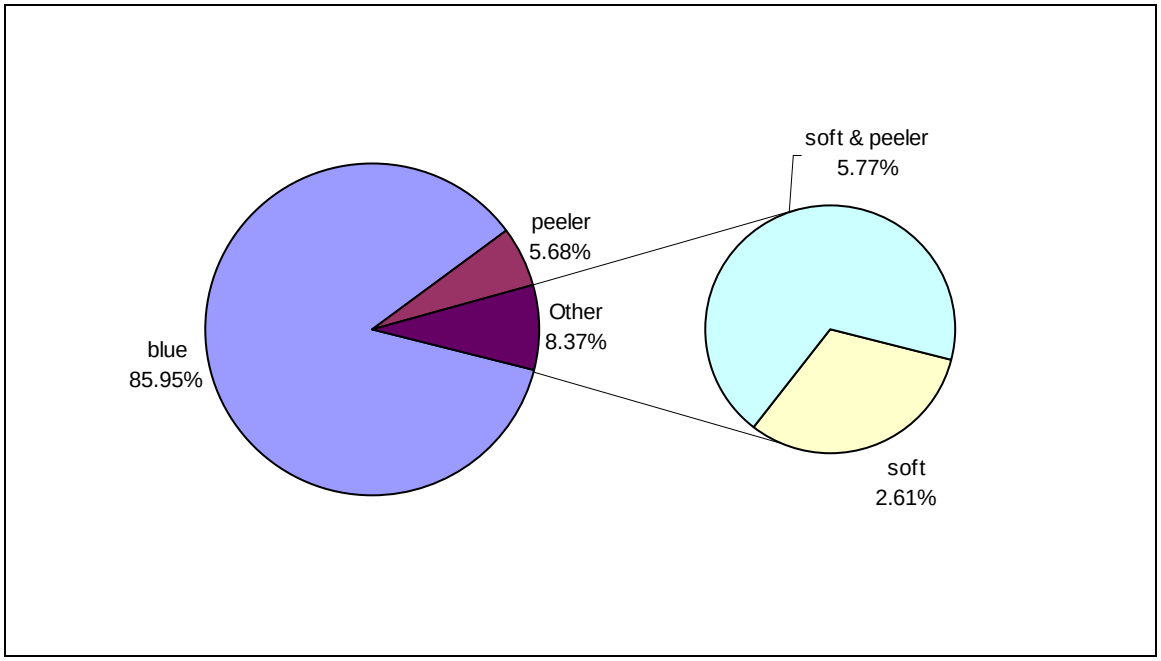

Figure 2: Percentage representing the average economic value of Atlantic blue crab marketed for sale during 2000 to 2004.

2.1.2 *Domestic Crab Landings*

Unlike the landings of imported crabmeat, which are based only on the weight of meat extracted from the crab, domestic landings are based on the weight of the whole crab. Therefore, a modification was made in order to be consistent with reports of import landings. Even though male crabs yield more meat than females, the amount of crabmeat yield, which varies depending on different processing and biological factors, has a range of 8.0% to 12% (Oesterling, 1998; Johnson, *et al*., 1998; Hong, 1991). For this reason, the domestic landings reported by NMFS (2006) were adjusted according to an average estimate of 10% yield.

The adjusted total blue (hard) crab landings from 1990 to 2004 (NMFS, 2006) are shown in Figure 3.

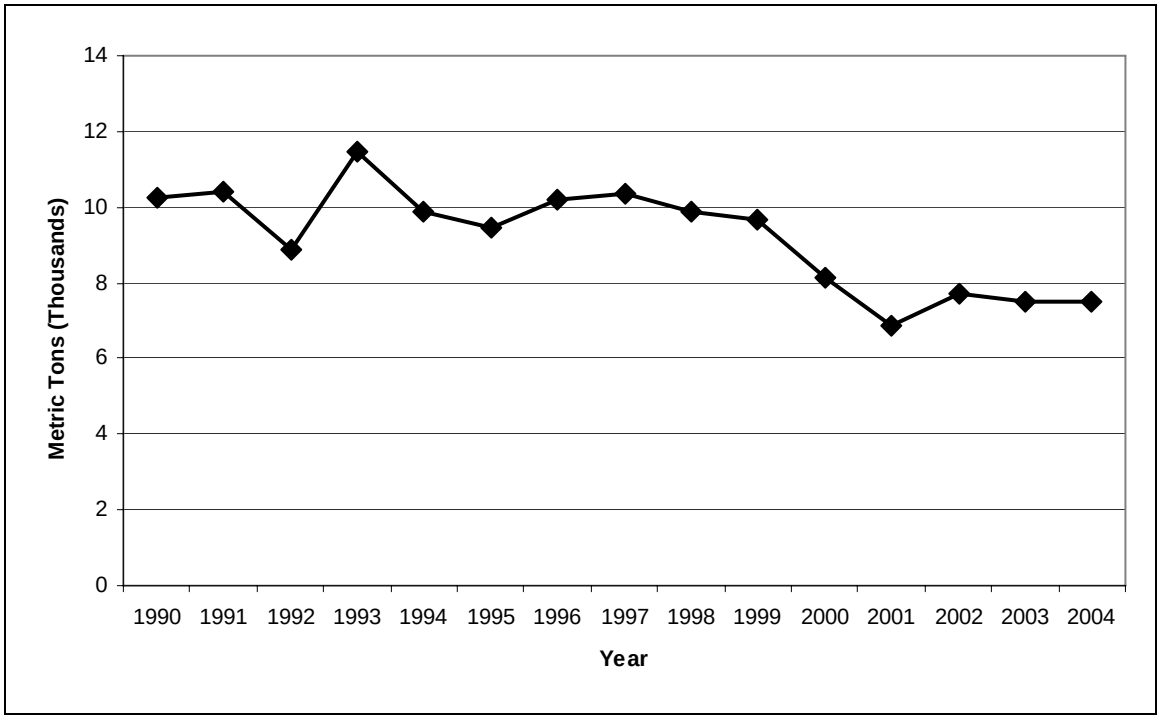

Figure 3: Annual metric tons (thousands) of *Callinectes sapidus* crabmeat landed in the United States.

Annual domestic landings have decreased drastically in the last few years, i.e. 2000 to 2004, compared to the early and middle 1990s, even though the associated annual economic value (Figure 4) has remained at a plateau between $120-140 M. during the equivalent time period (NMFS, 2006). Declines in the blue crab population, decreases in total catches, increases in domestic and worldwide demand, and increases in blue crab industry production costs have influenced the price of crabmeat. Hence, the economic value from 2000 to 2004 is still high even though it represents a period that includes the lowest domestic crab landings. Significant decreases in both blue crab populations and commercial landings, among other factors, also negatively affect the commercial fishermen who depend on crab operations to survive economically (Eggleston *et al*., 2002).

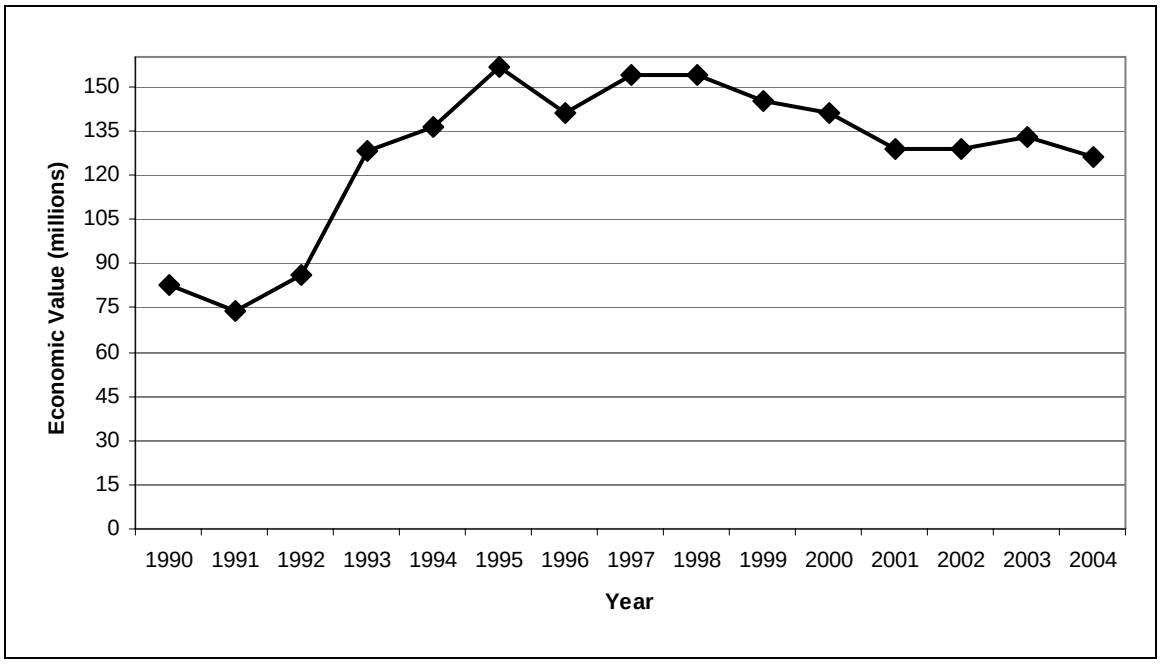

Figure 4: Annual economic value (millions) of *Callinectes sapidus* crabmeat harvested in the United States.

2.1.3 *Factors Affecting the Blue Crab Industry*

The primary concern of any hard processing operation is to "supply a sufficient quantity of high-quality crabs at a reasonable market price" (Johnson *et al*., 1998). This goal, however, has become challenging due to several factors and changes in fishery regulations. These include, but are not limited to, limits on number of crab pots permitted, restrictions on harvest time of year, crab size limitations, resurgence of predatory fish (such as the Atlantic striped bass), the impact of tropical storms and hurricanes on pristine waters, and the increasing competition of the crabmeat import market. Unfortunately, the growth of the domestic blue crab industry is further inhibited by variations in crab populations, steady consumer demand during a time of limited supply, unresolved quality issues, such as ammonia as an indicator of decomposition in crabmeat products, and the effect of water quality on crab survival (Chaves and Eggleston, 2003; Jahncke *et al*., 2000).

Due to these factors, the domestic blue crab industry has become dependent on imports of blue crab from other sources of the world, such as South America and the Mediterranean. In the last decade, crab imports have doubled causing the domestic processing industry a significant loss of market share. As much as 65% of crabmeat consumed in the U.S. is imported (Jahncke *et al*., 2000). In 1998, for example, national blue crab landings represented 98.8 thousand M.T., a value corresponding to $149.1 M. That same year, 11.8 thousand M.T. of fresh and frozen crabmeat, worth $116 M., was imported.

Table 1 shows the annual landings, in thousand M.T., and annual economic value, in millions of dollars, of domestic and imported crabmeat from 2000 to 2004 (NMFS,

2006). Although the annual landings and economic value of domestic crabmeat show an overall decline, the opposite is true for imported crabmeat. The annual landings and associated economic value of imported crabmeat have increased steadily and show little signs of slowing down.

Table 1: Annual landings, in thousand metric tons (M.T.), and economic value, in millions, of domestic versus imported crab meat.

	Domestic Crab Meat		Imported Crab Meat	
Year	M. T. (thousand)	Value (million)	M. T. (thousand)	Value (million)
2000	8.13	141.35	10.65	116.93
2001	6.88	128.96	13.59	171.26
2002	7.71	128.95	14.61	187.52
2003	7.50	133.12	14.43	179.18
2004	7.52	126.24	16.04	205.96

2.2 *Portunus pelagicus:* The Other "Blue Crab" Industry

Portunus pelagicus, referred to as "blue swimming crab" or "sand crab," is found throughout the Indo-Pacific region. More precisely, blue swimming crabs are distributed from the east African coast to Tahiti, Japan, the Philippines, throughout Southeast Asia, Indonesia, and are widely distributed in Eastern Australia and the Fiji Islands (Clarke and Ryan, 2004; Enzerross *et al*., 1997). This "blue crab" supports commercial fishermen from Africa to Australia and is the most prevalent crabmeat being imported into the United States (Oesterling, 1998).

In addition, competition from imported crabmeat products, such as those produced from *Portunus pelagicus*, which are less expensive and of poorer quality than those produced from *Callinectes sapidus*, also negatively affect the domestic blue crab industry (Oesterling, 1998). Figures 5 and 6 show the annual landings and economic value, respectively, of imported "blue crab" meat from years 2000 to 2004 (NMFS,

2006). The figures represent imported *Callinectes sapidus* (black histogram), imported *Portunus pelagicus* (gray histogram), and the total import value (series line).

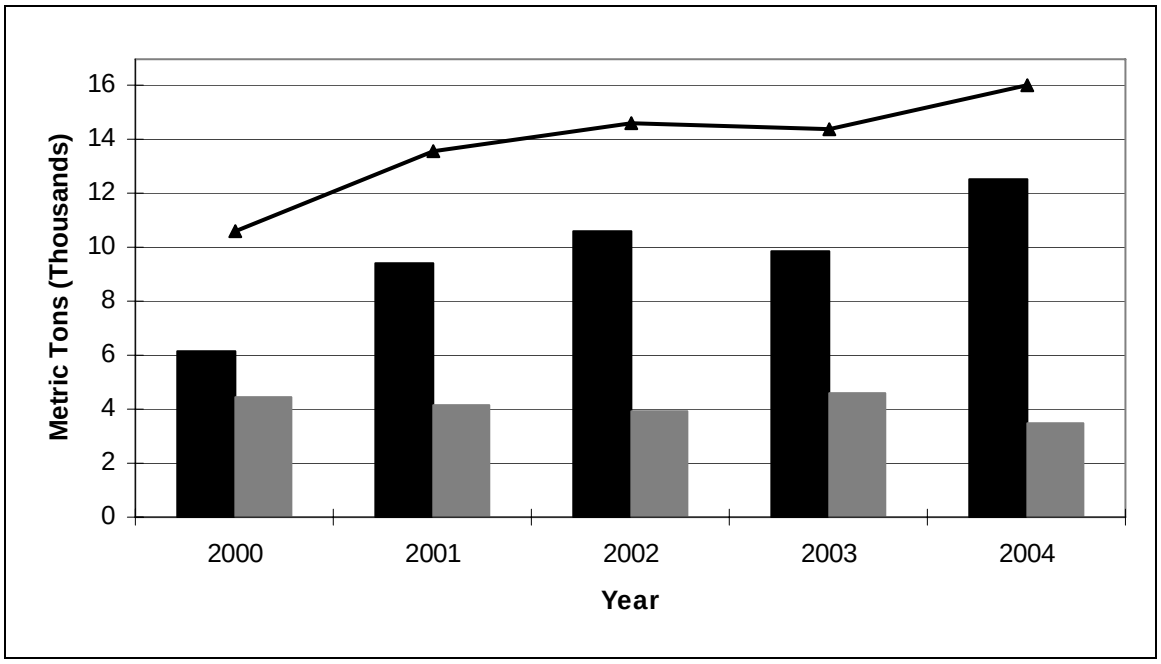

Figure 5: Annual metric tons (thousands) of imported "blue crab" meat. Histograms show totals of *Callinectes sapidus* and *Portunus pelagicus*, respectively. The series line shows total import "blue crab" meat landings (*Callinectes sapidus* and *Portunus pelagicus*).

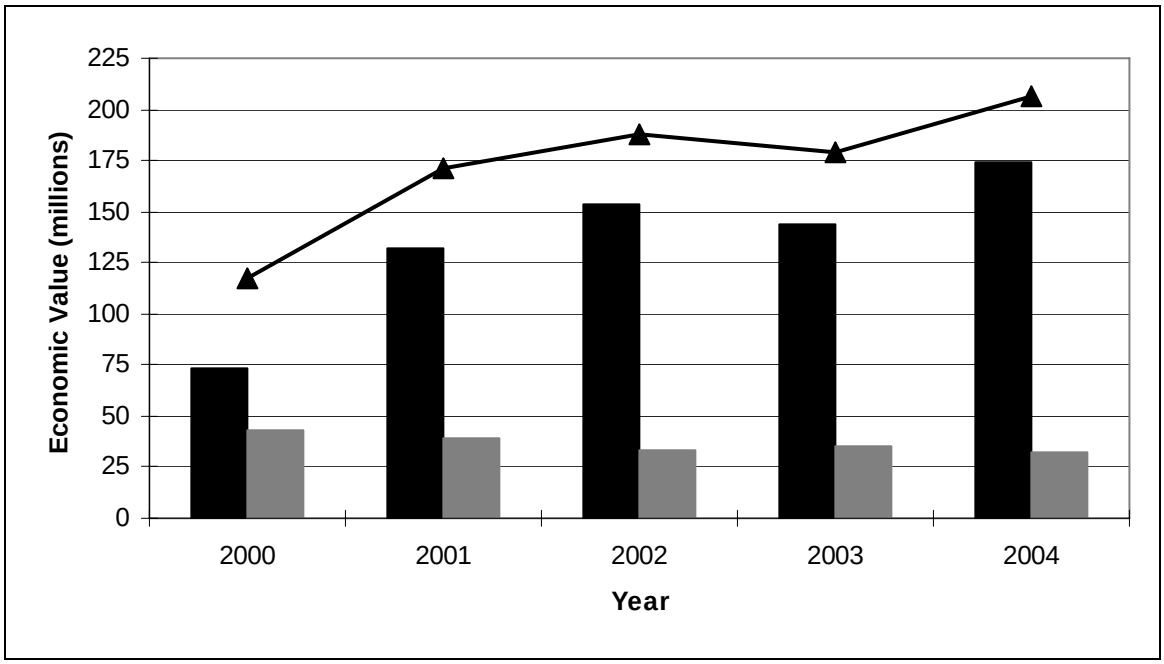

Figure 6: Annual economic value (millions) of imported "blue crab" meat. Histograms show totals of *Callinectes sapidus* and *Portunus pelagicus*, respectively. The series line shows total import "blue crab" meat landings (*Callinectes sapidus* and *Portunus pelagicus*).

The annual landings and economic value of the imported "blue" crabmeat have been increasing steadily. Even though the statistics of *Portunus pelagicus* imports are incomplete, i.e. NMFS record keeping is not available prior to the year 2000, annual imports from 2000 to 2004 have remained between 3.5 to 4.5 thousand M.T. with an associated economic value ranging from $31 to $43 million dollars (Figures 5 and 6, respectively). This steady import of blue swimming crabmeat, in addition to import of *Callinectes sapidus* from other areas of the world, has caused a significant market share loss for the domestic blue crab processing industry.

As opposed to domestic crabmeat landings, which break down totals according to marketing categories of crabs (soft, peeler, soft and peeler), further quantification of the amount of crabmeat imports is not possible because the meat is grouped generically as "crabmeat" totals (Oesterling, 1998).

2.3 Comparison of *Callinectes sapidus* and *Portunus pelagicus*

The quality of both domestic Atlantic blue and imported blue swimmer crab meat can be determined by combining sensory analyses with specific microbiological and chemical tests (Jahncke *et al*., 2000). Even though sensory profiles exist between seafood products of Atlantic blue and blue swimmer crabmeat composition, none have been characterized by a trained or consumer panel.

The chemical composition and nutritive value of crabmeat has been investigated for the two types of crab. Hong (1991) reported that the blue crab composition was 77.4% moisture, 19.8% crude protein, 2.1% ash, and 0.9-1.5% fat. A study on the blue crab showed that fatty acid profiles of claw meat and breast meat were statistically significant (Celik *et al.*, 2004). However, composition and nutritive value depends on region of the world, gender of crab, stages of molting, sexual cycles, seasonal factors, and diet (Tsai *et al.*, 1984; Reddy *et al.*, 1991).

A study done by Tureli *et al.* (2000) revealed that the ratio of crude protein, dry matter, and crude ash in the meat taken from male crabs was higher in swimmer crabs (*Portunus pelagicus*) than in blue crabs (*Callinectes sapidus*). Lipid levels, however, were not statistically different ($p>0.05$) between male crabs of the two species. Comparing female crabs, the same study revealed that the ratio of crude protein, dry matter, crude ash, and lipid levels were higher in swimmer crabs.

A different study revealed that no significant differences existed in the moisture, fat, and ash contents of claw and body meat between the two species, but protein content of the swimmer crab was significantly higher than that found in the blue crab (Gokoolu and Yerlikaya, 2003). Mineral content, in particular Na, K, Ca, Zn, Cu, Mg, and Fe, was

also investigated. Results indicated that there was no significant difference in the values of these minerals between the two crab species.

2.4 Economic Adulteration and Species Authenticity

The continued high demand of crabmeat for human consumption, along with the other challenges facing the domestic blue crab industry, has contributed to illegal practices within the U.S. blue crab industry. Due to the decline in the number of crabs harvested, some harvesters disregard size limitations in order to meet demand and still survive financially. In addition, the steady import of blue swimmer crabmeat from Asia and the Indo-Pacific facilitates products being incorrectly labeled as blue crabmeat or blended with domestic crabmeat with the intent of selling it as the higher valued domestic blue crabmeat (Johnson *et al*., 1998). Both of these acts, i.e. misbranding and economic adulteration, are illegal practices under FDA regulations and aim to gain financially via consumer fraud. The combination of these illegal practices not only raises the picking price of crabmeat for domestic processors, often to unprofitable levels, but also reduces the availability of good quality crabs.

Detection of economic adulteration and species authenticity has traditionally relied on expensive and time-consuming wet chemistry methods, such as gas chromatography and mass spectrometry (Downey, 1998). Even though some methods, such as protein electrophoresis, have been used to identify unprocessed seafood products, they have limited use in detecting economic adulteration. Variation in factors, such as temperature, pasteurization, and processing parameters, affect the results of these

methods. Therefore, a major concern with these traditional methods is reproducibility of identification results.

With the advances of biotechnology in the past decade, the Atlantic blue crab has been successfully identified using polymerase chain reaction. Dolan and colleagues (2002) developed a reproducible diagnostic tool that differentiated between meat of *Callinectes sapidus* and meat from *Portunus pelagicus*. Specific DNA sequences of both types of crab were targeted for amplification by designing site-specific primers. The PCR products were analyzed via DNA staining and gel electrophoresis. DNA electrophoresis results revealed unique bands in the Atlantic blue crab that make it possible to differentiate it from the blue swimmer crab. Based on these results, markers were developed, and are commercially available, to identify *Callinectes sapidus* and *Portunus pelagicus* crabmeat samples. However, these methods are expensive, require complex sample preparation to extract DNA, and are time consuming.

2.5 VIS/NIR Spectroscopy

2.5.1 Electromagnetic Radiation and Spectroscopy

Spectroscopy is the study of the interaction of light with matter (Smith, 1999). The basis of spectroscopy lies in the electric and magnetic properties of light. Light, also known as electromagnetic radiation, can be represented as simple harmonic motion of a traveling sine wave and is defined as:

$$Y = A \sin(\omega t) \qquad [1]$$

Where A is the radius (in rad) and ω is the angular velocity ω (in rad s^{-1}) and t (in s) is the time traveled.

Electromagnetic waves, however, are often characterized by the distance traveled during one complete cycle, i.e. their wavelengths, or the number of cycles per second, i.e. their frequency (Smith, 1999). Equation 2 shows the relationship between these two characteristics.

$$\lambda = c/\nu \qquad [2]$$

Where λ is the wavelength, c is the speed of light ($3x10^{10}$ cm/sec), and the frequency, $\nu = \omega/2\pi$

Alternatively, especially in the field of spectroscopy, electromagnetic radiation can be described in terms of its wavenumber, i.e. $\upsilon = 1/\lambda$ cm^{-1}. Transitional changes between energy levels can then be described using a proportionality between wavenumber and frequency, i.e. $\nu = c\upsilon$.

The effects of radiation on matter vary depending on the electromagnetic waves' frequency and wavelength. It can be viewed as a stream of photons, i.e. particles of light, whose energy is proportional to the frequency of radiation. This electromagnetic radiation can be absorbed, scattered, or emitted by atoms and molecules. These processes are associated with changes in energy levels that represent the interaction of the atoms or molecules with electromagnetic radiation (Hof, 2003). Transitions between two energy levels, E1 and E2, can occur provided that an appropriate amount of energy is absorbed or emitted. This change in energy is described by:

$$\Delta E = h\nu \qquad [3]$$

Where h = Planck's constant ($6.63x10^{-34}$ Joule-sec) and ν is the frequency (Hz)

In other words, if electromagnetic radiation containing different frequencies is directed onto a molecule at energy level E1, the molecule will absorb energy from the

radiation at a particular frequency. The increase in energy due to absorption will induce a transition from the current energy level, E1, to the next level, E2. Since the only reduction in the radiation intensity occurs at a particular frequency ($\nu=\Delta E/h$), all other intensities of different frequencies remain unchanged. As a result, an absorption spectrum of measured radiation intensity versus wavelength can be produced to qualify and quantify interactions at a molecular level.

2.5.2 The Beer-Lambert Law

The main concept behind the Beer-Lambert Law (often referred simply as Beer's Law) is that the spectral measurements at various wavelengths are linearly related to the analyte concentration (Kalivas and Lang, 1994). Due to the changes in energy caused by the electromagnetic radiation, spectrometers do not measure the analyte itself but rather its absorbance values (Smith, 2002; McClure, 1994). The absorbed electromagnetic radiation can then be used to determine analyte concentrations. The most common form of Beer's law is:

$$A = \varepsilon lc \qquad [4]$$

Where ε is the absoptivity coefficient, l is the pathlength of the sample, and c is the concentration. In other words, the amount of electromagnetic radiation absorbed by a sample depends on the sample's absorptivity, the thickness of the cuvette sample chamber, and the concentration of analyte.

Sometimes, spectra are recorded in transmittance mode. In this case, Beer's Law still holds due to the relationship between transmittance and absorbance:

$$A = \log\left(1/T\right) = \log\left(I_o/I\right) \qquad [5]$$

$$T = \log\left(I \middle/ I_o \right) \qquad [6]$$

Where I_o is the energy intensity striking the sample, I is the energy intensity passing through the sample, and T and A are the transmittance and absorbance, respectively.

For more complex samples, i.e. non-homogeneous and presence of scattering, diffuse reflectance spectroscopy is often used. In this type of setup, the energy shining on the sample, I_o, is reflected, and the value of I recorded. To comply with the Beer-Lambert law, the transmittance is replaced with reflectance (R) and absorbance is recorded as log(1/R). Even though both transmittance and absorbance modes are used in spectral analysis, absorbance values are more common in quantitative analysis because of the proportionality between absorbance and concentration stated by Beer's Law (Osborne *et al*., 1993). Unlike transmittance, which shows peaks pointing down, absorbance spectra generate peaks pointing upward which represent the wavelengths at which radiation was absorbed by the sample (Smith, 1999).

2.5.3 Visible Spectroscopy

The visible region is defined as the spectrum containing wavelengths 380 to 700 nm. Few biological compounds absorb light in this range, and absorption is usually function-related (Harris, 1996). Examples include the trapping of light energy to serve as an energy source (e.g., chlorophyll) or in pigmentation (e.g., melanin). Biochemical compounds, such as DNA and RNA, absorb light in the near ultraviolet range though absorption is unrelated to function. In this case, continued light exposure can lead to damage and mutations.

