Fall 2014

Pattern recognition methods for automated detection and quantification: applications to passive remote sensing and near infrared spectroscopy

Hua Yu
University of Iowa

PATTERN RECOGNITION METHODS FOR AUTOMATED DETECTION AND QUANTIFICATION: APPLICATIONS TO PASSIVE REMOTE SENSING AND NEAR INFRARED SPECTROSCOPY

by

Hua Yu

A thesis submitted in partial fulfillment
of the requirements for the Doctor of
Philosophy degree in Chemistry
in the Graduate College of
The University of Iowa

December 2014

Thesis Supervisor: Professor Gary W. Small

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Hua Yu

has been approved by the Examining Committee
for the thesis requirement for the Doctor of
Philosophy degree in Chemistry
at the December 2014 graduation.

Thesis Committee:

Gary W. Small, Thesis Supervisor

Mark A. Arnold

M. Lei Geng

Claudio J. Margulis

William E. Eichinger

To my Parents

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Gary W. Small for his guidance, encouragement and support throughout the entire time I pursue my PhD in his research group. I would also like to thank the members of my committees, Dr. Mark A. Arnold, Dr. Lei M. Geng, Dr. Claudio J. Margulis, and Dr. William E. Eichinger. For the invaluable discussion and suggestion they offered to shape this dissertation. Many thanks go to current group member, Dr. Sanjeewa Rasika K Ranasinghe Pathirajage, Brian Dess, and Ziqi Fan as well as former group members. Dr. Qiaohan Guo, Dr. Chamathca Priyanwada Kuda-Malwathumullage and Wei Wang. I would like to thank the Environmental Protection Agency for funding the research in this dissertation. Special acknowledgement goes to Dr. Robert Kroutil and Dr. Mark Thomas from EPA, who were in charge of data collection of remote sensing FTIR and gamma-ray spectra used in this research and offered offer their expertise in the field of remote sensing.

Last but not the least, my deepest and utmost appreciation goes to my parents, Xueying Chen and Rongfang Yu, who always encourage me to better myself and meet challenges. Without their unconditional love, inspiration and sacrifice, I won’t be able to complete the course of my graduate study, and become who I am today.

ABSTRACT

Pattern recognition has over past decades become a fast growing area of chemometrics. Accurate, user-friendly, and fast pattern recognition methods are desired to accommodate the increased capacity of automated instruments to obtain large-scale data under complex circumstances. It has found significant applications in diverse fields such as environmental monitoring and biomedical diagnostics. In this dissertation, the capabilities of pattern recognition methods in case studies related to environmental remote sensing and biomedical sensing are investigated.

For remote sensing applications, two types of airborne spectroscopic data, passive Fourier transform infrared (FTIR) and gamma-ray, are subject to analysis in order to develop automated classifiers for either ammonia vapor or the radioisotope cesium-137 in the open-air. Support vector machine (SVM) classification is the primary pattern recognition method used in this work. In order to overcome the limitation of available representative patterns associated with airborne data, and provide sufficient patterns presenting the analyte-active class for use in the training set, a spectral simulation protocol is employed to generate abundant patterns bearing both the signature of the target analyte and the background spectral profile. Signal processing procedures including segment selection and digital filtering are further used to extract the information most relevant to the target analyte out the acquired raw data. Also, to ease the computational demand from the SVM, an alternative pattern recognition method, piecewise linear discriminant analysis (PLDA) is applied to optimize signal processing conditions for final SVM classification. Process control techniques are applied to the SVM score profiles of prediction sets to improve pattern recognition performance by incorporating probabilities associated with every SVM score. Ammonia classifiers developed from this methodology result

in classification performance with high sensitivity and selectivity, and the cesium-137 classifiers developed from the same concepts exhibit excellent sensitivity to test data with very low signal strengths. Under the case of ammonia classification, the relationship between the concentration profile of the active patterns in the training set and the limit of detection of the corresponding classifier is investigated. Classifiers built to detect low concentrations of ammonia are developed and tested through this work.

For a glucose sensing application, studies are conducted to provide sound performance diagnostics for an established calibration model for glucose from near infrared spectroscopic data. Six-component aqueous matrixes of glucose in the presence of five other interfering species, all spanning physiological levels, serve as samples to be analyzed. A novel residual modeling protocol is proposed to retrieve the residual glucose concentrations, the concentration not being predicted by the calibration model, from the residual spectra, the portion of the raw spectra not being used by the calibration model. The recovered glucose concentration from the residual modeling can be used as a means, combined with process control techniques, to evaluate the performance of the established calibration model. Several modeling techniques are used for residual modeling, including PLS, support vector regression (SVR), a hybrid method, PLS-aided SVR, and an amplified version of the hybrid, amplified PLS-aided SVR. Through this work, a calibration updating strategy is developed which provides an effective way to monitor the established calibration model.

PUBLIC ABSTRACT

How to discover a wealth of information about chemical of interest, whether is qualitative or quantitative information, buried in a variety of large-scale spectral data is the broad-sense topic of this research. The set of tool used for such data-mining tasks is chemometrics, Chemometric methods such as pattern recognition, multivariate calibration and signal processing were collectively used here to extract chemical information from spectral data. Such information can be used for chemical identification, qualification, or monitoring changes of chemical over time. Four projects presenting different challenging issues, either from environmental monitoring or biomedical application, down to the roots are all keen on solutions for chemical qualification. How to solve the issues were showcased here. Based on airborne infrared spectra collected in the remote sensing measurement using natural occurring sunlight as light source, pattern recognition methods was able to establish an automated ammonia classifier helping to pinpoint hazard emission, further refined classifier prove to be able to provide quantitative information, such as limit of detection. The similar methodology applied to airborne remote sensing gamma-ray spectra, efforts was made to develop an automated classifier for radioactive isotope, cesium-137. In the last project, an innovative way to morning the calibration model itself was developed pattern recognition, while glucose level in aqueous samples will be predicted based on near infrared spectra by the multivariate calibration model. Successful implementation of chemo metrics to gain chemical information from spectra in four different stories vouches the methodologies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Chemometrics

As a sub-discipline of analytical chemistry, chemometrics features the extensive use of a collection of techniques drawn from mathematics, statistics and formal logic to design or select optimal measurement procedures, to provide maximum relevant chemical information by analyzing chemical data, and to obtain knowledge about chemical systems. [1]

The ongoing digital revolution that started 70 years ago has had profound influence on all aspects of how we live today, including how experimental chemistry is conducted. In the wake of the proliferation of computers in chemistry laboratories, the term “chemometrics” debuted officially in the early 1970s, as a response to the demand for improved methods to design and control experiments, and to analyze the wealth of collected digital data which had never been encountered in traditional instrumental analysis. The emergence of chemometrics also echoed the desire in that era to move analytical chemistry outside the controlled laboratory environment, and thus attempt to make measurements in complex environments such as industrial settings. [2-4] Since then, over the past 40 years, the scope of this chemical sub-discipline, across both academic research and industrial applications, has grown, evolved and been refined. This development has been driven by economic forces, enhanced computing capacities and increasingly sophisticated analytical instrumentation. [5-7]

To date, chemometric techniques, such as signal processing, multivariate calibration, pattern recognition, experimental design/optimization, have been widely investigated in academic settings or been adopted in industrial applications, such as chromatographic peak resolution, [8,9] online process control of matrix samples, [10-12] and automated identification of

chemicals. [13,14] This dissertation describes the development of chemometric methodology directed to the analysis of data from applications of infrared and gamma-ray spectroscopies.

1.2 **Spectroscopic data**

1.2.1 **Infrared spectroscopy**

Though chemometrics has been widely associated with a variety of analytical instrumental data, such as high performance liquid chromatography (HPLC) [15,16] and gas chromatography-mass spectrometry (GC-MS), [17,18] the most numerous examples of the applications of chemometrics have been with optical spectroscopic data. Among spectroscopic techniques, infrared (IR) measurements have received great benefit from the application of chemometric methods. [19-23]

Infrared spectroscopy involves the analysis of electromagnetic radiation in the IR region and how this energy interacts with chemical systems. As typically defined, the IR spectral region spans from 12,800 – 400 cm^{-1} in wavenumber (or 700 nm – 25 μm in wavelength). Depending on the energy level of the radiation, the IR region can be further divided into three major sub-regions: near-IR (12,800 – 4000 cm^{-1} or 0.78 – 2.5 μm), mid-IR (4000 – 200 cm^{-1} or 2.5 – 50 μm), and far-IR (200 – 10 cm^{-1} or 50 – 1000 μm). Infrared spectroscopy interrogates transitions between vibrational and associated rotational energy levels of molecules. [24] In this dissertation, the techniques of interest are vibrational spectroscopies in the mid-IR and near-IR regions.

To interact with IR radiation, a molecule must undergo a net change in dipole moment as it vibrates. Under this condition, covalent bonds within molecules can be envisioned as stiff springs that can oscillate in a variety of ways, such as bending or stretching, thereby generating quantized vibrational energy states, which are labeled by the vibrational quantum number, υ. When incident IR radiation carries the same level of energy, it can lead to a net uptake of energy, resulting in IR absorption.

Infrared spectroscopy is widely used for both qualitative and quantitative purposes. The frequency corresponding to a fundamental vibrational transition, a type of transition from the ground state ($\upsilon = 0$) to the first vibrationally excited state ($\upsilon = 1$), is characteristic of the specific functional group that contains the dipole that serves as the basis for the transition. The frequency thus provides useful information for molecular structural elucidation and chemical identification. On the other hand, though not obeying the IR selection rule but partially explained by anharmonicity, overtone bands, corresponding to transitions from the ground state ($\upsilon = 0$) to the second or higher excited states ($\upsilon \geq 2$), and combination bands, corresponding to transitions from the ground state ($\upsilon = 0$) to different first excited states ($\upsilon_i = 1$, $\upsilon_j = 1$) simultaneously, do exist. In fact, the weakness of overtone and combination bands allows longer optical pathlengths, which turns into advantages when IR (especially near-IR) is used for quantitative analysis. [25]

The majority of current IR instruments are of the Fourier transform (FT) type, many of which are based on the two-beam interferometer designed by Michelson. In the Michelson interferometer, a beamsplitter is used to divide an incident light beam and direct each half to either a fixed or moving mirror. Depending on the distance each beam has traveled, recombination at the beamsplitter results in optical interference. By translating the moving mirror at constant velocity, an alternating pattern of constructive and destructive interference is obtained.

For absorption spectroscopy, the resultant beam that is generated at the beamsplitter is passed through a sample and onto a detector. For emission measurements, the emission from the sample is directed into the interferometer and the resultant beam after recombination at the beamsplitter is sent to the detector. The detector signal produces an interferogram, a plot of

measured signal vs. displacement distance of the moving mirror. One interferogram corresponds to one traversal of the moving mirror along its track.

The interferogram demonstrates an alternately constructive or destructive interference pattern, modified by whatever frequency-dependent changes in intensity that result from vibrational transitions in the sample. The frequency of this interference pattern is proportional to the velocity of the moving mirror and the wavenumber of the incident light. For example, a mid-IR wavenumber of 1000 cm^{-1} produces an interference pattern oscillating in the kHz range if the mirror velocity is on the order of 0.5 cm/sec.

As a result, the Michelson interferometer uniquely encodes a very high frequency optical signal on the order of 10^{13} Hz into a corresponding periodic waveform varying in the kHz range. Whereas currently available detectors and electronics cannot respond quickly enough to see intensity oscillations at 10^{13} Hz, the kHz pattern can be observed easily. The interferometer can thus be viewed as a light modulator and the interferogram as a modulated waveform.

In the case of a polychromatic light source, the resulting interferogram is essentially the integral of the interferograms produced by each light frequency emitted by the source. When the moving and fixed mirrors are equidistant from the beamsplitter, constructive interference will occur and the interferogram generated for every light frequency will have a maximum. The integral of these signals will thus produce a maximum value in the overall composite interferogram. This is termed the interferogram centerburst. As the mirror moves away from this position, the composite interferogram will exhibit an exponentially damping signal due to the result of the summation of the interferograms produced by all the source frequencies.

As long as all the modulated frequencies are within the response time of the detector and associated electronics, the digitized interferogram can be analyzed to recover its component

frequencies and corresponding intensities. Through application of the FT, the single-beam spectrum (light intensity vs. wavenumber) can thus be obtained from the measured interferogram signal.

Fourier transform IR (FT-IR) offers throughput and multiplex advantages over dispersive IR spectrometers, owing to the contribution of every light frequency to every interferogram point and the ability to measure the spectrum with an open-beam optical geometry (i.e., no requirement for slits and typically fewer reflective optics). For mid-IR and near-IR frequencies that are not currently accessible with inexpensive multichannel detectors, the FT-IR design offers fast scanning with high optical throughput. [26] This provides significant advantages in measurement applications in which the presence of low light signals dictates the need for both high optical throughput and fast enough scanning to allow practical signal averaging. The dissertation research focuses on two such applications: the qualitative analysis of mid-IR emission data derived from passive remote sensing measurements and the quantitative analysis of near-IR spectra collected from highly absorbing aqueous samples.

1.2.2 **Gamma-ray spectroscopy**

Another type of spectroscopy under the scope of this dissertation is gamma-ray emission spectroscopy. Radioactive decay occurs when excited radionuclides of a radioactive isotope undergo relaxation leading to stable nuclides. Gamma rays represent one of the forms of emission of energy that results from radioactive decay. Emitted gamma rays (actually high-energy photons) can interact with matter via three main mechanisms: the photoelectric effect, Compton scattering, and pair production.

Over the typical gamma-ray energy range of 3 keV to 2MeV, Compton scattering is the major form of interaction. Photons striking the detectors are captured, converted and amplified to an electrical signal. The gamma-ray spectrum is recorded as a plot of the count of the gamma-ray

photons detected vs. the corresponding energy. The characteristic peaks for the various radioisotopes or their decay products are superimposed on a Compton continuum. [24,27] The dissertation research includes an application of the developed chemometric methodology to the automated identification of target radioisotopes from gamma-ray spectra acquired from a downward-looking spectrometer mounted on an aircraft platform.

1.3 **Applications**

1.3.1 **Environmental analysis based on infrared and gamma-ray spectroscopies**

Air pollution, caused by the emission of toxic pollutants, is considered one of the most challenging current environmental issues, leading to adverse health effects and environmental impact, both in the United States and globally. Stringent air quality standards have been established worldwide in the effort to reduce the emission of potential pollutants. [28,29] Following this trend, the need exists for analytical methods that are able to provide reliable and quick identification/qualification of air pollutants. [30-32]

As an example, for environmental applications in this category, remote sensing methods based on FT-IR spectrometry have gained popularity in applications such as the identification of volatile organic compounds (VOCs) in the ambient atmosphere. The portability of IR remote sensing instrumentation, combined with the molecular identification capacity of FT-IR measurements in the mid-IR region, especially at wavelengths in the atmospheric transmission window of 8-12 μm, have contributed to the success of the methodology for environmental monitoring applications.

Depending on the type of light source, field-based IR remote sensing can be divided into two modes: active or passive. An instrumental light source is employed in active mode remote sensing. Except for the use of open air as the sample and the presence of a longer and uncontrolled optical pathlength for the sample, active mode remote sensing has essentially the

same configuration as a laboratory IR measurement. As a consequence, active-mode measurements feature more sensitive detection or relatively lower limits of detection when compared to passive methods. The active mode suffers from poor flexibility and a lack of mobility of the experimental setup, however, because of the need to have either the light source or a retroreflector on the other side of the atmospheric sampling volume. There is no opportunity for mounting the instrument on a moving platform such as an aircraft.

Passive mode remote sensing, featuring the use of naturally occurring IR radiance as the light source, is much more compatible with mobile data acquisition. There is no need to position a source or retroreflector on the other side of the sampling volume. The IR background is simply the summation of the radiance sources within the field-of-view (FOV) of the spectrometer. The signatures of IR-active species within the FOV of the spectrometer will be superimposed on the background, either as absorption or emission bands.

The use of natural IR radiance as the light source creates its own limitations, however. First, according to the radiance model that governs the experiment, detection capacity is dependent on a temperature difference between the IR background and the target analyte within the instrumental FOV. This results in low analytical sensitivity when the temperature differential is very low. Second, if the spectrometer is placed on a moving platform, it is impossible to collect a stable and constant background radiance. The intensity of the naturally occurring light source and the background scenes within the instrumental FOV are constantly varying. For this reason, a spectrum computed in units of absorbance, calculated on the basis of the ratio of a sample spectrum to a stable background spectrum, is not available.

Another consequence of the collection of data from a moving platform is the expense and complexity of collecting data for use in calibrating any qualitative or quantitative analysis

method. For example, if the spectrometer is mounted in a downward-looking mode on an aircraft, there will be relatively few detections of any ground-level emissions of a target analyte because of the small relative area over which the emission is occurring. Most of the data collected from the air will be backgrounds that contain no analyte signatures. Limitations in sample size thus impede the development of reliable quantitative/qualitative analytical methods. [33-35]

In this dissertation, a series of chemometric techniques are implemented on airborne remote sensing passive FT-IR data in order to develop automated identification methods for target VOCs. The theory and instrumentation related to FT-IR data collection and passive remote sensing are described in Chapter 2. A detailed description of the chemometric techniques employed in this research, including pattern recognition, signal processing and multivariate calibration are provided in Chapter 3.

Chapters 4 and 5 are case studies related to developing an automated classifier for ambient ammonia based on passive FT-IR spectroscopic data. In Chapter 4, a full set of methodology designed specifically for this type of spectroscopic data, including supervised pattern recognition algorithms for binary classification, mathematically synthetic data as representative patterns, digital filtering and segment selection for prepossessing interferograms are developed and validated.

Based upon the framework developed in Chapter 4, Chapter 5 explores the relationship of the concentration profile of synthetic data used in developing the classification models and the estimated detection limits of the resulting classifiers. Unsupervised pattern recognition methods are used in this study. On the basis of this work, a more refined classifier with a lower limit of detection for ammonia is proposed.

Another application of interest in the field of environmental remote sensing lies in radioisotope detection. The proliferation of nuclear materials and the threat of terrorist attacks involving nuclear materials have led to increased interest in remote monitoring capabilities for detecting specific radioisotopes.

Gamma-ray spectroscopy is an appealing technique for use in detecting radioisotopes because each isotope typically produces a characteristic pattern of gamma-ray emissions. Furthermore, just as passive IR spectroscopy is amenable to implementation on a moving platform, so too are gamma-ray measurements. In Chapter 6, the signal processing and pattern recognition algorithms developed in Chapters 4 and 5 for use with passive FT-IR remote sensing data are applied to the automated detection of gamma-ray signatures. In this work, gamma-ray spectra collected with an airborne spectrometer are used for the detection of signatures of ^{137}Cs.

Biomedical applications

Besides environment applications, IR spectroscopy, especially near-IR spectroscopy, has found increased interest in biomedical applications. One example application is the development of noninvasive *in vivo* blood glucose monitoring based on near-IR spectroscopy. [36,37] The advantages of near-IR transmission measurements, such as requiring little or no sample preparation, reduced water absorption in the region, and relatively long optical path lengths, show promise for the implementation of non-destructive direct monitoring of glucose levels in the dermis. The disadvantages of near-IR spectroscopy, such as weak signals and highly overlapped spectral signatures, can be overcome by using chemometrics to implement multivariate calibration methods such as partial least-squares (PLS) regression. Through this approach, quantitative information related to glucose can be extracted from near-IR transmission data collected from complex samples. [25,26,38]

However, multivariate calibration models are subject to model degradation over time, caused by variations from difference sources, such as drift in the instrumental response, changes in environmental conditions associated with the data measurement, or sample-related spectral shifts occurring with time. It makes calibration updating a significant topic associated with near-IR multivariate calibration. [39-41]

In Chapter 7, near-IR transmission spectra of aqueous samples containing physiological levels of six biologically-relevant components are used to develop PLS calibration models for the quantitative determination of glucose. It was found that due to spectral variations that occur with time, there are substantial amounts of quantitative information remaining in the spectra after application of the PLS method. The work described in Chapter 7 relates to the extraction of this residual information and its use as both a diagnostic of calibration performance and as a calibration updating method.

In Chapter 8, based upon the research projects presented in Chapters 4 -7, overall conclusions are drawn and some of the potential directions for future work are discussed.

Chapter 2

INFRARED AND GAMMA-RAY MEASUREMENT TECHNIQUES

In this chapter, important principles of FT-IR instrumentation, associated raw data processing, background related to passive FT-IR based remote sensing measurements and near-IR transmission measurements will be discussed. In addition, theoretical and instrumental background related to the gamma-ray measurements will be provided.

2.1 Overview of FT-IR principles 42-44

Since the middle 1980s, FT spectrometers have rapidly replaced dispersive instruments and have become the major type of spectrometer used in the IR region. Compared to a dispersive instrument with a single-channel detector, the FT-IR design with its throughput and multiplex advantages can typically collect spectra with better SNRs, more precise wavenumber measurements, and shorter acquisition times.

The advantage of FT-IR spectrometry comes from its unique design. The key to every FT-IR design is an optical device called an interferometer. One of the most widely used interferometers is based on the original design of Michelson in 1891. A schematic diagram of a Michelson interferometer is presented in Figure 2.1. A collimated beam of radiation from the IR source is divided into two equal parts by a beamsplitter. The resultant half beams, one from transmission and the other from reflection occurring at the beamsplitter, then travel different optical paths. The reflected half-beam goes to a fixed mirror, and the transmitted half-beam goes to a mirror that has the ability to move along a track. When the two beams reflected by the mirrors return to the beamsplitter, they recombine. Interference between the two recombined

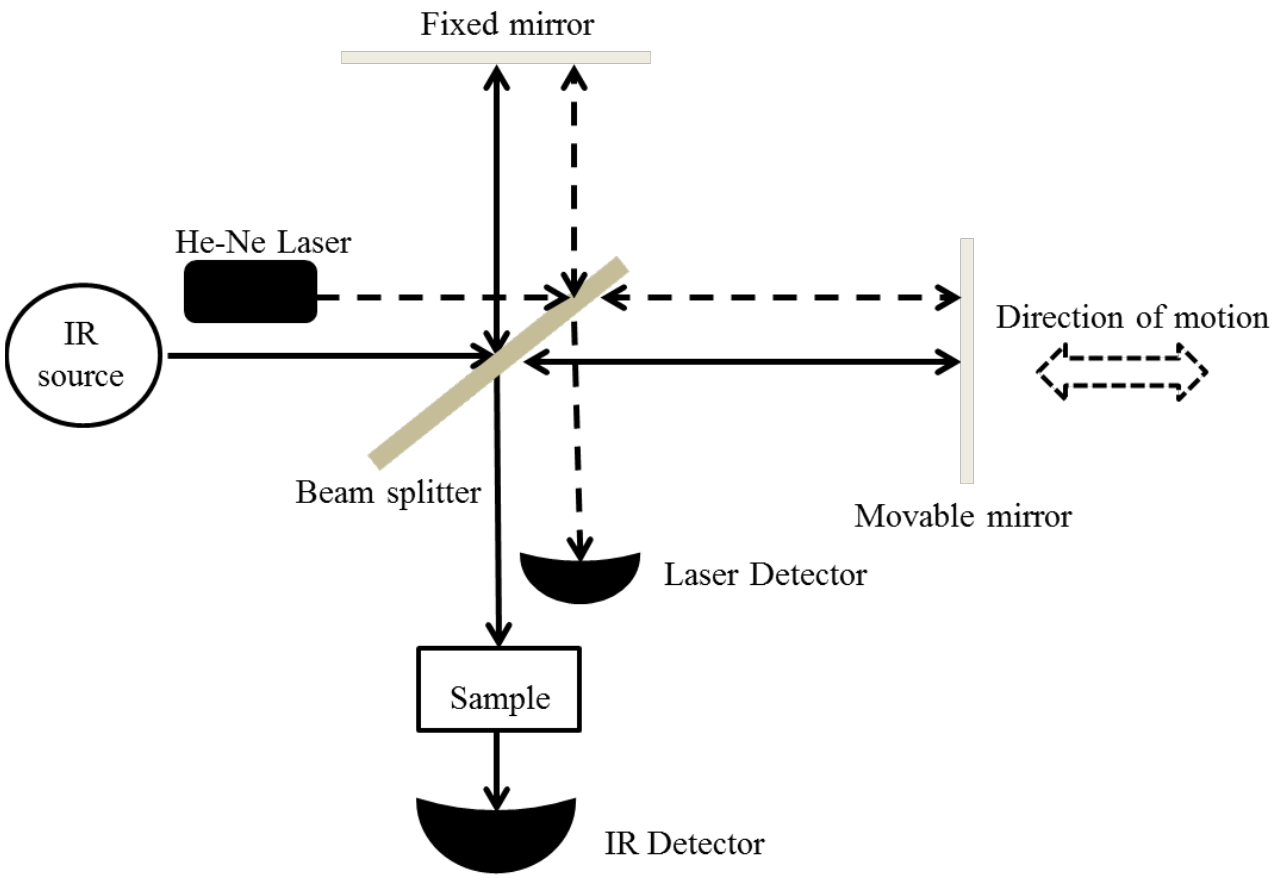


Figure 2.1 Schematic diagram of a FT spectrometer based on a Michelson interferometer. Light from the IR source is divided into two beams by a beamsplitter. The reflected beam travels to and is reflected back by the fixed mirror, while the transmitted beam travels to is reflected back by the movable mirror. When they recombine at the beamsplitter, the optical path difference between the two beams generates an interference pattern that propagates through the sample and is finally captured by the detector. The He-Ne reference laser is also directed through the interferometer, and its interference pattern is used as a clock signal to allow precise timing of the data acquisition.

beams occurs due to the difference between the distances travelled by the photons. This path difference is termed the retardation, δ. The recombined beam then reflects/transmits at thebeamsplitter, with half the beam directed to the detector. In a conventional laboratory transmission or reflectance measurement, the sample is placed between the interferometer and detector. For an emission experiment, the infrared emission enters the interferometer in place of the conventional light source.

For a monochromatic source with a wavelength of λ (or a wavenumber of $\bar{\upsilon} = \frac{1}{\lambda}$), the intensity of the beams arriving at the detector is composed of a constant (DC) component and a modulated (AC) component, shown as the first and second terms, respectively, in Eq. 2-1:

$$I'(\delta) = B(\bar{\upsilon}) + B(\bar{\upsilon})\, cos\, 2\pi\bar{\upsilon}\delta \qquad (2\text{-}1)$$

In the equation, $I'(\delta)$ denotes the overall detected beam intensity, which is retardation-dependent, and $B(\bar{\upsilon})$ is a wavenumber-dependent instrumental response term that combines the source intensity, detector sensitivity, and beamsplitter efficiency. As shown in the equation, the modulated component, usually termed the interferogram, $I(\delta)$, is recorded as a cosine function of the retardation.

The interferogram can be considered as a function of time as well. Essentially, it derives from the scanning of the movable mirror at a constant velocity, V, that introduces the optical path difference. The relation between retardation (cm), time (s) and mirror scanning velocity ($cm \cdot s^{-1}$) is given by

$$\delta = 2Vt \qquad (2\text{-}2)$$

Given that a cosine wave propagating over time takes the form of $\cos(2\pi ft)$, the oscillation frequency of the interferogram cosine wave, f, in units of s^{-1} or Hertz (Hz) can be obtained by substituting Eq. 2-2 into Eq. 2-1 and recognizing that

$$f = 2V\bar{\upsilon} \qquad (2\text{-}3)$$

From Eq. 2-3, it is clear that through modulation by the interferometer, the propagating high-frequency ($\bar{\upsilon}$) light can be carried by a cosine wave signal at a much lower frequency (f), which meets the limitations of detectors and electronics normally used in the interferometer. For example, to properly record IR signals at 4000 cm^{-1} without any light modulation, detectors and electronics would have to respond to the intrinsic optical frequency of 1.2×10^{14} Hz. This is beyond the capacities of current instrumentation. However, by use of an interferometer with a mirror velocity of 1 $cm\cdot s^{-1}$, an interferogram is generated with a frequency of 8.0×10^{3} Hz. Light oscillating in intensity at this frequency can easily be detected with current detectors and associated electronics.

When a broadband source is used, the interferogram can be represented by the integral of the individual interference patterns occurring at every light frequency, $d\bar{\upsilon}$

$$I(\delta) = \int_{-\infty}^{+\infty} B(\bar{\upsilon})cos(2\pi\bar{\upsilon}\delta)\cdot d\bar{\upsilon} \qquad (2\text{-}4)$$

This equation explains the attenuated wave pattern usually observed in the interferogram. At zero path difference (ZPD, $\delta = 0$), the cosine of zero is one, and the spectral intensities, $B(\bar{\upsilon})$, simply integrate across the entire wavenumber range. As a result, the intensity of the interferogram, $I(\delta)$, reaches its maximum. This interferogram location is known as the centerburst. When away from the centerburst (i.e., $\delta > 0$ or $\delta < 0$), the intensity of the

inteferogram damps as the individual cosine waves cancel. The broader the spectral bandwidth (i.e., the more $\bar{\upsilon}$ present), the faster the interferogram damps.

Equation 2-4 also represents one-half of a cosine FT pair, where the other half is

$$B(\bar{\upsilon}) = \int_{-\infty}^{+\infty} I(\delta)\, cos(2\pi\bar{\upsilon}\delta) \cdot d\delta \qquad (2\text{-}5)$$

This equation represents the corresponding single-beam spectrum in the frequency domain ($B(\bar{\upsilon})$) for a given interferogram. Since $I(\delta)$ is an even function about the ZPD, Eq. 2-5 can be rewritten as

$$B(\bar{\upsilon}) = 2\int_{0}^{+\infty} I(\delta)\, cos(2\pi\bar{\upsilon}\delta) \cdot d\delta \qquad (2\text{-}6)$$

Given modern computers and efficient computational algorithms, the FT, as represented by Eqs. 2-4 to 2-6, can be easily implemented to obtain spectral information from the collected interferogram data (Eq. 2-5) or to compute the interferogram corresponding to a spectrum (Eq. 2-4).

Absorption or emission peaks at characteristic frequencies corresponding to certain energy transitions can be used for structural elucidation, and therefore traditional analysis of FT-IR data focuses on transforming the collected interferogram data to the frequency domain. Even though the information pertaining to all spectral frequencies is present within the interferogram, the format of the data does not lend itself to visual interpretation. For automated (i.e., computer-based) analysis, however, direct use of the interferogram data has several advantages.

If a single-beam spectrum is considered to be the superposition of narrow analyte-specific features onto a broad IR background, the natural damping of the interferogram signal provides a simple way to separate background and analyte features by choosing an interferogram segment

displaced from the centerburst. The broad background feature damps quickly because of its large spectral bandwidth. Past a certain point in the interferogram, the background information has damped to zero and the remaining information pertains to narrower features such as those arising from chemical species present. Because these spectral features are narrow, their interferogram representations damp slowly, and thus their signals persist after the background information has damped completely. Frequency selectivity can be applied to a short interferogram segment by use of digital filtering rather than the FT. This allows an automated analysis to be implemented rapidly and requires only a short interferogram to be acquired.

Interferogram-based analysis can find potential applications in areas such as remote sensing, where the ability to extract specific analyte information from a constantly changing background using only short ranges of interferogram retardation is desirable. Chapters 4 and 5 describes the adoption of this approach to develop automated near real-time air monitoring methodology.

2.1.1 FT-IR related data processing [43] [26]

According to Eq. 2-6, a spectrum is an integration of the interferogram signal over infinite retardation. However, in reality, it is not feasible to obtain infinite retardation by scanning the movable mirror forever. In the acquisition of the digitized signal, a finite maximum retardation is applied to the FT of the true infinite waveform. This is mathematically equivalent to multiplying the infinite interferogram by a sampling function. This sampling function is called a boxcar truncation function, which is equal to "1" for the points within the sampling range and "0" elsewhere. The resulting spectrum is the convolution of the FT of the infinite interferogram with the FT of the boxcar truncation function (i.e., the sinc function, $\left(2\,\Delta\,\frac{\sin 2\pi\bar{\upsilon}\Delta}{2\pi\bar{\upsilon}\Delta}\right)$, where Δ is the finite sampling range).

The adoption of finite maximum retardation affects the spectrometer's resolution, which is defined as a measure of the capacity to distinguish two lines or bands exhibiting very close spectral frequencies. In the case of an infinite interferogram, a single sharp spectral line is observed for an individual frequency. In order to resolve two closely adjacent lines, $\bar{\upsilon}_1$ and $\bar{\upsilon}_2$, separated by a resolution element, $\Delta\bar{\upsilon} = \bar{\upsilon}_1 - \bar{\upsilon}_2$, the retardation has to increase so that two individual waves can complete a full cycle of phase change (i.e. in phase, out of phase, and back in phase). This, in other words, corresponds to

$$\boldsymbol{\Delta}\bar{\upsilon} = \bar{\upsilon}_{\mathbf{1}} - \bar{\upsilon}_{\mathbf{2}} = \frac{\mathbf{1}}{\boldsymbol{\delta}} \qquad \textit{(2-7)}$$

This indicates that to achieve finer resolution (i.e., smaller $\Delta\bar{\upsilon}$), greater retardation is needed. In the case of a finite interferogram, introduction of a boxcar function gives rise to a spectral band centered at $\bar{\upsilon}_1$ with the characteristic shape of a sinc function. In this context, if the full-width at half-height (FWHH) criteria is adopted, the spectral resolution becomes $\Delta\bar{\upsilon} = \frac{1.207}{\delta}$, where $\Delta\bar{\upsilon}$ is the spectral width at half height of the main lobe band (i.e., the main spectral peak).

Another issue associated with the sinc function is the presence of side lobes radiating outward from the central spectral peak. A sampling function other than the boxcar function can be chosen in order to suppress the magnitude of these oscillating side lobes. Such a process is called windowing or apodization of the interferogram. Instead of an unweighted sampling function like the boxcar function, another sampling function can be implemented by simply multiplying the collected interferogram point-by-point by a mathematically generated function that truncates the interferogram intensity in a smooth manner toward the edges. By eliminating sharp discontinuities at the edges of the interferogram, the effective sampling function can be changed and the side lobe artifacts can be suppressed.

A common apodization function, termed the triangular function, can be generated as:

$$A(\delta) = 1 - \left|\frac{\delta}{\Delta}\right| \quad for\ \delta \leq |\Delta|$$

$$A(\delta) = 0 \quad for\ \delta \geq |\Delta| \qquad (2\text{-}8)$$

The FT of $A(\delta)$ takes the form of $\Delta \frac{\sin^2(\pi\tilde{\upsilon}\Delta)}{(\pi\tilde{\upsilon}\Delta)^2}$, called the $sinc^2$ function. By introduction of weighting coefficients that are less than unity, triangular apodization can reduce the abrupt transitions at the sampling edge, and thereby reduce the amplitudes of the induced side lobes.

According to the Nyquist sampling theorem, to collect a continuous signal without any loss of information, the sampling frequency should be at least twice the highest frequency contained in the signal. In other words, for a cosine wave signal, the minimum sampling rate of two points per cycle is needed to preserve the signal frequency.

As one of the most important advantages of FT spectrometry, high reproducibility of the wavenumber axis from scan to scan is realized by synchronizing the position of the moving mirror with a clock signal. The clock signal is provided by a reference helium-neon (He-Ne) laser with a wavenumber of 15,802.8 cm^{-1}, which is directed through the interferometer. The modulated laser signal is detected with a photodiode and used as the timing control for the sampling frequency of the infrared detector. Circuitry is built to detect the zero-crossings of the cosine wave signal of the reference laser, and the sampling rate of the IR interferogram is specified as a point for every *n* zero-crossings of the reference signal. For a reference interferogram generated by a He-Ne laser, the spectral bandwidth is 15,802.8/*n* cm^{-1}.

The value of *n* is a user-specified parameter and its selection depends on the source bandwidth, spectral response of the detector, and the transmission properties of the beamsplitter,

For near-IR measurements, n=1 is typically selected. For the passive IR remote sensing work described in Chapters 4 and 5, no light is typically present at > 2000 cm^{-1}, and a value of n=8 is often used. This produces a maximum digitized spectral frequency of 1975 cm^{-1}. For a given resolution, $\Delta\bar{\upsilon}$, and maximum sampled frequency, $\bar{\upsilon}_{max}$, the number of sampled interferogram points, N_s, is given by

$$N_S = \frac{2\bar{\upsilon}_{max}}{\Delta\bar{\upsilon}} \qquad (2\text{-}9)$$

In theory, the interferogram should be a symmetric cosine function about the centerburst, but, in reality, existence of factors such as imprecision in retardation and frequency-dependent phase lag introduced by the electronics can cause phase shifts in one or more frequencies. This has the effect of making the phase angle, $2\pi\bar{\upsilon}\delta$, of the sampled interferogram non-zero. As a result, sine components are effectively added to the theoretical cosine waveform of the interferogram. Phase correction is a process of removing these phase shifts and restoring the symmetry of the interferogram.

Phase correction can be implemented either in the interferogram domain using the Forman method or in the spectral domain using the Mertz approach. [26]The Forman method was used as part of the data processing in Chapters 4 and 5, while the Mertz method was employed with the near-IR data discussed in Chapter 7.

Both methods employ a short double-sided interferogram segment around the centerburst and use this segment to estimate the phase angle at each spectral frequency. The short interferogram is extended to the full length of the original interferogram by zero-filling. Following a triangular apodization, a complex FT is applied to the extended interferogram. The resulting complex spectrum, $B(\bar{\upsilon})$, takes the form of

$$B(\bar{\upsilon}) = Re(\bar{\upsilon})\cos\theta(\bar{\upsilon}) + Im(\bar{\upsilon})\sin\theta(\bar{\upsilon}) \qquad (2\text{-}10)$$

where, at every wavenumber, $\bar{\upsilon}$, the phase angle, θ, can be calculated as the arctangent of the ratio between the imaginary spectrum, ($Im(\bar{\upsilon})$), and the real spectrum, ($Re(\bar{\upsilon})$).

In the Forman method, the computed phase spectrum is returned to the interferogram domain by use of the inverse FT and the resulting phase interferogram is convolved with the original interferogram. The convolution result is the phase-corrected interferogram. The FT can be applied to this corrected interferogram and only the cosine term needs to be used in computing the spectral intensity. In the Mertz approach, the FT is applied to the original interferogram and the computed phase angle at each frequency is used in the calculation of the spectral intensity at the corresponding frequency.

2.2 Near-IR transmission measurements [25]

The near-IR region spans the spectral range from 4000 to 12,800 cm^{-1}. Molecular vibrational transitions occurring in this region are mainly combination or overtone bands of the fundamental vibrations of C-H, N-H, and O-H bonds. The overtone or combination bands in the near-IR region have higher frequencies than the fundamental bands normally found in the mid-IR region, and also have much weaker signals (e.g., 10 -1000 times weaker). The weakness of the signals allows much longer optical pathlengths to be used (typically 1 -10 mm), and thereby makes less or no sample preparation possible as samples can often be measured directly without the need for dilution. Thanks to this major benefit, near-IR spectroscopy has been widely used as a quantitative analysis method and has been applied in a variety of fields, such as agriculture, food manufacture, pharmaceutical production, and environmental monitoring.

Near-infrared spectroscopy is also a very versatile technique. It can use samples with different physical phases: solid, liquid, or gas, or can be implemented under different governing

optical mechanisms such as transmission or diffuse reflectance. Chapter 7 in this dissertation describes an account of work based on near-IR transmission spectroscopy of aqueous samples.

The principle behind near-IR quantitative analysis based on transmission measurements is the Beer-Lambert law, which states the linear relationship between a measured quantity, absorbance, and the sample concentration as follows:

$$A(\bar{\upsilon}) = -log_{10} T(\bar{\upsilon}) = -log_{10} \frac{I(\bar{\upsilon})}{I_o(\bar{\upsilon})} = \varepsilon(\bar{\upsilon}) bc \qquad (2\text{-}11)$$

In Eq. 2-11, A, T, I_o, I, and ε represent absorbance, transmittance, the intensity of light incident on the sample, the intensity of light transmitted by the sample, and the molar absorptivity, respectively. These terms are all wavenumber dependent. Absorbance is linearly proportional to the product of molar absorptivity, the optical pathlength (b), and the concentration of analyte within the optical path (c). In general, A and T are dimensionless, ε is given in units of $M^{-1}cm^{-1}$, b is in cm, and c is in M. Deviations from this linear relationship may occur at high concentrations or extremely long pathlengths.

Typically, light sources used in near-IR measurements are broadband blackbody sources. The source materials are electrically heated to the temperature range of 1500 to 2200K. A continuum of radiation, approximately blackbody emission, is produced. The most commonly used near-IR light source is the tungsten-halogen lamp.

Depending on the spectral range to be measured and the desired sensitivity, several types of detectors are available in the near-IR region, such as lead sulfide (PbS), indium antimonide (InSb), and indium gallium arsenide (InGaAs). In Chapter 7, the near-IR source used was a tungsten-halogen lamp and the detector employed was a liquid nitrogen cooled InSb detector.

2.3 Passive FT-IR remote sensing [34,45,46]

Without physical contact, remote sensing is able to detect or classify target objects on the ground or target analytes in the open air. A common implementation of remote sensing is based on the use of optical measurements to probe a conical volume of the atmosphere. The sensor can view this atmospheric volume in either a horizontal or vertical geometry. For the work described in Chapters 4 and 5, a passive FT-IR remote sensor mounted in a downward-looking mode aboard an aircraft was used to probe the atmosphere below the aircraft.

Infrared spectroscopy has gained significant attention for use in remote sensing measurements because of the general IR transparency of the atmosphere. The two major atmospheric constituents, oxygen and nitrogen, are not IR active and thus provide no interference. Even though they are relatively minor constituents, water vapor and carbon dioxide are both strong IR absorbers. However, both exhibit no interference in the important fingerprint region of the IR between 8 and 12 μm (833 – 1250 cm^{-1}).

The measurement setup for passive IR sensing is illustrated in Figure 2.2. A downward-looking emission FT-IR spectrometer is mounted on a flying aircraft. The spectrometer carries no internal source, receiving instead the upwelling IR radiance emanating from the ground. The entrance optics employ a telescope to restrict the field-of-view (FOV) to approximately 0.5°. This minimizes the atmospheric volume sampled with a single interferogram scan. Within the FOV of the sensor, the incident radiance, excluding the part from the constant IR self-emission of the spectrometer itself, can be simply attributed to three perpendicular bottom-up layers of

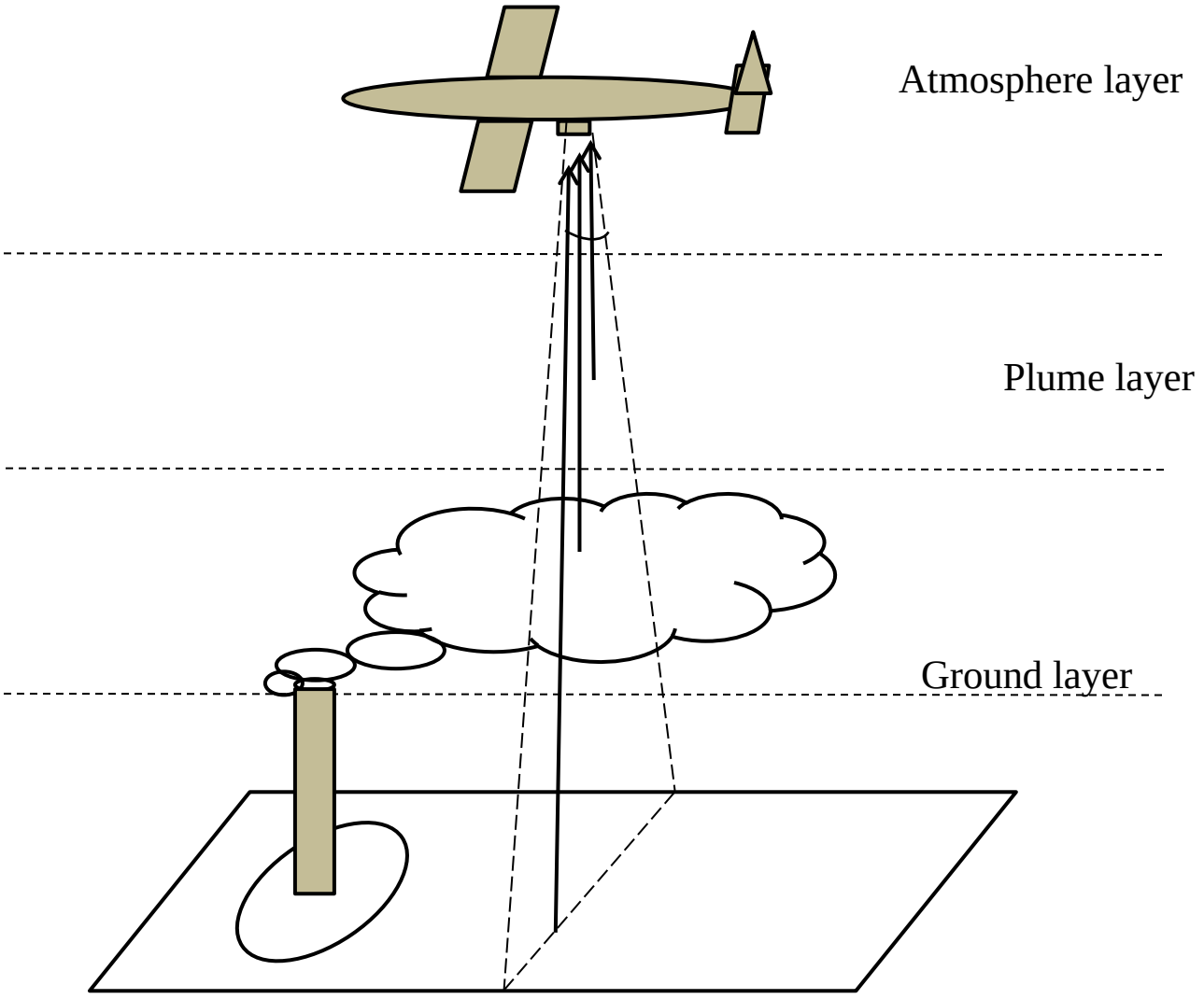


Figure 2.2 Schematic diagram of airborne remote sensing of stack emissions using passive FT-IR spectrometry. The radiance arriving at the downward-looking spectrometer mounted on a cruising aircraft can be considered from three sources: the ground layer, the plume layer and the atmosphere layer. Solid arrows represent the radiance from each layer.

space: the ground layer, the plume layer, and the atmosphere layer. In this model, the plume contains an analyte species whose detection is desired. A static air model is used to describe the composition of the radiance as shown in Eq. 2-12:

$$L_{obs} = \varepsilon_a B_a + \tau_a \varepsilon_p B_p + \tau_a \tau_p B_g \qquad (2\text{-}12)$$

In the equation, the overall radiance at a given wavelength arriving at the spectrometer, L_{obs}, can be seen as the sum of the radiance emanating from three layers. The first term on the right of the equation stands for the contribution from the atmospheric layer, which can be approximated as the radiance from a blackbody at the temperature of the atmosphere layer, B_a, modified according to its emittance, ε_a. The second term is from the plume layer, where the original plume radiance, represented by the blackbody radiance at the same temperature, B_p, is emitted at an emittance of ε_p and then is transmitted through the upper atmosphere layer with a transmittance of τ_a before entering the spectrometer. The third term, the contribution from the ground layer source, can be considered as the portion of ground radiance from the blackbody radiance at the temperature of the ground layer, B_g, which transmits up through both the plume layer and the atmosphere layer with the transmittance of τ_p and τ_a, respectively.