Absorption of visible light is due to a group of atoms called a chromophore (Harris, 1996). These groupings are usually conjugated double bonds and/or contain the presence of a transition metal. Absorption occurs when the upper and lower energy levels of a molecule are separated by an energy gap. The energy of electromagnetic radiation in the visible range induces excitation (and transitions) of the outer valence electrons shell of a molecule causing electrons to move between one molecular orbital (π) and another (π^*) (Clark *et al*., 1993).

The absorption of light by chromophores has been used to derive information about molecules and biological systems. Examples include the discovery of the cytochromes and the iron-sulfur proteins, two types of classes of mitochondrial electron transfer proteins.

2.5.4 Near-Infrared Spectroscopy

The near-infrared region (780 nm to 2500 nm) is dominated by overtone and combination bands arising from the unharmonic nature of molecular vibrations (Martin, 1992). When electromagnetic radiation is absorbed, the covalent bonds vibrate at a larger amplitude. Absorption in the NIR region arises from vibrational motion of molecules (Osborne *et al.*, 1993). Vibration involves: 1) a change in the interatomic distance along the bond between two atoms (i.e., stretching) or 2) a change in bond angle (i.e., bending). The majority of the absorption bands in NIR arise from overtones of hydrogenic stretching vibrations or combinations involving stretching and bending modes.

The molecular vibration is modeled by a mass, m, attached to a spring. The other end of the spring is fixed. Hooke's Law describes the mechanical system:

$$F = -ky \qquad [7]$$

Where F is the force exerted by the spring, k is the spring's constant, and y is the distance traveled from the equilibrium position. Applying Newton's second law of motion and substituting $\frac{d^2y}{dt^2}$ for acceleration, the following differential equation is derived:

$$m\frac{d^2y}{dt^2} = -ky \qquad [8]$$

Which has, as one of the solutions:

$$y = \sin(\alpha t) \qquad [9]$$

Where α is the positive square root of k/m. This equation is used to describe the simple harmonic motion of a traveling wave.

A transition from the ground state to the first vibrational state of excitation occurs at the fundamental absorption frequency, usually found in the mid-infrared region. In NIR spectroscopy, the term overtone is used to understand the relationship between the fundamental absorption of a chemical substance and its remaining absorptions. At twice the fundamental absorption frequency, known as the first overtone, a smaller number of molecules are induced to the second state of excitation. Similarly, the second overtone corresponds to three times the fundamental absorption frequency, and so on. Remembering the relationship between frequency and wavelength, the first and second overtones correspond to one half and one third of the wavelength.

Water content, for example, influences the shape of the spectra because water is a main component of biological materials and food products and absorbs strongly. The fundamental absorption of water occurs in the infrared region (~3840 nm) (Yee and Gimel'farb, 1999). Therefore, the first, second, and third overtones of water correspond

to one half, one third, and one fourth of the fundamental wavelength, or 1920 nm, 1280 nm, and 960 nm, respectively.

In reality, however, the overtones do not correspond exactly to multiples of the fundamental frequency. This is attributed to deviations in the behavior of molecules from the laws of simple harmonic motion and the lack of compliance with Hooke's Law (Osborne *et al*., 1993). As one atom approaches another one, the potential energy increases faster, due to repulsion of both atoms' nuclei, than predicted by harmonic approximation. Dissociation occurs, hence leveling off the potential energy of the molecule.

The NIR spectrum primarily shows bands of hydrogen stretching vibrations due to the influence of the small weight of the hydrogen atom on the large deviation from harmonic motion. Functional groups that involve hydrogen, such as C-H, N-H, and O-H, pre-dominate in this region. Due to the extent of the subject, which is beyond the scope of this work, chemical structures and types of stretching or predicted positions of overtones will not be explained here.

Evaluation of spectra in the NIR region, however, can be complicated due to the overlap of bands (Hof, 2003). The increase of interest in NIR spectroscopy arises from its advantages over traditional wet chemistry methods (Gusnanto *et al*., 2003). Little pre-treatment is needed to record spectra for solid and liquid samples and NIR spectroscopy can be used to predict physical and chemical parameters from a single spectrum. These attributes make it an effective technology for rapid characterization of samples (Blanco and Villarroya, 2002).

2.6 Chemometrics

Chemometrics, introduced in 1972, is a discipline that utilizes mathematical and statistical methods to evaluate chemical data and derive structure-activity relationships (Otto, 1999; Feudale *et al*., 2002). Even though chemometrics originated in chemistry, it has expanded into other fields, such as spectroscopy (Lavine and Workman, 2005). This section will focus on the most common chemometrics applications utilized in VIS/NIR spectral analysis: Principal Component Analysis (PCA), Multiple Linear Regression (MLR), Principal Component Regression (PCR), and Partial Least Squares (PLS). In addition, Genetic Algorithms (GAs) will also be described as a form of variable selection.

2.6.1 Principal Component Analysis (PCA)

Spectral data contains variables (wavelengths) that are highly correlated and contain redundant information, i.e. they are collinear (Lavine and Workman, 2005). Due to the high collinearity among variables, a new coordinate system based on variance can be found to represent the original data. PCA can easily be explained visually. The absorbance values of three wavelengths, i.e. variables p1, p2, and p3, can be plotted to reveal a scatter plot of each sample according to the values for each variable (Figure 7).

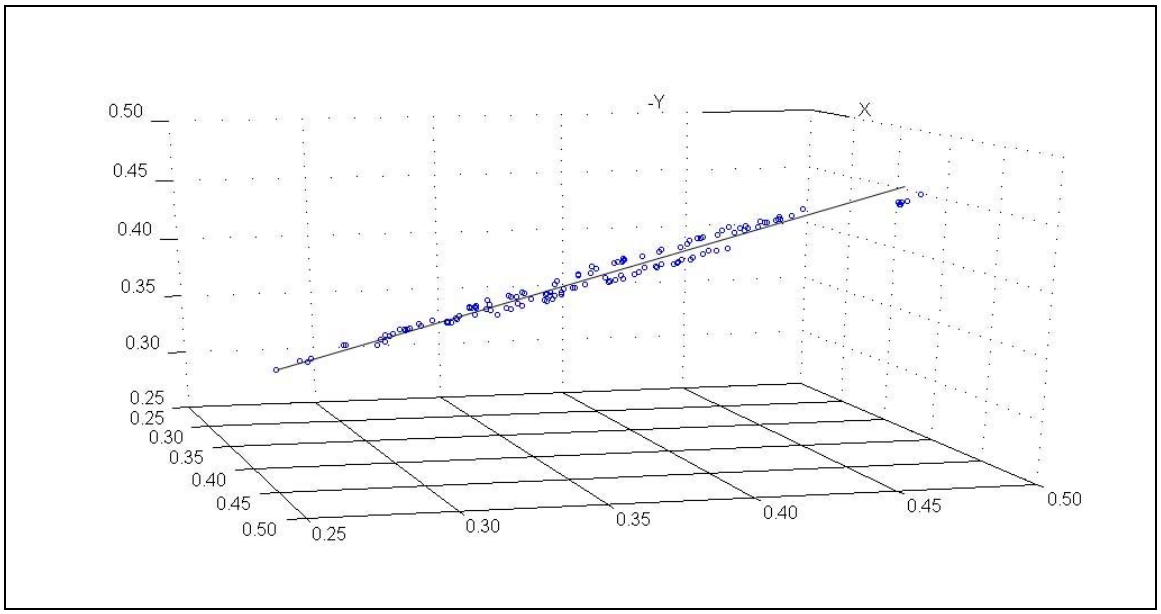

Figure 7: Scatter plot of data from three wavelengths. Central axis within the scatter plot reveals the Principal Component (PC).

Due to the correlation among the variables, a new axis, in the direction of maximum variability of the data, can then be found to describe the original three variables, thus reducing the dimension from 3 to 1 (Martin, 1992). This central axis, called the first Principal Component (PC), represents the linear relation among the variables and lies along the direction of maximum variance, i.e. spread. Alternatively, the axis can be found by minimizing the sum of squares of the residuals from each sample (Esbensen *et al*., 1994; Geladi and Kowalski, 1986). So, PCA can be thought of as an algorithm that finds PCs that either maximize variance or minimize the squared residuals of the samples.

Similar axes can be found for multidimensional data. The second PC will be orthogonal to the first and have the second largest variance, etc. The variance explained by each principal component, expressed in terms of its eigenvalue, is used to arrange the PCs in decreasing order of importance (Lavine and Workman, 2005). PCA will continue

to generate PCs, which are orthogonal to each other, that will describe a new coordinate system to replace the p-dimensional variable space.

The orthogonality property of the principal components removes collinearity among spectral variables without eliminating spectral information. The majority of the variation is contained in the first few components. Noise, though spread through all the components, is mainly contained in latter components (Martin, 1992). The maximum number of PCs, however, is limited by either the number of samples or the number of variables, whichever value is smaller, even though the full set of PCs, i.e. a full model, is seldom used (Esbensen *et al*., 1994). In order to determine the optimum number of PCs, one must examine the explained variance (high is good) or the residual variance (low is good). The interpretation of the PCs and the phenomena described, however, rests in the hands of the scientist.

In mathematical terms, PCA decomposes the spectral data matrix, X (n samples by p variables), into two parts (structure and noise), thus reducing the variable space, a p-dimensional coordinate system, by finding directions of maximum variance (Esbensen *et al*., 1994). The PC model is described by Equation 10:

$$X = \left(TP^{T}\right) + E = \left(structure + noise\right) + E \qquad [10]$$

Where T is the score matrix, P^T is the transpose of the loading matrix, and E, the residual matrix, is the part not explained by the model, i.e. used to measure lack of fit.

The linear combinations of variables in each PC contain p coefficients, called loadings, which describe the relationship between the p variables and the PCs (Esbensen *et al*., 1994). The loadings constitute what is referred to as a transformation matrix P. Graphs of loadings reveal a map of variables. Since spectral data sets often contain

several hundred and sometimes thousands of wavelength variables, loading graphs reveal a snake-like pattern that generates little information about the behavior of the variables (Figure 8).

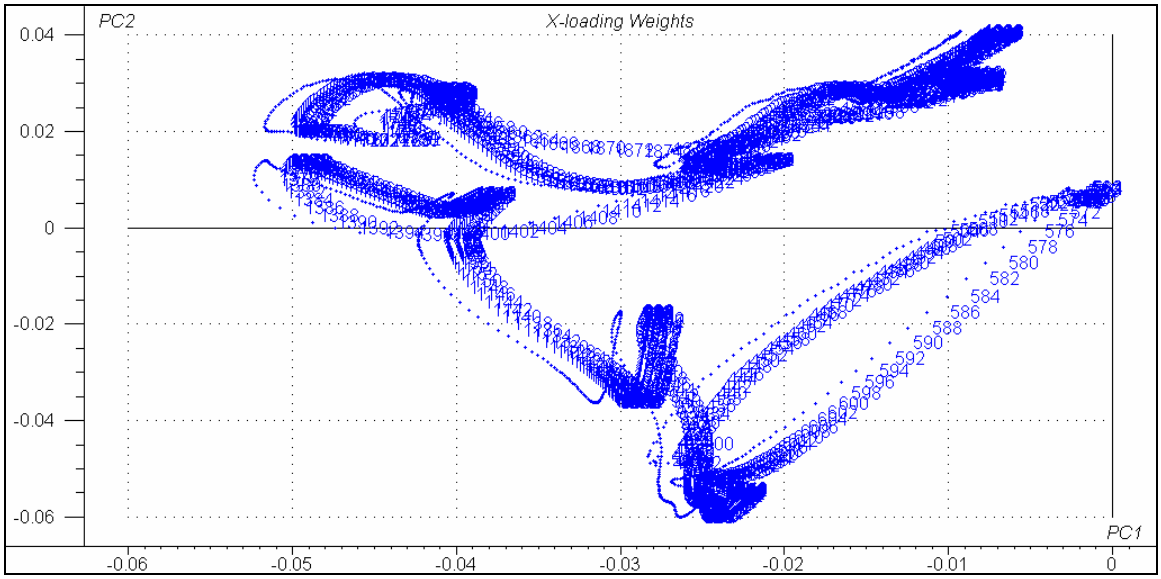

Figure 8: VIS/NIR loading weights of wavelengths 400 nm to 2498 nm (2 nm increments). PC1 and PC2 contain 87% and 5% explained variance, respectively.

Instead of loadings, it is often more useful to talk about PC scores, the perpendicular projection of each sample to each PC (Geladi and Kowalski, 1986). The scores represent the samples' coordinates in a PC coordinate system (Esbensen *et al.*, 1994). The scores for all the samples constitute the score matrix T, whose columns are orthogonal and represent the scores for one PC with n number of entries. The plotting of the score vectors, i.e. columns in T, yields a score plot, an important feature of PCA that is used to infer relationships among samples.

Even though there are as many score vectors as there are PCs, the score vectors of PC1 and PC2, i.e. representing the largest and second largest explained variance, are often used to determine relationship patterns among the samples (Esbensen *et al.*, 1994).

Score plots reveal maps of samples and are more beneficial in spectral data analysis because they are often used to identify groups of similar object or find trends to explain chemical behavior in the samples (Figure 9).

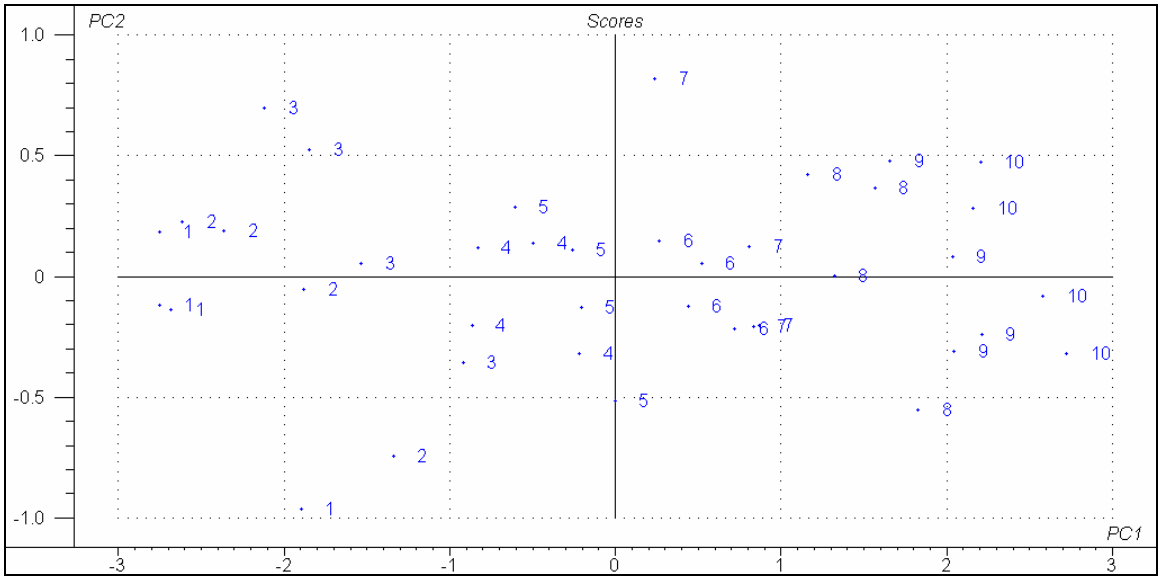

Figure 9: Score plots of samples in VIS/NIR spectral analysis. PC1 and PC2 contain 95% and 5% explained variance, respectively.

In summary, PCA takes p variables (X_1, X_2, …, Xp) for n samples and finds linear combinations of these variables to produce uncorrelated features called principal components (PCs) (Manly, 1986). The lack of correlation in the features enables the PCs to measure different dimensions within the data space. The PCs represent variance in the original data and are ordered in decreasing order of importance, i.e. var(PC1) > var(PC2) > … > var(PCp). The majority of the variation found in the original data is usually found in the first few PCs; however, additional components are often needed in classification and prediction.

2.6.2 Multivariate Calibration

Multivariate calibration is a regression relationship between two matrices, an X matrix containing the independent variables and a Y matrix containing the dependent variable(s), with the purpose of finding a model to use for future sample prediction (Geladi and Kowalski, 1986). This section will describe the most common calibration algorithms in spectral analysis: Multiple Linear Regression, Principal Component Regression, and Partial Least Squares.

2.6.2.1 Multiple Linear Regression (MLR)

Multiple linear regression, widely applied in spectral analysis, utilizes linear combinations of a set of X variables that have the most correlation to the corresponding response variable, Y. For one response variable, the model can be defined as:

$$Y = Xb + E \qquad [11]$$

Where Y = dependent variable, X is an n by m matrix of the independent variables, b is an m by 1vector of regression coefficients, and E is the error term. A least squares criteria is applied to estimate the vector of regression coefficients, b, so that the error term, E, is minimized (Geladi and Kowalski, 1986):

$$\hat{b} = \left(X^T X\right)^{-1} X^T Y \qquad [12]$$

Even with the wide applications of MLR in VIS/NIR spectroscopy, this method does not work well in situations where there is strong correlation and collinearity among the variables, high variable-to-sample ratio, and high noise level in X. The most frequent problem in MLR is that, when any of these scenarios occur, the inverse of $X^T X$ may not exist making the matrix inversion calculation equivalent to "dividing by zero" (Geladi

and Kowalski, 1986). To overcome this scenario, X is made full rank by deleting variables, thus decreasing the variable-to-sample ratio. Deletion of variables, however, is tricky because it is not often obvious which variables contain the most relevant information and which contain mostly noise.

Another constraint in MLR is the fact that the number of calibration samples must be greater than the number of variables. This hard-to-meet requirement limits the use of MLR in most VIS/NIR spectroscopy applications due to the large number of variables in spectral data sets. Various methods, based on wavelength selection strategies to pick out a small number of variables that show little or no covariance, have been successfully applied in spectral data analysis. Algorithms for MLR that evaluate a subset of variables by adding or deleting variables include forward and backward elimination and stepwise linear regression.

2.6.2.2 Principal Component Regression (PCR)

PCR is a two-step process that combines PCA and MLR. In the first step, PCA is done on the X matrix to derive the score vectors, T (Geladi and Kowalski, 1986):

$$T=XP \qquad [13]$$

In the second step, the T-matrix containing the score vectors, is substituted into Equation 11 of the MLR model:

$$Y = TB + E \qquad [14]$$

Which estimates B using the following equation:

$$\hat{B} = \left(T^T T\right)^{-1} T^T Y \qquad [15]$$

In this case, the transformation of the original variables in X, i.e. correlated data, into new variables, T, that are orthogonal, i.e. uncorrelated linear combinations of the original variables, guarantees that the inverse of T^TT exists (Geladi and Kowalski, 1986). In other words, the orthogonality of the scores and the exclusion of scores with small eigenvalues from the MLR step remove the problem of collinearity, thus making PCR one of the preferred methods for handling spectral data sets.

In addition, the use of PCA also enables interpretation of the results via score and loading plots. The main disadvantage of PCR, however, lies in the decomposition of the X matrix. There is no guarantee that the PCs used in MLR, i.e. usually the first several PCs, contain relevant information to predict the y variable (Esbensen *et al*., 1994). In other words, there might be information in the PCs used that is not related to Y or there might be valuable information in higher order PCs that never make it into the MLR stage. This weakness inhibits the optimization power of PCR.

2.6.2.3 Partial Least Squares (PLS)

Partial Least Squares, also known as Projection to Latent Structures, is the most common chemometric technique used in spectral data analysis. It is useful to predict dependent variables (Y) from a set of independent variables (X), especially when the number of independent variables in VIS/NIR spectroscopy, i.e. wavelengths, is large compared to the number of samples (Lazraq and Cleroux, 2001). In simple terms, PLS estimates latent structures (also referred to as principal components or PLS factors) as projections of the original X variables to relate X (n samples by number of wavelengths)

and Y (n samples by number of constituents) via regression (Wold *et al*., 2001). The regression model takes the form:

$$Y = XB + E \qquad [16]$$

Where E is the residual matrix and B, the regression coefficients, is calculated by Equation 17.

$$B = W\left(P^T W\right)^{-1} Q^T \qquad [17]$$

Two sets of X loadings are calculated in PLS calibration: the loadings, P, similar to those gathered in PCA and the loading weights (or PLS weights), W. The P-loadings describe the relationship between the data matrix X and its scores, T, whereas, the loading weights, W, connect the matrices Y and X via regression. In addition, the set of Y-loadings, Q, containing the regression coefficients relating the Y-variables to the T scores, are also used to estimate B.

Unlike PCR, PLS uses the y-variance to decompose the X matrix generating an optimal regression for prediction purposes and utilizing fewer PCs. PLS can be described as two dependent PCA analyses (one on Y and one on X) in which scores are interchanged. In other words, the X and Y space are modeled interdependently (Esbensen *et al*., 1994; Geladi and Kowalski, 1986).

$$Y = \sum UQ^T + F \qquad [18]$$

$$X = \sum TQ^T + E \qquad [19]$$

The u-score vectors derived from the Y matrix are the starting points for the t-score vectors in decomposition of the X matrix (Esbensen *et al*., 1994). This algorithm

lets the structure in the Y matrix guide the PCA decomposition of the X matrix. For example, t1 is substituted by u1 to guide PCA(X) decomposition. In a subsequent PCA(Y), u1 is substituted by t1 so the X matrix also influences decomposition of Y. Iteratively, u to t and t to u substitutions continue until there is a convergence and a final set of t,w,u,and q vectors are calculated for the current PLS component. This combination of Y and X decomposition makes PLS a powerful alternative to PCR because it reduces the variation in X that does not correlate with Y for prediction purposes.

2.6.3 Performance criteria of VIS/NIR Calibration

The performance of calibration models derived from chemometrics is usually measured via three parameters: standard error of calibration (SEC), standard error of prediction (SEP), and their associated coefficient of determination (R^2). These performance criteria are described by the following equations (Esbensen *et al*., 1994):

$$SEC = \sqrt{\frac{\sum_{i=1}^{C}(\hat{y}_i - y_i)^2}{C - n - 1}} \qquad [20]$$

$$SEP = \sqrt{\frac{\sum_{i=1}^{V}(\hat{y}_i - y_i - Bias)^2}{V - 1}} \qquad [21]$$

$$R^2 = 1 - \left(\frac{SEC}{SD_{range}}\right)^2 \qquad [22]$$

Where, $\hat{y}_i$ is the predicted Y-value, y_i is the measured Y-value, C is the number of samples in the calibration set, n is the number of regressor terms or wavelengths, *Bias* is

the difference between the predicted and measured Y-values averaged for all samples in the validation set, V is the number of samples in the validation set, and SD_{range} is the standard deviation of all the reference values in the calibration data set. Even though the formula in Equation 22 is derived for the calibration set, a similar value can be found for the validation set by using SEP, instead of SEC, and SD_{range} as the standard deviation for all values in the validation data set.

Proper validation is one of the key elements in chemometrics. In data sets with a small number of samples, cross-validation is the preferred method to produce reliable results. The standard error of cross validation (SECV) is described by:

$$SECV = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N-1}} \qquad [23]$$

Where $\hat{y}_i$ is the measured value, y_i is the predicted value, and N is the total number of samples. In cross validation, there are as many models as there are samples. Each time a model is created, one sample is left out and used for prediction. The final model is based on the performance of all models. However, Esbensen *et al*. (1994) state that test set validation, i.e. testing on a different data set, is the best type of validation. This usually entails twice as many samples split evenly into the calibration and validation data sets. In some cases, however, gathering so many samples might not be possible, especially when sample preparation is very difficult, dangerous, or expensive.

2.6.4 Genetic Algorithms

Genetic Algorithms (GAs), a subset of Evolutionary Algorithms, are artificial biological systems based on Darwin's "survival of the fittest" that use reproduction operators to produce offspring better suited to survive and, hence, optimize the solution(s) to a particular problem. GAs are used to determine optimal parameters, such as weight values of an artificial neural network, and for feature extraction. In order to derive to an optimal solution, the parameters or features are passed through a fitness function, such as PLS, that evaluates, usually in terms of an error, how well they perform in the problem. Unlike other algorithms, which search the error space one point at a time and generate a single solution, GAs perform in parallel, i.e. error space is searched simultaneously in different areas, and can create different solutions (Haupt and Haupt, 1998). In other words, the error space is searched in different areas simultaneously to generate either the optimal single solution or a number of solutions that perform similarly.

2.6.4.1 Anatomical Analogy

The theory and application behind GAs is based on genetic and reproductive properties of simple organisms in biological systems (Luke, 2003). For the purpose of explaining the analogy to biological systems, the following discussion, in simple terms, will just focus on organisms whose genetic information is contained in deoxyribonucleic acid (DNA). Each organism contains *chromosomes*, which contain genetic information encrypted in a string of DNA. Each position, known as a *locus*, in this string can contain one of four different nucleotides: adenine (A), guanine (G), cytosine (C), and thymine(T).

Within each chromosome there are regions, known as *genes*, which code for the formation of a specific protein, such as an enzyme. Alternative forms of the same gene occupying a specific locus, are referred to as *alleles*. The string of nucleotides (A, G, C, and T) is translated into amino acids until the desired protein is formed. The *genotype*, i.e. specific genetic makeup in terms of amino acid sequence, of an organism codes for its *phenotype*, i.e. the physical characteristics of the organism.

Like their biological counterparts, GAs contain all of their information in a set of chromosome-like structures referred to as *genetic vectors*. The main coding for each genetic vector is binary, i.e. 1 or 0, in which each locus represents a gene and each allele is either turned on (1) or off (0). The bit string encoding the data is the artificial organism's genotype whereas the ability to solve the problem, i.e. how well the data fits, is its phenotype. The *fitness* of each organism is defined as the performance and ability of each genotype, or chosen data, to solve the particular problem.

2.6.4.2 Exploration versus Exploitation

It is useful when talking about genetic vectors of offspring to talk about pattern(s) of identical values that form in the genetic vector of the new generation. The patterns, called schema, are discovered by focusing on specific loci in all the genetic vectors of the population (Luke, 2003). The following example aims to describe the purpose of the schema.

Assume that the population contains genetic vectors of length 10 and, after a few generations, the genetic vector in the population looks like (-,-,*,*,-,-,*,-,-,*), where each * indicates a fixed locus within the population, i.e. the same value among the genetic

vectors, and each - represents a variable locus, i.e. different values among genetic vectors. This pattern is what is referred to as a schema. The advantage, in this example, is that the search space is reduced from 10-dimensional, i.e. the full genetic vector, to 6-dimensional, i.e. just the variable loci. As long as the mutation operator, which inverts the value at a locus, does not change the fixed loci, the schema will be inherited by the offspring. Eventually, as the GA proceeds, more loci will become fixed, making the schema grow, until nearly all the population has identical genotypes creating a population of optimal answers to the problem at hand.

Exploitation, the heredity of genetic vector segments from parents to offspring, encourages schema formation. Therefore, the error is searched in a small area instead of the full search space. Exploration, on the other hand, is used to define the creation of offspring whose genotype varies greatly from the parents. When genetic vector segments are not preserved, via mating and mutation operators, schema formation is inhibited, and the offspring's genotype will be different from the parents. In this case, exploration of other areas within the error search space will occur.

As with most error search space techniques, care must be taken to avoid underfitting, i.e. the algorithm does not model the training data, and overfitting, i.e. the algorithm not only models the training data but also the noise within it making the algorithm perform poorly in the testing data set (Leardi, 2003). For this reason, it is important to have a good balance between exploitation and exploration.