Assuming there is no reflection, Kirchhoff's law applies here, which states that the emittance of a material, ε_a, is equal to its absorption, and the sum of absorption and transmittance is always equal to one. Accordingly, Eq. 2-12 can be rewritten as below,

$$L_{obs} = (1 - \tau_a) B_a + \tau_a (1 - \tau_p) B_p + \tau_a \tau_p B_g \qquad (2\text{-}13)$$

Furthermore, assuming the temperature difference between the plume layer and the atmosphere layer is negligible, the blackbody radiance levels representing the two layers are considered equal, $B_a = B_p$, which results in

$$L_{obs} = (1 - \tau_a \tau_p) B_p + \tau_a \tau_p B_g \quad (2\text{-}14)$$

The temperature differential between the ground layer and the plume layer plays a significant role in determining the degree to which analyte information is present in the observed radiance. According to Eq. 2-14, in the case that the temperatures of the ground and the plume layer are the same, i.e., $B_p = B_g$, it follows that $L_{obs} = B_p$, and the observed radiance becomes no different than the radiance profile from the plume layer. As a result, the spectral features of the target analyte distributed in the plume layer will not be observed.

By contrast, if a significant temperature difference exists between the ground and the plume layer, the blackbody radiance levels generated from the two layers are different, $B_g \neq B_p$, and thus, according to Eq. 2-14, the spectral signature encoded in the transmittance of the plume layer, τ_p, can be captured in the radiance reaching the spectrometer. Therefore, to observe spectral features of the target vapor present in the plume, a sufficiently large temperature difference between the ground and the plume layers is required.

In the case that the temperature of the ground layer is higher than the temperature of the plume, $B_g > B_p$, it results in a local minimum in radiance registered in the spectrometer. The observed radiance will appear with an absorption peak in the frequency range characteristic of the target analyte. On the other hand, if the temperature of the ground layer is lower than the temperature of the plume, $B_g < B_p$, the observed radiance will appear with an emission peak in the frequency range characteristic of the target analyte.

2.4 Airborne gamma-ray remote sensing [27,47,48]

When an excited radioisotope nuclide (generally, a by-product of alpha or beta radiation) relaxes to a more stable nuclear state, the surplus energy is radiated as a high-energy photon, termed a gamma-ray. The radioactive decay law is applicable to the gamma-ray as it is to other

radiation types. The rate at which the decay takes place is proportional to the number of nuclei present and a decay constant characteristic of a given radionuclide. As a measure of the radioactivity, the half-life, the time required for a given amount of radioactive material to reduce to one-half, is unique for each radioisotope. Sometimes radioactive decay occurs in a series resulting in a sequence of daughter products.

Like all forms of electromagnetic radiation, gamma-rays have no mass or charge, and lose energy slowly. Therefore, they are able to travel significant distances. Depending on the original energy, they can travel ten to hundreds of meters in air. A relatively long penetration distance makes airborne gamma-ray spectrometry a feasible method for aerial survey of radioisotopes existing on the ground or in the air. Airborne gamma radiation comes from a variety of sources, including naturally occurring sources such as ^{40}K, ^{238}U and ^{232}Th, the daughter products from decay of parent isotopes, such as Rn, background from the aircraft itself, cosmic background, and fallout products from man-made atomic explosions and nuclear accidents, such as ^{137}Cs. The observed radiation intensity can be expressed as follows:

$$I_{obs} = aI_K + bI_U + cI_{Th} + dI_{cos} + eI_{rn} + I_{air} + fI_{ground} \qquad (2\text{-}15)$$

In Eq. 2-15, I_K, I_U, and I_{Th}, represent gamma radiance generated by pure sources with unit concentrations of K, U, and Th, respectively. The terms, I_{cos}, I_{rn}, I_{air} and I_{ground} are radiance due to cosmic background, radon radiation, the aircraft itself, and ground-based nuclear reaction fallout. The parameters, *a, b, c, d, e,* and *f* are calibration coefficients for each individual source.

A commonly used gamma radiation detector, thallium activated sodium iodide scintillation crystals (NaI[Tl]) was used in this work. The detector was mounted on the same aircraft as mentioned in the IR-based remote sensing work. Incident gamma-ray photons within the detector produce fast electrons, whose energies are converted into photons of visible light.

The sum of all photons of visible light is proportional to the characteristic energy of the incident gamma-ray.

Besides the physical properties of the radioactive sources, there are many factors that need to be considered in airborne gamma-ray spectrometry, such as detector response, source-detector geometry, and environmental effects. In order to quantify a target ground-based gamma source, I_{ground} in Eq. 2-15, several calibration and data processing steps are needed. First, the measured multichannel spectra are corrected with constant instrument related factors, I_{air}, such as spectrometer's dead time and energy drift. Second, the rest of the radiation contribution, except the terrestrial radiation, I_{ground}, is corrected by subtracting spectra collected over a water area at the survey altitude, where the terrestrial radiation contribution, I_{ground}, is assumed to be negligible. Next, a channel interaction correction, or stripping, is conducted on multiple photopeaks to remove counts due to other radioisotopes. In this step, calibration coefficients for the individual radiation contributions, *a* through *e*, are determined. Finally, the background-corrected and stripped count rates are corrected for variation in the height of the detector and reduced to elemental concentrations on the ground. In Chapter 6, based on a set of acquired airborne gamma-ray spectra, a pattern recognition method is employed to establish an automated identification protocol for the target ground-based radioisotope, ^{137}Cs.

Chapter 3

SIGNAL PROCESSING, PATTERN RECOGNITION, AND DATA ANALYSIS

In this chapter, a general overview will be provided for methods used for signal processing, multivariate regression, and pattern recognition in this dissertation. Signal processing methods, primarily digital filtering, combined with signal segment selection, were applied to the FTIR, FT-NIR, and gamma-ray data, either in the interferogram or in the spectral domain. For the applications discussed in Chapters 4, 5, 6 and 7 focused on qualitative data analysis, following the signal processing step, pattern recognition methods, such as principal component analysis (PCA), K-nearest neighbor (KNN) analysis, piecewise linear discriminant analysis (PLDA), and support vector machines (SVM) were used to develop binary classifiers for the detection of target analytes. For the quantitative data analysis work in Chapter 7, multivariate regression methods, such as partial least-squares (PLS) and support vector regression (SVR) were applied to develop either calibration models or residual models for glucose information from FT-NIR spectra.

3.1 Signal processing

Within any form of data acquired by modern instrumental measurement techniques, besides the expected signal from the target analyte, there exists a variety of sources of interference, including unwanted signals contributed by non-analyte constituents within the sample matrix, noise arising from the uncertainty of the measurement, drift of the instrumental response caused by changes in environmental conditions, or artifacts introduced by data manipulation. In order to establish reliable qualitative or quantitative relationships between the target analyte and the measured data containing the characteristic signature of the analyte, it is necessary to preprocess the raw data in such a way that, with minimal change made to the signal of the analyte within in the data, the undesirable interferences in the data are suppressed.

Signal processing methods such as normalization and the calculation of derivatives are frequently used to correct drifts of response intensity and remove broad background features, but both of these approaches are somewhat inflexible (i.e., they lack a significant tuning capability to allow an optimal implementation). On the other hand, through modulation of the data with basis functions such as sin and cosine waveforms, methods such as digital filtering can target the specific frequency of a sine or cosine waveform associated with the analyte of interest, while suppressing frequencies associated with noise and unwanted components in the data.

Digital filtering can be applied to either true time-domain data, e.g., the interferogram acquired from an FT-IR measurement, or frequency-domain data, e.g., the spectrum obtained from application of the Fourier transform to the interferogram data.[49-52] In both cases, the mathematical operations treat the data as signal intensities equally sampled along a time axis. In this dissertation, digital filtering was applied to inteferogram data from passive FT-IR experiments and gamma-ray spectral data.

Designed digital filters can be classified into four categories depending on how the signal frequencies are selected: lowpass, highpass, bandpass and bandstop. [53] The raw data can be considered a combination of a series of sine and cosine waveforms with varying frequencies. Lower waveform frequencies represent data components that change slowly. Broad spectral features caused by baseline variation, and instrumentation drift are dominated in this region. High waveform frequencies model spectral components that change rapidly, e.g., fast-varying random noise. As illustrated by the filter names, a lowpass filter which allows low frequencies to pass while attenuating higher frequencies can be used to suppress random noise. Analogously, a highpass filter can be used to remove broad features such as baseline variation while retaining spectral components with higher frequencies. A bandpass filter is able to retain spectral

components of a certain frequency range while eliminating features outside this range, including both baseline variation and spectral noise. A bandstop filter, by contrast, is used to attenuate spectral components of a certain frequency range while passing features with frequencies outside this range. In this dissertation, bandpass filters were primarily used to extract a specific frequency range associated with the target analyte while suppressing noise and unwanted frequencies.

In the frequency domain, each digital filter can be characterized by the frequency response, in which a measure of the signal attenuation achieved by the filter as a function of frequency is plotted. The frequency axis is typically normalized to a scale of 0.0 to 1.0, where 1.0 corresponds to the maximum harmonic frequency of the data as determined by the sampling rate.

However, from the perspective of the Fourier transform, the frequency response of the digital filter cannot be applied directly to the raw data which is a quantized signal containing samples collected as a function of time. A representation of the frequency response in the time-domain, termed the impulse response of the filter, must be obtained before being applied to the time sequence data. The output of the convolution of the impulse response with the measured data corresponds to the filtered data in the time-domain, e.g., a filtered interferogram from an FT-IR measurement, or a filtered gamma-ray spectrum from a gamma-ray measurement. The FT of this convolution result generates output corresponding to filtered data in the time-domain, e.g., a filtered spectrum from IR spectrometry.

Depending on how the convolution is estimated, digital filters are classified into two categories: finite impulse response (FIR) and infinite impulse response (IIR). [53,54] The output of an FIR filter takes the sum of the convolution of the current input data points and the filter's

impulse response, while an IIR filter includes both current data points and the filtered output of past data points when computing the filtered output of the current point. For a given length of the impulse response, IIR filters will tend to achieve higher attenuation of unwanted frequencies and sharper transitions between the passband and stopbands of the frequency response. Only IIR filters were designed and investigated in this dissertation.

The implementation of FIR or IIR filters on an array of measured data is expressed in Eq. 3-1 or Eq. 3-2, respectively:

$$y_n = \sum_{k=0}^{N} c_k \, u_{n-k} \qquad (3\text{-}1)$$

$$y_n = \sum_{k=0}^{N} c_k u_{n-k} \; + \; \sum_{k=1}^{M} d_k y_{n-k} \qquad (3\text{-}2)$$

In Eq. 3-1, within a finite range defined by *k* points, for each quantized data point *n*, c_k is the filter coefficient of a discrete point of the impulse response function of the filter, $u_{n\text{-}k}$ denotes the corresponding input data, and the sum of the convolution, c_k * $u_{n\text{-}k}$, corresponds to the filtered signal at data point *n*, denoted as y_n. In Eq. 3-2, the d_k comprise a second set of filter coefficients of the IIR filter, and the $y_{n\text{-}k}$ denote the past filtered data *k* points prior to the current input data point.

The number of filter coefficients, termed the filter order, is a measure of the complexity of the filter design. The higher the filter order, the greater the computational demands to approximate the target frequency response. For the FIR filter, as shown in Eq. 3-1, the filter order is N. The order of the IIR filter is the maximum of (N, M). As noted above, by including past filtered data points to calculate the current point being filtered, generally IIR filters can either achieve a similar filtering effect as FIR filters but with a significantly lower filter order or an improved filtering effect with the same filter order.

Depending on how the impulse response of a target frequency response is estimated, widely used IIR filters include four major design types: Butterworth, Chebyshev Type I, Chebyshev Type II and Elliptic. [55,56] The Chebyshev Type II filter design was used in this dissertation. This filter design has a fast transition between the stopband and the passband. The resulting frequency response features a flat passband and rippled stopband, which offers great flexibility in frequency selection. [57]

3.2 Pattern recognition

3.2.1 Support vector machines

A support vector machine (SVM) is a supervised kernel-based learning method that can be applied to classification, regression, or ranking tasks. The SVM operates in a feature space with higher dimensionality than the input data, and thus attempts to address problems which are unsolvable in the original space of the input data. [58-60] In this dissertation, binary-class SVM classification was used as the pattern recognition algorithm to implement automated detection of analytes in FT-IR and gamma-ray spectroscopy data.

In the case of binary classification, a set of representative patterns labeled with known classes are used to compose the training set that is employed to build an SVM learning model as expressed below:

$$D(m,n) = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \cdots, (\boldsymbol{x_m}, y_m)\} \qquad (3\text{-}3)$$

In Eq. 3-3, D represents the training set with m data points. The i^{th} data point, D_i, is represented by an n-dimensional vector, $\mathbf{x_i}$, termed a pattern, and its corresponding class label, y_i, with value of either 1 or -1, denoting the binary class to which the data point belongs. In the applications discussed here, the classification problem revolves around detecting the presence (class 1) or absence (class 2) of an analyte species to be detected.

The linear classification function implemented by the SVM can be expressed as

$$f(\boldsymbol{x}) = (\boldsymbol{w} \cdot \boldsymbol{x}) + b \qquad (3\text{-}4)$$

where **w** is a weight vector that defines the classification model, (·) denotes the dot product, and b is a linear constant. The values of **w** and b are optimized by use of the data in the training set. For an input pattern, **x**, the computed value of $f(\mathbf{x})$ is the SVM score that is used to predict the class membership of the pattern. If the linear function (as defined by **w** and b) satisfies that $f(\mathbf{x}) > 0$ for all data points with class label $y_i = +1$ and $f(\mathbf{x}) < 0$ for all data points with class label $y_i = -1$, the training set, D, is considered linearly separable. The place where $f(\mathbf{x}) = 0$ is termed the separation hyperplane and defines the boundary between the data classes.

A principal goal of the design of SVM classifiers is to maximize their generalizability, i.e., their ability to operate successfully outside of the set of training data used in the optimization of the classification model. The high generalization capacity of SVM classifiers can be attributed to the use of a "margin" in the definition of the boundary between the data classes. The margin can be conceptualized as the thickness of the classification boundary. By maximizing the margin during the training of the classifier, the resulting classification boundary tends to be more centered between the data classes and thus positioned more generally for use with future data whose class membership is to be predicted.

Mathematically, the margin is defined as the distance from the separation hyperplane to the closest data points from either class. A normalized margin for a binary classifier can be expressed as

$$margin = \frac{2}{\|\boldsymbol{w}\|} \qquad (3\text{-}5)$$

Maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|$, the vector magnitude of **w**. For computational convenience, the minimum of $\frac{1}{2}\|\mathbf{w}\|^2$ can also be sought. A training problem in developing an SVM thus becomes a constrained optimization problem as defined below:

$$minimize: \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$subject\ to: y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) \geq 1 \qquad (3\text{-}6)$$

When all patterns in the training set are correctly classified by the linear function, it is called a hard margin. However, a hard margin with perfect linear separation between binary classes is rare for non-trivial classification problems. Most common is a soft margin, where a certain number of errors in classifying the training data points are allowed while still maximizing the overall margin. A slack variable, ξ_i, is introduced to measure the degree of misclassification. The constrained optimization problem can be modified as

$$\text{minimize: } Q_1(\mathbf{w}, b, \xi_i) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$$

$$subject\ to: y_i(\langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b) \geq 1 - \xi_i, \quad \xi_i > 0 \qquad (3\text{-}7)$$

where *C* is a regularization parameter that determines the tradeoff between the margin size and the amount of misclassification allowed in training. The larger the value of *C*, the greater the emphasis placed on minimizing the training error. By contrast, a smaller value of *C* places more emphasis on maximizing the margin at the expense of correctly classifying the training patterns. The optimal value of *C* is data-set dependent and must typically be discovered through systematic investigation.

The objective function, Q_1, in Eq. 3-7 is a primal problem based on a convex function of **w**, with constraints that are linear in **w**. To solve this constrained optimization problem, Lagrange multipliers, $\boldsymbol{\alpha}$, are introduced to construct a dual problem, which has the same optimal value as the primal problem. The solution is determined by the saddle point of a Lagrange function, a function of **w**, b, and $\boldsymbol{\alpha}$. This function has to be minimized with respect to **w** and b, and maximized with respect to $\boldsymbol{\alpha}$. Through partial differentiation and substitution, the constrained optimization problem described in Eq. 3-7 can be rewritten as

$$maximize\, Q_2(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \boldsymbol{\alpha_i}\boldsymbol{\alpha_j}\, y_i y_j \boldsymbol{x_i}\boldsymbol{x_j}$$

$$subject\ to:\ \sum_i \boldsymbol{\alpha_i} y_i = 0, \quad 0 \le \alpha \le C \qquad (3\text{-}8)$$

The optimization depends on a set of dot products of paired patterns in the training set. In Eq. 3-8, the vector, $\boldsymbol{x_i}$, represents the i^{th} data point in the training set, and the vector, $\boldsymbol{x_j}$, represents the j^{th} data point. Once the optimal Lagrange multipliers, denoted as $\boldsymbol{\alpha^*}$, are determined, the optimal weight vector, $\mathbf{w^*}$, can be written as

$$\boldsymbol{w}^* = \sum_i \alpha_i{}^* y_i \boldsymbol{x_i} \qquad (3\text{-}9)$$

According to the Karush-Kuhn-Tucker condition of optimization theory, [61] the optimal Lagrange multipliers, $\boldsymbol{\alpha^*}$, must satisfy the following condition

$$\alpha_i{}^* \{y_i\, (\boldsymbol{w}^* \cdot \boldsymbol{x}_i + b) - 1 + \xi_i\} = 0 \qquad (3\text{-}10)$$

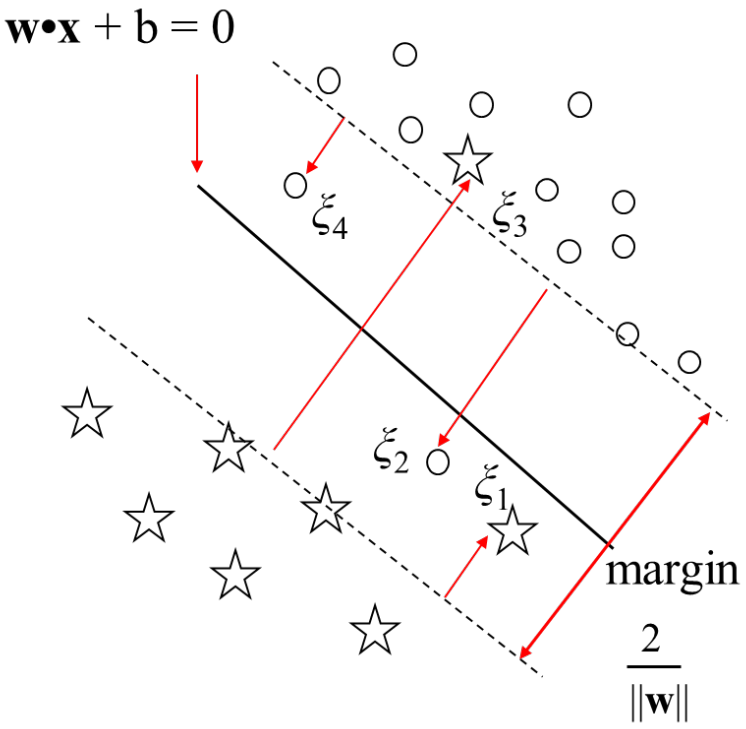


Figure 3.1 Schematic diagram of support vector machines. Data points from binary classes, denoted with stars or circles, respectively, are classified by a separation plane at $w \cdot x + b = 0$, with a maximum margin of $\frac{2}{\|w\|}$. Support vectors are those data points, circle or star, located at the boundary of either side of the margin. Slack variables, $\xi 1$ to $\xi 4$, are the distances from the incorrectly classified data points to the boundary of its correct class.

where either α_i^* or the constraint, $y_i\,(\mathbf{w}^* \cdot \mathbf{x_i} + b) - 1 + \xi_i$ must be nonzero. This condition implies that only when the constraint is equal to zero will its corresponding coefficient, α_i, be nonzero, or further, nonnegative if the constraint in Eq. 3-8 is taken into consideration. In other words, the data vector, $\mathbf{x_i}$, whose corresponding coefficient, α_i, is zero will have no influence to the optimal weight vector, **w***. Thus, the optimal weight vector, or the margin, will only depend on the support vectors whose coefficients are nonnegative. The concept of SVM classification is illustrated in Figure 3.1.

So far, what has been discussed pertains to linear separation. If the classes of the training data are not linearly separable, there is no hyperplane that can separate the binary classes as shown in Figure 3.1. To solve the nonlinear classification problem, the input vectors can be transformed nonlinearly into a higher dimensional feature space where a linear separation can be sought as described above. A linear discriminant function in the high-dimensional space is equivalent to a nonlinear function in the original space.

In the implementation of SVM, a kernel function is used to map vectors in the original space to the feature space with higher dimension. The utility of the kernel function is that the inner product in the feature space can be replaced with a kernel function of vectors in the original input space, demonstrated as

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_i}) \cdot \phi(\boldsymbol{x_j}) \qquad (3\text{-}11)$$

In Eq. 3.11, the output of the kernel function, *K*, corresponds to the dot product of pattern vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ that have been transformed by the nonlinear function, φ, into the higher dimensional feature space. Therefore, using the kernel function, the constrained optimization problem in the dual form, as shown in Eq. 3-8, can be rewritten as

$$maximize\, Q_3(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j\, y_i y_j \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j})$$

$$subject\ to: \quad \sum_i \alpha_i y_i = 0, \quad 0 \le \alpha \le C \qquad (3\text{-}12)$$

The types of commonly used kernel functions include polynomials, radial basis functions (RBFs), and sigmoid functions. In this dissertation, the RBF was used as shown below, where the kernel parameter, γ, and the regularization parameter mentioned earlier, *C*, are the two parameters required to configure the classifier

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = e^{-\gamma \|x_i - x_j\|^2} \qquad (3\text{-}13)$$

Typically, the sign of the SVM score (Eq. 3-4) is used to determine the class membership for an individual data point. In Chapters 4 and 5, however, a new classification rule is developed for data sets derived from a time-based data acquisition. The specific case addressed is one in which one class is dominating while the other is in the minority. Detection of members of the minority class is the focus of the analysis. For this situation, we introduce the use of a control chart on top of the SVM score profile of the time-series data. Use of the control chart helps to account for variation in the SVM scores with time and allows more confident identification of the minority class.[48]

The construction of a control chart involves calculation of the average score and the standard deviation associated with the range of scores under investigation. An upper control limit (UCL) and lower control limit (LCL) are normally calculated as 2-3 times the standard deviation about the average. According to the assumption mentioned above, within an SVM score profile, data points from the dominant class tend to remain within the range between the UCL and LCL. However, a data point from the minority class is typically recognized by a score exceeding the

UCL. Use of UCL and LCL values derived from the current data set under investigation allows a tuning of the classification threshold to help minimize both false and missed detections of the minority class.

3.2.2 Piecewise linear discriminant analysis

As an extension of the technique of linear discriminant analysis (LDA), piecewise linear discriminant analysis (PLDA) approximates a nonlinear separation among data objects (patterns) from binary classes. [62,63] The linear discriminant function for LDA takes the same form as the linear SVM classification (Eq. 3-4). The output of the linear discriminant function, $f(\boldsymbol{x})$, is termed the discriminant score. The weight vector, **w**, and the linear bias, b, are sought so that the discriminant hyperplane lies where $f(\boldsymbol{x}) = 0$. With this orientation, a discriminant score of $f(\boldsymbol{x}) > 0$ will identify that data object, **x**, belongs to one class, while $f(\boldsymbol{x}) < 0$ will specify that **x** belongs to the other class. As with the SVM classifier described previously, a training set of representative patterns from each data class is used in conjunction with an iterative optimization procedure to compute the discriminant functions.

To address the nonlinear classification, LDA is used as the basis to construct a piecewise linear discriminant, where multiple single-sided linear discriminants are computed in a stepwise manner. The term, single-sided, specifies that patterns from only one class lie on one side (the "pure" side) of the discriminant boundary. Taken together, the set of discriminant functions proscribes a piecewise linear approximation to a nonlinear classification boundary. As depicted in Figure 3.2 for a compound detection application, data from two classes, analyte-active patterns

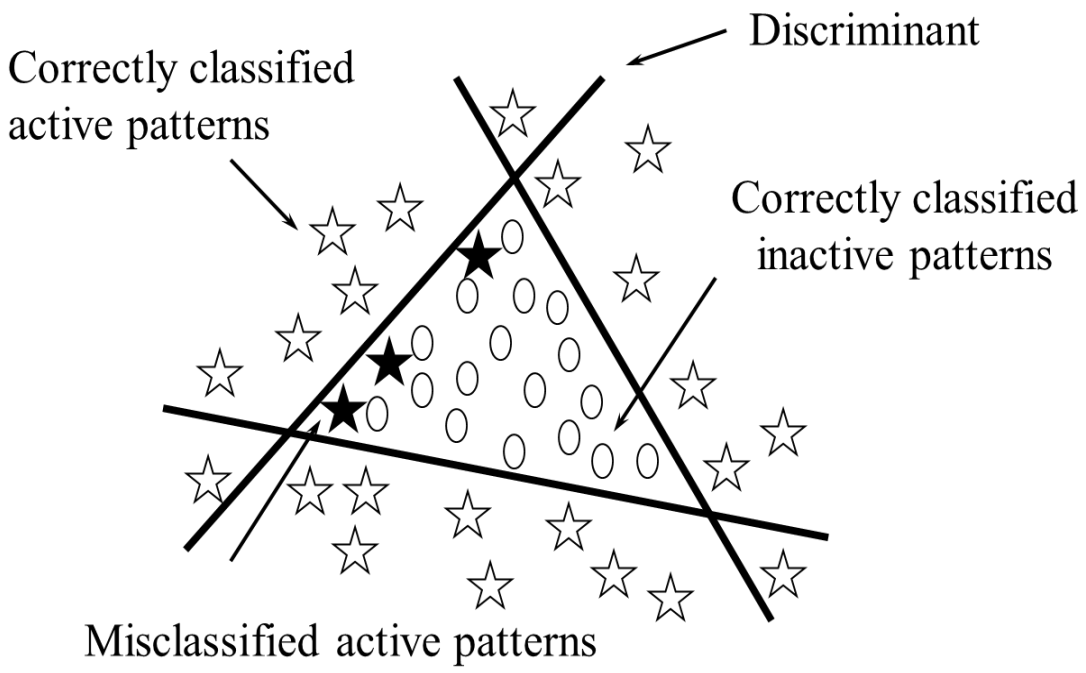


Figure 3.2 Schematic diagram of piecewise linear discriminant analysis. In PLDA, to separate analyte-active (star) and inactive patterns (oval), multiple single-sided linear discriminants (black solid line) are developed, which in combination form a nonlinear separating plane between the two classes. For each pattern, PLDA generates a discriminant score which is a signed distance from the patterns to the nearest boundary of the separating plane. A negative discriminant score for an analyte-active pattern places the data point (filled star) on the opposite side of a cluster of active patterns. This pattern is considered a misclassified active pattern or missed detection of the analyte.

denoted by star symbols and analyte-inactive patterns specified with ovals, are separated by a combination of three single-sided discriminant functions (black lines).

In the implementation of PLDA used in this work, a Bayes classification method is used to compute the initial discriminant function and the simplex optimization algorithm is subsequently employed to refine the positioning of the discriminant boundary.[64] Each discriminant is optimized to place the maximum number of analyte-active patterns from the training set on the pure side of the discriminant boundary. The remaining mixed group of inactive patterns and unseparated active patterns serve as inputs to the calculation of the next discriminant function. This stepwise calculation continues until no further separation can be found.

Once the set of discriminant functions is established, PLDA can be used to classify unknown patterns. For the analyte-active and analyte-inactive classification problem, a pattern is classified as analyte-active if it is located on the pure side of any of the computed discriminant functions. Otherwise, the pattern is judged to be analyte-inactive. In Figure 3.2, the filled stars illustrate cases in which active patterns are classified on the opposite side of the boundaries from active group. These would be examples of missed detections. In Chapter 4 of this dissertation, PLDA is used for model selection prior to SVM classification.

3.2.3 Principal component analysis

Principal component analysis (PCA) is a latent variable based method, where original correlated variables are replaced by orthogonal latent variables which are linear combinations of the original variables. A set of *n* responses, each consisting of *p* elements, can be reformulated as a function of latent variables as below

$$\boldsymbol{R} = \boldsymbol{T}\boldsymbol{V}^{T} + \boldsymbol{E} \qquad (3\text{-}14)$$

In Eq. 3-14, **R** is the $n \times p$ matrix of responses, **V** is a $p \times h$ matrix of loadings (latent variables, factors) with each column corresponding to an individual loading, and **T** is an $n \times h$ score matrix with each column corresponding to the projection of the responses onto one of the h loadings. The $n \times p$ product matrix, $\mathbf{TV}^T$, can be considered a reconstruction of the response matrix where only certain components of the original responses have been retained. The remaining information (i.e., not included in $\mathbf{TV}^T$) is contained in the $n \times p$ matrix of residuals, **E**.

Due to the orthogonality of the latent variables, there is an efficient separation of information derived from the original response matrix. Those latent variables that carry a large fraction of the information in the responses will tend to represent major features in the data, while those that carry smaller amounts of information will represent minor components or noise. For data derived from instrumental measurements, it is common that $h << p$, where h denotes the number of latent variables deemed to carry useful information. The PCA technique can thus be used as a means to compress data.

A general approach to implement PCA starts with subtracting the mean response from the rows of **R** in order to set aside a dominant portion of the total data variance represented by the mean. Next, the loadings, **V**, are computed from the mean-centered response matrix. With **V** and the mean-centered **R** substituted into Eq. 3-14, the scores, **T**, are calculated next. As a rearranged and more effective representation of the responses in the orthogonal latent variable space, the PCA scores can be used alone as patterns for pattern recognition, or be used as independent variables to develop quantitative models for some property variable such as concentration. When used for quantitative modeling, the technique is termed principal component regression (PCR). In Chapter 4 and 5 of this dissertation, scores from PCA were plotted to explore grouping across different classes of patterns.

In practice, the set of orthogonal loadings, **V**, can be computed through the application of eigenvector decomposition methods such as singular value decomposition (SVD) and nonlinear iterative partial last squares (NIPALS).[65,66] In this dissertation, SVD was applied to the mean-centered response, **R**, as below

$$\boldsymbol{R = UDV^T} \qquad (3\text{-}15)$$

Columns in the resulting orthogonal $p \times p$ matrix **V** are the eigenvectors of $\mathbf{R^TR}$, and the singular values in the first p (assuming $p < n$) entries on the main diagonal of matrix **D** correspond to the square root of the eigenvalues of $\mathbf{R^TR}$. Since the eigenvalues encode the sum of squares (i.e., squared magnitude) of the corresponding eigenvectors, and the ratio of each eigenvalue to the sum of the all eigenvalues represents the fraction of variance in **R** explained by the corresponding eigenvector, the most significant eigenvectors can be identified as those with the largest eigenvalues. These are termed principal components (PCs).

3.2.4 K-Nearest neighbor pattern recognition

The K-nearest neighbor (KNN) technique is a supervised pattern recognition method where an unknown object is classified by the majority vote from its k nearest neighbors with known class memberships.[67,68] The Euclidean distance, d_{ij}, is calculated for every two data points, $\mathbf{x}_i$ and $\mathbf{x}_j$, in n-dimensional space

$$d_{i,j} = \sqrt{\sum_{m=1}^{N} \left(x_{i,m} - x_{j,m}\right)^2} \qquad (3\text{-}16)$$

For each data point, the distances to all other data points are sorted from smallest to largest. The first k adjacent data points with the closest Euclidean distances to the unknown sample are chosen as the training samples. The class membership of the unknown sample is

determined by the majority vote from these training samples with known class memberships. For purpose of simplification, *k* is normally an odd integer. For example, if *k* = 5 nearest neighbors are chosen, if the Euclidean distance calculation results in 3 or more out of 5 of the neighbors belong to the class +1, then the sample is predicted to belong to class +1. If the predicted class and actual class match, the prediction is considered a success, otherwise, an incorrect classification results. The overall successful classification rate, calculated over the entire set of data, is a measure of the degree of clustering by data class across all the samples. In Chapter 5 of this dissertation, KNN was employed as a quick pattern recognition method to facilitate the selection of training sets prior to the implementation of more complex pattern recognition methods such as SVM.

3.3 Multivariate regression

3.3.1 Partial least-squares

Partial least-squares (PLS) is a bilinear latent variable based approach widely used to compute calibration models for quantitative analysis. Other than the latent variable based PCR method dealing with a single linear regression, PLS incorporates the covariance between two matrices into the computation of latent variables. In the case of quantitative analysis, a joint decomposition of the response matrix, **R** (*n*×*p*) and concentration vector **c** (*n*×*1*) is performed, aiming to maximize the covariance of **R** and **c** explained by the model.

$$\boldsymbol{R} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{T}} + \boldsymbol{E}$$

$$\boldsymbol{c} = \boldsymbol{T}\boldsymbol{q} + \boldsymbol{e} \qquad (3\text{-}17)$$

The first equation in Eq. 3-17 has the same form as shown previously for PCA in Eq. 3-14. This is a linear regression for **R** in terms of a set of *h* spectral loadings, **P**. The second

equation is the linear regression for **c**, where **q** and **e** in the second equation are concentration loadings and the residual concentration not explained by the model, respectively. Both linear models share the $n \times h$ score matrix, **T**, which is the key to the joint covariance maximization the occurs in PLS.

There are a variety of related PLS algorithms.[69] The PLS 1 algorithm is the one used in this dissertation. To develop a calibration model, first, the loading weights for the i^{th} latent variable, $\mathbf{w_i}$, are computed, as shown below, to identify the covariance between the columns of **R** and **c**.

$$\boldsymbol{w_i} = \frac{\boldsymbol{R^T c}}{\|\boldsymbol{R^T c}\|} \qquad (3\text{-}18)$$

Next, scores of the first latent variable, $\mathbf{T_1}$, are obtained by projecting the mean-centered **R** to the first loading weight, $\mathbf{w_1}$, as shown below:

$$\boldsymbol{T_1} = \boldsymbol{R w_1} \qquad (3\text{-}19)$$

Once $\mathbf{T_1}$ available and substituted into Eq. 3-17, the dual equations can be solved by least-squares to obtain the spectral loading, $\mathbf{p_1}$, concentration loading, $\mathbf{q_1}$, spectral residual, $\mathbf{E_1}$, and concentration residual, $\mathbf{e_1}$, which represent the first PLS factor. To obtain the next factor, **R** is replaced by $\mathbf{E_1}$ and **c** by $\mathbf{e_1}$, and steps from Eq. 3-17 to 3-19 are repeated.

To establish the calibration model, the most significant factors are selected as independent variables, and the vector of linear regression coefficients, **b**, is calculated as

$$\boldsymbol{b} = (\boldsymbol{T^T T})^{-1} \boldsymbol{T^T c} \qquad (3\text{-}20)$$

Parameters such as the response mean, loading weights, spectral loadings and regression coefficients help to construct the calibration model, which can subsequently be used to predict the concentration of an unknown sample outside the calibration set.

Through a similar recursive procedure to that used to establish the calibration model, the response matrix from a set of unknown samples, $\mathbf{R}_{pred}$, is first centered with the mean response from the calibration data and then used together with the previously computed **P** and **W** to compute a corresponding set of PLS scores, $\mathbf{T}_{pred}$. For example, the scores along the first PLS latent variable are computed as:

$$\boldsymbol{t}_{1,pred} = \boldsymbol{R}_{pred}\boldsymbol{w}_{\mathbf{1}} \qquad (3\text{-}21)$$

Then, the first spectral loading computed from the calibration data is used to remove the information from $\mathbf{R}_{pred}$ that was captured into $\mathbf{t}_{1,pred}$. The resulting residual matrix is then used with $\mathbf{w}_2$ to compute $\mathbf{t}_{2,pred}$ in a manner analogous to that shown in Eq. 3-21. The linear combination of all the scores with the previously calculated calibration regression coefficients yields the predicted concentrations for the unknown samples, $\hat{\mathbf{c}}$, as shown in Eq. 3-22 below.

$$\hat{\boldsymbol{c}} = \boldsymbol{T}_{\boldsymbol{pred}}\boldsymbol{b} \qquad (3\text{-}22)$$

The PLS method was used as one of the multivariate regression methods in Chapter 7 of this dissertation, along with a cross-validation (CV) procedure to facilitate the selection of the parameters associated with the PLS model. A series of PLS models derived from the same calibration set but with varying numbers of latent variables or different ranges of the input responses was investigated.

The CV procedure is based on withholding samples from the calibration data set, building a model with the remaining calibration samples, and then predicting the samples withheld. The calculation cycles until every calibration sample has been withheld once. Once the full cycle is complete, the cross-validation standard error of prediction (CV-SEP) can be calculated, as shown below, to evaluate the intrinsic performance of each calibration model.

$$CV - SEP = \sqrt{\frac{\sum_{i=1}^{m}(c_i - \hat{c}_i)^2}{m}} \qquad (3\text{-}23)$$

In Eq. 3-23, the value of CV-SEP is computed on the basis of m calibration samples, where c_i denotes the true concentration of sample i and $\hat{c}_i$ denotes the corresponding predicted concentration.

For the work described in this dissertation, the optimal PLS model was taken as the model based on the smallest number of latent variables whose CV-SEP value was not statistically different from the model that produced the overall lowest CV-SEP value.

3.3.2 Support vector regression

As an extension of SVM classification, support vector regression (SVR) can be applied to quantitative analysis by the introduction of a precision-insensitive loss function. [59, 61] In SVR, the training set can be presented with the same form as in SVM classification (Eq. 3-3), but the label, y_i, for each data point in the training set is a quantitative value associated with its pattern vector, $\mathbf{x}_i$, (e.g., the analyte concentration) instead of a class label.

The loss function, L(y), can be expressed as

$$L(\boldsymbol{y}) = \begin{cases} 0 & if\ |y - f(\boldsymbol{x})| < \varepsilon \\ |y - f(\boldsymbol{x})| - \varepsilon & otherwise \end{cases} \qquad (3\text{-}24)$$

where f(**x**), taking the form of Eq. 3-4, is a linear regression estimation of *y*, and ε is the precision associated with the difference between the estimated and real concentrations. Except for the initial setting mentioned above, the principles of resolving the constrained optimization problem in SVM classification, such as maximizing the margin, use of the kernel function, and the dual form of the optimization problem remain the same for SVR.

Similarly, the solution for a nonlinear SVR solution can be expressed in a manner analogous to Eq. 3-12:

$$maximize\, L(\alpha) = \sum_i(\alpha_i - \hat{\alpha_i}) - \varepsilon\sum_i(\alpha_i + \hat{\alpha_i}) - \frac{1}{2}\sum_i\sum_j(\alpha_i - \hat{\alpha_i})(\alpha_J - \hat{\alpha_j})\ \boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j})$$

$$subject\ to: \quad \sum_i(\alpha_i - \hat{\alpha_i}) = 0, \quad 0 \le \alpha \le C, 0 \le \hat{\alpha} \le C \qquad (3\text{-}255)$$

In Eq. 3-25, $\hat{\alpha}$ is an additional Lagrange multiplier introduced by the loss function. In Chapter 7 of this dissertation, as a presentation of nonlinear regression, SVR was employed to retrieve concentration information in comparison to or in combination with PLS regression.

3.4 Summary

This chapter provided a necessary background regarding the key signal processing, pattern recognition and multivariate regression methods used in the work of this dissertation. Their applications can be found in Chapters 4 through 7.

Chapter 4

ROBUST CLASSIFIER FOR THE AUTOMATED DETECTION OF AIRBORNE AMMONIA BY PASSIVE FOURIER TRANSFORM INFRARED SPECTROMETRY

4.1 Introduction

Passive Fourier transform infrared (FT-IR) spectrometry has proved to be a promising data collection technique for a variety of environmental monitoring applications. [70-75] Radiance from a naturally occurring IR source, across a long open-air path length, is acquired by an emission FT-IR spectrometer. This instrument can be placed in a stationary position or mounted on a moving platform. The instrumental field-of-view (FOV) is directed to an appropriate background target. For example, if the spectrometer is placed on a flying aircraft, the entrance optics are directed downward and the terrestrial scene below the aircraft serves as the IR background. This implementation serves as the basis for the work described in this chapter.

Depending on the temperature difference between the ground level and the air effluent layer, volatile organic compounds (VOCs) of interest, existing in the pathway between the light source and spectrometer will absorb or emit light at their characteristic vibrational frequencies. Thus, their vibrational (or rotational-vibrational) signatures will be superimposed on the background radiance arriving at the spectrometer. These signatures can be extracted from the background radiance with appropriate data analysis protocols and used for both qualitative[75] and quantitative analysis.[33]

The focus of the work described in this chapter is the development of automated compound identification methodology for passive FT-IR remote sensing data. In the airborne monitoring application described above, an *in situ* near real-time identification regarding the presence or absence of targeted VOCs is desired. It demands an automated classifier compatible with the characteristics of this specific data measurement technique. As the aircraft flies at a

speed of ~150 knots and collects 80 interferograms per second, it is impossible to manually inspect every scan and pinpoint those that exhibit signatures of targeted VOCs in a near real-time manner.

For this scenario, supervised pattern recognition methods are useful tools to develop the needed classifiers. With pre-classified data as input patterns, a classifier is "trained" by a pattern recognition algorithm. Once developed, a classifier is able to predict the class identification of an individual scan as the classifier output. Thus, a set of classifiers can be implemented in real time to detect the characteristic signatures of the targeted VOCs as the data are acquired.[76] In the application targeted for this work, the goal is emergency response monitoring of releases of potentially hazardous VOCs as a consequence of incidents such as natural disasters, chemical plant accidents, train derailments, and the like.

There are several complicating issues associated with the development of compound classifiers based on passive FT-IR data. The signature of a targeted VOC can typically be visually identified in frequency-domain spectra, but in this case, due to the nature of the data collection, there is no stable background single-beam spectrum available in order to generate frequency-domain absorbance or transmittance spectra. Furthermore, since the compound identification procedure is to be performed with a computer-based algorithm rather than visually, there is little benefit to even performing the FT to transform the interferogram data into the frequency domain. For this reason, the work described here focuses on the application of pattern recognition methods directly to the collected interferogram data.

In the time-domain interferogram, the signature of a targeted VOC is not easily recognizable by visual analysis, but from the perspective of computer-based pattern recognition, direct analysis of the interferogram is feasible due to the fact that each VOC will yield a

characteristic signature induced by the effect of its spectral response on the interference pattern observed in the interferogram.

As discussed in Chapter 2, the spectrally broader character of the IR background causes its interferogram representation to decay much faster than the corresponding representations of narrower spectral features such as the absorption/emission bands of VOCs. This difference in decay rates provides a simple means to separate background and analyte signatures and suggests that a simple windowing of the interferogram (i.e., selection of an appropriate segment) can be an effective preprocessing step to remove the effects of the IR background.

While the windowing step removes the effects of broad background features, narrower features such as those arising from the spectral bands of atmospheric constituents will all contribute to the windowed interferogram segment (i.e., all spectral frequencies contribute to each interferogram point). This is problematic, as it decreases the selectivity of the analyte signature within the segment. To improve selectivity, a bandpass digital filter can be applied to the windowed segment to isolate a range of frequencies that are optimal for the selective detection of the analyte. While both the windowing and filtering steps require optimization of several relevant parameters (i.e., segment size and location, filter bandpass response and degree of attenuation), once these parameter values have been established, the ability exists to create selective input patterns for use in implementing binary (i.e., yes/no or active/inactive) classification decisions regarding the presence of a target analyte signature within the acquired data.[77]

Another issue associated with the development of automated classifiers with airborne passive FT-IR data is the sparseness of field data in which the signatures of target VOCs are present. Most compounds of interest for monitoring purposes are not naturally occurring.