2.6.4.3 Artificial reproduction

As in living organisms, a new generation of offspring are chosen based on the initial parent population. Parents are selected to mate based on Darwin's "survival of the fittest," and parts of the genetic information from their chromosome are exchanged via mating and mutation operators to form the chromosome of the offspring. Usually, the population size is kept constant throughout all the generations, so those individuals with a higher fitness value replace the ones with lower fitness.

Three genetic operators (selection, recombination, and mutation) guide the search of the error space (Golovkin *et al.*, 2002). For each generation (i.e. iteration), the strings are evaluated in terms of a fitness function. String selection depends on the fitness value of the current chromosome relative to the rest and is given a higher probability for future selection. Crossover (i.e. reproduction), usually of high probability (50-90%), is applied to generate new offspring. Mutation with low probability (~1%) is also applied to a small number of strings. String evaluation, selection, and recombination continues for each generation. The main steps in a Simple Genetic Algorithm are as follows:

1. Creation of initial population and calculation of fitness values.
2. Random selection of parents based on fitness values.
3. Reproduction between parents, via mating and mutation operator, to produce a new generation.
4. Calculation of fitness values of members in the new generation.
5. Replacement of the weakest member(s) in the population with some from the new generation.

Repetition of steps 2-5 continues until a termination criterion, usually a defined number of generations, is met.

References

Blanco, M., and Villarroya, I. 2002. NIR spectroscopy: a rapid-response analytical tool, *Trends in analytical chemistry*, vol. 21, no. 4, pp. 240-250.

Chaves, J.C., and Eggleston, D.B. 2003. Blue Crab Mortality in the North Carolina Soft-Shell Industry: Biological and Operational Effects, *Journal of Shellfish Research*, vol. 22, no. 1, pp. 241-249.

Clark, B.J., Frost, T., and Russell, M.A. 1993. UV Spectroscopy Techniques, instrumentation, data handling, Chapman & Hall: New York, pp. 1-15.

Clarke, K., and Ryan, S. 2004. Ecological Assessment of the Queensland Blue Swimmer Crab Pot Fishery, Department of Primary Industries and Fisheries - Queensland Fisheries Service, p.9-11.

Celik, M., Tureli, C., Yanar, Y., Edem, U., and Kucukgulmez, A., 2004. Fatty acid composition of the blue crab (*Callinectes sapidus*) in the north eastern Mediterranean, *Food Chemistry*, vol. 88, pp. 271-273.

Dolan, M.C., Bennet, S., Medina-Bolivar, F., Applewhite, L.D., and Green, D.P. 2002. Species identification of pasteurized crabmeat by a discriminating PCR-based diagnostic, presented at the 2002 Annual Meeting and Food Expo, Anheim, CA.

Downey, G. 1998. Food and Food Ingredient Authentication by Mid-Infrared Spectroscopy and Chemometrics, *Trends in Analytical Chemistry*, vol. 17, no.7., pp. 418-424.

Eggleston, D., B., Hightower, J., and Johnson, E., 2002. Population dynamics of the blue crab Callinectes sapidus in North Carolina. Fishery Resource Grant Report 00FEG-11.

Enzerross, R., Enzerross, L., and Bingel, F. 1997. Occurrence of blue crab, *Callinectes sapidus* (Rathbun, 1896) (Crustaceae, Brachyura) on the Turkish Mediterranean an Adjent Aegean Coast and its size distribution in the Bay of Iskenderun, *Turkish Journal of Zoology*, vol. 21, pp. 113-122.

Esbensen, K., Schonkopf, S., and Midtgaard, T. 1994. Multivariate analysis in practice. CAMO AS, Wennbergs Trykkeri AS, Trondheim, Norway.

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., and Ferre, J. 2002. Transfer of multivariate calibration models: a review, *Chemometrics and Intelligent Laboratory Systems*, vol. 64, pp. 181-192.

Geladi, P., and Kowalksi, B.R. 1986. Partial Least-Squares Regression: A tutorial, *Analytica Chimica Acta*, vol. 185, pp. 1-17.

Gokoolu, N., and Yerlikaya, P. 2003. Determination of proximate composition and mineral contents of blue crab (*Callinectes sapidus*) and swim crab (*Portunus pelagicus*) caught off the Gulf of Antalya, *Food Chemistry*, vol. 80, pp. 495-498.

Golovkin, I.E., Mancini, R.C., Louis, S.J., Lee, R.W., and Klein, L. 2002. Analysis of X-ray spectral data with genetic algorithms, Journal of Quantitative Spectroscopy & Radiative Transfer, vol. 75, pp. 625-636.

Gusnanto, A., Pawitan, Y., Huang, J., and Lane, B. 2003. Variable selection in random calibration of near-infrared instruments: ridge regression and partial least squares regression settings, *Journal of Chemometrics*, vol. 17, pp. 174-185.

Harris, D.A. 1996. Light Spectroscopy, BIOS Scientific Publishers Ltd.: Oxford, pp. 1-36.

Haupt, R.L., and Haupt, S.E. 1998. Practical Genetic Algorithms, John Wiley & Sons, Inc., New York, pp. 1-48.

Hong, G.P. 1991. A Technology Analysis of the U.S. Atlantic Blue Crab (*Callinectes sapidus*) Processing Industry. Ph.D. Dissertation, Virginia Institute of Technology and State University.

Hof, M. 2003. Basics of optical spectroscopy, In, Gauglitz, G., and Vo-Dinh, T. (Eds.), Handbook of spectroscopy, WILEY-VCH, Weinheim, pp. 39-46.

Jahncke, M., Green, D., Rippen, T., Hicks, D., Reilly, T., Collette, R., Fernandes, C., Huang, Y., and Fisher, R. 2000. A plan to improve the competitiveness of the Blue crab processing industry and the quality of crab meat products in the U.S. marketplace, Sensory Working Group, VSGCP-G-00-001.

Johnson, J.A., Jr., Green, D.P., and Martin, R.E. 1998. Industry Perspectives: The Hard Blue Crab Fishery – Atlantic and Gulf. *Journal of Shellfish Research*, vol.17(2), pp. 371-374.

Kalivas, J. H., and Lang, P. M. 1994. Mathematical analysis of spectral orthogonality, Marcel Dekker, Inc.: New York, pp. 1-74.

Lavine, B. K., and Workman, J., Jr. 2005. Chemometrics: past, present, and future, In, Lavine, B. K. (Ed.), Chemometrics and Chemoinformatics, American Chemical Society, ACS Symposium Series 894, pp. 1-14.

Lazraq, A., and Cleroux, R. 2001. The PLS multivariate regression model: testing the significance of successive PLS components, *Journal of Chemometrics*, vol. 15, pp. 523-536.

Leardi, R. 2003. Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets, In Leardi, R. (Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier: New York, pp. 169-196.

Luke, B. T. 2003. Genetic algorithms and beyond, In Leardi, R. (Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier: New York, pp. 3-48.

Manly, B.F.J. 1986. Multivariate statistical methods: a primer, Chapman and Hall, New York.

Martin, K.A. 1992. Recent advances in near-infrared reflectance spectroscopy, *Applied spectroscopy reviews*, vol. 27, no. 4, pp. 325-383.

McClure, W.F. 1994. In, Wilson, R.H., Spectroscopic techniques for food analysis, VCH Publishers, Inc, New York, pp. 13-58.

NMFS. 2006. Personal communication from the National Marine Fisheries Service, Fisheries Statistics Division, Silver Spring, MD.

Osborne, B.G., Fearn, T., and Hindle, P.H. 1993. Practical NIR spectroscopy with applications in food and beverage analysis, Longman Scientific & Technical, John Wiley & Sons, Inc., New York, pp. 13-36.

Osterling, M.J. 1998. "Blue" Crab Resources in Other Countries: Implications for the U.S. Industry, *Journal of Shellfish Research*, vol. 17, no. 2, pp. 375-378.

Otto, M. 1999. Chemometrics: Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, New York, pp. 1-11.

Reddy, N.R., Flick, G.J., Argnosa, G.C., and Young, R.W. 1991. Composition and Pesticide and Herbicide Residue Analysis of Fresh and 40-year-old Pasteurized Blue Crab (*Callinectes sapidus*) Meat. *Journal of Food Protection*, vol. 54, no. 4, pp. 298-301.

Rugolo, L.J., Knotts, K.S., and Lange, A.M., 1998. Historical profile of the Chesapeake Bay Blue Crab (*Callinectes sapidus* Rathbun) Fishery, *Journal of Shellfish Research*, vol. 17, no. 2, pp. 383-394.

Smith, B. 1999. Infrared spectral interpretation – a systematic approach, CRC Press, New York, pp. 1-29.

Smith, B. 2002. Quantitative spectroscopy: theory and practice, Academic Press, New York, pp. 1-41.

Sutherland, D. 2006. Personal communication via email. National Oceanic and Atmospheric Administration (NOOA): National Marine Fisheries Service – Fisheries Statistics Division, Silver Spring, MD.

Tsai, D.E., Chen, H.C., and Tsai, C.F. 1984. Total lipid and cholesterol content in the blue crab, *Callinectes sapidus*, *Comparative Biochemistry and Physiology*, vol. 78B, No. 1, pp. 27-31.

Tureli, C., Celik, M., and Erdem, U. 2000. Comparison of meat composition and yield of blue crab (*Callinectes sapidus*) and sand crab (*Portunus pelagicus*) caught in Iskenderun Bay, North-East Mediterranean, *Turkish Journal of Veterinary and Animal Science*, vol. 24, pp. 195-203.

Ward, D.R. 2000. Processing blue crab, shrimp, and king crab, In Martin, R.E., Carter, E.P., Flick, G.J. (Jr.), and Davis, L.M. (Eds.), Marine & Freshwater Products Handbook, Technomic Publishing Company, Inc, Lancaster, Pennsylvania, pp. 299-307.

Wold, S., Sjotrom, M., and Eriksson, L. 2001. PLS – regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems,* vol. 58, pp. 109-130.

Yee, N., and Gimel'farb, G. 1999. Multi-band spectral measurement for moisture content, Computer Science Department of the University of Auckland, CITR at Tamaki Campus, New Zealand, CITR-TR-43.

3.0 MANUSCRIPT #1

Quantitative Analysis and Detection of Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy

Abstract

Visible and Near-Infrared (VIS/NIR) spectroscopy has been used to detect economic adulteration of crabmeat samples. Atlantic blue and blue swimmer crabmeat samples were adulterated with surimi-based imitation crabmeat in 10% increments. Waveform evaluation revealed that the main features seen in the spectral data arise from water absorptions with a decrease in sample absorbance with increasing adulteration level. Prediction and quantitative analysis was done using untreated data, a 15-point smoothing average, first derivative, second derivative pre-treated data, and a 150 wavelength spectral data gathered from a correlogram. Regression analysis included Partial Least Squares (PLS) and Principal Component Regression (PCR). Both models were able to perform similarly in predicting crabmeat adulteration. The best model for both PLS and PCR used the first derivative spectral data gathered from the correlogram, with a standard error of prediction (SEP) of 0.252 and 0.244, respectively. The results suggest that VIS/NIR technology can be successfully used to detect adulteration in crabmeat samples adulterated with surimi-based imitation crabmeat.

Keywords: Adulteration, VIS/NIR Spectroscopy, Crabmeat, PLS, PCR

1. Introduction

Economic adulteration of food products involves the substitution of cheaper and inferior, i.e. lower quality, ingredients for high-cost ingredients. Even though economic adulteration rarely presents a health hazard, it is an issue because it defrauds the consumer and undercuts legitimate industry prices. For this reason, there is a need for improved detection to address the problem of economic adulteration in the food industry, especially for "value-added" products, such as juices (orange and apple), honey, olive oil, and seafood.

Most of the initial research to detect adulteration has focused on detailed and expensive methodologies involving gas chromatography and mass spectrometry to identify unique chemicals that distinguish one ingredient from another (FDA, 1999). However, Visible and Near-Infrared (VIS/NIR) spectroscopy is an objective tool that can be very powerful when properly applied and requires minimal sample preparation and destruction. VIS/NIR technology has been used to determine the authenticity of olive oil that was adulterated with vegetable oils (Tay *et al*., 2002; and Lai *et al*., 1995), to detect, along with chemometrics, honey adulterated with sugar solutions (Kelly *et al*., 2004), and apple adulteration in diluted and sulfited strawberry and raspberry purees (Downey and Kelly, 2004). While previous research has been done to detect adulteration in honey, juices, and olive oil, little research has focused on seafood products, such as crabmeat.

Seafood is one of the many high-value products that are targets of intentional adulteration and economic fraud. Examples include overbreading shrimp and overglazing lobster tails, thus providing inaccurate net weights, and species substitution (FDA, 1993). Due to the reduced number of Atlantic blue crabs and the increase in

foreign imports, the Atlantic blue crabmeat market has become a target for economic adulteration. The greatest concentration of the blue crab industry is in the Chesapeake Bay, an area that once provided more crabs than anywhere else in the world (Johnson *et al.*, 1998). In 2002, 36.4 million pounds of crabs, valued at $29.3 million, were harvested in North Carolina, one of the top blue crabs producers in the nation (NFMS, 2002). A decrease in the blue crab population, several major hurricanes, and an increase in crabmeat imports, have resulted in a dramatic reduction in the number of commercial crabbers because of the substantial negative economic impact (Hampton, 2003). In order to continue in the business and increase revenues, some companies have gone as far as economically adulterating their crabmeat products. One such example is Miss Sally's Stuffed Crabs from Sam's Club membership stores (FDA, 1995). The packages contained pictures of crab shells stuffed with chunks of white meat, and the product claimed to have "more crabmeat than ever." Upon FDA inspection, however, instead of being stuffed with crabmeat, the shells were found to contain surimi-based imitation crabmeat.

The main objective of this study was to use VIS/NIR spectroscopy to determine the level of adulteration with surimi-based imitation crabmeat in two types of crabmeat, Atlantic blue **(***Callinectes sapidus***)** and an imported pasteurized frozen blue swimmer crabmeat **(***Portunus pelagicus***)**. The imitation crabmeat was chosen as the adulterant product due to its low cost, similar flavor and consistency, and widespread commercial availability. In addition, several data pre-treatments (a 15 point moving average, first derivative, and second derivative), as well as the use of a smaller data set showing high correlation with adulteration level, were used to determine effect on model performance.

2. Materials and Methods

2.1. Sample Preparation

Three pounds of Atlantic blue crabmeat were obtained from a regional crabmeat supplier (Sea Safari, LTD., Belhaven, NC), as well as three pounds of an imported pasteurized frozen blue swimmer crabmeat. A total of six one-pound packages (three for each type of crabmeat) were stored in a refrigerator (4˚C). Six pounds of surimi-based imitation crabmeat (Emerald Sea™) were obtained from a local supermarket and also stored at 4˚C until sample preparation and analysis.

Prior to analysis, pure and imitation crabmeat samples (3 lbs. each) were tempered to room temperature (25˚C) in a water bath. The three pounds of pure crabmeat were then pooled in a clean container and thoroughly mixed with gloved hands. The gloves were washed with water to get rid of powdery residue and dried with paper towels to remove excess moisture prior to hand mixing the pooled crabmeat. The same method was used to mix imitation crabmeat in a different container. Pure and imitation crabmeats were each weighed and mixed to obtain a total sample weight of 75g. Adulteration of pure crabmeat was completed in 10% increments (imitation crabmeat weight/total sample weight). For example, samples in the 10% adulteration class consisted of 67.5g of pure crabmeat and 7.5g of surimi-based imitation crabmeat for a total of 75g. Each pure-to-imitation weight combination corresponded to one of 11 adulteration classes, i.e. class 0: 0% adulteration, class 1: 10% adulteration, and so forth, until class 10: 100% adulteration.

Once weighed, each adulteration sample was placed in a blender and mixed for five 1-s intervals. Homogenized samples were placed in polyethylene bags, labeled, and

stored at room temperature (25˚C) until analysis. Samples were prepared in triplicate for each of the 11 classes: pure crabmeat (0), 9 classes of adulteration (1-9), and imitation crabmeat (10), giving a total of 33 samples (3 samples/ adulteration class) per type of crabmeat. In order to minimize the effect on the percentage of adulteration, the samples were prepared sequentially, from lowest to highest, according to adulteration class. The blender container was cleaned before samples composed entirely of pure or imitation crabmeat were prepared.

2.2. Data Analysis

Spectra from 400 to 2498 nm at 2 nm intervals (a total of 1050 wavelengths) were recorded in log(1/R) units at ambient room temperature (25°C) using a NIRSystems 6500 spectrometer (FOSS NIRSystems, Silver Spring, MD) equipped with a rectangular sample chamber. For each crabmeat sample, the recorded spectra consisted of the average of 32 sample scans. Each sample was randomly scanned four different times. The four replicates per sample were averaged to obtain a new data set consisting of 33 sample scans (3 sample scans / adulteration class) per type of crabmeat.

Exploratory data analysis, model development and verification, and calibration and validation were performed using Unscrambler software version 7.6 (CAMO Software Inc., OR, USA). Preliminary waveform evaluation of the untreated data set was performed using a correlogram, i.e. a plot between the wavelength data and adulteration class in terms of correlation coefficients. Correlograms are useful tools for determining wavelength importance in terms of a given attribute (McClure, 1994). Based on this waveform evaluation, a smaller data set (150 wavelengths), corresponding to a

continuous wavelength range exhibiting the highest correlation coefficient ($-0.941<R^2<-0.982$) between wavelength and adulteration class, was also selected for quantitative analysis.

Quantitative VIS/NIR analysis was developed using Partial Least Squares (PLS) and Principal Component Regression (PCR). The full spectrum data set was divided into a training set (two thirds) for calibration and a testing set (one third) for prediction with each set containing 44 samples (22 for each crabmeat type) and 22 samples (11 for each crabmeat type), respectively, across the range of adulteration classes. Three data pre-treatments (a 15-point smoothing moving average, first derivative, and second derivative) and the untreated data were used to determine the effect of data pretreatment on regression model prediction. The smaller data set obtained from the correlogram, was also used to determine effect on model prediction. The accuracy of the prediction models was examined in terms of a low standard error of calibration (SEC), a low standard error of prediction (SEP), and a high coefficient of determination (R^2).

3. Results and Discussion

3.1. Spectra

The average absorbance spectra for the 11 classes of adulterated Atlantic blue and blue swimmer crabmeat samples are shown in Figures 1 and 2, respectively. The main features of the spectra arise from water absorptions. High water absorptions are found in the NIR range. At 20°C, pure water has maxima at 970, 1190, 1450, and 1940 nm (Segtnan *et al*., 2001; Curcio and Petty, 1951). According to previous studies, three overtones, 1450, 970, and 760 nm, correspond to the first, second, and third overtones of

water, respectively. A combination of O-H stretching and bending also occurs around 1940 nm, and the 1190 nm maximum is attributed to second overtones of C-H stretching. With the exception of the third water overtone, which was not seen, the peaks of the crabmeat spectra (Figures 1 and 2) are consistent with published literature concerning water absorption. Even though the main features correspond to water absorption, VIS/NIR spectra usually describe many other interactions that give high absorptions at specific wavelengths, such as combination bands of C-H in carbohydrates, N-H in proteins, as well as double bonds of C=0 and C=C. Biomolecules often contain these types of bonds, and many biological materials contain a high water composition; therefore, these absorptions are to be expected. Tables of chemical assignments to spectral peak absorptions are easily found in NIR application books, such as Osborne *et al*., 1993.

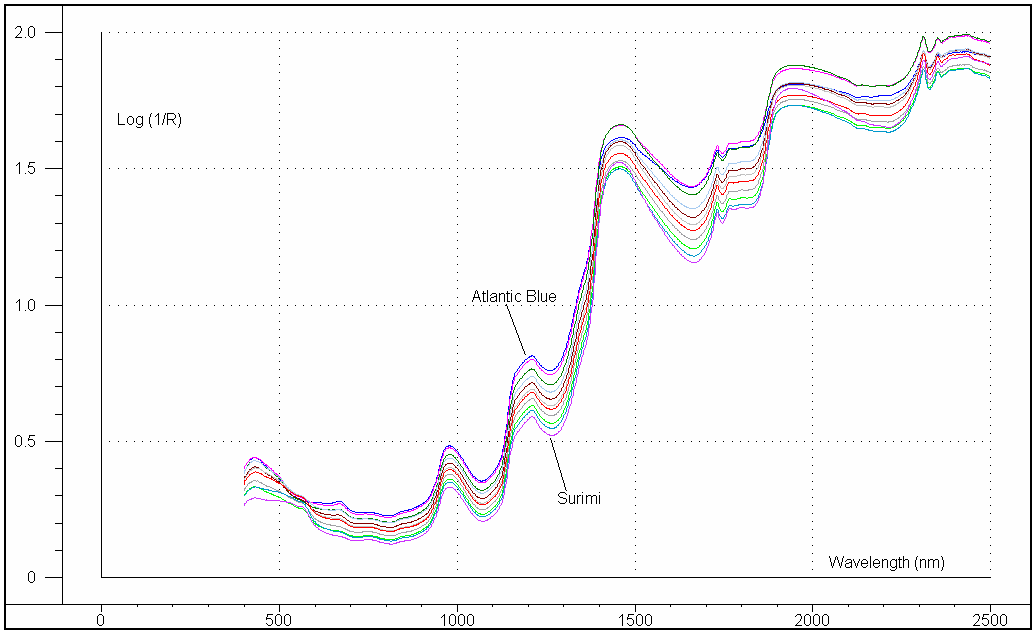

Figure 1: Average absorbance spectra of adulteration samples of Atlantic blue crabmeat. Top spectra represents class 0 (0% adulteration), bottom spectra class 10 (100% adulteration), and classes 1-9 (10%-90% adulteration) are in order from top to bottom.

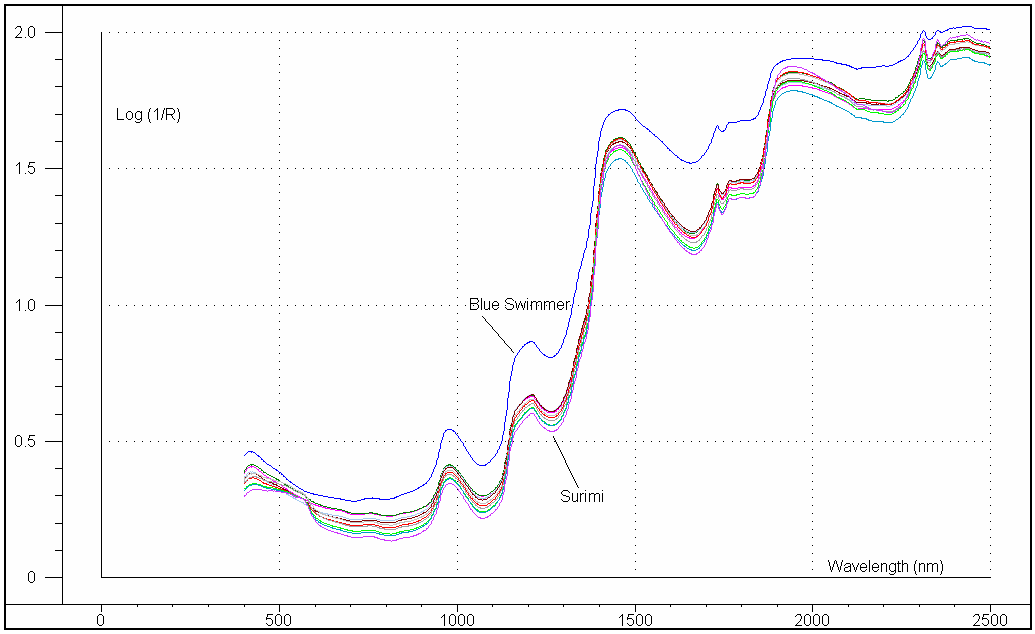

Figure 2: Average absorbance spectra of adulteration samples of blue swimmer crabmeat. Top spectra represents class 0 (0% adulteration), bottom spectra class 10 (100% adulteration), and classes 1-9 (10%-90% adulteration) are in order from top to bottom.

In both cases, the spectra of the samples pertaining to 100% crabmeat (0% adulteration), have a higher absorbance than the adulterated or surimi-based imitation crabmeat samples. As the level of adulteration increases, the absorbance decreases at every wavelength. The one exception occurs for the Atlantic blue crabmeat (Figure 1), which shows no visible distinction between the spectra of the 0% adulteration class and the spectrum for 10% adulteration. The blue swimmer crabmeat (Figure 2), however, reveals a large gap between the sample with 0% adulteration (class 0) and that with 10% adulteration (class 1). This difference in spectra can probably be attributed to the fact that the blue swimmer crabmeat was pasteurized while the Atlantic crabmeat did not undergo this processing treatment. Previous research has shown that NIR analysis is able to detect pasteurized versus un-pasteurized crabmeat (Requena, 1998).

Even though pasteurization might explain the observed spectral difference between types of crabmeat, it does not explain why there is the large gap between the 0% adulteration sample and the 10% adulteration sample of the blue swimmer crabmeat. It is possible this was due to interactions between the surimi-based imitation crabmeat and the blue swimmer pasteurized crabmeat. During sample preparation, it was noted that the blue swimmer crabmeat contained more moisture than the Atlantic blue crabmeat. This is to be expected because the blue swimmer crabmeat was an imported frozen food item. Most non-dried imported foods contain a higher moisture level, a post harvest food safety measure, to ensure that the food quality is maintained while in storage. The imitation crabmeat, however, is relatively free of moisture; therefore, it probably absorbed water from the blue swimmer crabmeat, reducing the amount of moisture in the overall sample. After the 10% adulteration sample, however, the spectra resemble the pattern seen in the Atlantic blue crabmeat samples.

Figure 3 shows the correlogram between the untreated wavelength data and adulteration class. Overall, the correlation is negative with a high negative correlation to adulteration level in the 600 to 1400 nm region. In the case of VIS/NIR applications, the use of a small set of optimal wavelengths that contains the most important information is desired. From this correlogram, a smaller data set containing only 150 wavelengths (602-900 nm), i.e. those with higher negative correlations ($-0.941<R^2<-0.982$), was chosen for quantification analysis.

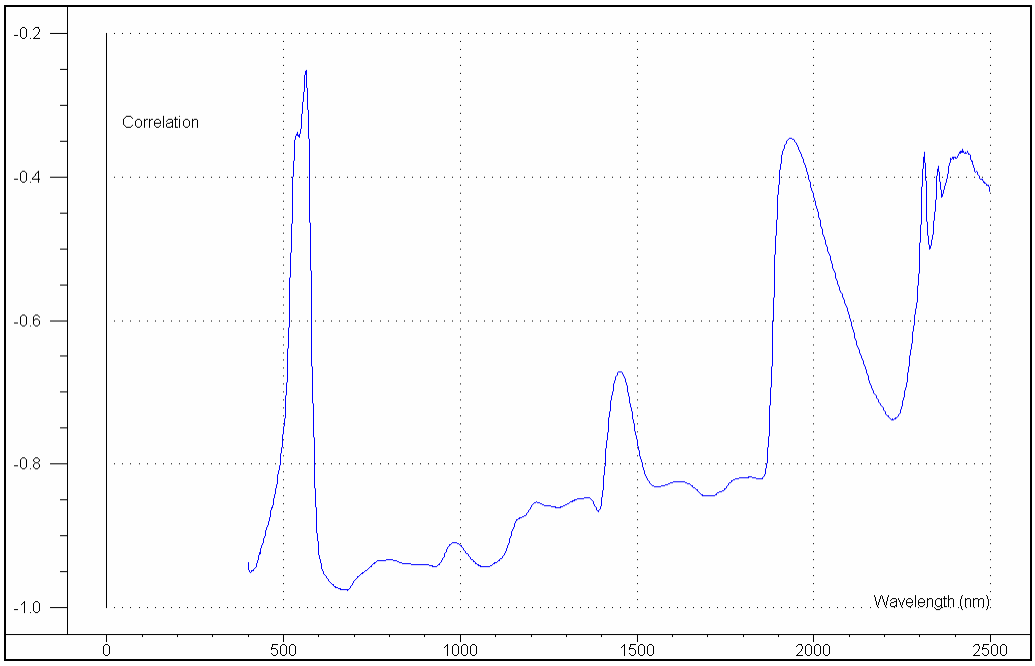

Figure 3: Correlogram between wavelength data and adulteration class for blue swimmer and Atlantic blue crabmeat data set.