Collection of field analyte-active data thus requires labor-intensive and costly experiments in which controlled releases of target analytes are conducted in coordination with overflights by the aircraft. Toxic compounds cannot be released in such experiments, and even with less toxic compounds the release plumes generated will necessarily be small. When the small FOV of the entrance optics of the spectrometer is considered (e.g., < 0.5°), an aircraft flying at even a relatively low altitude (e.g., 3000 ft) will produce few spectra in which the analyte signature is present. Thus, a large number of airborne backgrounds are available to define the analyte-inactive data class, but relatively few are available to represent analyte-active patterns.

This issue of the availability of airborne VOC-active interferograms limits the development of classification algorithms solely based on airborne interferograms. This suggests the use of mathematically synthesized interferograms,[76] which bear the spectral signature of the target analyte superimposed on a field-collected airborne background. Data of this type were used in this research to define the analyte-active class, while field-collected backgrounds represented the analyte-inactive class.

In this work, ambient ammonia was the targeted VOC that served as the focus of the methods development. Anhydrous ammonia is an important industrial chemical and is used widely in agricultural applications. As such, it is both transported routinely and stored at industrial facilities. Compounds derived from ammonia can be explosive, and numerous chemical plant accidents have occurred in which significant amounts of ammonia have been released.

In the work described in this chapter, the interferogram synthesis protocol described above was combined with signal processing and pattern recognition analysis to develop an

automated classifier for ammonia. The developed classifier was tested with airborne data sets that involved monitoring of ammonia releases during several emergency response incidents.

4.2 Experimental

4.2.1 Instrumentation and data collection

Interferogram data used in this work were collected by our research collaborators at the United States Environmental Protection Agency. A Bomem Model MR254/AB spectrometer (ABB Bomem, Quebec City, Que., Canada) was employed to acquire passive FT-IR interferograms. The spectrometer was equipped with a liquid-nitrogen cooled Hg:Cd:Te (MCT) detector operating over the 800 – 1200 cm^{-1} (MWIR) region. The FOV of the spectrometer was restricted to 0.3^{o} with a Cassegrain telescope. The resulting raw interferograms were double-sided single scans consisting of 512 points. Points were sampled at every eighth zero-crossing of the HeNe reference laser to produce a spectral bandwidth of 1975 cm^{-1}. Forman phase correction using 64 points on each side of the centerburst of the raw interferogram and normalization to a vector magnitude of 1.0 were applied prior to further processing.

To collect airborne data, the spectrometer mentioned above was mounted in a downward-looking configuration on a twin-engine high-wing aircraft (Aero Commander 680 FL, Aero Commander, Culver City, CA). The aircraft flew at altitudes of 2000 – 3000 ft. with a cruising speed of 100 – 150 knots. Open-air passive-mode interferograms were acquired at an approximate rate of 80 interferograms per second while the aircraft was flying over ground locations of interest. Four airborne remote sensing experiments (*A*, *B*, *C* and *D*) were conducted at two different sites. Three data sets, *A*, *B*, and *D*, were collected in the Houston, Texas area at three different times during the aftermath of Hurricane Rita in September, 2005. Data set *C* was collected in February, 2005 from a location in Missouri where emission of an unknown

concentration of ammonia was reported. These data sets were used as prediction or test sets for assessing the performance of the developed pattern recognition methodology.

In addition to the test sets described above, a database of >500,000 airborne background interferograms collected with the same instrumentation and aircraft described above were available for use in the synthesis of ammonia-active interferograms and as the inactive data class in training the ammonia classifiers. The subset selection algorithm developed by Carpenter and Small[78] was used to obtain representative samplings of this pool for use in the data analysis steps described below.

4.2.2 Data analysis implementation

Digital filters were designed with the aid of the Filter Design and Analysis Tool (FDATool) provided with the Signal Processing Toolbox (Version 5) of Matlab Version 7.4 (The MathWorks, Natick, MA). The pattern recognition methodology was based on support vector machine (SVM) classifiers implemented with the public-domain package, SVM^{light} (Version 6.01, http://svmlight.joachims.org). Digital filtering, SVM classification and all other data analysis tasks were performed on a Dell Precision 490 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Red Hat, Inc., Raleigh, NC). In-house software written in FORTRAN 77 was used for training and testing of classifiers based on piecewise linear discriminant analysis (PLDA). Analysis of variance (ANOVA) calculations were implemented under Minitab Version 16 (Minitab, Inc., State College, PA).

4.3 Results and Discussion

4.3.1 Overview of Methodology

As the target analyte for classifier development in this work, ammonia (NH_3) features a doublet with peaks at 964 and 933 cm^{-1} in its pure-component spectrum, corresponding to the symmetric H-N-H deformation band (v_2).[79] Figure 4.1, A displays a laboratory reference

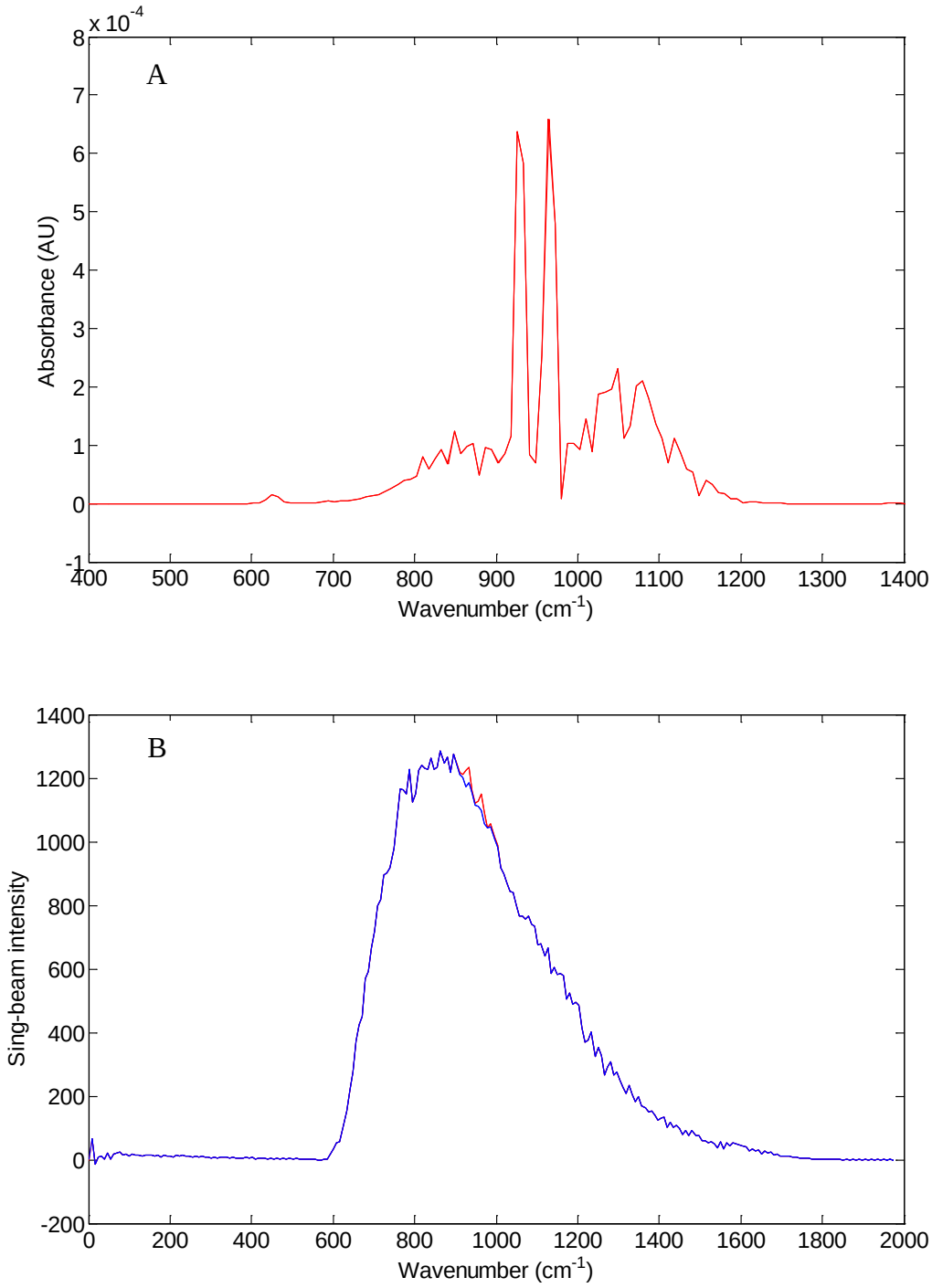
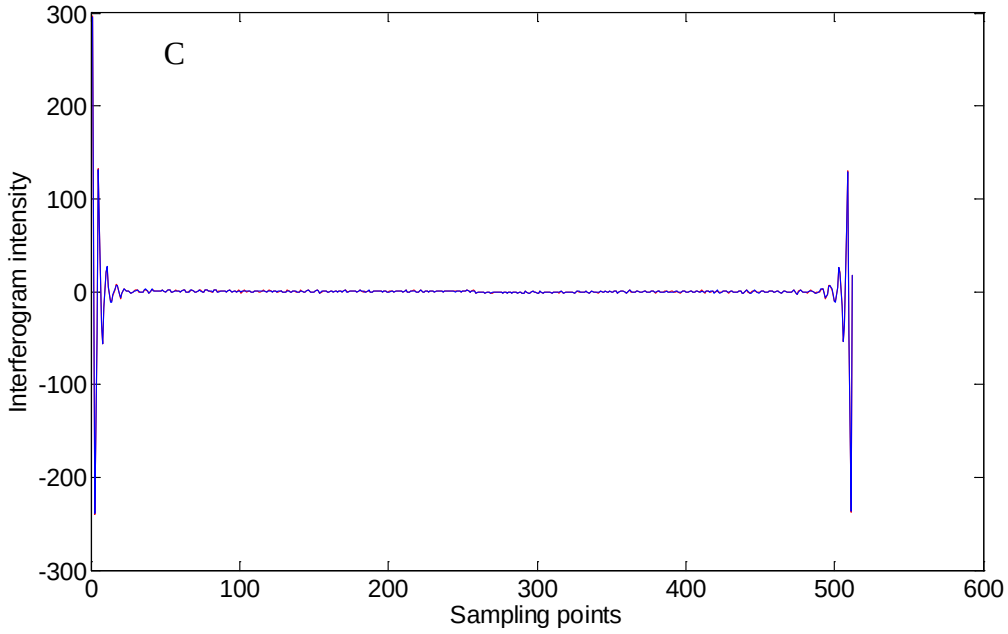
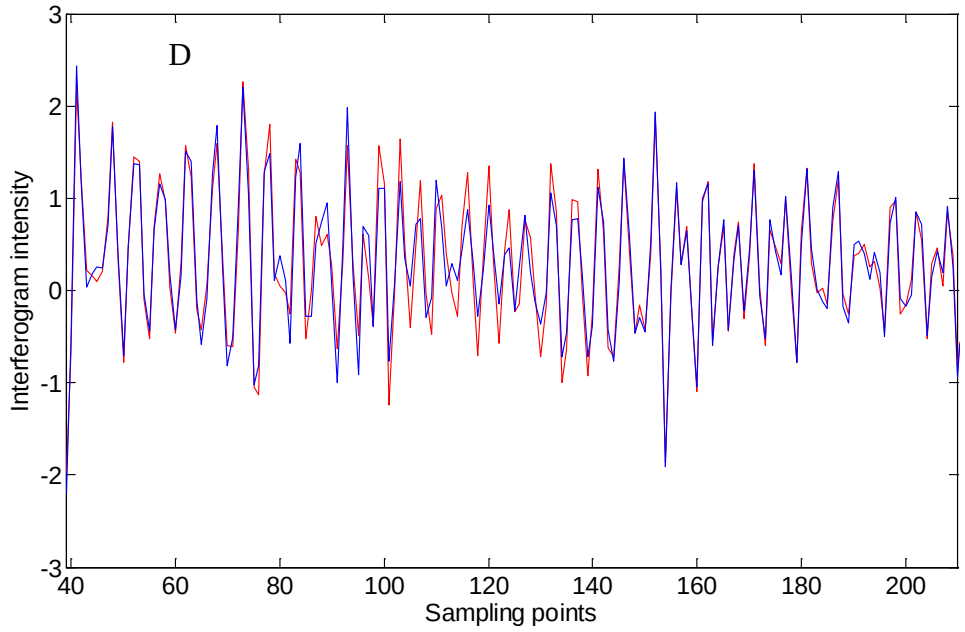


Figure 4.1 Scheme of the generation of synthetic data. (A) Pure-component absorbance spectrum of ammonia from the PNNL reference library [81] (1 ppm-m). (B) Aircraft background single-beam spectrum (blue line) and synthetic (ammonia-active) single-beam spectrum (red line). (C) Interferograms corresponding to the spectra in panel B. (D) Expanded view of C. The presence of ammonia become apparent starting around point 80.





(Figure 4.1 continued)

spectrum of ammonia that has been deresolved to 16 cm^{-1} to match the resolution of the Bomem spectrometer used to collect the remote sensing data.

This research focused on the development of a robust classification algorithm capable of implementing continuous monitoring for the presence of contaminating levels of ammonia in open air. The key to the development of such a classifier is the ability to extract the characteristic NH_3 spectral signature from the complex spectral background.

To develop a binary classifier for a determination of the presence or absence of ambient ammonia, PLDA and SVMs were employed as pattern recognition algorithms. As mentioned in Chapter 3, both PLDA and SVM are supervised pattern recognition methods. For supervised pattern recognition, a training set consisting of data (patterns) with known classification is required. There must be patterns representative of the analyte-active data class as well as patterns representative of the opposite (analyte-inactive) class within the training set. A successful classifier, after being "trained" how to classify patterns in the training set, will be expected to accomplish the same classification task on external prediction sets whose patterns do not have known class identities.

The PLDA and SVM methods are both examples of non-linear discriminant analysis, but they approach the non-linear separation of the data classes differently. As the name suggests, PLDA is essentially an extension of linear discriminant analysis. By combining multiple single-sided linear dicriminants, eventually PLDA can form an approximation to a nonlinear classifier to classify input patterns into the analyte-active or analyte-inactive data classes.

The SVM method achieves non-linear separation of the data classes by performing a non-linear mapping of input patterns in the original data space into a higher dimensional feature space, where a linear separation in the feature space is equivalent to a non-linear separation in

the original space. Compared to PLDA, the SVM algorithm is mathematically more potent and tunable, but the cost of the tunability is a high computational load involved in optimizing parameters related to the SVM architecture. The SVM method is difficult to implement in the context of a large-scale experimental design in which data processing parameters (e.g., digital filter or interferogram segment specifications) are being studied.

Instead, PLDA, with a relatively rigid training algorithm, can produce reasonable models efficiently in the context of studying other data processing parameters. Therefore, in this work, PLDA was used as the pattern recognition method for model selection. The optimal filter and interferogram segment selected by PLDA were then employed by the SVM algorithm to develop a pool of classifiers by varying the SVM configuration parameters. The optimal SVM classifier, selected by evaluating the performance results of the pool of classifiers when applied to the same data used with PLDA, was subsequently applied to the external prediction sets. [46,80]

As noted above, a great number of classifiers can be developed by varying data processing parameters such as the digital filter configuration and interferogram segment location. Another data subset, called the monitoring set, was formed in order to assist in the selection of the optimal classifier(s) from the pool of developed classifiers. The monitoring set was formed as a subset of the training set employing procedures to be detailed in a later section. Thus, the classifications of the analyte-active and analye-inactive patterns in the monitoring set were known.

In the search for the optimal classifier, parameters were adjusted, a classifier was developed on the basis of these parameters using the training subset, and the resulting classifier was applied to predict the classifications of the patterns in the monitoring set. Because the patterns in the monitoring set were withheld from the training step, they served as an

independent test set for use in evaluating the computed classifier. The classifier that produced the best classification results for the monitoring set was chosen as optimal. To be successful, this protocol assumes that the training subset is globally representative of future data that will be encountered and that the monitoring set is representative of the subsequent prediction data to which the classifier will be applied.

4.3.2 Generation of synthetic data

For absorption of background radiance by an analyte, the calculation of synthetic spectra was based on the premise that a single-beam spectrum is a resultant of the transmittance of the target analyte, convolved with the relevant instrumental and environmental responses. This can be represented by

$$\boldsymbol{I}(\bar{\upsilon}) = \boldsymbol{I_o}(\bar{\upsilon})\boldsymbol{T}(\bar{\upsilon}) \qquad (4\text{-}1)$$

where $I(\bar{\upsilon})$ corresponds to the synthetic single-beam spectral response, $I_o(\bar{\upsilon})$ is the environmental or instrumental profile and $T(\bar{\upsilon})$ represents the overall transmittance of the pure-component sample of the analyte. All of these parameters are functions of wavenumber, $\bar{\upsilon}$. In this work, the $I_o(\bar{\upsilon})$ were defined by infrared backgrounds collected by the airborne spectrometer when no analyte was present.

If multiple analytes are present within the FOV, Eq. 4-1 can be extended by simply including an additional $T(\bar{\upsilon})$ term for each analyte present. For the corresponding case in which the analyte is emitting at its characteristic vibrational frequencies, the transmittance is replaced by the emittance.

If we assume the concentration of the target analyte is relatively low, that intermolecular interactions have a negligible effect on the absorption and the refractive index of the sample is

constant with respect to $\bar{\upsilon}$, the transmittance for a particular sample can be rewritten in accordance with the Beer-Lambert law by

$$\boldsymbol{I}(\bar{\upsilon}) = \boldsymbol{I_o}(\bar{\upsilon})\mathbf{10}^{-\boldsymbol{A}(\bar{\upsilon})\boldsymbol{s}} \qquad \textit{(4-2)}$$

where $A(\bar{\upsilon})$ is the spectral response in absorbance mode for a given sample. The absorbance can be further expanded to be the product of absorptivity, path length, and concentration, illustrating the manner in which the depth of the sample along the optical axis and the analyte concentration contribute to the response.

For the current application, the pure-component absorbance spectrum of ambient ammonia is available from the PNNL quantitative infrared database[81] at a path-length integrated concentration of 1 ppm-m. This spectrum is shown in Figure 4.1, A. The magnitude of absorbance at unit concentration in ppm units over a 1 m path length is on the order of 10^{-4} absorbance units (AU).

To generate synthetic absorption or emission spectra with varying ammonia concentrations, a scaling factor, *s*, was added to the synthetic formula in Eq. 4-2, and was used to adjust the absolute intensity of the ammonia signature being superimposed and to encode whether the signal represented absorption or emission. For example, a scaling factor of +10 indicates an absorbance spectral feature of 10 ppm-m of ammonia will be superimposed, while a scaling factor of -20 means an emission feature of 20 ppm-m of ammonia will be superimposed in the generation of the synthetic spectrum.

In this case, interferograms were randomly picked from a pool of 40,528 airborne ammonia-free interferograms and were converted into single-beam spectra through application of the FT. As noted above, the resulting ammonia-inactive single-beam spectra were used to define the background spectral profiles, $I_o(\bar{\upsilon})$. A value of *s* randomly selected from a defined range was

applied as shown in Eq. 4-2 along with the $A(\bar{\upsilon})$ obtained from the 1 ppm-m reference spectrum of ammonia. This produced a synthetic ammonia-active single-beam spectrum.

Finally, the corresponding interferogram was computed for the synthetic single-beam spectrum through an inverse FT as

$$\boldsymbol{f(t)} = \frac{\mathbf{1}}{\mathbf{2\pi}} \int_{-\infty}^{\infty} \boldsymbol{I}(\bar{\upsilon}) \boldsymbol{e}^{-\boldsymbol{i}\bar{\upsilon}\boldsymbol{t}}\, \boldsymbol{d}\bar{\upsilon} \qquad \textit{(4-3)}$$

where $f(t)$ corresponds to the interferogram computed by application of the inverse FT to the synthetic single-beam spectrum, $I(\bar{\upsilon})$.[82] This process was repeated to produce each synthetic ammonia-active interferogram.

The process of interferogram simulation as described above is depicted in a stepwise manner in Figure 4.1. Starting with the pure-component absorbance spectrum of ambient ammonia (Figure 4.1, A) and a source single-beam spectrum (Figure 4.1, B, blue line), a synthetic single-beam spectrum is generated according to Eq. 4-2. This spectrum combines the features of the IR background and the ammonia signature. The scaling factor employed in this example was +29. As shown in Figure 4.1, B, in the single-beam mode, the synthetic spectrum (red line) is highly overlapped with the source single-beam spectrum (blue line). This is a reflection of the weak 29 ppm-m ammonia signature that has been added.

The corresponding interferogram representation in Figure 4.1, C, also exhibits a profile that is visually indistinguishable from the background interferogram. However, the expanded view in Figure 4.1, D, reveals a clear difference in patterns between the synthetic and background interferograms. As expected from the discussion above, the dominance of the initial section of the interferogram by the representation of the broad IR background features causes the early part of the interferogram to be virtually identical between the background and ammonia-

active data. However, points 80 to 140 relative to the centerburst begin to show clear differences in patterns. This is the information that will serve as the basis for the classification models discussed below.

4.3.3 Data set assembly and partitioning

As mentioned earlier, to develop classifiers based on supervised pattern recognition algorithms, interferograms were partitioned into independent training, monitoring and prediction sets. Each data set was composed of a representative number of interferograms selected from a larger pool. Each of the data subsets is assumed to be representative of the breadth of data that will be encountered in field measurements.

The training set was used to develop classifiers designed to detect ammonia signatures in the passive IR data acquired from the air. The monitoring set was used as an internal test set to optimize parameters associated with the filter design, interferogram processing and, if applicable, the SVM construction. Both PLDA and SVM used the same data subset as the monitoring set. The final selected classifier was applied to the four independent external prediction sets. The classifier's performance with the prediction sets will serve as a measure of the robustness of the classifier.

Each interferogram was treated as a pattern and was classified into binary classes, either ammonia-active or ammonia-inactive. Those interferograms placed in the analyte-inactive class were known to contain no ammonia on the basis of their corresponding ground locations. The synthetic ammonia-active interferograms were known to contain ammonia signatures. This left the prediction data sets in which it was expected that ammonia was present on the basis of the time and location of the data collection, but exactly which scans contained ammonia signatures was unknown.

For these data sets, each interferogram was classified as either ammonia-active or ammonia-inactive by visual inspection. Interferograms were converted through application of the FT to single-beam spectra and normalized to unit area. Within the same data collection session, a group of interferograms, which were confirmed ammonia-inactive (i.e., they were collected territorially further away from the suspected analyte source), were processed similarly, and the average of the resulting single-beam spectra was taken as an approximate background spectrum. This spectrum was subtracted from each single-beam spectrum in the data set. The corresponding difference spectra were then visually inspected for the signature doublet peaks of ammonia. If the feature was present, the interferogram was classified into the ammonia-active group, while if it was clear that the expected feature was absent, the interferogram was placed into the ammonia-inactive class. For cases in which the presence of the compound signature was indeterminate, the corresponding interferograms were excluded from further use in order to keep the task of classification as clear-cut as possible.

The data partitioning results are summarized in Table 4.1. For the prediction sets, the number of interferograms in the two data classes was less than the total number of interferogams in the data set due to the exclusion of the ambiguous interferograms. Also, as discussed previously, the population of active patterns is much sparser in actual field data due to the occurrence of small emission sources, the small FOV of the spectrometer, the high speed of the aircraft, and imprecision in the flight path relative to the targeted location. This accounts for the small number of confirmed ammonia-active interferograms in the prediction sets.

Table 4.1 Data partitioning and composition

Data subsets	No. ammonia-active patterns	No. ammonia-inactive patterns	No. of active + inactive patterns	Total No. patterns in test file
Training set	3000	12000	15000	15000
Monitoring set	2897	17367	20264	20264
Prediction set *A*	102	1935	2037	2127
Prediction set *B*	54	1898	1952	2039
Prediction set *C*	8	1789	1797	1800
Prediction set *D*	21	1126	1147	1274

In this work, the training set and monitoring set were formed in a similar manner. The training set contained 3000 ammonia-active patterns and 12,000 ammonia-inactive patterns, while the monitoring set consisted of 2897 ammonia-active patterns and 17,367 inactive patterns. Active patterns for both training and monitoring sets were obtained from the pool of synthetic interferograms. No field-collected ammonia-active interferograms were used in the training and monitoring sets because of the small number of these interferograms available and the need for the development of methodology that did not depend on the availability of such data. The imbalance in the sizes of the data classes reflected the need for the inactive class to span as much variety as possible in the IR backgrounds that may be encountered during field measurements. This helped to prevent the occurrence of false detections (false positives) when the classifiers were implemented.

Two sets of scaling factors, *s*, with different statistical characteristics were used in Eq. 4-2 to generate synthetic interfrograms with different ammonia intensities for the training and monitoring sets. The absorbance range (mean ± standard deviation) of the designed ammonia-active patterns was a normally distributed range of $(-3.3\pm4.8)\times10^{-3}$ for the training set, and $(-3.3\pm2.4)\times10^{-3}$ for the monitoring set. The distribution of the designed absorbance ranges in both the training and monitoring sets is displayed in Figure 4.2. These ranges were set on the basis of the knowledge of the absorbance at 1 ppm-m to be on the order of 10^{-4} AU and the fact that this absorbance magnitude is reflective of an active (rather than passive) spectroscopic measurement.

As illustrated in Chapter 2, the analyte signal observed in a passive experiment is

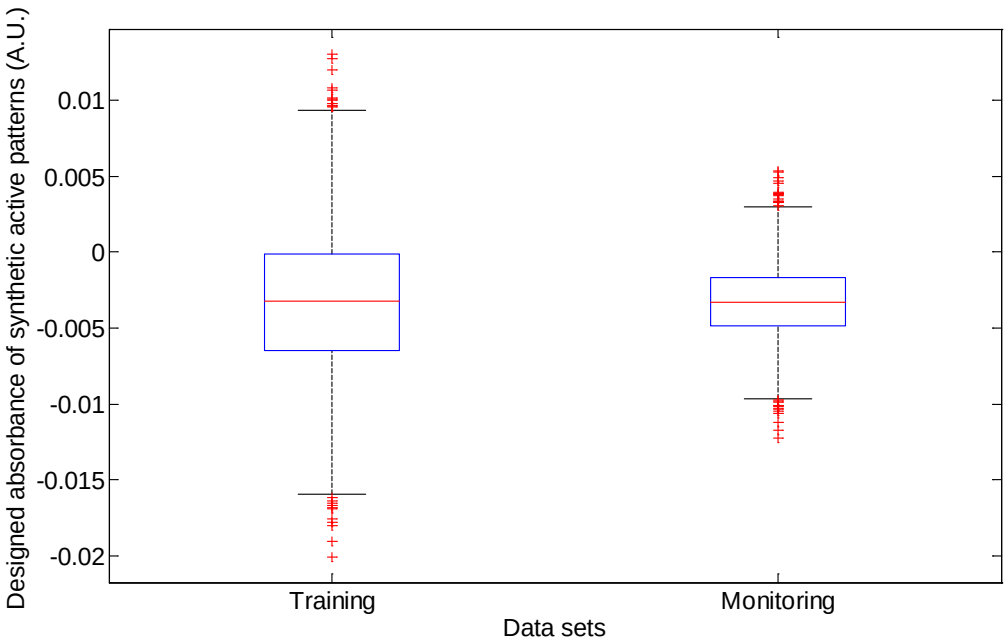


Figure 4.2 Boxplots showing the distribution of designed absorbance values of the synthetic data. For each distribution, the horizontal line within each box specifies the median absorbance, while the box limits denote the upper and lower quartiles. Whiskers are drawn to the extreme data values that are not outside three times the semi-interquartile range. Points outside these limits are plotted individually and would be considered outliers if the distribution were Gaussian (i.e., outside of approximately 99.7% of the distribution).

dependent not only on the product of absorptivity, path length, and concentration (i.e., $A(\bar{\upsilon})$ in Eq. 4-2) but also on the temperature differential between the analyte and background. Through several preliminary investigations, the absorbance ranges noted above appeared to produce classifiers that had both good sensitivity to ammonia and good generalizability to operate outside of the training data. All of the generated data was weak in terms of the magnitudes of the ammonia signals present visually in the spectra.

Ammonia-inactive patterns for both training and monitoring sets were from previously inspected airborne interferograms. As noted previously, the subset selection method of Carpenter and Small[78] was used to select the interferograms in the training and monitoring sets from a pool of >500,000 background interferograms assembled from multiple aircraft flights at a variety of locations.

As shown in Figure 4.3, A, principal component analysis (PCA) revealed the similarity between the synthetic active patterns in the training set (green) and confirmed active patterns in the combined prediction set (red), even with raw interferograms as input patterns. However, the separation between the active (green and red) and inactive (black and blue) classes across the data sets was not clear.

Further signal processing steps on these raw interferograms, including interferogram segment selection and digital filtering, improved the distinguishability between the ammonia-active and ammonia-inactive classes, as shown in Figure 4.3, B, even though the binary

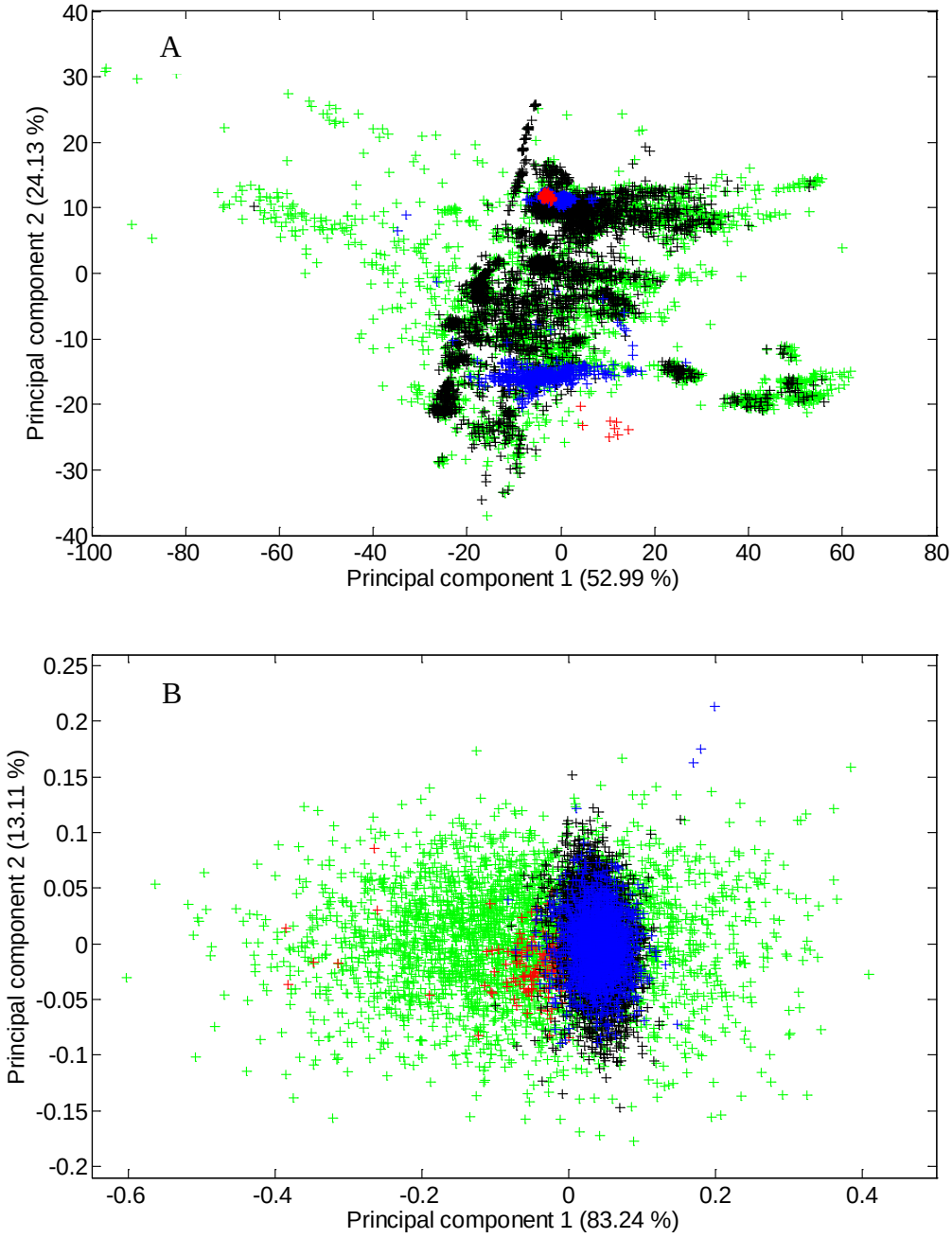


Figure 4.3 Principal component score plots illustrating the degree of clustering and separation among different pattern groups. Data plotted include airborne ammonia-active patterns in the combined prediction set (red), airborne ammonia-inactive patterns in the combined prediction set (blue), synthetic ammonia-active patterns in the training set (green), and synthetic ammonia-inactive patterns in the training set (black). The entire raw interferogram was used to generate plot A, while the optimal preprocessing parameters were used in plot B. A 120-point interferogram segment starting at point 88 (relative to the centerburst) was used. The digital filter used was a Chebyshev Type II IIR filter centered at 964 cm-1, with a passband FWHM of 28 cm-1 and a stopband attenuation of 40 dB.

separation is still visually marginal in the principal components space. Details regarding signal processing and the optimization of data set assembly, respectively, are discussed later in this chapter and in Chapter 5.

4.3.4 Characterization of signal strengths

The spectral signal strengths of ammonia in all data subsets were characterized by the estimated signal-to-noise (S/N) ratio on the basis of computed difference spectra. The procedure was similar to that described in previous work.83,84 The ammonia-inactive single-beam spectra corresponding to the interferograms in the training set were used to define a pool of possible backgrounds for use in computing difference spectra.

For each ammonia-active spectrum whose S/N ratio was sought, the three best matching background spectra were selected on the basis of the highest correlation coefficients computed over the combined spectral range of 800-910 and 980-1200 cm-1. Each selected background was then used as the independent variable in a two-parameter (i.e., slope and intercept) linear regression fit to the ammonia-active single-beam spectrum. Together with the background spectrum, the least-squares slope and intercept determined from the spectral range noted above were used to predict the spectral baseline in the ammonia peak region of 910-980 cm-1. By subtracting the predicted baseline from the ammonia-active spectrum, a baseline-corrected spectrum was obtained. The ammonia signal level was then taken as the absolute value of the difference between the maximum and minimum values in the 910-980 cm-1 region of the baseline-corrected spectrum. This value was used in the numerator of the calculation of the S/N ratio.

To compute a noise level for use in the denominator of the S/N ratio, 2000 random pairs of the single-beam spectra in the pool of backgrounds were selected and the same procedure described above was used to compute ammonia signal values for cases in which it was known

that no ammonia was present. This calculation produced a mean signal level of 0.0137 and corresponding standard deviation of 0.0100. This standard deviation was used as the noise value in the calculation of the S/N ratio corresponding to each signal value.

Figure 4.4 is a series of box plots showing the distribution of the computed S/N ratios for the training set, monitoring set, and the combined prediction sets. The median values of the ammonia S/N ratios aligned at similar levels for the training set, the monitoring set, and the combined prediction set, but following this order, increasingly narrower ranges of S/N ratio values were observed. This was consistent with the trend of designed absorbance ranges for the training set and monitoring set as mentioned earlier. The low values of the computed S/N ratios also confirm the presence of very weak ammonia signals in all the data sets.

4.3.5 PLDA-based model selection: effects of segment conditions, filtering parameters and potential interaction effects

As noted previously, the interferogram representation corresponding to the broad IR background decays much faster than the corresponding representation of the narrower features of ammonia. The difference in decay rates makes it possible and necessary to select an interferogram segment which contains relatively small contributions from the background, while most of the ammonia information is retained. Analysis based upon shorter interferogram segments is particularly favorable for rapid airborne passive remote sensing, where fast data acquisition and real-time analysis are desired.

Although dominant broad background signals are removed by proper segment selection, there are still some unwanted spectral features in the selected interferogram segments, such as

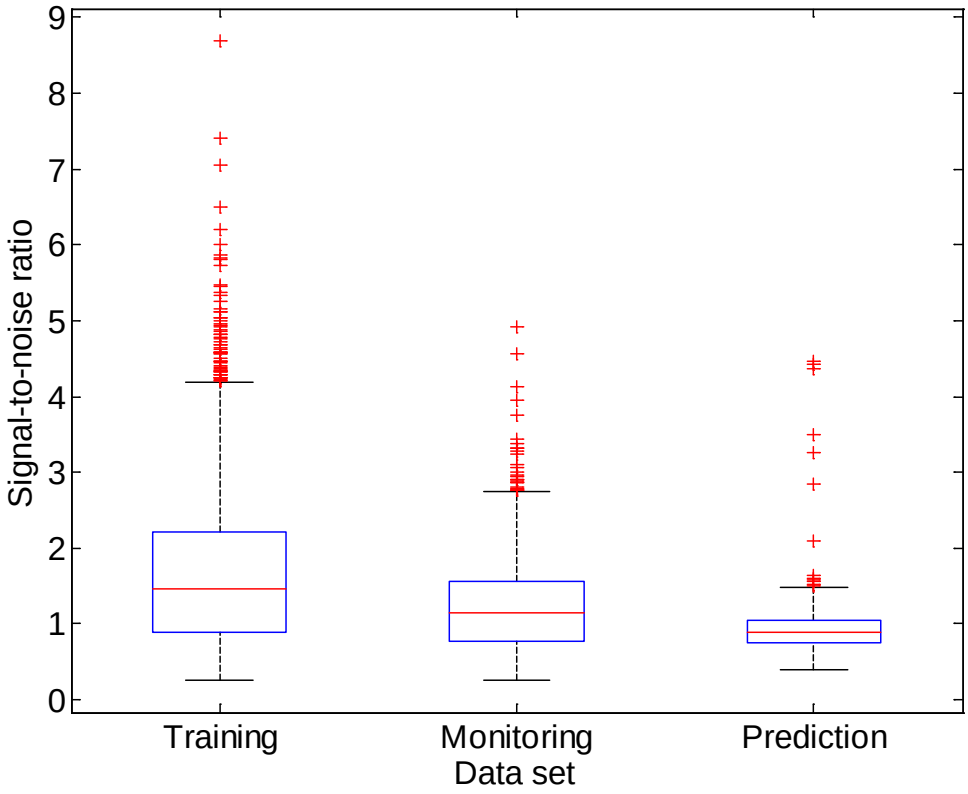


Figure 4.4 Boxplots characterizing the distribution of spectral S/N ratio values computed from the training, monitoring, and combined prediction sets. For each distribution, the horizontal line within each box specifies the median S/N ratio, while the box limits denote the upper and lower quartiles. Whiskers are drawn to the extreme data values that are not outside three times the semi-interquartile range. Points outside these limits are plotted individually and would be considered outliers if the distribution were Gaussian (i.e., outside of approximately 99.7% of the distribution).

those associated with spectral bands of interferences or narrow noise features from the instrument. To improve the selectivity for the analyte, a bandpass digital filter was applied to the selected interferogram to further suppress contributions from unwanted frequencies. Also, the fact that the implementation of the bandpass filter will not only extract the analyte peak, but also truncate the IR background emission to the same width and shape as the filter gives rise to concern about the interaction between segment and filter effects (i.e., the question of whether the optimal filter parameters depend on which interferogram segment has been selected). For this reason, the design of the model selection scheme should include investigation over the major effects as well as potential interactions between these effects. A joint optimization of the segment and filter specification is thus required.

A full factorial experimental design, as detailed in Table 4.2, was developed to include five factors at different levels. These included three levels of filter passband center, six levels of filter passband width, two levels of filter stopband attenuation, four levels of segment length, and six levels of segment starting point. The resulting combinations with varying conditions were employed with the training set to generate a total of 864 classifiers based on PLDA. Each PLDA classifier was based on two linear discriminant functions. Once developed, all PLDA classifiers were applied to predict the classifications of patterns in the monitoring set. The classification performance with the monitoring set, balanced with the results from the training set, was used to evaluate each classifier, and eventually to determine the optimal classifier and its corresponding segment and digital filter parameters.

Table 4.2 Parameters for PLDA model selection

Parameters		Levels	Values
Filter	Passband center (cm^{-1})	3	933, 949, 964
	Passband width, FWHM (cm^{-1})	6	10, 16, 28, 40, 52, 64
	Stopband attenuation (dB)	2	20, 40
Segment	Starting point	6	16, 40, 64, 88, 112, 136
	Segment length	4	60, 80, 100, 120

Infinite impulse response (IIR) filters were chosen for this work. The IIR method allows filtered points prior to the point being filtered to contribute to the approximation of an overall filtered interferogram. As a result, IIR filters with a smaller filter order (number of coefficients) can achieve a similar filtering effect as conventional finite impulse response (FIR) filters with higher filter order.

Chebyshev Type II filters were the type of IIR filters used in this work. Their frequency response features a flat passband and rippled stopband, which is preferred for interferogram processing in order to remove unwanted frequencies while preserving the ammonia signals in an unaltered manner.

One of the implementation issues that must be considered with IIR filters is the need for the filter to stabilize. Because the filter employs previously filtered points in its calculation of the filtered intensity at point *i*, the filter needs to operate long enough so that previously filtered points are available and the action of the filter becomes consistent. This translates to the need for a gap or offset between the starting point of the interferogram segment and the first point filtered.[85,86] In this work, the gap between the segment starting point and filter starting point was set to 14. For example, if the interferogram segment started at point 50, the filter began operating at point 36.

All filters were designed based upon the spectral features of ammonia in the region of 900 to 1000 cm^{-1}. Three levels of filter center were chosen at the center of the left peak, the middle of the doublet, and the center of the right peak. Four levels of filter passband width were chosen so that the overall range of widths spanned gradually from slightly narrower than an individual peak to wider than the doublet peaks. Plus, the stopband attenuation evaluated at two levels added the ability to study the extent to which the signal at unwanted frequencies was

suppressed by the filter. This design scheme provided opportunities to study the relationships between the filter design specifications, spectral features to be extracted and the corresponding background profile. The frequency responses of all 36 filters, grouped by three varying filter factors are plotted in Figure 4.5.

Four performance measures were used to evaluate the classifiers: (1) the training set separation rate (i.e., the percentage describing the fraction of ammonia-active patterns being classified correctly), (2) the monitoring set missed detection rate (i.e., the percentage describing the fraction of ammonia-active patterns classified as ammonia-inactive), (3) the monitoring set false detection rate (i.e., the percentage describing the fraction of ammonia-inactive patterns classified as ammonia-active), and (4) the difference between the monitoring set missed detection rate and the false detection rate. The false detection rate for the training set was not considered because the PLDA method implemented in this work is heavily weighted against false positives in the training set. The results obtained from PLDA in terms of the first three performance measures are summarized in Figure 4.6.