3.2. Quantification of Adulteration Level

In order to quantify the level of adulteration, sixteen models were developed for the data set that contained both the Atlantic blue and the blue swimmer crabmeat. The untreated data, data from the three pre-treatments (a 15-pt smoothing moving average, the first derivative, and second derivative), and the correlogram data set composed of only the 602-900 nm wavelength data were used to develop regression models using PLS and PCR.

The number of factors, i.e. principal components, used in each model was chosen to optimize model performance and minimize model errors, such as those caused by underfitting and overfitting the data, by looking at the residual variance. Figure 4 shows a typical plot of the residual variance in terms of the number of factors. For this example, five factors were required to reach a stable minimum residual validation variance; hence,

five factors were used in the PLS model. Choosing more factors than required often leads to overfitting the data. In this case, even with a low SEC, the model will be unable to accurately predict unseen samples and will generate a high prediction error.

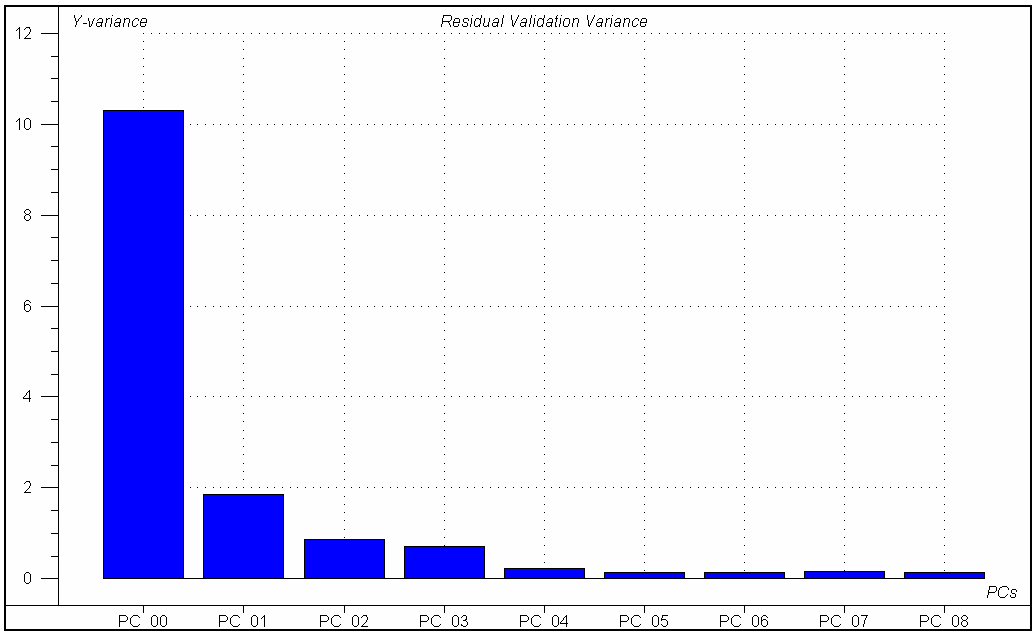

Figure 4: Plot of adulteration level residual variance versus number of factors used in model calibration.

A summary of the PLS results obtained from the quantification analysis is given in Table 1. Overall, the SEC of the models is low (<0.460) with the exception of the second derivative data of the correlation data set (SEC=0.609) and all have a high (>0.980) coefficient of determination. The SEP is higher than the SEC for each model but still shows, with the exception of the second derivative data sets, an error less than 0.57. The 15-point smoothing average did not optimize model performance (SEC=0.396, SEP=0.480) over models with the untreated data (SEC=0.386, SEP=0.468). A slight decrease in both SEC and SEP was noted in the smoothing average model using the

correlation data set; however, due to the small difference, this type of data pre-treatment was not useful in enhancing the model's ability to predict adulteration level.

Table 1: Summary results of quantitative analysis performed with PLS. The best model is indicated in bold.

		Calibration		Prediction	
Data Treatment	Factors	SEC	R^2	SEP	R^2
None	5	0.386	0.993	0.468	0.989
None + Corr	5	0.459	0.990	0.564	0.985
15-pt	6	0.396	0.992	0.480	0.989
15-pt + Corr	6	0.449	0.990	0.549	0.986
1st der	5	0.483	0.989	0.545	0.986
1st der + Corr	**5**	**0.251**	**0.997**	**0.252**	**0.997**
2nd der	10	0.112	0.999	1.926	0.820
2nd der + Corr	5	0.609	0.982	0.838	0.966

Generating a low SEC does not necessarily imply that a model is adequate and able to predict future unseen data samples. An example is the model performance of the second derivative pre-treatment data set. Even though the model generated the lowest SEC (0.112), it also generated the highest SEP (1.926). It is possible that this model overfitted the data; hence, the low error in the calibration data set. Verification of the model's performance, however, is done by predicting samples that were not used in the model development during calibration. The SEP is indicative of the model's ability to accurately predict samples in terms of adulteration class. The high SEP indicates that the model is not adequate or effective in accurately predicting samples according to level of adulteration. In VIS/NIR quantitative analysis, 2*SEP is regarded as a 95% confidence interval (Kelly *et al*., 2004; Esbensen *et al*., 1994). Therefore, this model is able to predict adulteration within ±1.926 classes. Since each class corresponded to a 10% increase in adulteration, the true estimate of the sample lies within almost ±2 classes (or

±20% adulteration). In other words, this model cannot accurately estimate samples below 20% adulteration. In addition, the model will falsely classify a sample, i.e. a sample actually in class 4 (40%) classified as a class 6 (60%); hence, it is not an accurate model for determining crabmeat adulteration.

The best model used the first derivative data of the correlation data set, as data pre-treatment, and generated the lowest SEC and SEP, 0.251 and 0.252, respectively, while also generating the highest coefficient of determination for both calibration and prediction data (R^2=0.997). The errors associated with this model imply that future unseen samples will be able to be predicted within ±0.252 of a level or 2.52% of the true adulteration percentage.

A plot of the first two principal components (96% of the total variation) of the first derivative correlation data (Figure 5) shows an interesting relationship not only in the adulteration level but also in the type of crabmeat. There is a visible trend (left to right) describing level of adulteration, with samples belonging to 0% adulteration on the left and 100% adulteration to the right. In addition, there is another trend, one for each type of crabmeat, which describes adulteration level. Based on this observation, the first factor (PC1: 73% variance) contains information that describes adulteration level; whereas, the second factor (PC2: 23% variance) seems to contain information describing crabmeat type since it almost separates the two types of crabmeat. The horizontal linear trend (top part of the graph) corresponds to samples from the Atlantic blue crabmeat, while the almost 45˚ linear trend (bottom right) corresponds to the blue swimmer crabmeat. Even though the two types of crabmeat are differentiated at the lower adulteration levels, they seem to converge at a common region in the upper right quadrant

of the graph, which occurs at the samples containing 100% adulteration, i.e. surimi-based imitation crabmeat. Hence, the first two factors contain enough information to be able to determine a relationship not only for adulteration in general, but also for adulteration in terms of each type of crabmeat. The convergence of the surimi-based imitation crabmeat samples is to be expected because all the samples in this adulteration class should depict a similar spectral pattern.

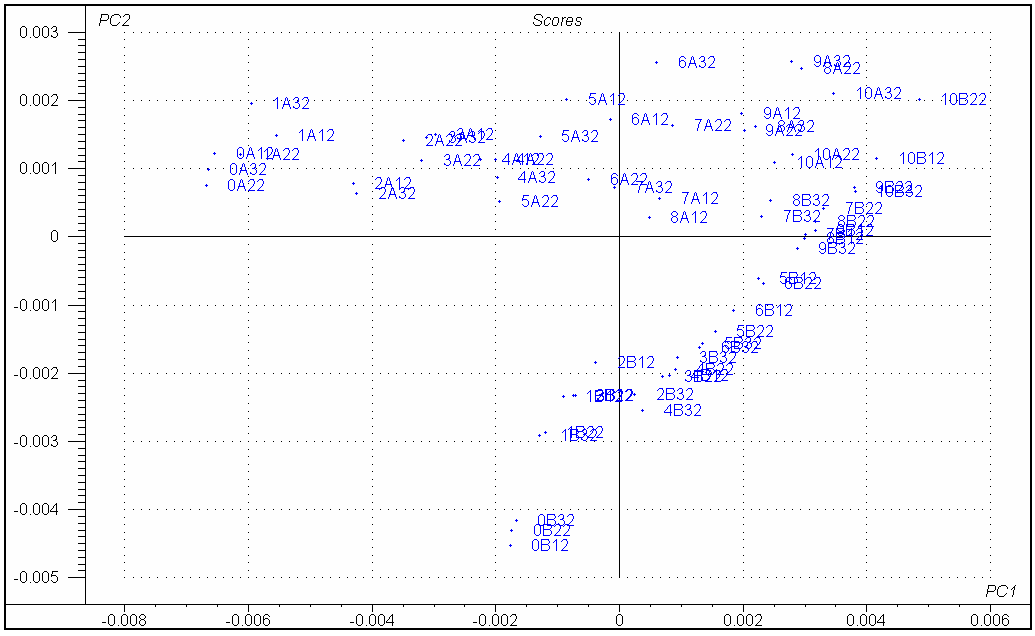

Figure 5: Plot of the first two principal components of PLS model with first derivative correlation data. Explained variation in PC1 and PC2 were 73% and 23%, respectively. Samples are named according to adulteration level (from 0, i.e. no adulteration, to 10, i.e. 100% adulteration), crabmeat type (A = Atlantic blue, B = blue swimmer), and sample number (1-3). Due to repetitions being averaged, the last number is constant throughout the samples.

Figure 6 shows the actual versus predicted adulteration level for calibration and prediction corresponding to the first model of Table 1 (i.e. no data treatment and 5 factors). The model has a SEC=0.386 (R^2=0.993) and a SEP=0.468 (R^2=0.989), giving a very good data fit, not only for the calibration but also for the prediction. The high

coefficient of determination of the SEP is a good indicator that the results of the model are reliable and the model is able to detect adulteration level. For the PLS models, the first derivative of the correlation data set gave the best performance (boldface in Table 1).

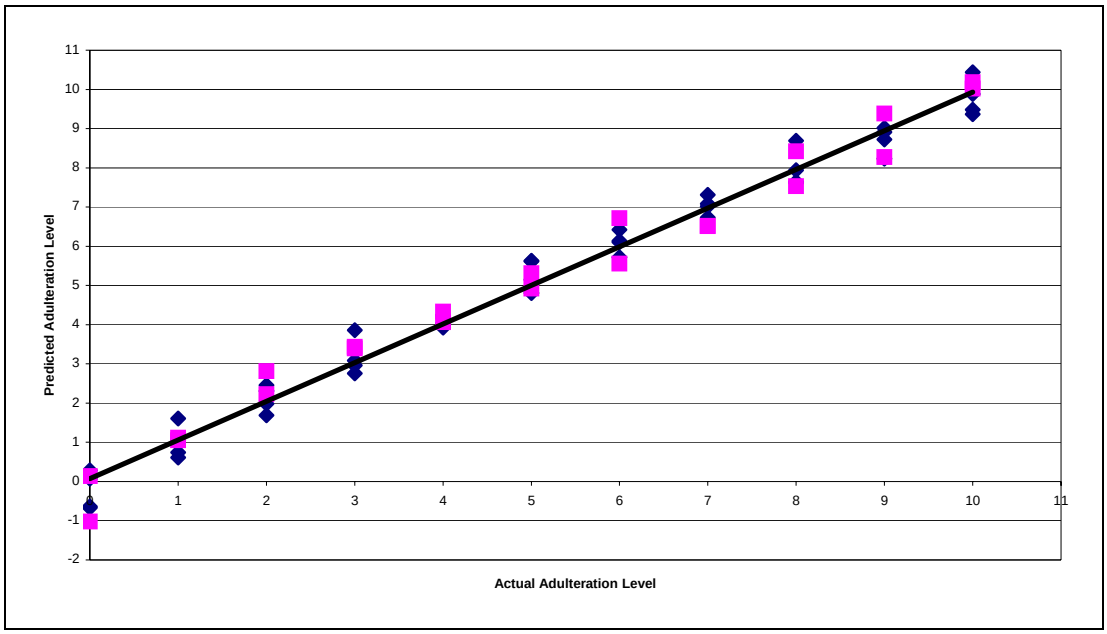

Figure 6: Actual versus predicted adulteration level of the untreated data. PLS model with 5 factors showing the calibration data (◊) and the prediction data (□).

The second derivative data alone generated the lowest SEC (0.112) but was the worst model in terms of SEP (1.926), and required a higher number of factors. The correlation data using the second derivative, on the other hand, produced lower errors of calibration and prediction (Table 1). A plot of the first two principal components of the second derivative data of the correlation data (Figure 7) indicates that the variation found in the data set was explained by the difference in type of crabmeat, not necessarily adulteration class. While there is a visible adulteration pattern in the Atlantic blue crabmeat samples, the blue swimmer crabmeat samples seem to have three main groups

in which adulteration classes are mixed (0 to 2, 3 to 5, and 6 to 10) indicating a lack of a sequential relationship of adulteration in the blue swimmer crabmeat samples. Even though this model (2nd der + Corr) gives errors less than 0.9 for both SEC and SEP and corresponding high coefficient of determination (>0.96), it is not likely to be an acceptable method for detecting level of adulteration since it cannot accurately predict adulteration of blue swimmer crabmeat.

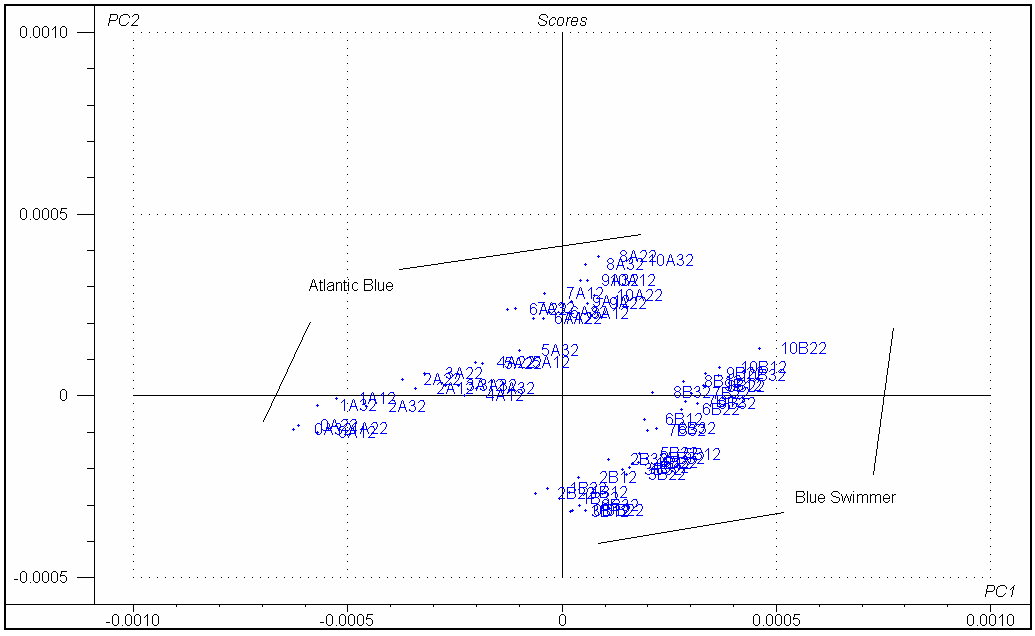

Figure 7: Plot of the first two principal components of PLS model with the 2nd derivative correlation data. Explained variation in PC1 and PC2 were 58% and 18%, respectively. Samples are named according to adulteration level (from 0, i.e. no adulteration, to 10, i.e. 100% adulteration), crabmeat type (A = Atlantic blue, B = blue swimmer), and sample number (1-3). Due to repetitions being averaged, the last number is constant throughout the samples.

A summary of the PCR results obtained from the quantification analysis is given in Table 2. Of the eight models, the one with the first derivative of the correlation data set (with 6 factors) also resulted in the lowest SEC (0.278) and SEP (0.244) and the highest coefficient of determination for both the calibration and prediction data sets

(R^2>0.996). Based on the SEP, PCR is able to predict the true value of an unseen sample within a level of ±0.244, or 2.44% adulteration. Graphical representation of the first two factors shows results similar to those in Figure 5.

Table 2: Summary results of quantitative analysis performed with PCR. The best model is indicated in bold.

		Calibration		Prediction	
Data Treatment	Factors	SEC	R^2	SEP	R^2
None	8	0.383	0.993	0.414	0.992
None + Corr	8	0.329	0.995	0.488	0.989
15-pt	8	0.390	0.993	0.420	0.992
15-pt + Corr	7	0.323	0.995	0.476	0.990
1st der	6	0.610	0.982	0.576	0.984
1st der + Corr	**6**	**0.278**	**0.996**	**0.244**	**0.997**
2nd der	10	1.393	0.900	2.211	0.770
2nd der + Corr	10	0.751	0.972	0.912	0.960

Overall, PCR generated a lower SEC and SEP than PLS in all the models, with the exception of the model using the first derivative data and any model that used the second derivative data. A big difference between PLS and PCR is the number of factors that were required to achieve a constant minimum variance, with PCR requiring more factors than PLS. The number of factors for PLS was relatively constant (5-6), with the exception of the second derivative, whereas the factors for PCR range from 6 (for the first derivative) to 10 (for the second derivative). The worst models are also those developed from the second derivative data, though, as seen with PLS, the correlated data gives better results than the model with just the second derivative data.

A plot of the first two principal components of the second derivative data (Figure 8) reveals two clusters, one pertaining to Atlantic blue crabmeat (left cluster) and the other to the blue swimmer crabmeat (right cluster). While PLS was able to model the

level of adulteration of the Atlantic blue crabmeat (Figure 7), PCR, using the same data set, was not able to model the samples according to adulteration class though it is able to group them into type of crabmeat. For the purpose of this experiment, which was to determine amount of adulteration with surimi-based imitation crabmeat, this type of model was found to be inadequate and cannot predict adulteration, regardless of the type of crabmeat.

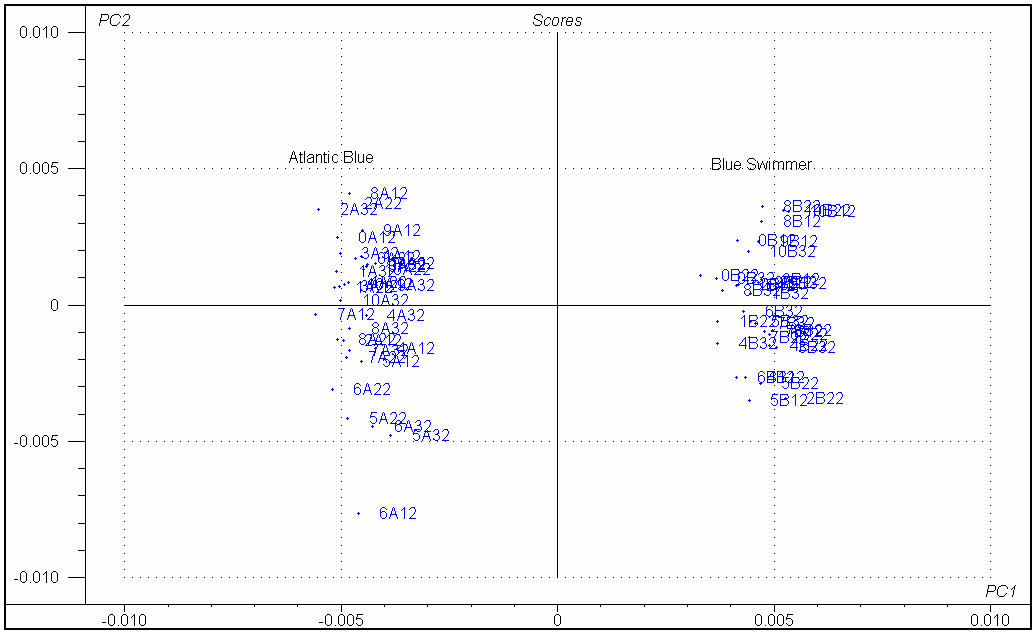

Figure 8: Plot of the first two principal components of PCR model with the 2nd derivative data. Explained variation in PC1 and PC2 were 28% and 8%, respectively. Samples are named according to adulteration level (from 0, i.e. no adulteration, to 10, i.e. 100% adulteration), crabmeat type (A = Atlantic blue, B = blue swimmer), and sample number (1-3). Due to repetitions being averaged, the last number is constant throughout the samples.

4. Conclusions

Adulteration of high quality and high priced food products, such as crabmeat, is a commercial and economic problem. The widespread availability of surimi-based imitation crabmeat makes it an ideal adulterant for crabmeat products. In this study,

VIS/NIR spectroscopy was used to detect adulteration level in two types of crabmeat. PCR generated lower SEC and SEP for the models generated with the untreated data and the 15-point smoothing moving average data. Overall, however, the PLS models were preferred because they used less factors in model development than PCR. Data pre-treatments are often used to optimize model performance. Based on the results from this study, a 15-point smoothing average did not enhance the model's ability to predict samples and created no advantage over models derived with the untreated data set. Even though PLS using the second derivative data generated lower calibration and prediction errors than PCR, this data pre-treatment was not beneficial to optimize the models due to the large error associated with the prediction (SEP) of new adulterated samples. Additionally, the second derivative data sets seemed to contain more information about the type of crabmeat rather than adulteration class. The results here indicate that using the first derivative is more successful in modeling adulteration than the second derivative or the 15-point smoothing average. A smaller set of correlation data was found to enhance model performance, especially with the first derivative data, since it contained information that pertained to both type of crabmeat and also adulteration class. The results from this study indicate that it is possible to detect adulteration using VIS/NIR analysis in crabmeat samples adulterated with surimi-based imitation crabmeat.

References

Curcio, J.A., and Petty, C.C. 1951. The near infrared absorption of liquid water, *Journal of Optical Society of America*, vol. 41, pp. 302-304.

Downey, G., and Kelly, J.D. 2004. Detection and quantification of apple adulteration in diluted and sulfited strawberry and raspberry purees using visible and near-infrared spectroscopy, *Journal of Agricultural and Food Chemistry*, vol. 52, pp. 204-209.

Esbensen, K., Schonkopf, S., and Midtgaard, T. 1994. Multivariate analysis in practice. CAMO AS, Wennbergs Trykkeri AS, Trondheim, Norway.

FDA. FDA Consumer. 1993. "Is Something Fishy Going On?" U.S. Food and Drug Administration, FDA Consumer September 1993, vol. 27, no. 7. Online. http://www.fda.gov/fdac/

FDA. FDA Consumer – Investigators' Reports. 1995. "Seafood Maker Fined for Misbranding," U.S. Food and Drug Administration, FDA Consumer November 1995, vol. 29, no. 9. Online. http://www.fda.gov/fdac/

FDA. FDA Consumer. 1999. "Fake Food Fight," U.S. Food and Drug Administration, FDA Consumer March-April 1999, vol. 33, no. 2. Online. http://www.fda.gov/fdac/

Hampton, J.S. 2003. "Hard-crab harvest in N.C. waters skyrocket; the take's worth for 2002 is valued at $29.3 million,' *The Virginian-Pilot*, Landmark Communications, Inc. Online.

Johnson, J.A., Green, D.P., and Martin, R.E. 1998. Industry perspectives: the hard Blue crab fishery - Atlantic and Gulf. *Journal of Shellfish Research*, vol. 17, No. 2, pp. 1-4.

Kelly, J.F.D., Downey, G., and Fouratier, V. 2004. Initial study of honey adulteration by sugar solutions using midinfrared (MIR) spectroscopy and chemometrics, *Journal of Agricultural and Food Chemistry*, vol. 52, pp. 33-39.

Lai, Y.W., Kemsley, E.K., and Wilson, R.H. 1995. Quantitative analysis of potential adulterants of extra virgin oil using infrared spectroscopy, *Food Chemistry*, vol. 53, pp. 95-98.

McClure, W.F. 1994. Near-infrared spectroscopy, In Wilson, R.H. (ed.) Spectroscopic Techniques for Food Analysis, VCH Publishers, Inc., New York, New York, pp. 13-52.

NFMS. 2002. National Marine Fisheries Service. Personal communication from the National Marine Fisheries Service, Fisheries Statistics and Economics Division, Silver Spring, MD.

Osborne, B.G., Fearn T., and Hindle, P.H. 1993. Practical NIR spectroscopy with applications in food and beverage analysis, 2nd ed. Longnam Scientific & Technical, Harlow, Essex, England.

Requena, D.C. 1998. VIS/NIR multivariate spectroscopic quality determination of blue crab meat. Ph.D. Dissertation: North Carolina State University, Raleigh, NC.

Segtnan, B.H., Sasic S., Isaksson, T., and Ozaki, Y. 2001. Studies on the structure of water using two-dimensional near-infrared correlation spectroscopy and principal component analysis, *Analytical Chemistry*, vol. 73, pp. 3153-3161.

Tay, A., Singh, R.K., Krishnan, S.S., and Gore, J.P. 2002. Authentication of olive oil adulterated with vegetable oils using Fourier transform infrared spectroscopy, *Lebensm.-Wiss. Technology*, vol. 35, pp. 99-103.

4.0 MANUSCRIPT #2

Detection and Quantification of Species Authenticity and Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy

Abstract

Seafood processing often removes morphological properties of seafood species that enable the consumer to distinguish one type of organism from another. For this reason, species substitution is the most common form of economic adulteration in the seafood industry. Visible and Near-Infrared (VIS/NIR) spectroscopy has been used to detect and quantify species authenticity and adulteration in crabmeat samples. Atlantic blue crabmeat was adulterated with blue swimmer crabmeat in 10% increments. Water absorption bands dominated the main features in the crabmeat spectra, with a decrease in sample absorbance with increasing adulteration percentage. Several data pre-treatments, i.e. moving average, combing, first and second derivatives, and multiplicative scatter correction, in addition to the untreated data, were investigated for prediction and quantitative data analysis using Partial Least Squares (PLS). In addition, quantitative analysis was done using the full spectrum and a sequential approach in which 50 wavelengths were added sequentially to determine a new model and find an optimal solution. The results suggest that VIS/NIR spectroscopy is a suitable technology that can be applied to detect and quantify species authenticity and adulteration in crabmeat.

Keywords: Species Authenticity, Adulteration, VIS/NIR Spectroscopy, PLS

1. Introduction

The issue of food authenticity has been around since the early 1800s and is mainly related to improper labeling and economic adulteration (E.A.), i.e. the substitution, in part or whole, of cheaper and inferior food products for high-cost foods in order to defraud the consumer (Hargin, 1996). An authentic food, defined as a food that "conforms to the description provided by the producer or processor," includes the process history of a product or ingredient, its geographic region of origin, or the species or variety of the ingredient (Downey, 1998). Though rarely a health hazard, E.A. is driven by the demand for higher value goods, global trading, and price fluctuations, factors that provide an opportunity for illegal profits. For these reasons, food processing industries and regulatory agencies have pushed for analytical methods to confirm food product authenticity (Lees, 1994).

The development of world-wide high-seas fishing vessels, improvement in food processing and storage, and the establishment of fishing industries in developing countries have increased the variety of seafood species, both fresh and processed, currently available in markets (Sotelo and Perez-Martin, 2003). These factors have contributed to an increase in total catches from fisheries and, thus, seafood consumption worldwide. The demand for a year round seafood supply, however, has negatively impacted the number of some valued and appreciated species due to exploitation. Therefore, some have turned to illegal practices in order to meet the high demand for these valued and appreciated seafood products.

Since most consumers are not very familiar with the taxonomical and morphological characteristics of seafood species, such as skin pattern, body appearance and size, eyes,

shape and number of fins, they are subject to being defrauded by buying a seafood product that is not what it claims to be (Mackie, 1996). In addition, the processing of seafood products, which often requires the removal of significant morphological characteristics, hinders species recognition. Due to these reasons, species substitution has become the main form of adulteration in the seafood processing industry.

Detection of food authenticity, by focusing on food adulteration, has traditionally relied on wet chemistry analyses by determining the amount of compounds in a food product and comparison of the values obtained with known, i.e. previously documented, values for authentic products (Downey, 1998). Early methods for species identification have relied on using proteins as species markers. The field of electrophoresis, for example, made it possible to obtain water-soluble protein patterns, which have become a reference method for species identification to differentiate genetically related fish species (AOAC, 1990). However, due to protein denaturation at high temperatures, these techniques are not effective in determining species authenticity for processed seafood products (Sotelo and Perez-Martin, 2003). Additional disadvantages include the large range of compounds needing quantification, discrepancies in protein patterns among members of the same species, and the fact that wet chemistry techniques are time-consuming and expensive. These drawbacks, plus the fact that food adulterers are applying more sophisticated techniques to adulterate their food products, have increased the application of suitable technologies into the field of food adulteration and authenticity.

The need for fast, reliable, and on-line methods to detect species authenticity and adulteration has increased interest in the application of the use of spectroscopic research.

Visible and Near-Infrared spectroscopy (VIS/NIR) is a non-invasive and non-specific technique that has been used to measure spectra of different types of foods (Osborne *et al*., 1993). VIS/NIR spectroscopy can be used to distinguish among biological samples through spectra describing the sample's biological composition, such as fat, protein, lipid, and water. In addition, it has been used to identify different species though research has focused on meat products (McElhinney *et al*., 1999).

The aim of this paper was to investigate the application of VIS/NIR spectroscopy to address the issue of species authenticity and adulteration in crabmeat. The use of several spectral data pre-treatments, i.e. moving average smoothing, first and second derivatives, combing, and multiplicative scatter correction, was explored to determine the effects on model performance. In addition, two spectral approaches were studied to build the calibration models: full spectrum and a sequential spectrum in which 50 wavelengths were added sequentially to the previous model to determine the best optimal solution.

2. Materials and Methods

2.1. *Sample Preparation*

Crabmeat was obtained from a local supermarket for two species of crabs, Atlantic blue (*Callinectes sapidus)* and blue swimmer (*Portunus pelagicus)*, and stored at 4°C. The imported blue swimmer crabmeat was chosen as the adulterant due to its year round availability and reduced cost. Prior to sample preparation and analysis, the crabmeat was tempered individually to room temperature (25°C) in a water bath, pooled in a separate clean container, and thoroughly mixed with gloved hands. Samples represented authentic crab meat species, i.e. Atlantic blue and blue swimmer, and a range of adulterated

samples (10%-90%) containing different amounts of both crab species in 10% increment according to weight. Crab meat of *Callinectes sapidus* and *Portunus pelagicus* was weighed individually and homogenized using a blender (5-second interval) to obtain fifty-five samples (total sample weight of 70g). Homogenized samples were then divided into two equal parts (one for calibration, the other for validation), each consisting of 35g, generating a total of ten samples for each of the 11 classes, i.e. class 0: 100% Atlantic blue (0% blue swimmer), class 1: 90% Atlantic blue (10% blue swimmer), and so forth until class 10: 0% Atlantic blue (100% blue swimmer). The 110 samples were placed in labeled polyethylene bags and stored at room temperature (25°C) until spectral analysis.

2.2. Visible and Near-Infrared Spectroscopy Analysis

Spectra (400 to 2498 nm at 2 nm intervals) were recorded in log(1/R) units, for a total of 1050 wavelengths, using a NIRSystems 6500 spectrometer (FOSS NIRSystems, Silver Springs, MD) equipped with a rectangular sample chamber. Crabmeat samples, at room temperature (25°C), were randomly scanned, and the spectrum, an average of 32 scans, was recorded per sample to obtain a total of 110 spectra. Spectral analysis, model development, calibration, and validation, were performed using The Unscrambler®, software version 7.6 (CAMO Software, Inc., OR, USA).

2.3. Spectral Data Pre-treatment Methods

Because spectral data often contain noise and extra information irrelevant to the problem at hand, an appropriate model is necessary to extract the relevant information for the prediction of the response variable, i.e. adulteration percentage in this study. Several

spectral pre-treatment methods – derivatives, combing, smoothing, and multiplicative scatter correction (MSC) – have been used and compared in this paper in order to asses the best pre-treatment and regression model combination for determining species authenticity.

2.3.1. Derivatives

First and second derivatives are used to reduce peak overlap and remove constant and linear baseline drift, respectively (Osborne *et al*., 1993). Differentiation was done using the Savitzky-Golay algorithm in which a moving average is applied to the spectra prior to differentiation. Care should be taken when choosing an appropriate window for the moving average. The wider the window segment, the greater the noise reduction as well as the greater the distortion of the signal (Zeaiter *et al*., 2005). Five windows composed of 1, 5, 15, 30, and 50 consecutive wavelengths were used to compute the moving average prior to obtaining the polynomial approximation for differentiation. Table 1 shows the different data pre-treatments and parameters used in the convolution intervals for each algorithm.

Table 1: Data Pre-treatments and Parameters.

Data Pre-treatment	Window Segment or algorithm
1st Derivative	1, 5, 15, 30, 50
2nd Derivative	1, 5, 15, 30, 50
Combing	2, 4, 16, 32, 64
Smoothing	none, 5, 15, 30
MSC	Common offset, common amplification, full MSC

2.3.2. Data Combing

Combing, a data reduction technique, is used to choose a user-defined number of data points at equal intervals from each spectrum for subsequent data analysis (McClure and Crowell, 1996). Five combing intervals (2, 4, 16, 32, and 64 points wide) were chosen to examine effect on model performance. Each data set contained 525, 263, 66, 33, and 17 data points, respectively.

2.3.3. Smoothing

Smoothing modifies the magnitude of absorption peaks and shifts the position of asymmetric absorption bands (McClure and Crowell, 1996). A moving average was used to replace each wavelength spectral data with an average of adjacent values. Four convolution intervals (none, 5, 15, and 30 points wide), applied to each spectrum using a moving average, were chosen to compare effect on model predictability.

2.3.4. Multiplicative Scatter Correction (MSC)

MSC compensates for additive (offset) and/or multiplicative (amplification) effects in VIS/NIR spectroscopy and reduces the likelihood that these effects are dominating factors in the spectral data (Osborne *et al*., 1993). Three types of corrections were investigated in this paper: common offset (additive effects), common amplification (amplification effects), and full MSC (additive and multiplication effects).

2.4. Model Development: Calibration and Validation

Multivariate calibration relates two data sets, X (containing the independent variables, i.e. spectral data) and Y (containing the dependent variable, i.e. adulteration percentage), via regression with the purpose of using the model for prediction. Validation, on the other hand, is used to test the model's prediction ability on a new data set, which has not been used in the model development.

For each model, the crabmeat samples were divided equally into a training set (55 samples) for calibration and a testing set (55 samples) for validation. Two different approaches were investigated in each data pre-treatment to determine which method produced optimal models. The full spectrum approach used the complete wavelength range, i.e. 400-2498 nm (1050 data points). In the sequential approach, a 100 nm window was added to the previous consecutive spectrum window to determine a new model. In other words, the first model was developed based on the 400-500 nm wavelength range. The remainder models were developed by adding a 100 nm wavelength range at a time, i.e. 50 wavelengths at a time, so the second model corresponded to 400-600 nm, the third corresponded to 400-700nm, etc., until model 20, which used the full spectral data range, i.e. 400-2498 nm.

There has been much debate as to the importance of finding those few wavelengths that contain significant information for optimal model development, thus reducing the number of wavelengths, variables, and model complexity. Recently, however, research has found the importance of combining some wavelengths, i.e. synergic, though not necessarily significant by themselves, to those containing problem-dependent information, i.e. descriptive wavelengths, to improve model performance (Leardi, 2003).

The spectral data range was broken down in this sequential fashion in order to observe wavelength relevance on model prediction and investigate synergistic and descriptive relationships among consecutive wavelength regions.

2.4.1. Partial Least Squares (PLS)

The chemometric analysis used to calibrate the model was PLS. This method was used instead of Principal Component Regression (PCR) because it focuses on the dependent variables and uses fewer latent variables to reach the optimal solution (De Jong, 1993). This projection method uses both the X and Y matrices, i.e. independent and dependent variables, to find the latent variables in X that will be useful in predicting the latent variables in Y. The term latent variable, principal component (PC), or PLS factors are often used synonymously in literature to define the features used to extract the relevant information and reduce the dimensionality of the data. For the purpose of this paper, the extracted features will be referred to as PCs.

The residual Y-variance plot was examined to determine the number of PCs used in each model. PLS with the full spectrum and sequential spectrum approaches was examined to determine the optimal model for predicting species authenticity. The effects of the data pre-treatment methods on the performance of PLS calibration models were evaluated in terms of the standard error of calibration (SEC), i.e. how well the model fit the training data, the standard error of prediction (SEP), i.e. the error expected when the calibration model is used in future predictions, and corresponding coefficient of determination (r^2), i.e. the relationship between the measured and expected values.

3. Results and Discussion

3.1. Spectra of Crabmeat Species

The absorbance spectra for *Callinectes sapidus* and *Portunus pelagicus* crabmeat samples (Figure 1) are dominated by water absorption bands at 970 nm (O-H bond stretching and second water overtone), 1450 nm (O-H bond stretching and first water overtone), and 1940 nm (O-H bond stretching and bending). The water content of biological samples, such as crabmeat, poses a limitation in the use of VIS/NIR spectroscopy because water absorbs strongly and contributes to a significant amount of light scattering (Bechmann and Jorgensen, 1998). In addition, water absorption bands can interfere with the spectral features of the chemical parameter of interest, such as an adulterant. Overall, however, a higher absorbance is visible for samples containing 100% Atlantic blue crabmeat, whereas the lowest absorbance values represent samples containing 100% blue swimmer crabmeat. The sample absorbance decreases at every wavelength as the percentage of adulteration increases.

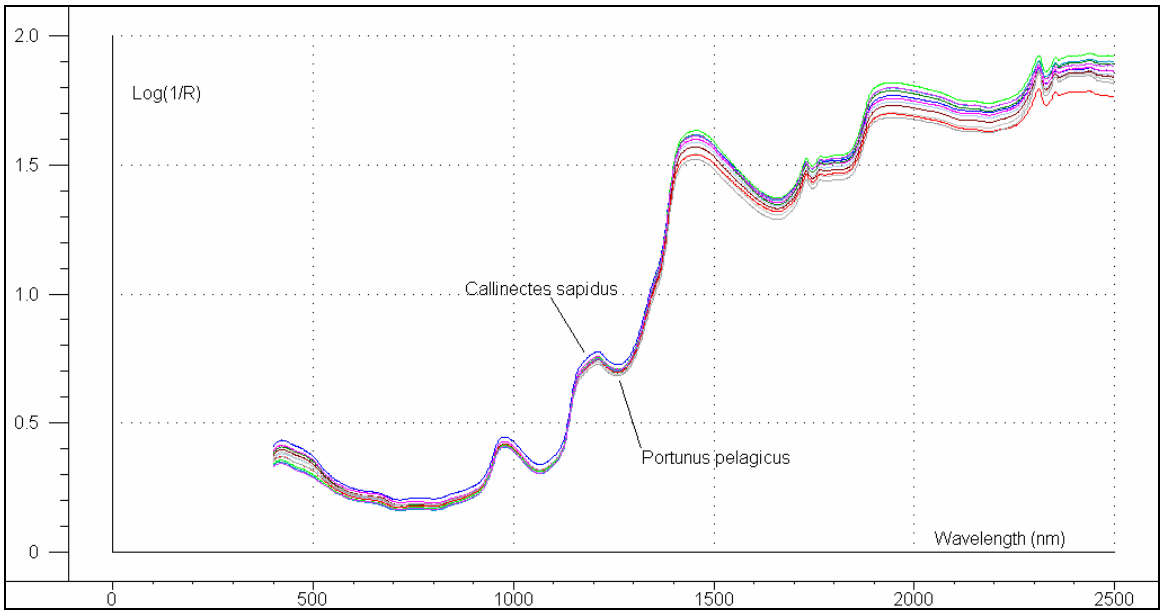

Figure 1: Average absorbance spectra of crabmeat samples. Top spectra represents class 0 (100% Atlantic blue crabmeat), bottom spectra class 10 (100% blue swimmer crabmeat), and classes 1-9 (10-90% adulteration) sequentially from top to bottom.

3.2. Determination of PCs in Model Development

PLS focuses on the Y matrix (the expected values) to decompose the X matrix (the spectral data). Usually, the first few PCs will describe the majority of the variation found in Y; however, this does not guarantee that these first PCs contain the problem-dependent information needed to accurately predict the desired constituent, i.e. adulteration level in this study. Principal Component Analysis (PCA), a data reduction and feature extraction statistical analysis, decomposes the data along directions of maximum variances. Directions of large variance in the first several components usually correspond to structure, i.e. problem-dependent information, and the directions of small variance in later components correspond mainly to noise. If the full set of PCs is used, there is no clear distinction between the structure part and the noise.

The number of PCs to use in the PLS model is very important because too few components will generate an underfitted model, i.e. fits loosely the data structure

(Esbensen e*t al*., 1994). Using too many, on the other hand, generates an overfitted model, one which fits parts of the noise of the calibration set, thus generating a low SEC but performing poorly in the validation set. The optimum number of PCs will then decompose the X matrix between the structure and noise. For this reason, evaluation of the variance plots is needed to determine which PCs describe most of the residual variance of the Y matrix in order to determine the optimal number of PCs to use in the regression model.

Even though the calibration variance is a measure of the model fit, i.e. how well the model fits the training data, it may not be useful to rely on it solely to determine optimal number of PCs. To ensure that the model is able to describe and predict new data, the validation variance must also be taken into account. Figure 2 shows the residual calibration and validation variances of the untreated data. Observation of both variances indicates that the first minimum is visible using 5 PCs; hence, for this particular model the optimum number of PCs used was five. The PCs for the remainder models were chosen similarly by focusing on the Y-variance plot of the calibration and validation data sets.

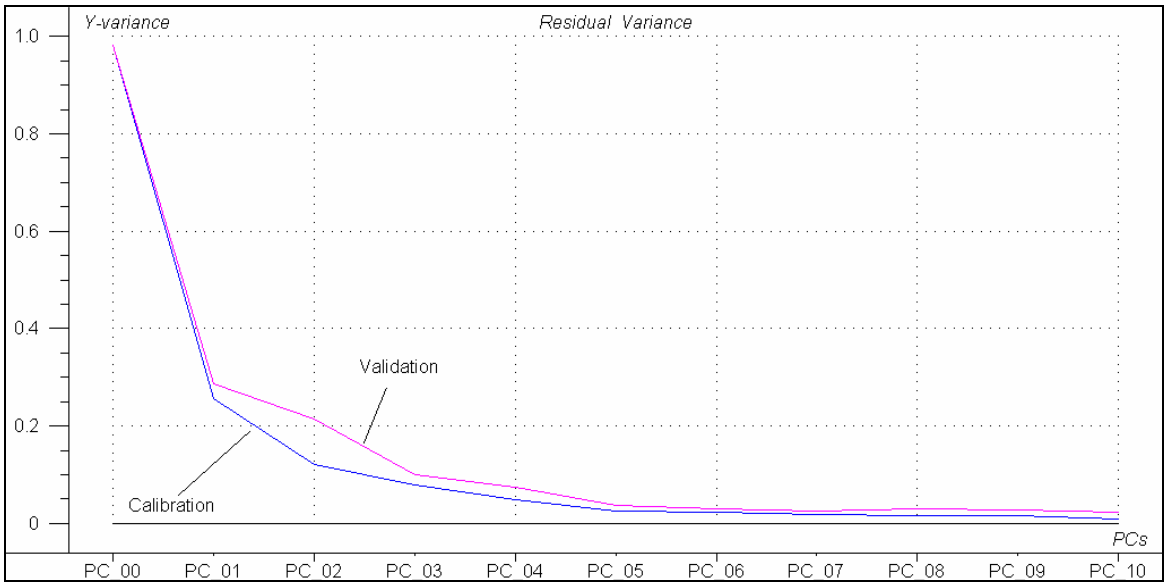

Figure 2: Residual variance of the calibration and validation data sets as a function of PLS components.

3.3. Quantification of Species Authenticity

Ideally, a correlation must exist between the set of variables measured instrumentally and the property to be estimated. This can be studied using a correlogram, a tool used to determine wavelength importance in terms of a given attribute (McClure, 1994). The correlogram describing the correlation between the spectral range and adulteration percentage is seen in Figure 3.

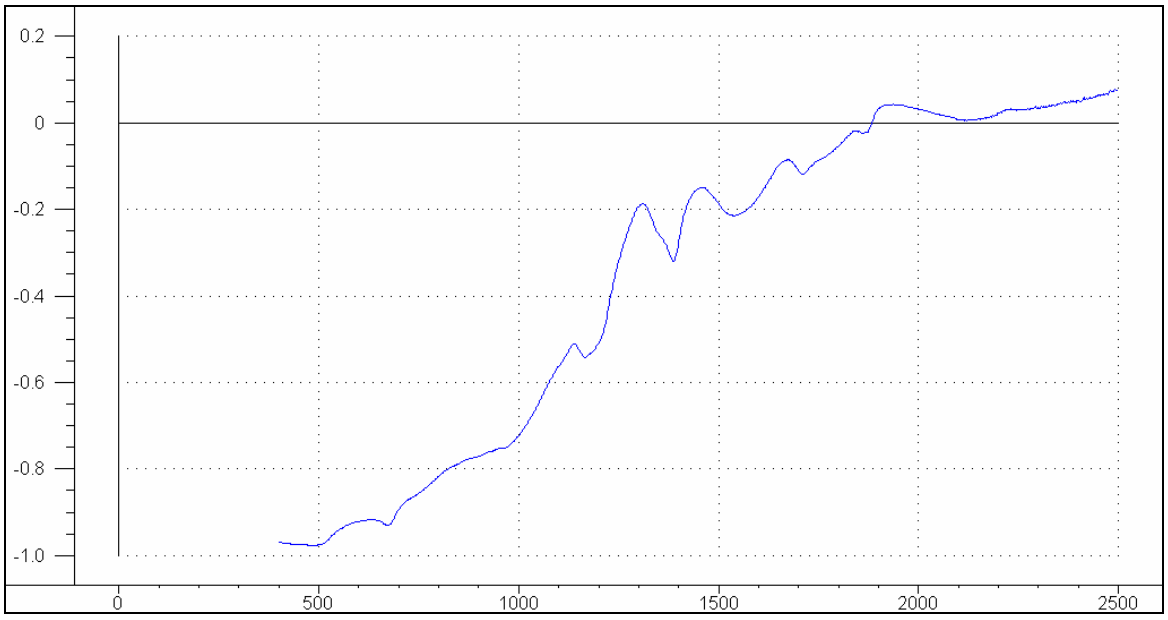

Figure 3: Correlogram showing the correlation (y-axis) between adulteration percentage and VIS/NIR spectral data (x-axis).

There is a high negative correlation ($-0.95 < R^2 < -0.80$) in the visible range (400-800 nm) and slowly decreases to a plateau around zero correlation in the higher NIR wavelengths (1900-2500 nm). Wavelength data can be classified into two categories: 1) predictive, in which the information is useful for modeling the relationship between the spectral data (X) and the response variable (Y), and 2) synergic, in which the information contained within the wavelengths does not improve the model but adding them to descriptive wavelengths enhances the model's predictive ability (Leardi, 2003). Interestingly, there is a slight increase in the correlation coefficient at the 1400 nm wavelength ($R^2 = -0.3$). Although small, this increase in correlation can be attributed to the presence of either synergic or descriptive wavelengths in the 1300-1400 nm region.

The results, based on SEC and SEP, of all the models of the different data pre-treatments were compared. The best model, in terms of SEC and SEP, for each data pre-treatment method was then compared with the results of the untreated data (Figures 4 and

5). The smoothing pre-treatment did not improve model performance when comparing the SEC and SEP values to those gathered from the untreated data. Using the whole spectrum, the best results of the smoothing, those gathered using a 5-pt moving average, generated an SEC and SEP of 5.45 and 5.85, respectively, whereas the untreated data generated 5.43 and 5.84. A 2*SEP is regarded as a 95% confidence interval in spectral quantitative analysis (Kelly *et al*., 2004; Esbensen, *et al*., 1994). Therefore, these models were able to detect adulteration using blue swimmer crabmeat within ±5.85% for the moving average and ±5.84% for the untreated data. Similar results were generated using the sequential spectrum approach.

Contrary to studies involving MSC, none of the methods (common offset, common amplification, and full MSC) improved model performance over the untreated data. This can be attributed to the fact that MSC is usually performed on samples that are not homogeneous and contain particles of different sizes. Since the crabmeat samples were blended, thus reducing the chances of light scattering effects due to particle size, it is possible that this data pre-treatment actually increased the signal-to-noise ratio. The best SEC and SEP, 5.83 and 6.23, respectively, was gathered by using a common offset MSC. Since these data pre-treatments performed equally (for smoothing) and worse (for MSC), further discussion of the results will focus on the untreated data and the remaining data pre-treatment methods, i.e. first and second derivative and combing (Table 2).

Table 2: Optimal Convolution Intervals for Remaining Data Pre-treatments.

Data Pre-Treatment	Optimal Convolution Interval
1st Derivative	30 pt.
2nd Derivative	5 pt.
Combing	64 pt.

Figure 4 shows the SEC, in the sequential spectral approach, of the remaining data pre-treatment methods and the untreated data across the sequential adding of spectra in 100 nm increments. Wavelengths on the figure (x-axis) represent endpoints of the spectral ranges used to derive PLS models. For example, at 1400 nm the model was generated based on the information contained in the 400 – 1400 nm spectral range.

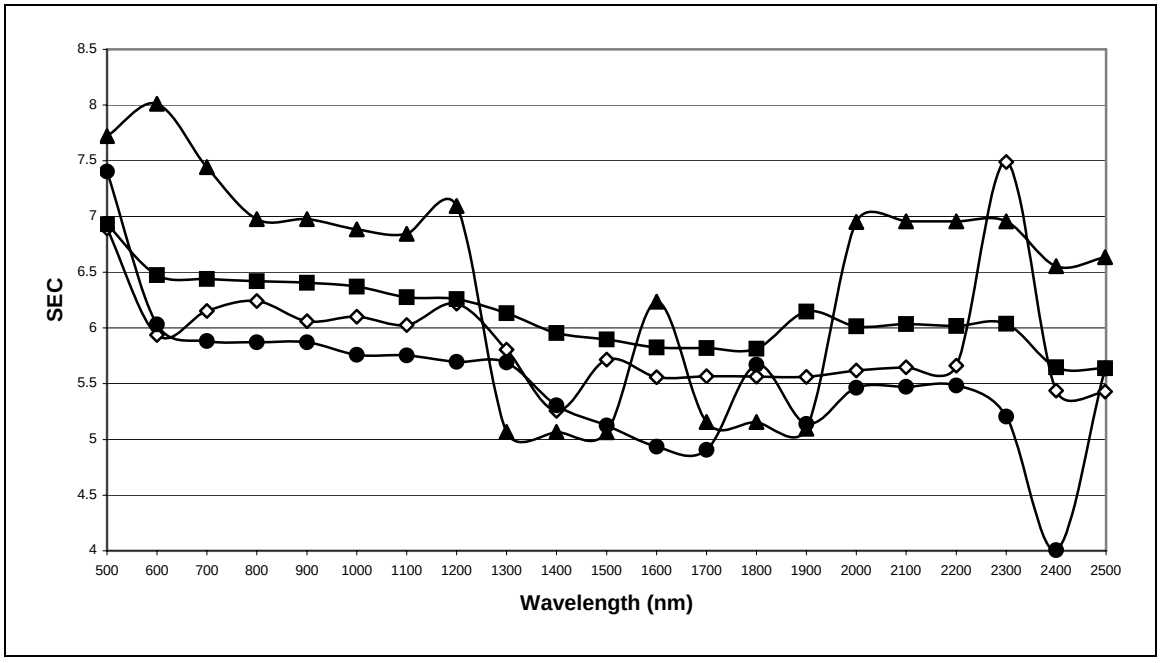

Figure 4: SEC of the best model for each data pre-treatment, i.e. 64-pt combing (▲), 30-pt first derivative (■), 5-pt second derivative (●), and the untreated (◊) data.

In general, the SEC of the models ranges from 4.00 (5-pt second derivative of a model using the 500-2400 nm range), to 8.00 / 7.50 (64-pt combing at 500-600 nm and untreated data at 500-2300 nm, respectively). Choosing an appropriate data-pretreatment

is problem dependent and, as verified in Figure 4, depends on the wavelength range used in the model calibration. If the complete spectrum is used, both the 30-pt first derivative and the 5-pt second derivative perform equally in terms of SEC (5.64), and slightly higher than the untreated data SEC (5.43). The lowest SEC (4.00) across the spectrum was generated using a 5-pt second derivative and a wavelength range of 400-2400 nm. With the exception visible in the 2200-2400 range, the untreated data generated a lower SEC than the 30-pt first derivative data and equal to the 5-pt second derivative data in the 700-1300 and 2000-2200 ranges.

Upon closer inspection of Figure 4, several regions are distinguishable according to the SEC values. The SEC of the untreated, first derivative, and second derivative data is initially between 7.0 and 7.5 and slowly decreases as more wavelengths are added sequentially to determine the next regression model. Models developed with wavelengths of two regions, i.e. 400-1300 nm and 400-2200 nm, generate an SEC between 5.5 - 6.5 and 5.4 – 6.0, respectively. Models generated by including 1400 - 1900 nm produced an SEC between 4.8 and 5.8 and those generated after including the 2200 nm region produced varying SEC values (between 4.0 and 7.5).