The four performance measures were used as response functions in an ANOVA study to evaluate all of the PLDA classifiers. The factors in the ANOVA study were the interferogram segment and digital filtering parameters under study. The results from the ANOVA study are presented in Figure 4.7 as a series of main effects and two-way interaction plots for each of the four response measures noted above. The main effects for each factor specify the average response across the levels of the factors. A larger range of response across the levels of the factor

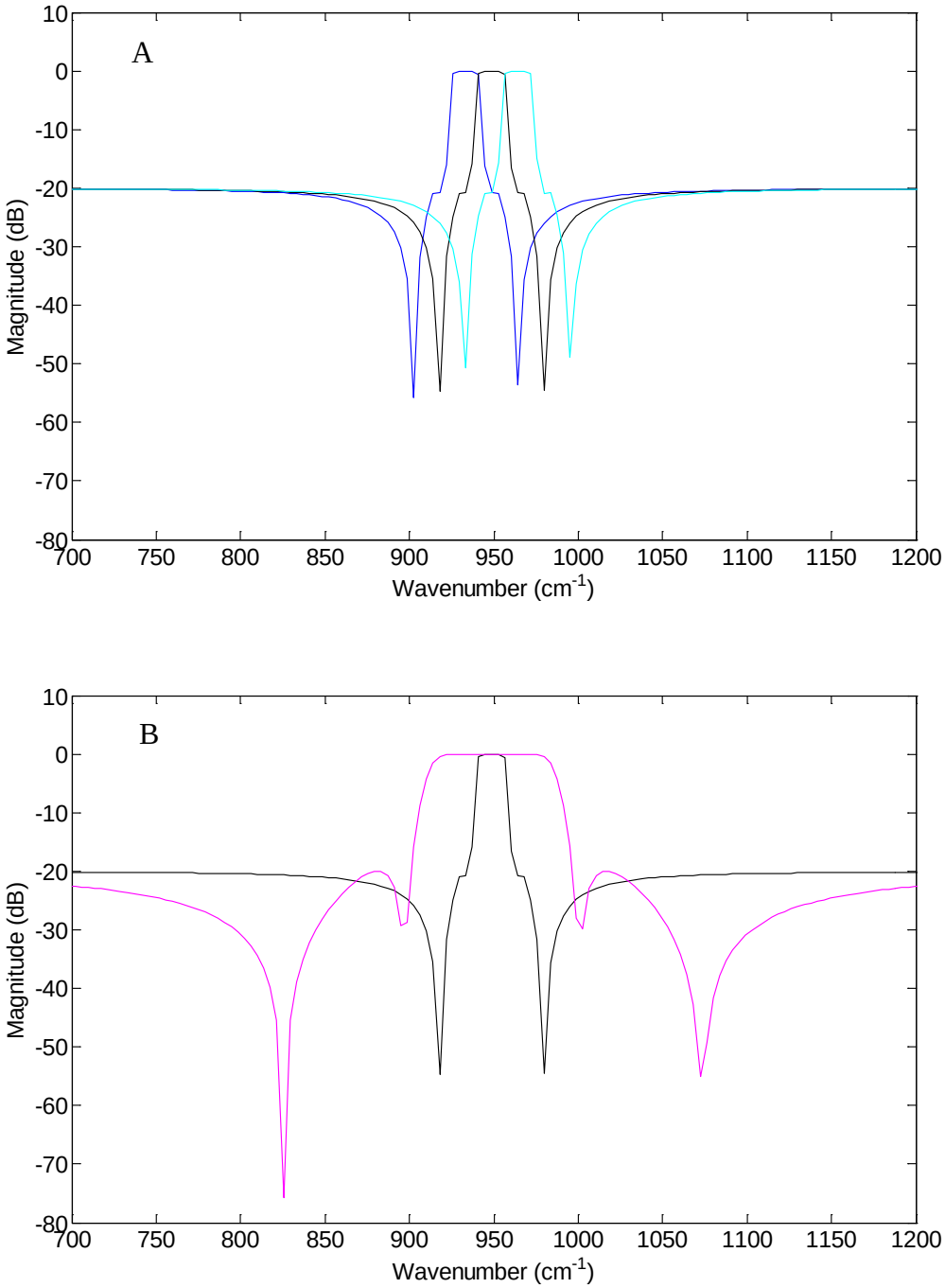
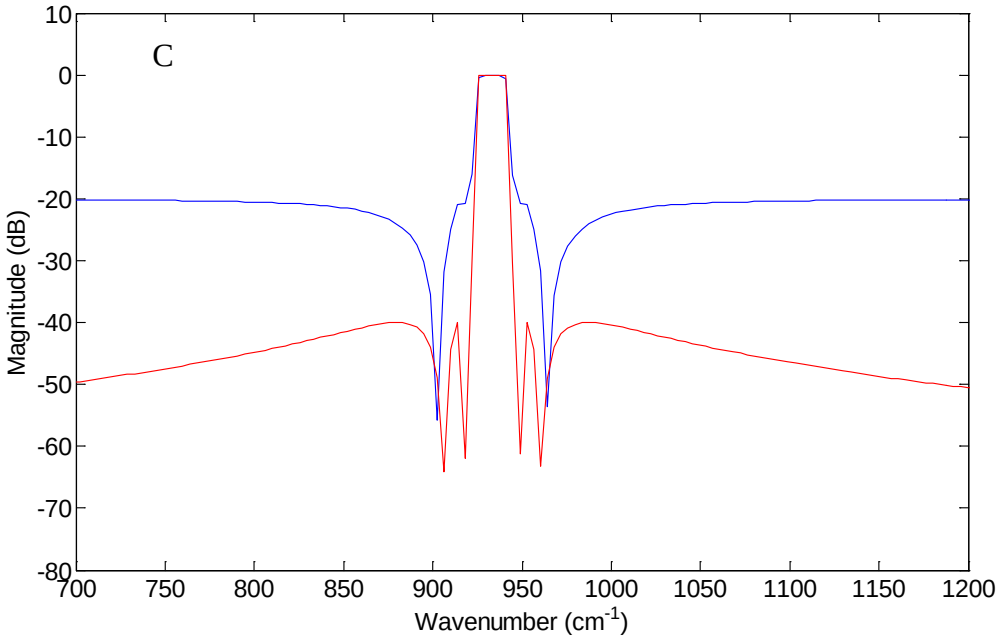


Figure 4.5 Frequency responses of designed IIR digital filters. (A) Three filters have fixed passband width (expressed as width at half-maximum (FWHM)) of 28 cm^{-1} and stopband attenuation of 20 dB, but varying passband centers. Blue: center at 933 cm^{-1}; black: center at 949 cm^{-1}, cyan: center at 964 cm^{-1}. (B) Two filters have fixed passband centers of 949 cm^{-1} and stopband attenuation of 20 dB, but varying passband width (FWHM). Black: FWHM width of 28 cm^{-1}, magenta: FWHM width of 64 cm^{-1}. (C) Two filters have fixed passband centers of 949 cm^{-1} and passband width of 28 cm^{-1}, but varying stopband attenuation. Blue: dB = 20, red: dB = 40.



(Figure 4.5 continued)

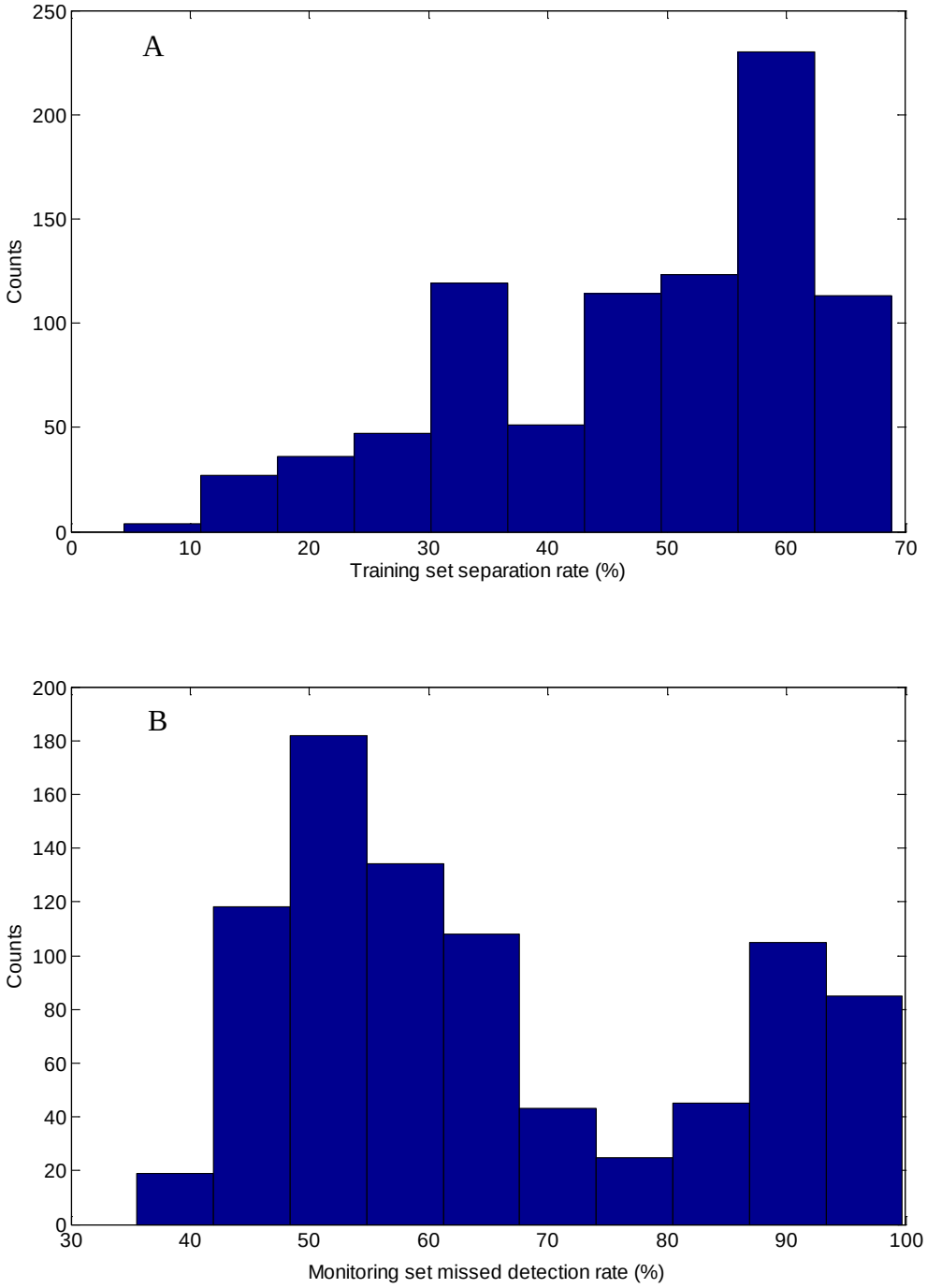
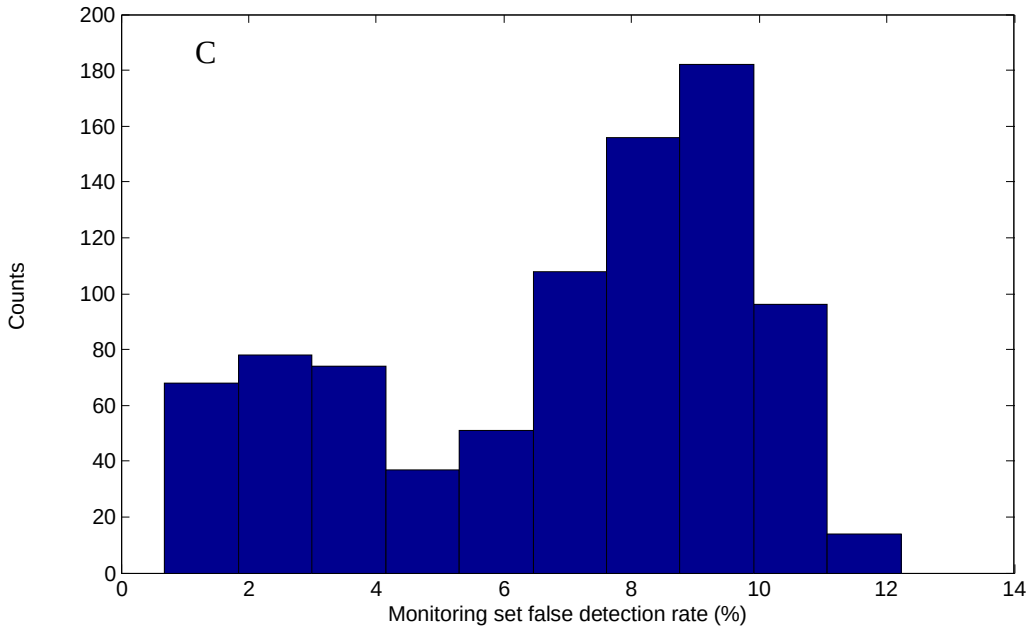


Figure 4.6 Summary of classification results of 864 PLDA models for model selection. (A) Distribution of training set separation rates. (B) Distribution of monitoring set missed detection rates. (C) Distribution of monitoring false detection rates.



(Figure 4.6 continued)

suggest the factor is having a significant effect on the response. The interaction plots characterize the two-way interactions among the factors. For each pair of factors, deviations of the groups of traces from parallel signal that the levels of one factor are having an effect on the way in which the levels of the other factor affect the response.

The classifier to be selected was expected to have as small as possible values for each of the four responses. As indicated in Figure 4.7, the segment starting point and the segment length were the two most significant main effects among the four responses. By contrast, the digital filtering parameters were much less significant. Meanwhile, two of the responses, the training set separation rate and the monitoring set false detection rate, had similar patterns in term of relationships with all the factors, while the other two responses, the monitoring set missed detection rate and the (missed – false) rate difference for the monitoring set, seemed to share similar trends in terms of both response and relationships to the factors. There was also a mirror image relationship between the two sets of responses. Also, among all possible combinations of two-way interaction effects, the ANOVA study revealed that the interaction between the segment starting point and the segment length was the only one really evident as being significant.

The importance of segment selection can be attributed to the high degree of overlap between the ammonia-active and ammonia-inactive interferograms observed earlier. It is critical to choose the most discriminating segment of the interferogram, including the length of the segment, the segment starting point, and the interaction between the length and the starting point, as patterns for classification.

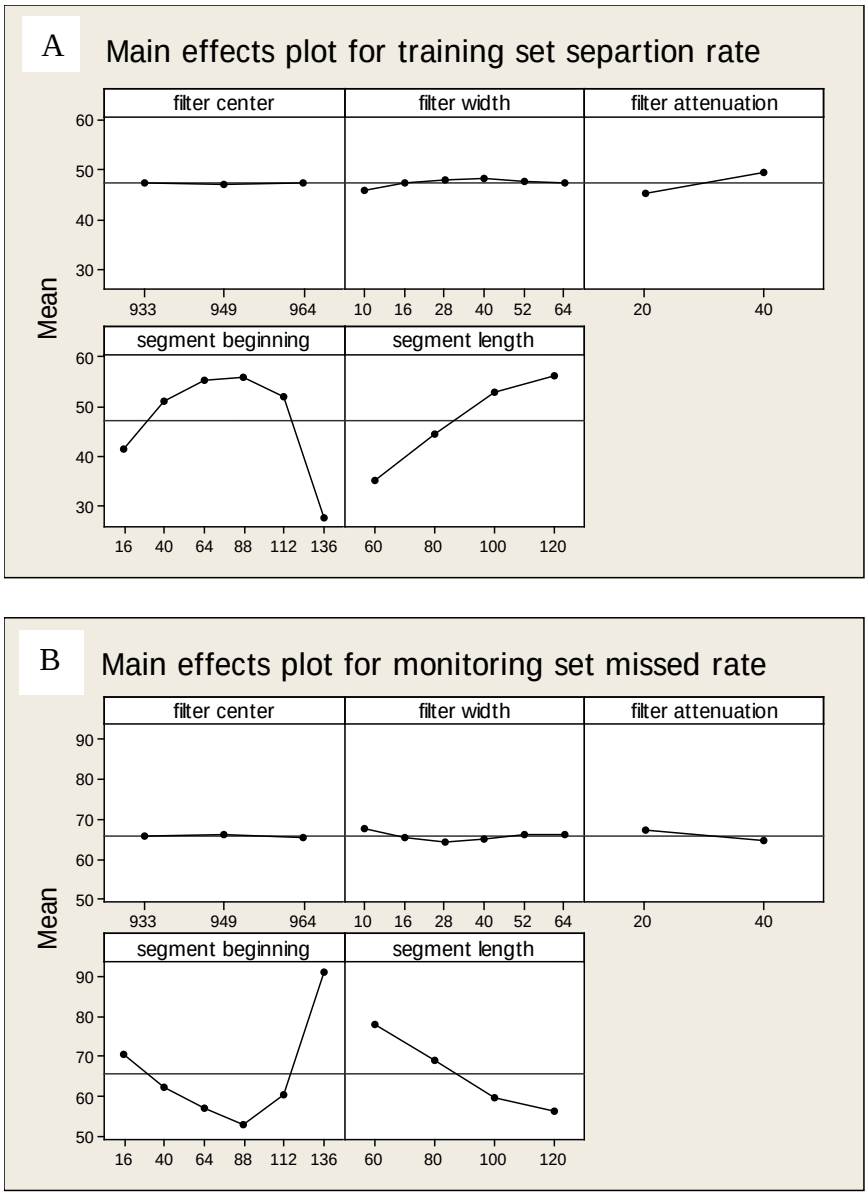
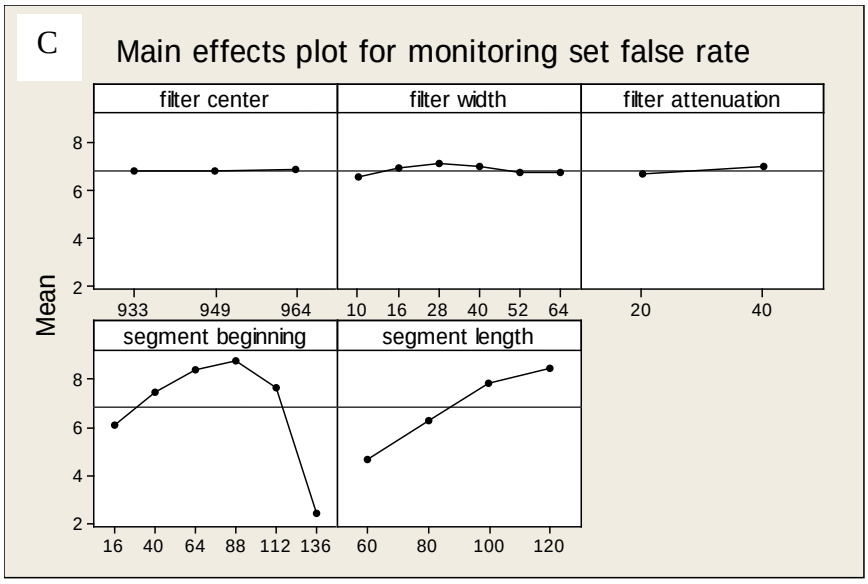
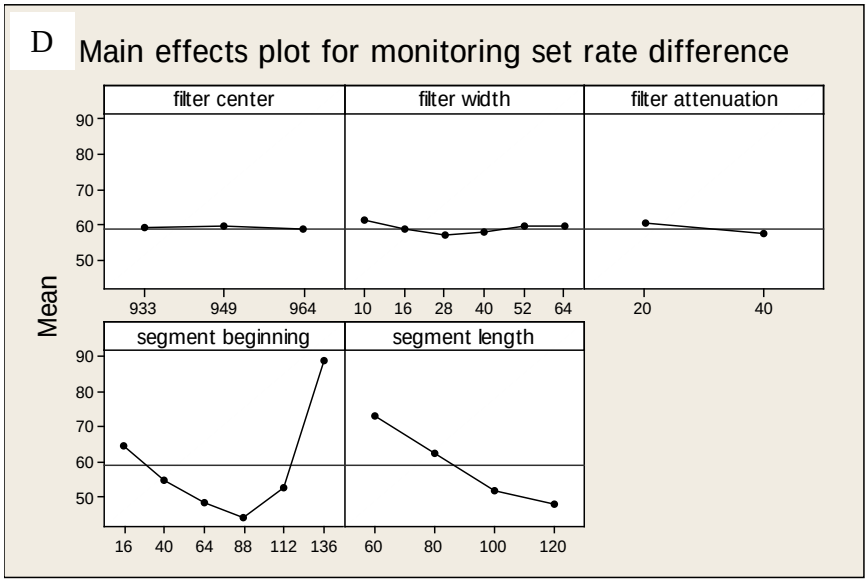
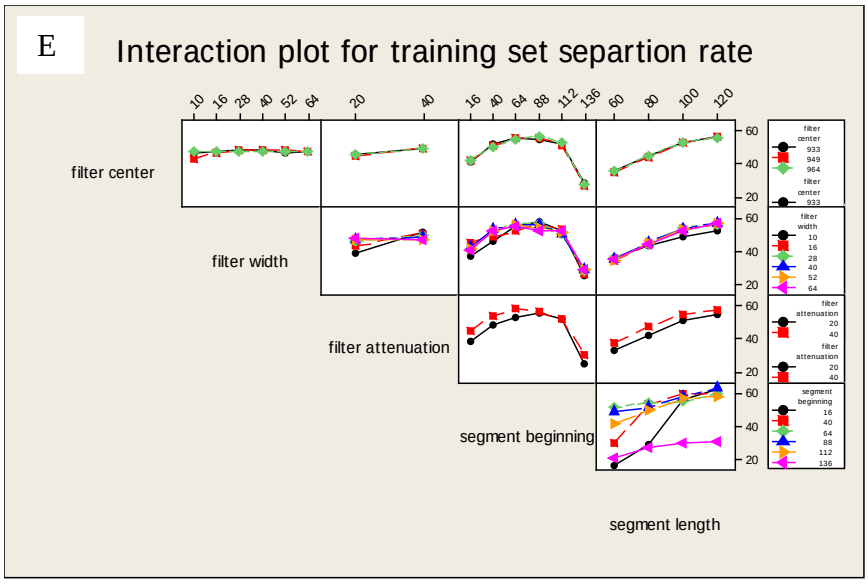


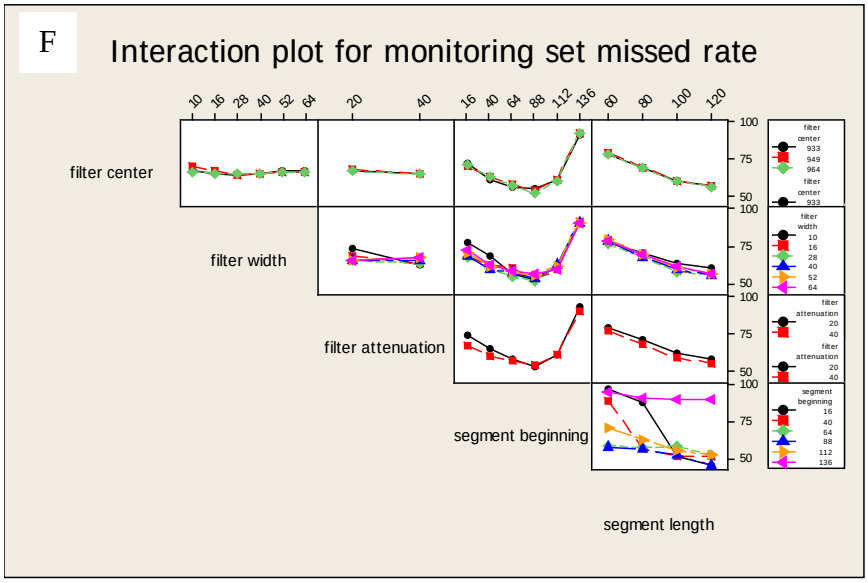
Figure 4.7 Results obtained from application of ANOVA to the PLDA classification results in which five factors were studied: segment length, segment stating point, Chebyshev Type II filter passband center, filter passband width and filter stopband attenuation. (A) Main effects plot with the training set separation rate as the response. (B) Main effects plot with the monitoring set missed detection rate as the response. (C) Main effects plot with the monitoring set false detection rate as the response. (D) Main effects plot with the gap between the missed and false detection rates for the monitoring set as the response. (E) Two-way interaction plot among the five factors with the training set separation rate as the response. (F) Two-way interaction plot among the five factors with the monitoring set missed detection rate as the response. (G). Two-way interaction plot among the five factors with the monitoring set false detection rate as the response. (H) Two-way interaction plot among the five factors with the difference between the missed and false detection rates for the monitoring set as the response.