It is interesting to note that the 64-pt combing data pre-treatment, though considerably higher in the 500-1200 nm and 2000-2500 nm ranges, has a minimum SEC of 5.0 in the 1300-1500 nm region, a region that coincides with an increase in negative correlation visible in Figure 3. Similarly, the untreated data also has a minimum SEC of 5.0 at 1400 nm. In addition, the combing data pre-treatment shows a second low SEC value in the 1700-1900 nm region. This phenomenon could help explain the importance of synergic wavelengths in model development.

The SEP of the models using the testing set is shown in Figure 5. The SEP of the 64-pt combing data pre-treatment shows a similar trend to that seen in Figure 4. Overall, the model's performance, in terms of SEP, is higher than any other pre-treatment with the exception of the models generated using wavelength ranges from 400-1300/1500 and even 1800/1900. The second derivative data gives a lower SEP in the 400-1200 data range than the first derivative data but performs worse than the first derivative in the models that include the 2000-2500 wavelength data.

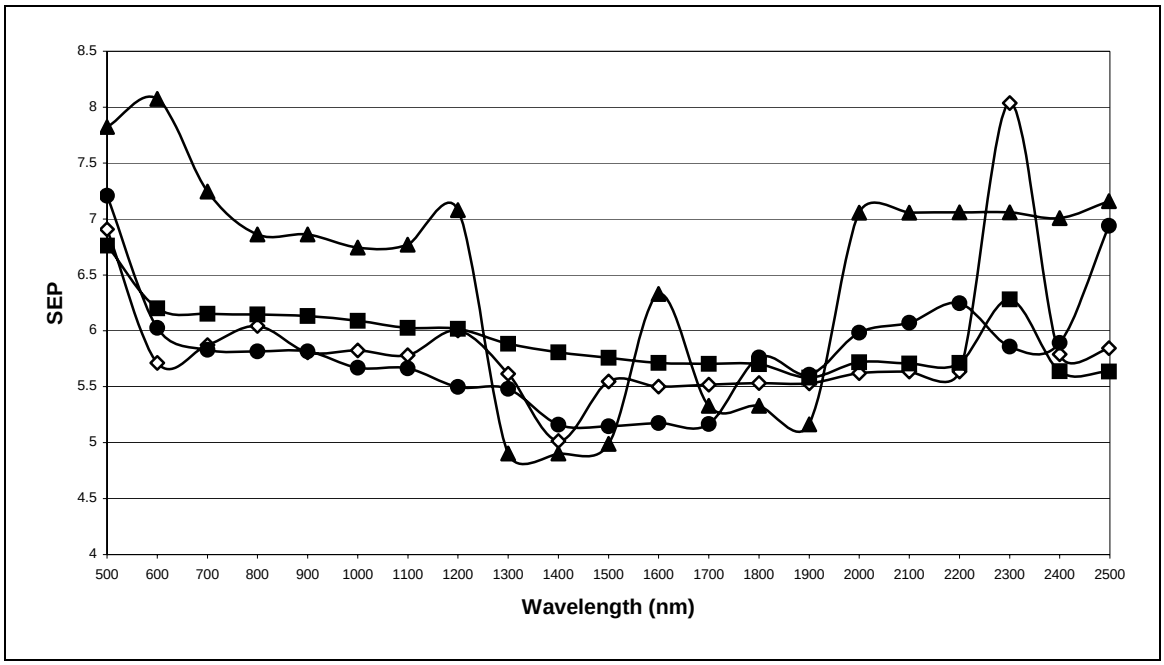

Figure 5: SEP of the best model for each data pre-treatment, i.e. 64-pt combing (▲), 30-pt first derivative (■), 5-pt second derivative (●), and the untreated (◊) data.

Table 3 shows the best model performance in terms of SEC, SEP, and coefficient of determination for each data pre-treatment in addition to the untreated data. Utilizing the full spectral wavelength data, i.e. 400-2500 nm, the 5-pt first derivative data gives the best model performance in terms of SEC (5.64), SEP (5.64), and the number of factors

(3) used in the model. Even though the untreated data generates a lower SEC, a higher number of factors are required, and there is a slight increase in the SEP. Therefore, the true adulteration of a sample, using a 5-pt second derivative data pre-treatment, would be predicted to be within ± 5.64%.

Table 3: Model performance for full and partitioned spectrum wavelength data for each data pre-treatment at optimal convolution intervals. Best models are in bold.

		Full Spectrum			Partitioned Spectrum			
Data Pre-treatment	Data Range	PCs	SEC (r2)	SEP (r2)	Data Range	PCs	SEC (r2)	SEP (r2)
None	400-2500	6	5.43 (0.985)	5.88 (0.983)	400-1400	5	5.26 (0.986)	5.02 (0.988)
1st Derivative	**400-2500**	**3**	**5.64 (0.984)**	**5.64 (0.984)**				
2nd Derivative	400-2500	4	6.63 (0.978)	7.16 (0.975)	**400-1700**	**3**	**4.91 (0.988)**	**5.17 (0.987)**
64-pt Combing	400-2500	5	5.64 (0.984)	6.94 (0.979)	400-1400	6	5.07 (0.987)	4.90 (0.988)

The best model for the sequential spectral approach, however, is generated by using the second derivative on the 400-1700 data range, not only in terms of SEC and SEP, but also in terms of the number of factors (Table 3). The first derivative data did not have a better model performance using a sequential approach so the best SEC and SEP were generated using the full spectrum. This exception aside, models using a partitioned spectrum generated lower SECs and SEPs than those using the full spectrum. Two of the models, i.e. untreated and 64-pt combing, gave the optimal performance when including data contained in the 1300-1400 range; whereas, the 5-pt second derivative gave the optimal performance when including data in the 1600-1700 range. This phenomenon suggests that there are synergic wavelengths in these regions that, when sequentially combined with predictive wavelengths, enhance the model's ability to predict species authenticity and adulteration. Even though the untreated data and the 64-pt combing data pre-treatment generated a lower SEP (5.02 and 4.90, respectively), the 5-pt second

derivative data is preferred because it generated a lower SEC (4.91) by using fewer PCs in order to achieve a similar SEP (5.17).

3.4. PCA Analyses

PCA analyses were done to graphically determine grouping patterns in an effort to classify samples according to adulteration percentage. Figure 6 shows the graph of the first two components of the first derivative data using the full spectrum. The explained variances in the first and second principal component are 22% and 62%, respectively. Based on just two components, a visual relationship among the samples can be gathered. The PCA plot shows a trend from samples of authentic Atlantic blue crabmeat (left) to samples of authentic blue swimmer crabmeat (right). This horizontal trend indicates that the first component contains information that pertains to species authenticity and adulteration percentage. As the percentage of blue swimmer increases, the samples are plotted towards the right side of the graph.

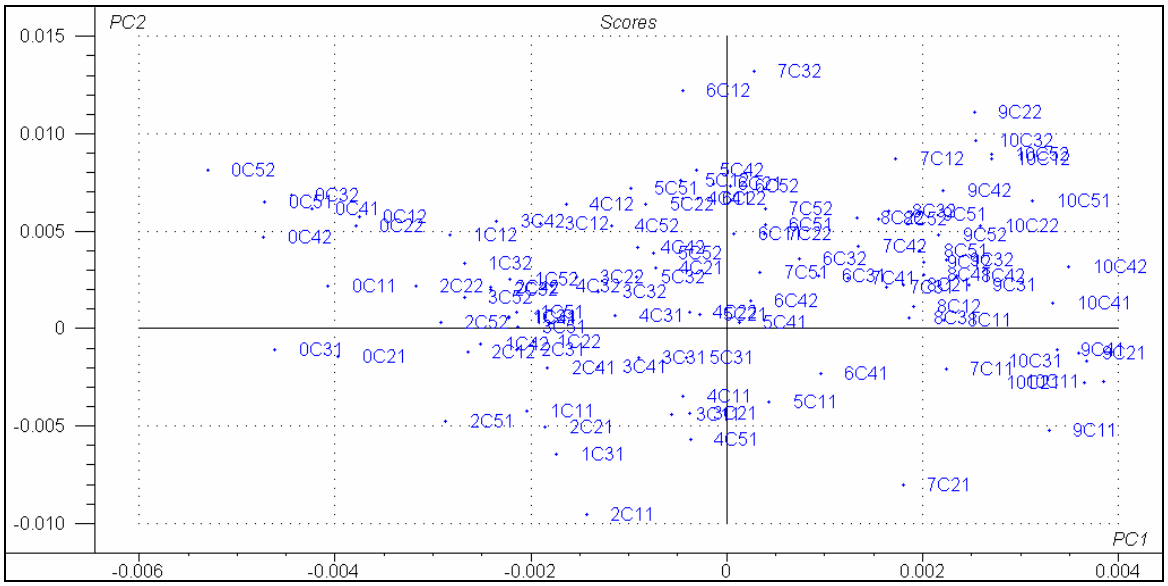

Figure 6: PCA plot of the first two principal components of the first derivative data using the full spectrum. Explained variation in PC1 and PC2 were 22% and 62%, respectively. Samples are named according to crabmeat species (0 = 100% Atlantic blue, 10 = 100% Blue swimmer, and 1-9 according to increasing adulteration percentage), a constant letter "C", sample number (1-5), and whether sample is in the calibration (last digit = 1) or validation (last digit = 2).

Figure 7 is the PCA plot of the second derivative data of the sequential spectrum approach, which shows a similar trend to that seen in Figure 6. The explained variances for the first and second PC are 38% and 34%, respectively. Based on both the model performance, in terms of SEC, SEP, associated coefficients of determination, and reduced complexity, in addition to the PCA plots, the best model to detect species authenticity uses the sequential spectrum approach (PLS for spectral data in the 400-1700 nm wavelength range) and a 5-point second derivative as a data pre-treatment.

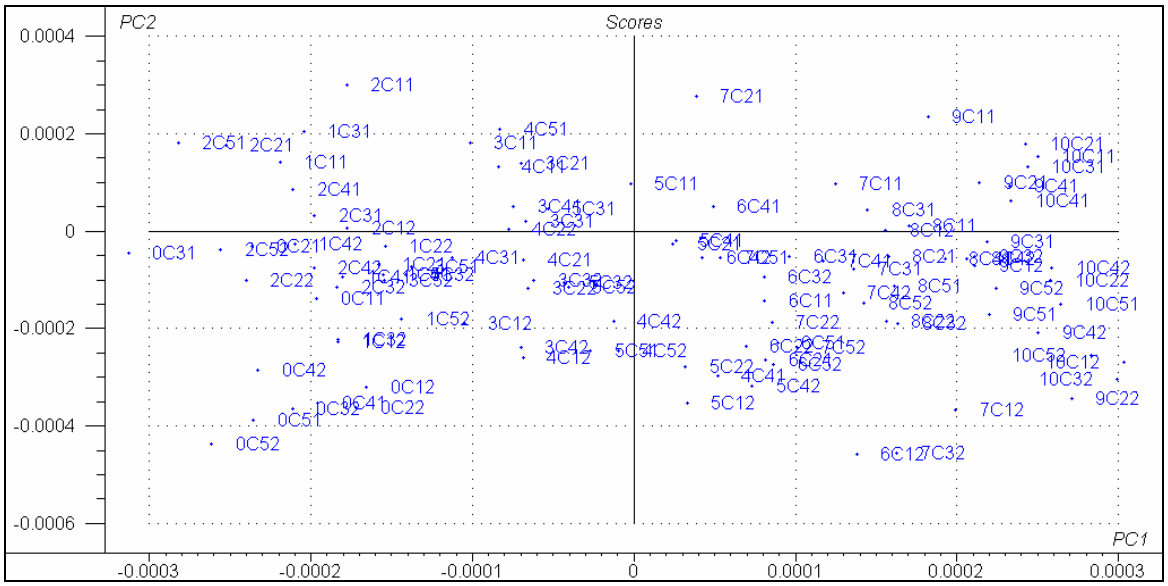

Figure 7: PCA plot of the first two principal components of the second derivative data using a partitioned spectrum (400-1700 nm). Explained variation in PC1 and PC2 were 38% and 35%, respectively. Samples are named according to crabmeat species (0 = 100% Atlantic blue, 10 = 100% Blue swimmer, and 1-9 according to adulteration percentage), a constant letter "C", sample number (1-5), and whether sample is in the calibration (last digit = 1) or validation (last digit = 2).

4. Conclusions

Adulteration, especially species substitution, of high quality and high priced food products, such as crabmeat, is an existing problem in the seafood industry. Seafood processing, which often removes morphological properties that enable consumers to distinguish one species from another, makes it easier for species substitution to occur. Results from this study encompass the possibility of using VIS/NIR spectroscopy to detect species authenticity and economic adulteration of crabmeat. The 30-point first derivative data generated the lowest error for the full spectral approach (SEC=5.64 and SEP=5.64) of the different parameters investigated for spectral data pre-treatments. A problem in determining species authenticity using VIS/NIR spectroscopy, however, is the large number of data points per spectrum. Even though multivariate analyses, such as principle component analysis, are used to reduce the dimensionality, it is important to

distinguish the wavelength regions and features that contain relevant information from the ones that do not, thereby, reducing sources of noise and creating more robust regression models. Utilizing a sequential approach, a data set containing the 400-1700 nm spectral range with a 5-point second derivative data generated the best model to determine species authenticity and adulteration (SEC=4.91 and SEP=5.17). Regardless of the type of approach to analyze the spectral data, i.e. full or sequential, used in generating the model, results from this study indicate that it is possible to detect species authenticity and adulteration with less than 6% error.

References

AOAC. 1990. Official Methods of Analysis. Association of Official Analytical Chemists, Washington, DC., pp. 883-889.

Bechmann, I. E. and Jorgensen, B.M. 1998. Rapid Assessment of Quality Parameters for Frozen Cod Using Near Infrared Spectroscopy, *Lebensm.-Wiss. U.-Technol*., vol. 31, pp. 648-652.

De Jong, S. 1993. PLS Fits Closer Than PCR. *Journal of Chemometrics*, vol. 7, pp. 551-557.

Downey, G. 1998. Food and Food Ingredient Authentication by Mid-Infrared Spectroscopy and Chemometrics, *Trends in Analytical Chemistry*, vol. 17, no. 7, pp. 418-424.

Esbensen, K., Schonkopf, S., and Midtgaard, T. 1994. Multivariate analysis in practice. CAMO AS, Wennbergs Trykkeri AS, Trondheim, Norway.

Hargin, K.D. 1996. Authenticity Issues in Meat and Meat Products, *Meat Science*, vol. 43, no. S, pp. S227-S289.

Kelly, J.F.D., Downey, G., and Fouratier, V. 2004. Initial study of honey adulteration by sugar solutions using midinfrared (MIR) spectroscopy and chemometrics, *Journal of Agricultural and Food Chemistry*, vol. 52, pp. 33-39.

Leardi, R. 2003. Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. In, Leardi, R (Ed) Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier, New York, 2003.

Lees, M. 1994. 2nd European Symposium on Food Authenticity – Isotope Analysis and Other Advanced Analytical Techniques, *Trends in Food Science & Technology*, vol. 51, pp. 200-203.

Mackie, I.M. 1996. Authenticity of Fish, In, In, Ashurst, P.R. and Dennis, M.J. (Eds.). 1996. Food Authentication, Blackie Academic & Professional; Chapman and Hall, New York.

McClure, W.F. 1994. Near-infrared Spectroscopy, In Wilson, R.H. (ed.). Spectroscopic Techniques for Food Analysis, VCH Publishers, Inc., New York, pp. 13-52.

McClure, W.F. and Crowell, B. 1996. Quantitative Information in Near Infrared Spectra. Part 1. Effects of Smoothing and Combing., *Journal of Near Infrared Spectroscopy*, vol. 4, pp. 129-137.

McElhinney, J., Downey, G., and O'Donnell. 1999. Quantification of Lamb Content in Mixtures with Raw Minced Beef Using Visible, Near and Mid-infrared Spectroscopy, *Journal of Food Science*, vol. 64(4), pp. 587-591.

Osborne, B.G., Fearn T., and Hindle, P.H. 1993. Practical NIR spectroscopy with applications in food and beverage analysis, 2nd ed. Longnam Scientific & Technical, Harlow, Essex, England.

Sotelo, C.G. and Perez-Martin, R.I. 2003. Species Identification in Processed Seafoods, In, Lees, M. (Ed.). 2003. Food Authenticity and Traceability, Woodhead Publishing Limited: CRC Press, New York.

Zeaiter, M., Roger, J.-M., and Bellon-Maurel, V.B. 2005. Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods, *Trends in Analytical Chemistry*, vol. 24, no. 5, pp. 437-445.

5.0 MANUSCRIPT #3

Detection of Economic Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy and a Genetic Algorithm for Wavelength Variable Selection.

Abstract

Visible and Near-Infrared (VIS/NIR) spectroscopy has been used in combination with a genetic algorithm and chemometrics to detect economic adulteration of crabmeat. Spectral data sets contain hundreds, even thousands, of variables (wavelengths) per sample. Partial Least Squares (PLS) and Principal Component Regression (PCR), latent-based methods, have been applied effectively in the field of VIS/NIR spectroscopy. Multiple Linear Regression (MLR), on the other hand, has limited applications in spectroscopy due to the high autocorrelation among the spectral variables. An effective variable selection, via genetic algorithm, was able to determine relevant variables in terms of economic adulteration for the untreated, first, and second derivative pretreated data. Of the chemometric methods, MLR, using the relevant variables selected by the genetic algorithm in the untreated data, outperformed the other methods based on latent structures. The results suggest that with VIS/NIR spectroscopy, a genetic algorithm for variable selection, and chemometrics is a suitable combination of technologies that can be used to detect economic adulteration in crabmeat.

Keywords: PLS, PCR, MLR, GA, variable selection, VIS/NIR Spectroscopy

1. Introduction

Traditional chemometric analyses in Visible and Near-Infrared (VIS/NIR) spectroscopy, such as Partial Least Squares (PLS) and Principal Component Regression (PCR), focus on the whole spectrum containing hundreds and even thousands of absorbance values describing combination bands and overtones of functional groups (Feudale *et al*., 2002; Janne *et al*., 2001). Basing regression on latent variables (LVs), i.e. linear combinations of the original variables, these methods were initially thought to be insensitive to noise and efficient at extracting relevant information; hence, their popularity in spectroscopy (Wold *et al*., 2001; Carlin *et al*., 1994). Research, however, has shown the importance of an effective variable selection to reduce the model's complexity and improve its predictive ability (Du *et al*., 2004; Broadhurst *et al*., 1997).

Unlike latent variable methods, Multiple Linear Regression (MLR) uses the original variables in the spectra. One advantage of MLR is the ease of model interpretation (Kompany-Zareh and Mirzaei, 2004). Unfortunately, the requirement that the number of samples be greater than the number of variables makes MLR unsuitable for spectral analyses because of the high variable-to-sample ratio. Effective variable selection methods, however, address this problem and enable MLR to perform satisfactorily (Hemmateenejad *et al*., 2002). Other methods with forward/backward elimination algorithms, such as stepwise multiple linear regression, have also been effectively applied in the field of spectroscopy.

One of the main problems in variable selection is the determination of variables, wavelengths in spectral data, containing relevant information that can be used to achieve optimal model performance (Abrahamsson *et al*., 2003). Due to the overlapping of

absorption bands and effects of experimental conditions on spectra, determination of important variables is not easy. An efficient variable selection method is needed to determine important variables and provide robust models. Improving the robustness of a model will decrease the impact of changes in environmental conditions on prediction ability.

The most basic approach on variable selection focuses on a correlogram, where variables with the greatest correlation to the response variable are chosen for analysis, as a simple form of variable selection (Gayo and Hale, 2006; Huang *et al*., 2002). One problem with this approach is its reliance on visual observation of regression coefficients and estimation of spectral variables. Variable selection is still "guess" work rather than being based on factual information.

Genetic algorithms (GAs) represent artificial biological systems that simulate evolution based on Darwin's theory of natural selection. The simultaneous search of different areas of the error space make them more effective at finding optimal solutions in comparison to traditional gradient-descent algorithms, which search the error space one point at a time and are more likely to find and get stuck in a local minimum or maximum (Wehrens and Buydens, 1998). An additional advantage is that GAs are not limited by search space assumptions, such as continuity and smoothness, so they can optimize problems with unknown gradient information.

A GA contains a population of chromosomes (binary strings) whose genes (each position on the string) encode for the inclusion or deletion of each variable. For example, a '1' on the string means inclusion of variable for further analysis whereas a '0' means that the variable is discarded. The best chromosomes in terms of a fitness value are

allowed to survive and reproduce to produce offspring. During reproduction, via a crossover operator, two fit chromosomes exchange information to produce new chromosomes (offspring). Small random changes in genes invert the bit on a position in the chromosome. This is termed mutation and is an important process that ensures diversity within the population.

GAs have been effectively utilized for variable selection in the field of VIS/NIR spectroscopy (Leardi and Norgaard, 2004). They have been used to find the relevant wavelengths in VIS/NIR spectroscopy to predict: 1) moisture, oil, and protein contents of soy flour (Leardi, 2000); 2) octane numbers of gasoline samples (Hageman *et al*., 2003; Goicoechea and Olivieri, 2003; Leardi, 2000); 3) moisture and protein content of wheat samples (Hageman *et al*., 2003; Leardi, 2000); and 4) Cd, Zn, clay and organic matter contents of floodplains (Hageman *et al*., 2003).

Economic adulteration (E.A.), i.e. the substitution of cheaper and inferior ingredients for high-cost ingredients, of food defrauds the consumer and undercuts legitimate industry prices. Initial attempts to detect E.A. have focused on wet chemical analysis, gas chromatography, and mass spectroscopy. Unfortunately, these methods are expensive, require trained personnel, and are very time consuming making them unsuitable for "on-line" monitoring. VIS/NIR spectroscopy is an affordable, powerful, and objective tool which is non-invasive, non-selective, and quickly generates results while requiring minimal sample preparation (Gusnanto *et al*., 2003). It has been used to detect adulteration in many high-value foods, including apple and orange juice, wine, honey, and meat products. Previous research, however, has not focused on economic adulteration of seafood products, such as crabmeat. Other studies have described the

potential of spectroscopy for on-line process monitoring (Jorgensen *et al*., 2004; Mertens *et al*., 2002; Miller, 2000). The objective of this study was to evaluate the application of VIS/NIR spectroscopy for detecting economic adulteration of crabmeat by investigating the prediction ability of chemometric techniques using the spectral variables selected by a genetic algorithm.

2. Materials and Methods

2.1. Sample Preparation

Crab meat was obtained from a local supermarket for two species of crabs, Atlantic blue (*Callinectes sapidus)* and blue swimmer (*Portunus pelagicus)*, and stored at 4°C. The imported blue swimmer crab meat was chosen as the adulterant due to its year round availability, reduced cost, and prevalence as an imported crabmeat from the Indo-Pacific region (Oesterling, 1998). Prior to sample preparation and analysis, the crab meat was tempered individually to room temperature (25°C) in a water bath, pooled in a separate clean container, and thoroughly mixed with gloved hands. Samples represented authentic crab meat species, i.e. Atlantic blue and blue swimmer, and a range of adulterated samples (10%-90%) containing different amounts of both crab species in 10% increments according to weight. Crab meat of *Callinectes sapidus* and *Portunus pelagicus* was weighed individually and homogenized using a blender (5-second interval) to obtain fifty-five samples (total sample weight of 70g). Homogenized samples were then divided into two equal parts (one for calibration, the other for validation), each consisting of 35g, generating a total of ten samples for each of the 11 classes, i.e. class 0: 100% Atlantic blue (0% blue swimmer), class 1: 90% Atlantic blue (10% blue swimmer), and so forth

until class 10: 0% Atlantic blue (100% blue swimmer). The 110 samples were placed in labeled polyethylene bags and stored at room temperature (25°C) for immediate spectral analysis.

2.2. Spectral Analysis

A NIRSystems 6500 spectrometer (FOSS NIRSystems, Silver Springs, MD) equipped with a rectangular sample chamber was used to record absorbance spectra (400 to 2498 nm at 2nm intervals, i.e. 1050 wavelengths) of crabmeat samples at room temperature (25˚C). Crabmeat samples were scanned randomly and each spectrum was recorded to obtain a data set of 110 spectra. Variable selection was done using a genetic algorithm programmed in Matlab®, version 5.3 (The Mathworks Inc., Natick, MA). Regression analysis and model development, i.e. calibration and validation, were performed using The Unscrambler®, software version 7.6 (CAMO Software, Inc., OR).

2.3. Spectral Data Pre-treatment Methods

Spectral data often contains noise and unnecessary information due to light scattering (Janne *et al.*, 2001). Data pre-treatments are done to emphasize relevant information and reduce noise. Based on previous work, first and second derivatives, using the Savitzky-Golay algorithm in which moving average of a fixed convolution interval was applied prior to fitting a polynomial, were chosen for data pre-treatment prior to variable selection (Gayo and Hale, 2006). The optimal convolution interval used in the moving average for the first and second derivative was 30 and 5 points, respectively.

2.4. The Genetic Algorithm

The PLS-genetic algorithm toolbox, developed by Dr. R. Leardi, was used for variable selection. It combines the strength of PLS as a regression analysis technique with a genetic algorithm. The following is a summary of important concepts taken into account in developing the genetic algorithm (Leardi *et al*., 2002; Leardi, 2000). Table 1 lists the parameters and associated values used in the genetic algorithm to obtain optimal results. Deviation from these parameters yielded worse results; hence, suggested values were used for simulation.

Table 1: Optimal parameter values for PLS-genetic algorithm.

GA parameters
Population size: 30 chromosomes
Average number of variables/chromosome: 5 (30 maximum)
Probability of cross-over (reproduction): 50%
Probability of mutation: 1%
Maximum number of components: 15 (determined by cross-validation of full model, i.e. with all wavelengths)
Stepwise elimination: backward every 100 evaluations
Number of runs: 100
Smoothing interval for variable frequency of selection: 3

2.4.1. Chromosome Fitness

The fitness value of each chromosome is usually evaluated in terms of a criterion. This genetic algorithm used PLS in determining chromosome fitness. The bit values of each chromosome represented the variables (initially 5 on average per chromosome) that underwent PLS regression. Via the crossover operator, the initial number of variables could increase though the maximum number of variables per chromosome was set to thirty. PLS used both the X (spectral variables) and Y (response variable) to find latent

variables (LVs) in X that will be useful in predicting the latent variables in Y. The LVs, arranged in decreasing order of importance, represent a percentage of the total variance in the original data.

Regression of the data contained in each chromosome yields the number of latent variables used in the model, root mean squared error of cross validation (RMSECV), and an explained residual variance (%). The fitness function of each chromosome was evaluated in terms of the percent variance of cross-validation (%CV) of the response variable (economic adulteration) and the number of latent variables used. The best chromosome was protected, i.e. could not be discarded through the generations, unless another produced a better fitness value with an equal or lesser number of features. The population of chromosomes remained constant throughout the simulation with better chromosomes replacing inferior ones. In addition, a twin chromosome was discarded to reduce the chance of convergence to a single solution. Chromosome subsets were also checked to ensure diversity within the population.

2.4.2. Overfitting: Exploration versus Exploitation

The longer the GA runs and the higher the number of variables, the higher the risk of overfitting. Two implementations helped avoid this problem: 1) original spectral variables were reduced by applying a six-wavelength smoothing interval (1050 wavelengths/6 = 175 variables) and 2) a large number of short, independent, simulations were done to influence the final model (Leardi, 2003). In order to obtain a good response, balance between exploitation, i.e. parent-to-offspring heredity of parts of the genetic vector, and exploration, i.e. offspring whose genetic vector differs greatly from

parents, must be achieved (Luke, 2003). Exploration, achieved by the mutation operator, in which one bit in the chromosome is inverted, emphasizes a random search; whereas, exploitation, achieved via selection and crossover (reproduction), emphasizes current exploration of the search space (Wehrens and Buydens, 1998).

Hill-climbing/steepest descent techniques, such as Artificial Neural Networks (ANNs), are characterized with high exploitation but low exploration. In other words, these techniques find a local maximum and get stuck. GAs can have high exploration but low exploitation, i.e. detect the areas around local maxima but find it hard to climb up. Research has shown that it is advantageous to use stepwise regression techniques, such as forward selection and backward elimination, in genetic algorithms (Chong and Jun, 2005). To improve the performance of the GA, a backward stepwise selection was done on the best chromosome. This entailed the removal of a variable from the chromosome. If, after backward elimination, a better chromosome was found, it substituted the original one.

2.4.3. Constraints on the Initial Population

Unlike a simple genetic algorithm (GA) in which variables of the initial population have a 50% probability of being selected, the PLS-Genetic Algorithm restricts the probability of selection to n/v, i.e. n variables will be selected from the total, v. If too many variables are selected, PLS will still yield a good model because of its robustness. Limits on the probability of selection and the maximum number of selected variables allowed ensure that variables yielding an unsatisfactory model will be discarded, thus, favoring selection of relevant variables. Additionally, this restriction also facilitates

investigation of combinations of more variables. The population size, restricted to a low number, also enhances exploitation (reduces exploration) because selection of a chromosome for reproduction (crossover) is biased towards the best chromosome.

2.4.4. Considerations of Spectral Data

Because spectral wavelengths have high autocorrelation, i.e. wavelength x is highly correlated with wavelengths x – 1 and x + 1, spectral analysis usually focuses on spectral regions formed by a specified number of adjacent wavelengths rather than single wavelengths (Abrahamsson *et al*., 2003; Leardi, 2000). Variable selection is a complex procedure and needs appropriate validation to decrease the risk of overfitting the model, especially with a high variable to sample ratio. Autocorrelation in spectral wavelengths can be used as an advantage by forming new variables from the average of several adjacent variables. This not only reduces the number of variables without losing important information but also diminishes the chance of overfitting (Leardi, 2003).

Due to the high autocorrelation among wavelengths, a moving average was applied to the probability vector after the initial generation and was updated until the simulation terminated. In other words, if wavelength x is selected, x-1 and x+1 should also be selected, and hence relevant; therefore, the moving average increases the adjacent wavelengths' probability of selection as well. This yields a model with contiguous wavelengths rather than ones dispersed throughout the spectrum.

2.4.5. Obtaining the Final GA Model

The final model after a high number of GA runs represents a stepwise methodology in which the variables enter the model according to the frequency of selection (higher frequency first, then next higher, and so on). Since a single run can yield a model with irrelevant variables, the combination of the different runs, in which each run influences the next emphasizes the inclusion of relevant variables in the final model. The GA simulation, evaluated in terms of the RMSECV and the number of selected variables, yields two models: one containing the global minimum (minimum RMSECV) and a suggested model according to an F-statistic criterion. The suggested model, which contains the smallest number of variables, has an RMSECV that is not significantly different according to an F test ($p<0.1$) from the global minimum.

2.5. Calibration and Validation of Models

Using the calibration set (55 samples), the genetic algorithm was used to determine the relevant spectral variables and corresponding number of components used in model development. Validation of the GA was accomplished with cross-validation (CV). Since leave-one-out CV, often yields very optimistic results because of overfitting, validation was done using 5 deletion groups, i.e. each group (made of 11 samples) was left out at a time. Using this approach, the samples in the deletion group were used to evaluate the model generated by the rest of the samples.

After variable selection, three regression methods (PLS, PCR, and MLR) were used to develop new models to predict economic adulteration. In these methods, the calibration set was used to relate the independent variables (X), i.e. relevant variables

selected by the GA, and the dependent variable (Y), i.e. adulteration percentage, via regression. The optimal number of components for PLS and PCR were determined by observing the residual Y-variance. The validation set was used to test each model's predictive ability when exposed to new data not used to develop the regression model. Extreme outliers were removed from the data sets based on influence and leverage information. The effects of variable selection on the performance of each calibration model were evaluated in terms of standard error of calibration (SEC) and standard error of prediction (SEP). Calibration and validation of models were formulated with untreated data and derivative (1^{st} and 2^{nd}) data.

3. Results and Discussion

3.1. Spectra

Representative spectra of the adulterated crabmeat samples after applying a moving average of 6 wavelengths to achieve variable reduction (from 1050 wavelengths to 175 wavelengths) are shown in Figure 1. Water absorption bands (at 970 nm, 1450nm, and 1940 nm) influenced the main features of the absorbance spectra, as described in a previous study (Gayo and Hale, 2006). Absorbance values decreased with increasing adulteration percentage.

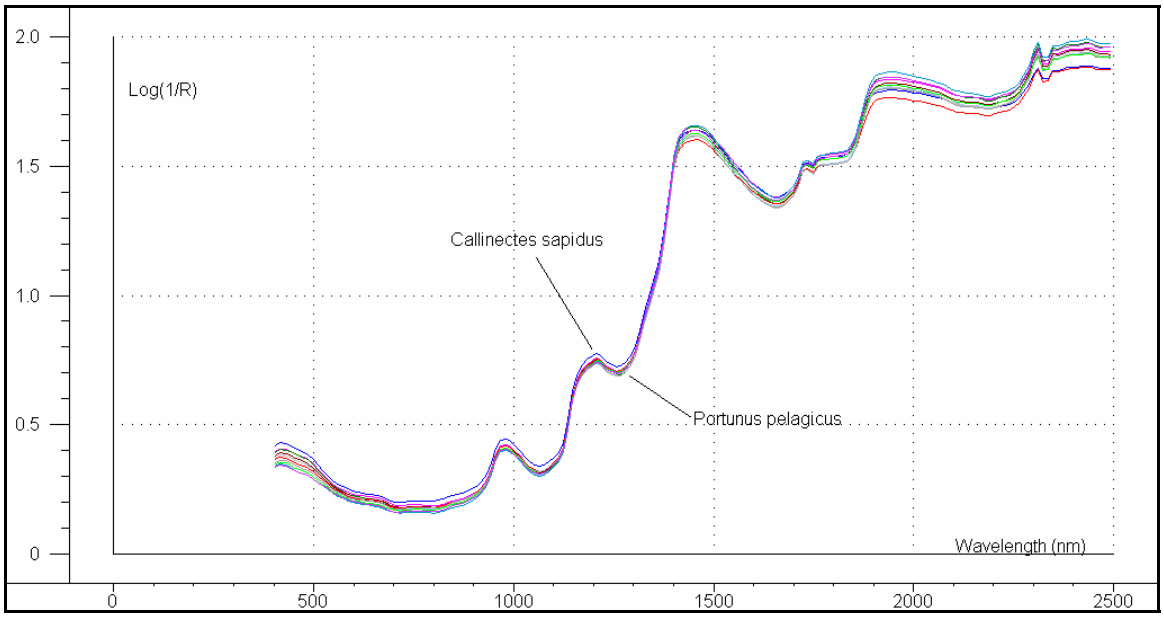

Figure 1: Absorbance spectra of crabmeat samples after a 6 wavelength moving average. Top spectra represents class 0 (100% Atlantic blue crabmeat) and bottom spectra class 10 (100% blue swimmer crabmeat). Spectra in between represent samples with 10 to 90% adulteration.

3.2. Variable Selection: Genetic Algorithm

Since genetic algorithms are stochastic in nature, five different simulations were done to compare selection of variables. Percentages of the cross-validation variance (%CV) in conjunction with the RMSECV, as a function of the number of included variables, were used to determine the optimal number of variables. Figure 2 shows the smoothed frequency of selection for simulation 1 of the untreated spectral data. The bottom line represents the model with the global minimum (21 variables and RMSECV = 4.427) and the top line represents the suggested model according to the F-statistic (5 variables and RMSECV = 4.892). Similarly, the variables chosen as relevant for the additional simulations were recorded. It is important to note that the variables chosen by the GA represent wavelength ranges. Because the original data were reduced by a smoothing interval, each variable selected represented the average of six continuous spectral wavelengths.

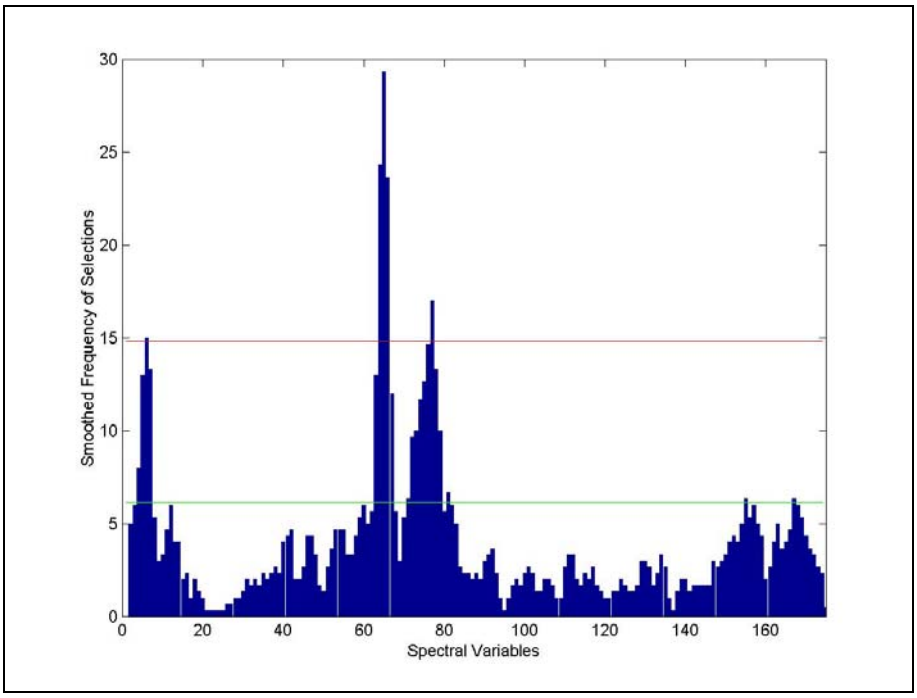

Figure 2: Smoothed frequency of variable selection for simulation 1. Bottom line represents the global model (21 variables and minimum RMSECV= 4.427) and top line the suggested model according to the F-criterion (5 variables and RMSECV= 4.892).

Since the purpose of variable selection was to reduce the number of variables to an acceptable low value, the variables in the suggested model (F-statistic) were chosen as relevant for regression analyses. GAs are stochastic in nature; therefore, five independent GA simulations were done on each data set (untreated, first derivative, and second derivative). Figure 3 shows the variables selected as relevant by the five different simulations of the untreated spectral data. The five broken lines below the absorbance spectra represent the spectral variables chosen as relevant by the five GA simulations. The RMSECV of the five GA simulations (from the top to the bottom of the broken lines) is 4.892 (five variables), 5.322 (8), 5.433 (9), 5.716 (5), and 5.630 (6).

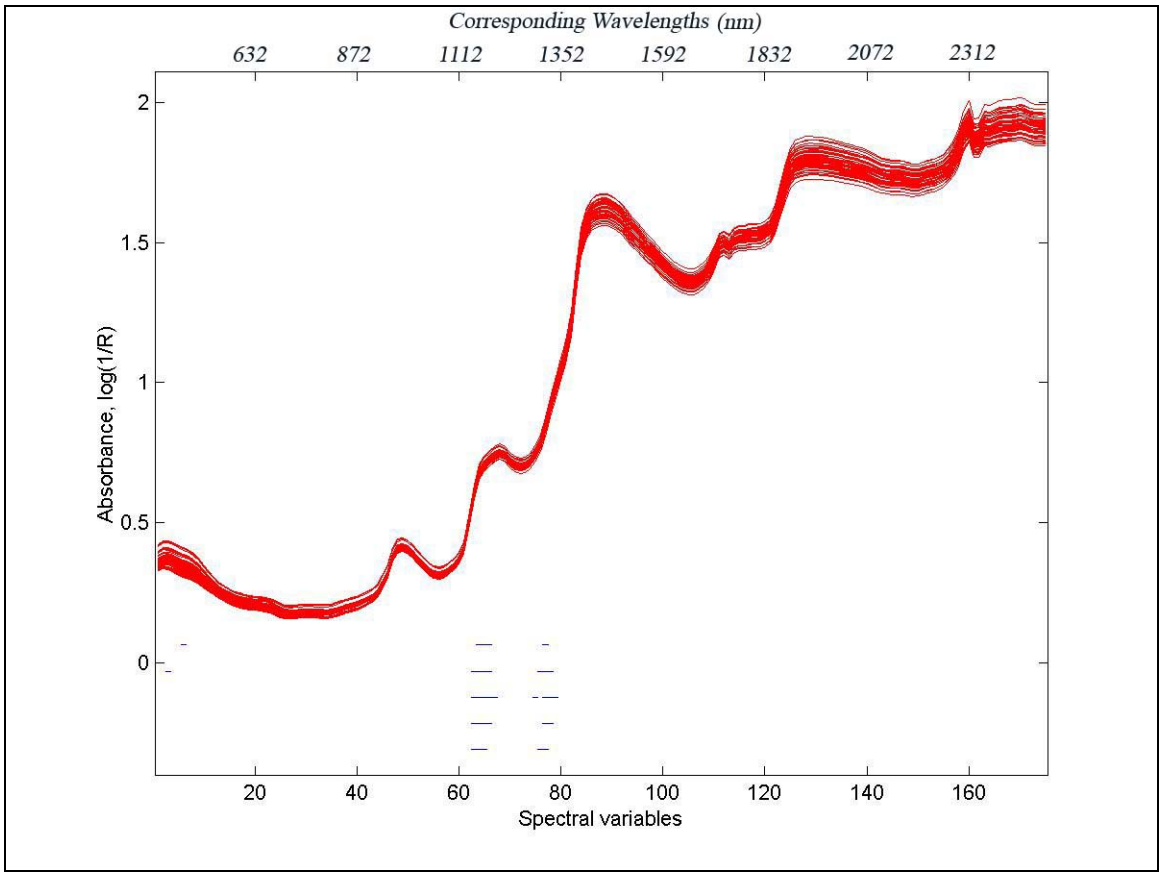

Figure 3: Variables selected from 5 separate simulations of the calibration set (untreated spectral data). Top axis contains corresponding wavelengths selected by the genetic algorithm.

The genetic algorithm selected two spectral areas of similar variables as relevant in detecting economic adulteration of blue crabmeat: 63-67 and 75-79, areas corresponding to 1144-1202 nm and 1290-1346 nm, respectively. Only the first two simulations selected an additional area (variables 3 and 6, corresponding to 424-434 nm and 460-470 nm, respectively).

Table 2 shows the selected variables, corresponding wavelengths, RMSECV, and percentage variance of cross-validation for the untreated data in the calibration set for each GA simulation. Tables 3 and 4 show the results for the first derivative data (30

point convolution interval) and second derivative data (5-point convolution interval), respectively.

Table 2: Variables selected in each simulation for the untreated spectral data. Mean RMSECV and %CV were 5.399 and 97.077, respectively.

Simulation	*Selected Variables*	*Corresponding Wavelengths (nm)*	*RMSECV*	*%CV*
1	6, 64-66, 76	460-470, 1156-1190, 1300-1310	4.892	97.606
2	3, 63-66, 76-78	424-434, 1144-1190, 1300-1334	5.322	97.167
3	63-67, 75, 77-79	1144-1202, 1288-1298, 1312-1346	5.433	97.048
4	63-66, 77	1144-1190,1312-1322	5.716	96.732
5	63-65, 75-77	1144-1178, 1288-1322	5.630	96.830

GA results from the untreated data yield PLS models containing from five to nine selected variables. The corresponding models had an average RMSECV and %CV of 5.399 and 97.077, respectively. Based on these results, the genetic algorithm, in conjunction with PLS, was able to detect economic adulteration of crabmeat, using cross validation, to an average of 5.4%.

Table 3: Variables selected in each simulation for the first derivative spectral data. Mean RMSECV and %CV were 5.585 and 96.881, respectively.

Simulation	*Selected Variables*	*Corresponding Wavelengths (nm)*	*RMSECV*	*%CV*
1	4, 64	496-506, 1216-1226	5.593	96.872
2	4, 63-64	496-506, 1204-1226	5.576	96.891
3	4, 63-65	496-506, 1204-1238	5.521	96.951
4	4, 64	496-506, 1216-1226	5.593	96.872
5	5, 63-64	508-518, 1204-1226	5.640	96.819

Results from the first derivative data yield models with fewer selected variables in comparison to the untreated data (2-3 versus 5-9), higher RMSECV average (5.585 vs.

5.399), and lower %CV average (97.077 vs. 96.861). In agreement with results in Table 2, two areas of spectral variables (4-5 and 63-65) were selected as relevant variables in detecting economic adulteration. Using the first derivative data, the genetic algorithm was able to detect economic adulteration to an average of 5.6%.

Table 4: Variables selected in each simulation for the second derivative spectral data. Mean RMSECV and %CV were 5.661 and 96.786, respectively.

Simulation	*Selected Variables*	*Corresponding Wavelengths (nm)*	*RMSECV*	*%CV*
1	1-3, 7-12, 40-42, 48	410-444, 482-552, 878-912, 974-984	5.482	96.994
2	1-2, 7-11, 15-17, 27, 41-42, 47-48	410-432, 482-540, 578-612, 722-732, 890-912, 962-984	5.067	97.432
3	1-4, 7-11, 63-64	410-456, 482-540, 1154-1176	5.751	96.692
4	1-3, 7-12, 65	410-444, 482-552, 1178-1188	6.078	96.306
5	1-2, 7-15, 26-27, 41-42, 48, 86	410-432, 482-588, 710-732, 890-912, 974-984, 1430-1440	5.926	96.506

Results from the second derivative data (Table 4) yielded models with a greater number of selected variables in comparison to the untreated and first derivative data and a larger RMSECV average (5.661). Unlike the GA results of the raw and first derivative data (Tables 2 and 3), the second derivative data selected variables from different areas of the spectrum (Figure 4). The RMSECV of the five GA simulations (from the top to the bottom of the broken lines) is 5.482 (thirteen variables), 5.067 (15), 5.751 (11), 6.077 (10) and 5.926 (17). The dominating region from which variables were selected as relevant was found in the range of variables 1-17 (wavelength range between 400 nm and 612nm). Several other scattered variables were selected, but very few corresponded to the areas chosen as relevant by the GAs that used the untreated and first derivative data.

The second derivative data generated the highest average error (5.66%) for detection of economic adulteration of crabmeat.

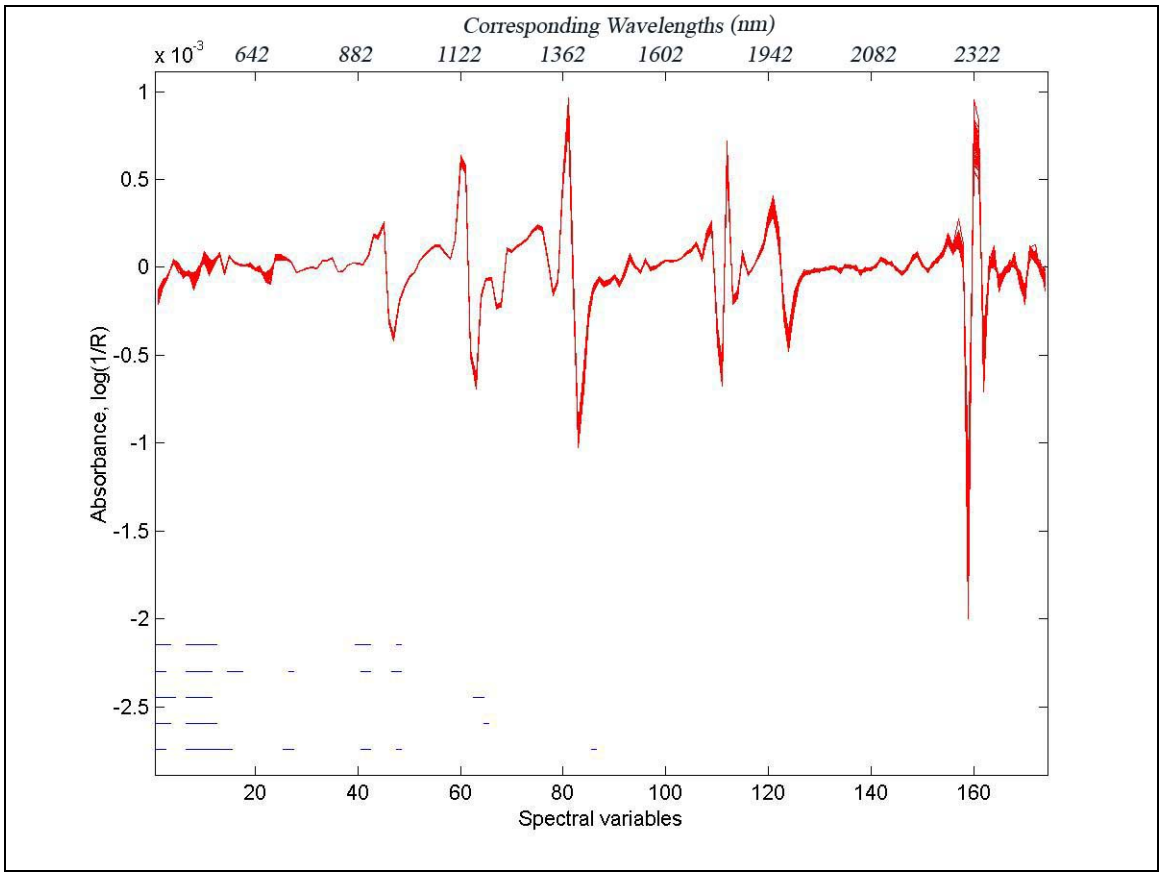

Figure 4: Variables selected from 5 separate simulations of the calibration set (2nd derivative data). Top axis contains the corresponding wavelengths selected by the genetic algorithm.

The selected spectral variables in Tables 2-4 were used in chemometric analyses using PLS, PCR, and MLR.

3.3. Chemometric Analysis

Tables 5-7 show the results from the three chemometric techniques for the untreated, first derivative, and second derivative data, respectively. Because there were five independent GA simulations, the variables selected in each simulation were used to

derive a regression model. The tables show the results, in terms of number of latent variables, SEC, and SEP of each individual simulation and the average values.

Table 5: Chemometric results using relevant variables selected by the genetic algorithm and the untreated data.

	PLS			PCR			MLR		
Simulation	LVs	SEC	SEP	LVs	SEC	SEP	LVs	SEC	SEP
1	2	5.893	5.186	3	5.553	5.435	N/A	4.258	4.179
2	2	4.608	4.650	3	5.266	4.834	N/A	3.991	3.897
3	2	5.566	5.220	3	6.177	5.454	N/A	3.976	4.065
4	2	5.034	4.170	3	5.656	5.391	N/A	4.252	3.653
5	2	5.011	4.157	3	5.025	4.162	N/A	4.588	4.114
Average	2	**5.222**	**4.677**	3	**5.535**	**5.055**	N/A	**4.213**	**3.981**

Of the regression methods used, MLR performed better than PLS and PCR in terms of SEC (average 4.213) and SEP (average 3.981). This chemometric method is able to detect economic adulteration to within 4.2% in the calibration set and to less than 4% in the validation set. This is superior to the model performance found previously (Gayo and Hale, 2006). In order to build stable models, research suggests that the number of samples must be at least five times the number of latent variables (Du *et al*., 2005). Therefore, the number of samples in the calibration (and validation) set was large enough to build stable PLS and PCR models since the number of LVs used was less than 11. Of the two latent-based methods, PLS performed better than PCR in terms of number of latent variables used, SEC, and SEP, as observed previously (Gayo *et al.*, 2006).

Table 6: Chemometric results using relevant variables selected by the genetic algorithm and the first derivative data.

	PLS			PCR			MLR		
Simulation	LVs	SEC	SEP	LVs	SEC	SEP	LVs	SEC	SEP
1	2	5.472	4.746	2	5.472	4.746	N/A	5.463	4.787
2	2	5.755	5.144	2	5.762	5.152	N/A	5.4501	4.845
3	2	5.965	4.944	2	5.973	4.949	N/A	5.205	4.483
4	2	5.472	4.746	2	5.472	4.746	N/A	5.463	4.787
5	2	5.723	5.095	2	5.731	5.232	N/A	5.454	4.832
Average	2	**5.677**	**4.935**	2	**5.682**	**4.965**	N/A	**5.407**	**4.747**

Results from the first derivative data again show MLR to be the best model, though the difference between its performance and PLS/PCR is lower. PLS and PCR performed similarly in terms of the number of latent variables (2 for both), SEC (5.677 versus 5.682, respectively), and SEP (4.935 vs. 4.965, respectively). Even though first derivative data is often used to remove constant background effects (Zeaiter *et al*., 2005), it can influence the genetic algorithm's selection of relevant variables. Due to the small number of variables chosen by the algorithm, it is possible that additional variables are needed to enhance the model's predictive ability to detect economic adulteration. Two types of spectral wavelengths exist in spectral analysis: descriptive, containing relevant problem-dependent information, and synergic. Though not relevant by themselves, synergic wavelengths enhance model performance by being combined with descriptive wavelengths (Leardi, 2003). Since the aim of GAs is to find relevant variables, i.e. descriptive wavelengths, it is possible that other less relevant variables, i.e. synergic wavelengths, not chosen by the GA, exist. Although it would increase the number of variables in the chemometric models, a combination of descriptive and synergic wavelengths could enhance the predictive ability of the models.

Chemometric results of the second derivative data are shown in Table 7. As in previous research (Gayo *et al*., 2006), models using selected variables of the second derivative data generated poorer results than those using either first derivative or untreated spectral data.

Table 7: Chemometric results using relevant variables selected by the genetic algorithm and the second derivative data.

	PLS			PCR			MLR		
Simulation	LVs	SEC	SEP	LVs	SEC	SEP	LVs	SEC	SEP
1	2	5.007	6.168	3	7.038	7.658	N/A	230.143	231.724
2	2	5.461	5.405	3	5.647	5.545	N/A	201.602	208.886
3	2	6.395	6.314	3	8.259	8.088	N/A	177.993	184.765
4	2	5.893	6.919	3	7.245	7.453	N/A	277.615	279.246
5	2	6.017	6.107	3	6.210	6.261	N/A	263.612	268.472
Average	2	**5.755**	**6.183**	3	**6.880**	**7.001**	N/A	**230.193**	**234.619**

Surprisingly, MLR, using the variables selected by the genetic algorithm, was not able to model the data (calibration or validation) as is evidenced by the extremely high SEC and SEP values. PLS and PCR, though able to model the data, generated higher SEC and SEP values than those generated by the raw and first derivative data. Second derivative data pre-treatment removes additive effects and baseline linear slope variations (Zeaiter *et al*., 2005). However, second derivative spectra are noisier, making it possible that PLS and PCR are modeling noise rather than adulteration percentage. In addition, it is possible that the convolution interval used in Savitzky-Golay derivation was not wide enough to remove collinearity among the remaining spectral variables in the second derivative data, resulting in high values of error in calibration and validation (Kompany-Zareh and Mirzaei, 2004). High collinearity among the spectral variables would yield an unstable estimator of the regression coefficients resulting in a poor prediction ability.

Visualization of the samples' scores was done in order to understand this unexpected result. A plot of the first two latent variables reveals a pattern of the samples that represent the adulteration percentage (Figure 5).

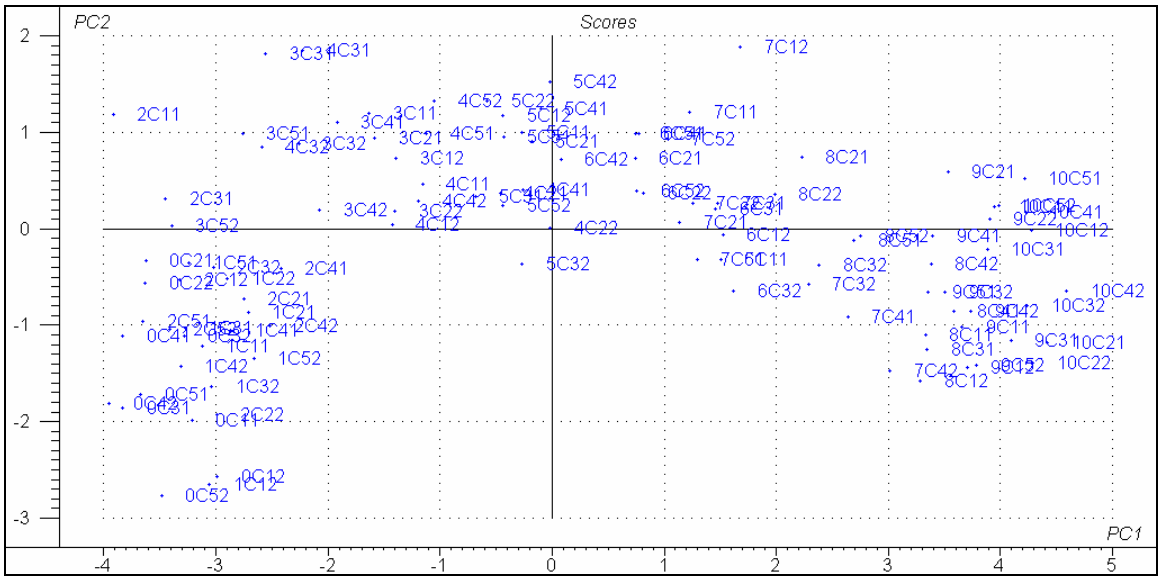

Figure 5: PCA plot of the first two principal components of the second derivative data. Explained variation in PC1 and PC2 were 60% and 9%, respectively. Samples are named according to economic adulteration (0 = 100% Atlantic blue, 10 = 100% blue swimmer, 1-9 according to increasing adulteration percentage), constant letter "C", sample number (1-5), and whether sample was in calibration (last digit = 1) or validation (last digit = 2).

Samples pertaining to 0% adulteration (100% Atlantic blue) are grouped in the lower left-hand quadrant. As adulteration percentage increases, samples reveal a convex upward relationship until the middle of the graph (upper half) where the samples belong to 50% adulteration. Further increase in adulteration reveals a linear relationship from the top middle of the graph down to the lower right-hand side quadrant where samples are 100% adulterated (100% blue swimmer). According to the PCA plot, PC1 and PC2 seem to explain economic adulteration. However, a different phenomenon, such as noise,

or an unknown experimental condition could also be explained by the two principal components and, thus, be modeled by PLS and PCR.

Research fails to demonstrate the improvement of a particular data pretreatment over the raw data in terms of quantitative analysis because it is not possible to know in advance which pretreatment-regression method combination will yield the best results over time (McClure, 1994). This is due to the fact that the performance of the calibration methods depends on the data structure of the current data set. Hence, superior performance of one method cannot be generalized to other data sets (Huang *et al*., 2002).

4. Conclusions

Adulteration of high valued food products, such as those made of crabmeat, is an existing problem in the seafood industry. Due to the large number of variables in spectral data sets, a genetic algorithm was used to select the relevant variables using untreated, first derivative, and second derivative pre-treated spectral data. The relevant variables were then used to derive calibration models using three common chemometric techniques (PLS, PCR, and MLR). Of the three models, MLR with the untreated data generated the lowest errors (average SEC = 4.213 and average SEP = 3.981). Of the two latent-based methods, PLS performed better than PCR in terms of the number of latent variables used and average values of errors (SEC and SEP). First derivative data yielded similar results though with closer average error values among the three regression methods. The second derivative data generated the highest error average for both PLS and PCR. MLR, using the selected variables from the second derivative data, was not able to model the data. Even though MLR did not work for second derivative data, results from this study

indicate that the combination of VIS/NIR spectroscopy, an efficient variable selection algorithm, and MLR can be used to detect economic adulteration of crabmeat to within 4.2%.

References

Abrahamsson, C., Johansson, J., Sparen, A., and Lindgren, F. 2003. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, *Chemometrics and Intelligent Laboratory Systems*, vol. 69, pp. 3-12.

Broadhurst, D., Goodavre, R., Jones, A., Rowland, J., and Kell, D. 1997. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression with applications to pyrolysis mass spectrometry, *Analytica Chimica Acta*, vol. 348, pp. 71-86.

Carlin, M., Kavli, T., and Lilekjendlie, B. 1994. A comparison of four methods for non-linear data modeling, *Chemometrics and Intelligent Laboratory Systems*, vol. 23, pp. 163-177.

Chong, I-G., and Jun, C-H. 2005. Performance of some variable selection methods when multicollinearity is present, *Chemometrics and Intelligent Laboratory Systems*, vol. 78, pp. 103-112.

Du, Y.P., Liang, Y.Z., Jiang, J.H., Berry, R.J., and Ozaki, Y. 2004. Spectral regions selection to improve prediction ability of PLS models by changeable size moving partial least squares and searching combination moving window partial least squares, *Analytica Chimica Acta*, vol. 501, pp. 183-191.

Feudale, R.N., Woody, N.A., Tan, H., Myles, A.J., Brown, S.D., and Ferre, J. 2002. Transfer of multivariate calibration models: a review, *Chemometrics and Intelligent Laboratory Systems*, vol. 64, pp. 181-192.

Gayo, J., Hale, S.A., and Blanchard, S.M. 2006. Quantitative analysis and detection of adulteration in crab meat using visible and near-infrared spectroscopy, *Journal of Agriculture and Food Chemistry*, vol. 54 (no. 4.), pp. 1130-1136.

Gayo, J., and Hale, S.A. 2006. Detection and quantification of species authenticity and adulteration in crab meat using visible and near-infrared spectroscopy, (Not Published)

Goicoechea, H.C., and Olivieri, A.C. 2003. A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy, *Journal of Chemometrics*, vol. 17, pp. 338-345.

Gusnanto, A., Pawitan, Y., Huang, J., and Lane, B. 2003. Variable selection in random calibration of near-infrared instruments: ridge regression and partial least squares regression settings, *Journal of Chemometrics*, vol. 17, pp. 174-185.

Hageman, J.A., Streppel, M., Wehrens, R., and Buydens, L.M.C. 2003. Wavelength selection with tabu search, *Journal of Chemometrics*, vol. 17, pp. 427-437.

Hemmateenejad, B., Miri, R., Akhond, M., and Shamsipur, M. 2002. QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. An application of genetic algorithm for variable selection in MLR and PLS methods, *Chemometrics and Intelligent Laboratory Systems*, vol. 64, pp. 91-99.

Huang, J., Brennan, D., Sattler, L., Alderman, J., Lane, B., and O'Mathuna, C. 2002. A comparison of calibration methods based on calibration data size and robustness, *Chemometrics and Intelligent Laboratory Systems*, vol. 62, pp. 25-35.

Janne, K., Pettersen, J., Lindberg, N.O., and Lundstedt, T. 2001. Hierarchical principal component analysis (PCA) and projection to latent structure (PLS) technique on spectroscopic data as a data pretreatment for calibration, *Journal of Chemometrics*, vol. 15, pp. 203-213.

Jorgensen, P., Pedersen, J.G., Jensen, E.P., and Esbensen, K. 2004. On-line batch fermentation process monitoring (NIR) – introducing 'biological process time,' *Journal of Chemometrics*, vol. 18, pp. 81-91.

Kompany-Zareh, M., and Mirzaei, S. 2004. Genetic algorithm-based method for selecting conditions in multivariate determination of povidone-iodine using hand scanner, *Analytica Chimica Acta*, vol. 521, pp. 231-236.

Leardi, R. 2000. Application of genetic algorithm – PLS for feature selection in spectral data sets, *Journal of Chemometrics*, vol. 14, pp. 643-655.

Leardi, R., and Norgaard, L. 2004. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, *Journal of Chemometrics*, vol. 18, pp. 486-497.

Leardi, R., Seasholtz, M.B., and Pell, R.J. 2002. Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data, *Analytica Chimica Acta*, vol. 461, pp. 189-2000.

Luke, B.T. 2003. Genetic algorithms and beyond, In Leardi, R. (Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier: New York, pp.3-54.

McClure, W.F. 1994. In, Wilson, R.H., Spectroscopic techniques for food analysis, VCH Publishers, Inc, New York, pp. 13-58.

Mertens, B.J., O'Donnell, C.P., and O'Callaghan, D.J. 2002. Modeling near-infrared signals for on-line monitoring in cheese manufacture, *Journal of Chemometrics*, vol. 16, pp. 89-98.

Miller, C.E. 2000. Chemometrics for on-line spectroscopy applications – theory and practice, *Journal of Chemometrics*, vol. 14, pp. 513-528.

Osterling, M.J. 1998. "Blue" Crab Resources in Other Countries: Implications for the U.S. Industry, *Journal of Shellfish Research*, vol. 17, no. 2, pp. 375-378.

Wehrens, R., and Buydens, L.M.C. 1998. Evolutionary optimization: a tutorial, *Trends in analytical chemistry*, vol. 17, no.4, pp. 193-203.

Wold, S., Trygg, J., Berglund, A., and Antti, H. 2001. Some recent developments in PLS modeling, *Chemometrics and Intelligent Laboratory Systems*, vol. 58, pp. 131-150.

Zeaiter, M., Roger, J.-M., and Bellon-Maurel, V. 2005. Robustness of models developed by multivariate calibration. Part II: The influence of pre-processing methods, *Trends in Analytical Chemistry*, vol. 24, no. 5, pp. 437-445.

6.0 SUMMARY AND CONCLUSIONS

A base for the determination of economic adulteration of blue crabmeat was established. The adulterants in this study included surimi-based imitation crabmeat and blue swimmer crabmeat. Visible and Near-Infrared (VIS/NIR) spectroscopy was evaluated as a fast and efficient technology for determination of species authenticity and adulteration level in blue crabmeat. Performance of several chemometric techniques was evaluated in terms of ability to predict economic adulteration. In addition, wavelength variable selection, via a genetic algorithm, was performed to determine the most relevant wavelengths for prediction of adulteration of crabmeat. The objectives of this research of detecting economic adulteration of crabmeat were achieved, and have been presented in three manuscripts.

The first manuscript, titled "Quantitative Analysis and Detection of Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy," was an initial study used to examine the feasibility of VIS/NIR spectroscopy to detect crabmeat adulterated with surimi-based imitation crabmeat. Two chemometric analyses (PLS and PCR) were evaluated for prediction ability of economic adulteration of crabmeat. Basic wavelength selection was also investigated using a correlogram. Principal component analysis revealed a pattern of samples showing grouping according to level of economic adulteration. Both PLS and PCR performed similarly, but PLS was preferred due to the lesser number of latent variables in the model. It was found that first derivative spectral data from the correlogram and PLS generated the best results in terms of prediction of economic adulteration (SEC and SEP of 2.52%).

The second manuscript, titled "Detection and Quantification of Species Authenticity and Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy," evaluated different data pre-treatments (first and second derivative, combing, smoothing, and multiplicative scatter correction) for PLS performance in detecting species authenticity and economic adulteration of crabmeat using VIS/NIR spectroscopy. Blue swimmer crabmeat, the most prevalent type of crabmeat imported into the United States, was the adulterant for this study. In addition, two spectrum approaches (full and partitioned) were investigated in terms of predictive ability of PLS for detecting species authenticity and economic adulteration. Using the full spectrum, the first derivative data generated the best results (SEC and SEP of 5.64 and 5.64, respectively). Using a partitioned approach, the second derivative data produced a better model with errors of less than 5.2%.

The third manuscript, titled "Detection of Economic Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy and a Genetic Algorithm for Wavelength Variable Selection," determined the minimum number of relevant wavelength spectral variables necessary to detect economic adulteration of crabmeat. Three calibration models (PLS, PCR, and MLR) were evaluated using the relevant wavelength variables for VIS/NIR spectroscopy performance in predicting economic adulteration. Two data pretreatments (first and second derivative) were evaluated in addition to the untreated data. The genetic algorithm was able to select between 5 to 9 spectral variables (out of 175) as relevant for detection of economic adulteration. Wavelength variables selected by the genetic algorithm improved regression model performance. It was found that, using untreated data, MLR outperformed the other models and generated the lowest SEC

and SEP values (4.21 and 3.98, respectively) to detect economic adulteration of crabmeat using blue swimmer as the adulterant.

The overall findings of this research of the predictive ability to detect economic adulteration of crabmeat adulterated with surimi-based imitation crabmeat, the detection of species authenticity and economic adulteration of crabmeat adulterated with blue swimmer crabmeat, and the advantage of an efficient wavelength variable selection on chemometric model performance provide a baseline for detection of economic adulteration of crabmeat. Therefore, it can be concluded that VIS/NIR spectroscopy is a fast, reliable, and accurate technology for on-line monitoring of economic adulteration at the processing level. Spectral wavelength ranges could be used to design and develop an instrument for fast detection of economic adulteration. Although more research is needed before implementing this technology, these results represent an important application for the blue crab industry. In summary, the main objective of finding a fast and reliable way to detect economic adulteration of crabmeat by using VIS/NIR spectroscopy has been achieved.

APPENDIX

The genetic algorithm was used to select the relevant spectral variables for the purpose of detecting economic adulteration of crabmeat samples adulterated with blue swimmer (*Portunus pelagicus*). Results from variable selection of the study are presented and discussed in the third manuscript, titled "Detection of Economic Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy and a Genetic Algorithm for Wavelength Variable Selection." It was demonstrated that MLR, utilizing the selected spectral variables to undergo regression, yielded the best model in terms of SEC and SEP. The aim of this appendix is to discuss the chemometric results gathered when the GA was used for selecting the relevant spectral variables to detect economic adulteration of crabmeat samples adulterated with surimi-based imitation crabmeat. The author encourages the reader to refer to manuscript 1 ("Quantitative Analysis and Detection of Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy") for description of sample preparation and analysis and to manuscript 3 ("Detection of Economic Adulteration in Crabmeat Using Visible and Near-Infrared Spectroscopy and a Genetic Algorithm for Wavelength Variable Selection") for the description of the procedure for variable selection and optimal parameters used in the GA. The results reported here are based on the untreated spectral data only.

For the benefit of the reader, the spectra for the Atlantic blue crabmeat and the blue swimmer are shown in Figures 1 and 2.

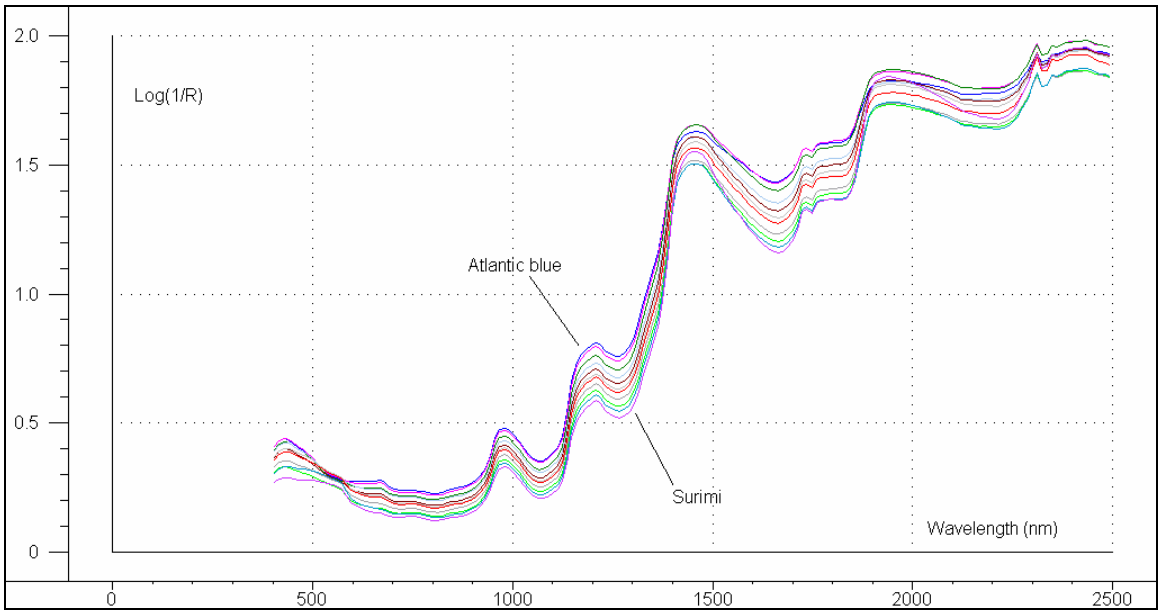

Figure 1: Average absorbance spectra of adulteration samples of Atlantic blue crabmeat. Top spectra represents class 0 (0% adulteration), bottom spectra class 10 (100% adulteration), and classes 1-9 (10%-90% adulteration) are in order from top to bottom.

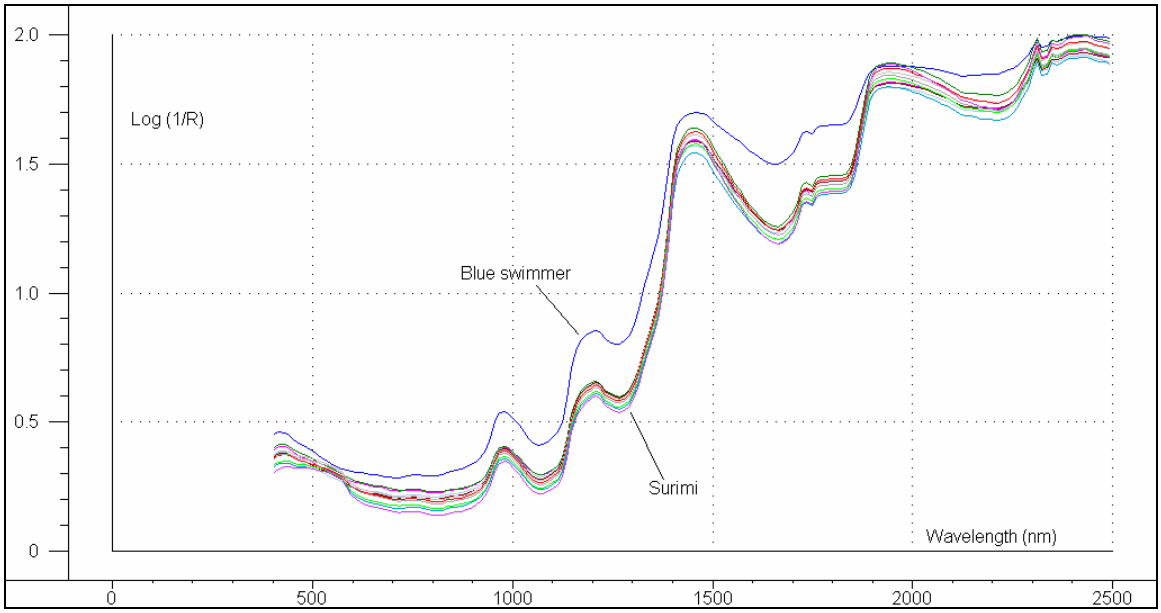

Figure 2: Average absorbance spectra of adulteration samples of blue swimmer crabmeat. Top spectra represents class 0 (0% adulteration), bottom spectra class 10 (100% adulteration), and classes 1-9 (10%-90% adulteration) are in order from top to bottom.

Table 1 shows the selected variables, corresponding wavelengths, RMSECV, and percentage variance of cross-validation, for the untreated data in the calibration set for each simulation.

Table 1: Variables selected in each simulation for the untreated spectral data. The number of variables selected by each simulation is 10, 13, 13, 12, and 15, respectively.

Simulation	*Selected Variables*	*Corresponding Wavelengths (nm)*	*RMSE CV*	*%CV*
1	47-51, 82-84, 102-103	952-1010, 1372-1406, 1612-1634	1.290	99.834
2	35-36, 47-52, 82-83, 103-105	808-830, 952-1022, 1372-1394, 1624-1658	1.340	99.825
3	47-50, 59, 83-85, 102-106	952-998, 1096-1106, 1372-1394, 1612-1670	1.209	99.854
4	47-51, 83, 100-102, 104, 123-124	952-1010, 1384-1394, 1588-1622, 1636-1646, 1864-1886	1.286	99.835
5	47-51, 61-63, 101-107	952-1010, 1120-1152, 1600-1682	1.510	99.772

None of the wavelengths corresponding to the selected spectral variables lie in the visible range (400-700 nm). Figure 3 shows the spectral variables selected as relevant by the five different simulations using the untreated spectral data. The five broken lines below the absorbance spectra represent the spectral variables chosen as relevant by the five GA simulations. The RMSECV of the five GA simulations (from the top to the bottom of the broken lines) is 1.290 (ten variables), 1.340 (13), 1.209 (13), 1.286 (12), and 1.510 (15).

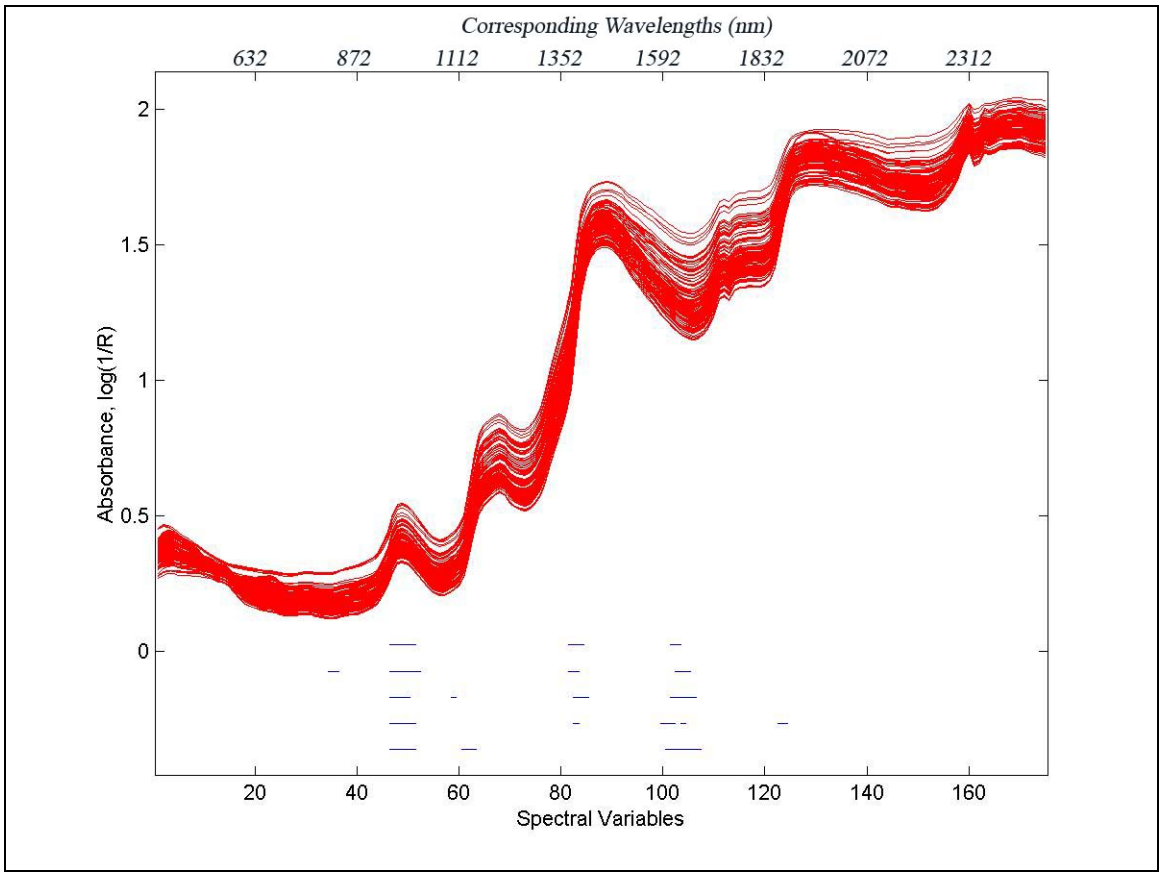

Figure 3: Variables selected from 5 separate simulations of the calibration set (untreated spectral data). Lines showing selected spectral variables represent the simulations from number one (top) to number five (bottom).

Based on the graph, two areas (47-51 and 100-104) were selected by the GA simulations. These areas corresponded to ranges of 952-1010 nm and 1588-1658, respectively. A third (smaller) spectral area was also chosen by four of the five simulations (82-84 corresponding to 1376-1406 nm).

Table 2 shows the results in terms of number of latent variables, SEC, and SEP, of each individual simulation for the three chemometric techniques for the untreated data. The average value of the five simulations is marked in bold.

Table 2: Chemometric results using the relevant variables, selected by the genetic algorithm, and the untreated data. Average values in bold.

	PLS			PCR			MLR		
Simulation	LVs	SEC	SEP	LVs	SEC	SEP	LVs	SEC	SEP
1	5	2.605	3.158	6	3.073	3.878	N/A	0.957	1.062
2	5	2.720	3.088	6	2.761	3.214	N/A	0.933	1.109
3	5	3.189	3.061	6	2.736	3.365	N/A	0.861	1.004
4	5	2.773	3.100	6	2.033	2.636	N/A	0.963	1.074
5	5	2.637	2.965	6	2.736	3.002	N/A	1.008	1.199
Average	5	**2.785**	**3.074**	6	**2.668**	**3.219**	N/A	**0.944**	**1.090**

Both PLS and PCR performed similarly in terms of SEC (2.785 vs. 2.668) and SEP (3.074 vs. 3.219) though PCR utilized one more latent variable. MLR, with the selected spectral variables, not only generated the best results but is superior to the model previously found in manuscript 1. The previous model (PLS with first derivative data from a correlogram) generated a SEC and SEP of 2.51 and 2.52, respectively. Utilizing the spectral variables selected by the GA, MLR generated an average SEC and SEP of 0.944 and 1.090, respectively. Since 2*SEP is regarded as the 95% confidence interval, this model is able to detect economic adulteration of crabmeat adulterated with surimi-based imitation crabmeat to ±1.1%. The results from this research emphasize the importance of appropriate spectral variable selection in optimizing the model's ability to predict economic adulteration of crabmeat.