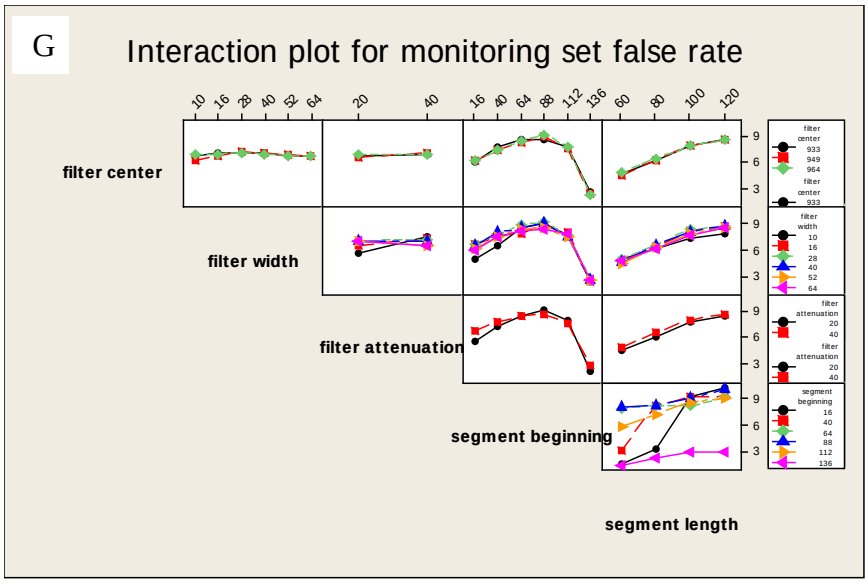


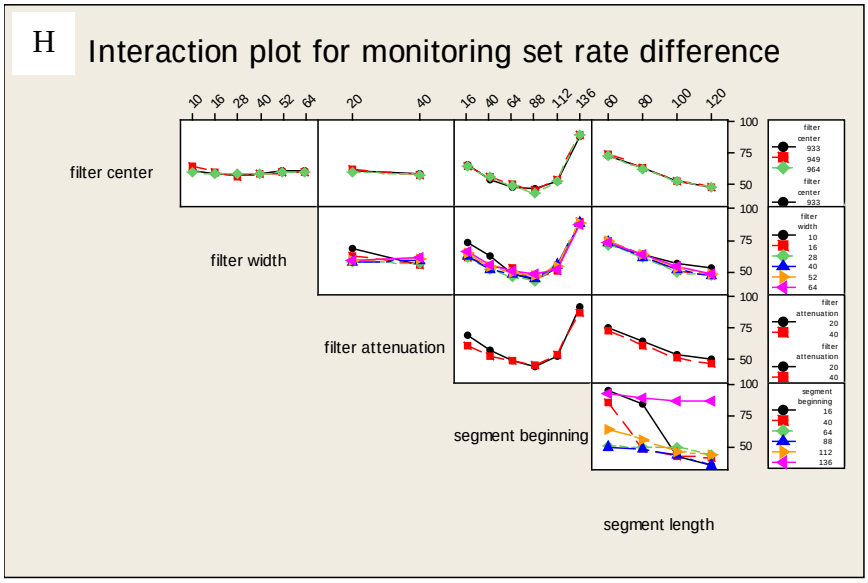
(Figure 4.7 continued)





(Figure 4.7 continued)





(Figure 4.7 continued)

The optimal segment starting point and segment length for the monitoring set missed detection rate were 88 and 120, respectively. As before, point specifications are relative to the assignment of point 1 as the interferogram centerburst. This approximately matches the visual assessment made earlier in the inspection of Figure 4.1, D that the segment from points 80 to 140 looked to be the most discriminating for ammonia.

However, from the perspective of the successful detection of the ammonia-inactive patterns (i.e., the false detection rate), a shorter segment, such as a length of 60 and a starting point either closer to the centerburst (point 16) or further from the centerburst (point 136) are suggested. In this case, the closer segment includes more background information and thus better characterizes the variation in the background signatures across the monitoring set. The segment further from the centerburst eliminates the region where the ammonia signals are the strongest and thereby appears to capture information related to when the ammonia signature decays to zero (i.e., becomes indistinguishable from the background).

In terms of absolute magnitudes, the false detection rate was much lower than the missed detection rate, which can be explained by the fact that both the training set and the monitoring set had a dominant number of ammonia-inactive patterns and relatively small numbers of ammonia-active patterns (about a 5:1 ratio). The same fact also accounted for why the trends for the training set separation sided with the monitoring set false detection rate, and trends for the rate difference in the monitoring set were similar to the false detection rate in the monitoring set.

The fact that the filter-related factors didn’t seem to play significant roles might be due to the characteristic spectral signature of ammonia. In this work, the difference in decay rates relative to the background caused by the narrow widths of the two spectral features of ammonia might have been sufficient for effective discrimination of the ammonia signature.

Correspondingly, there may have been no bands of similar width to the ammonia features and thus no need for the filters to suppress the intensities of frequencies outside of the ammonia region.

In terms of classifier selection, it was clear that a tradeoff needed to be made between the missed and false detection rates. To effect this compromise, a segment starting point of 88, a segment length of 120, a Chebyshev Type II filter centered at 964 cm^{-1}, with a passband full-width at half-maximum (FWHM) of 28 cm^{-1}, and a stopband attenuation of 40 dB, were chosen as the optimal conditions. This combination of parameters achieved 64.7 % separation of the training set and produced missed and false detection rates of 44.7 and 10.2 % respectively, when the PLDA classifier was applied to the monitoring set. The low separation rates for the training set and high missed detection rates for the monitoring set are again a reflection of the weak ammonia signals generated in the data synthesis procedures.

4.3.6 SVM classifiers

The optimal conditions selected by PLDA were subsequently applied to develop SVM classifiers. Two configuration parameters, the kernel parameter, γ, at 10 levels, and the regularization parameter, *C*, at 20 levels, were used to construct a pool of 200 SVM classifiers (Table 4.3). Three criteria were considered to evaluate the efficiency of the classifiers and their degree of generalization to operate outside of the training set: (1) the difference between the missed and false detection rates for the training set, (2) the difference between the missed and false detection rates for the monitoring set, and (3) the number of support vectors produced in the optimization of the SVM classifier. In an SVM classifier, a support vector is a pattern lying

Table 4.3 Parameters for SVM model selection

Parameters	Levels	Values
Kernel parameter (γ)	10	0.02, 0.04, 0.08, 0.16, 0,32, 0.64, 1.28, 2.56, 5.12, 10.24
Regularization parameter (C)	20	3.0×10^{3}, 1.5×10^{4}, 2.7×10^{4}, 3.9×10^{4}, 5.1×10^{4}, 6.3×10^{4}, 7.5×10^{4}, 8.7×10^{4}, 9.9×10^{4}, 1.11×10^{5}, 1.23×10^{5}, 1.35×10^{5}, 1.47×10^{5}, 1.59×10^{5}, 1.71×10^{5}, 1.83×10^{5}, 1.95×10^{5}, 2.07×10^{5}, 2.19×10^{5}, 2.31×10^{5}

within the interface region (termed the “margin”) between the data classes. A maximum in the number of support vectors corresponds to a classifier that has been positioned effectively to model the interface between the data classes.

The top five SVM classifiers selected from the pool of developed classifiers are listed in Table 4.4. These five classifiers were tied in terms of each of the three criteria noted above. For the monitoring set, the missed detection rate was ~ 33 %, the false detection rate was ~2 %, and the difference in rates was ~ 31%. An improved classification performance is noted relative to the corresponding PLDA classifier described previously. As noted in the discussion of the PLDA results, the seemingly low ammonia detection percentages are a reflection of the weak ammonia signals generated in the data synthesis procedure.

4.3.7 Performance of SVM classifiers with prediction sets

Each classifier listed in Table 4.4 was applied to all four prediction sets. The SVM output scores were plotted for the unsorted prediction sets, producing score profiles that were similar in terms of shape but different in magnitude. This plot orientation can be considered a time trace of the data collection, thereby corresponding to the flight track of the aircraft. Classifier 3 in Table 4.4, with $\gamma = 1.28$, $C = 8.7\times10^4$, produced SVM score profiles with the largest gaps between active and inactive patterns. This classifier was selected for further study.

The SVM score profiles produced by Classifier 3 for the four prediction sets are shown in Figure 4.8. The SVM score is the output of the SVM linear discriminant function in the feature space for an input pattern. The conventional classification rule is that a positive SVM score

Table 4.4 Top SVM classifiers

Model #	Kernel parameter (γ)	Regularization parameter (C)	Number of support vectors	Trn. missed rate (%)	Trn. false rate (%)	Mon. missed rate (%)	Mon. false rate (%)	Trn. rate difference (%)	Mon. rate difference (%)
1	10.24	3.0×10^{3}	2112	24.8	0.38	32.8	1.7	24.4	31.0
2	2.56	3.9×10^{4}	2119	25.1	0.39	32.8	1.7	24.7	31.1
3	1.28	8.7×10^{4}	2128	25.2	0.46	32.9	1.8	24.8	31.1
4	2.56	2.7×10^{4}	2123	25.2	0.42	32.9	1.7	24.8	31.2
5	2.56	1.5×10^{4}	2144	25.6	0.42	33.1	1.7	25.2	31.4

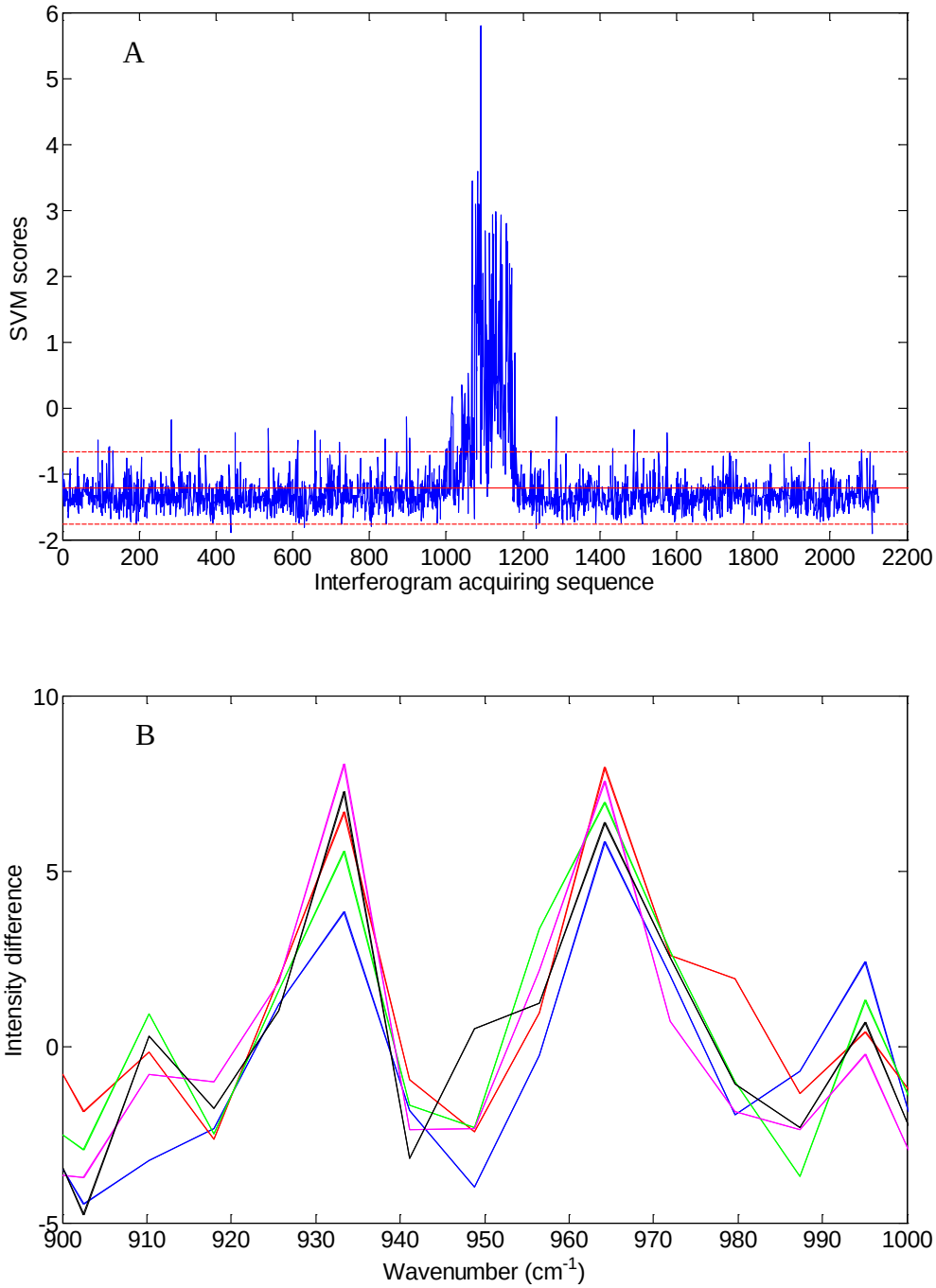
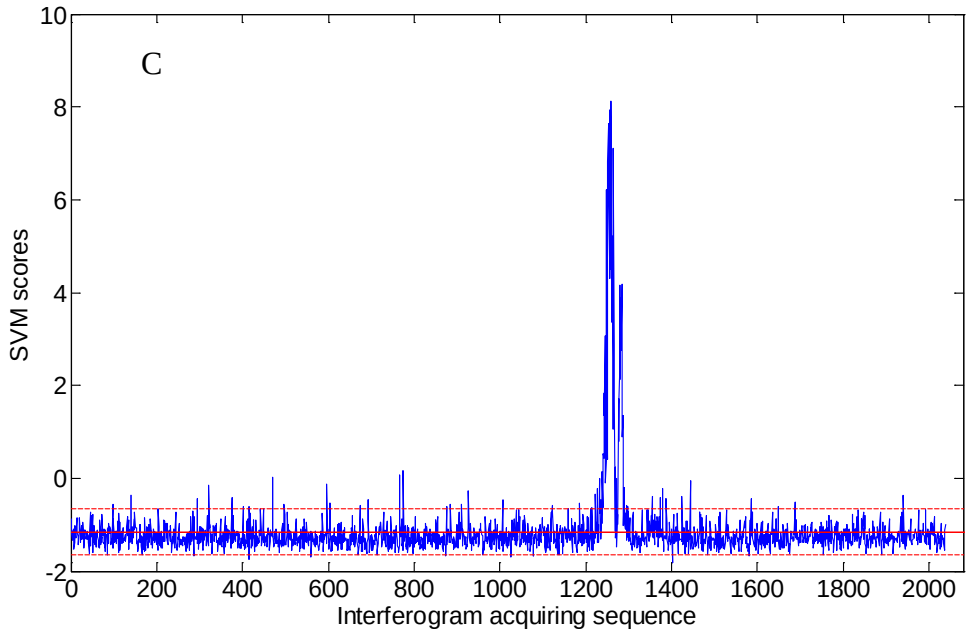
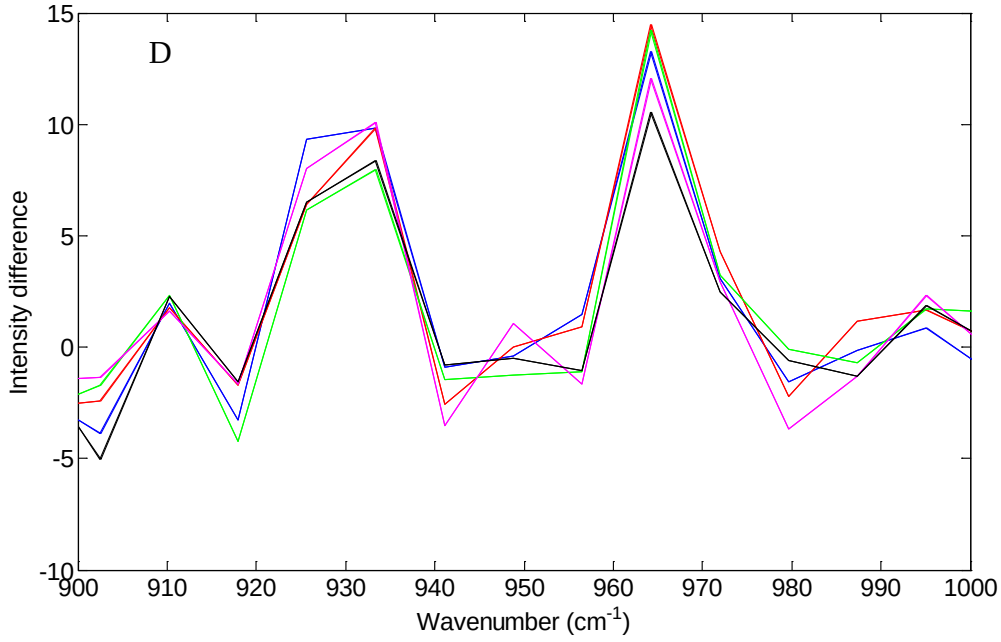


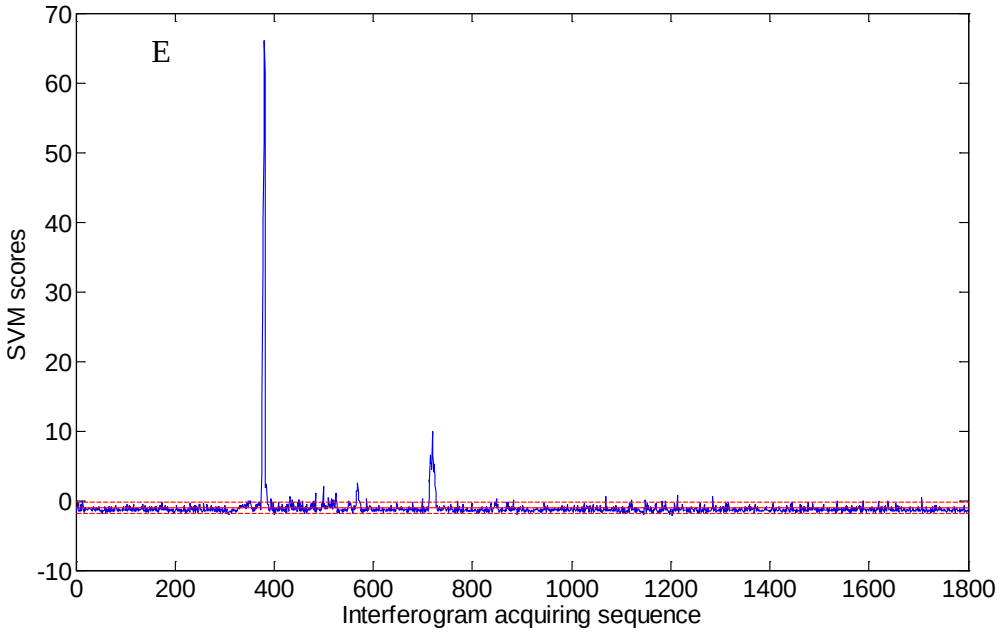
Figure 4.8 Score profiles obtained from the optimal SVM classifiers and corresponding difference spectra at the score maximum for the four prediction sets. (A) Score profile with calculated control limits for prediction set *A*. (B) Difference spectra corresponding to interferograms #1067, #1077, #1081, #1089, and #1091. (C) Score profile with calculated control limits for prediction set *B*. (D) Difference spectra corresponding to interferograms #1253, #1256, #1258, #1260, and #1265. (E) Score profile with calculated control limits for prediction set *C*. (F) Difference spectra corresponding to interferograms #378, #379, #380, #381, and #382.

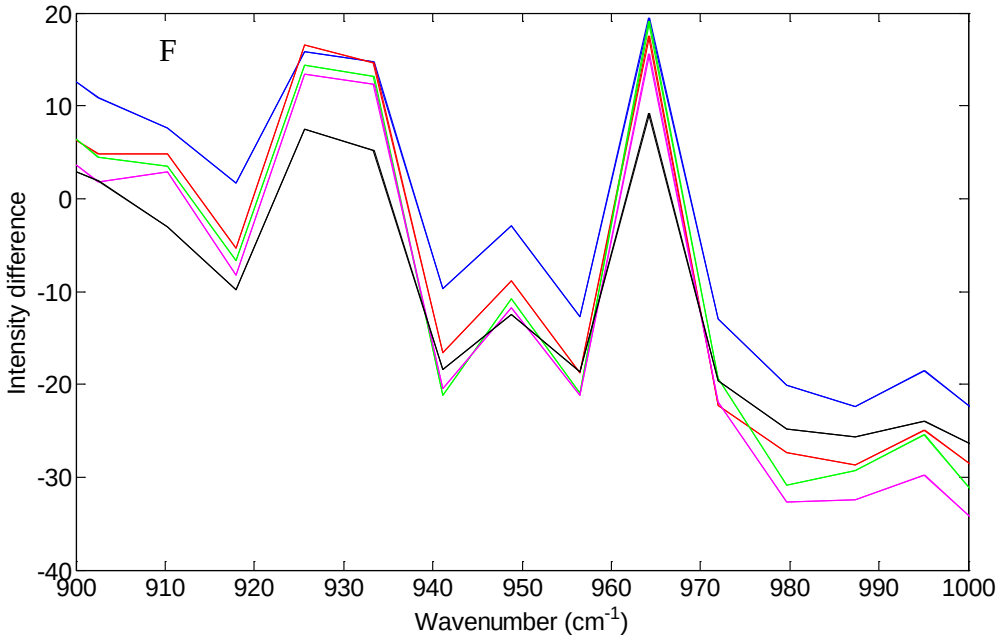
(G) Score profile with calculated control limits for prediction set *D*. (H) Difference spectra corresponding to interferograms #375, #376, #378, #379, and #380. Plotted control limits for each run are based on the mean SVM score for the run.



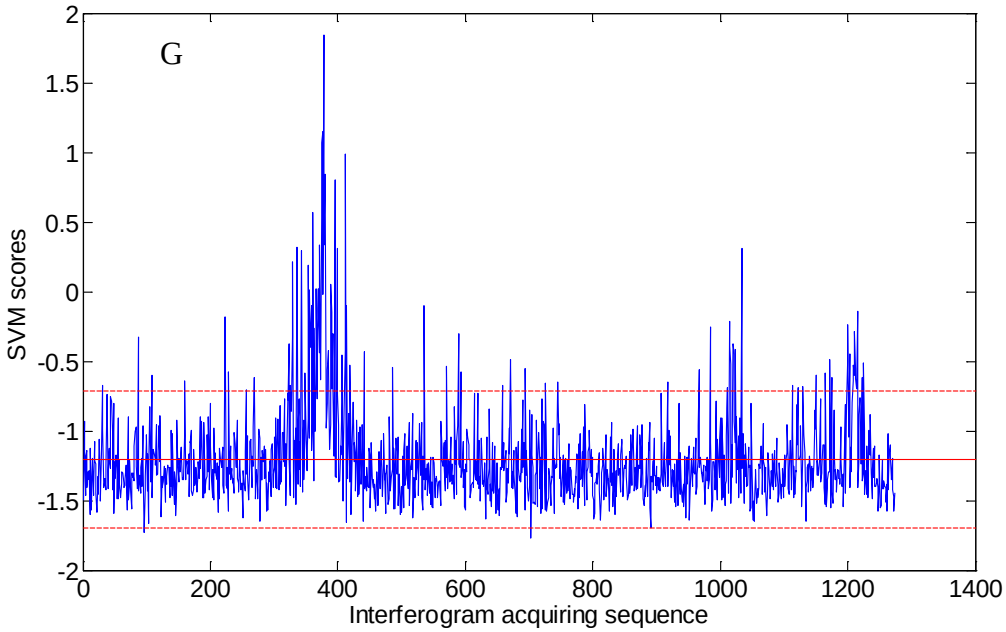


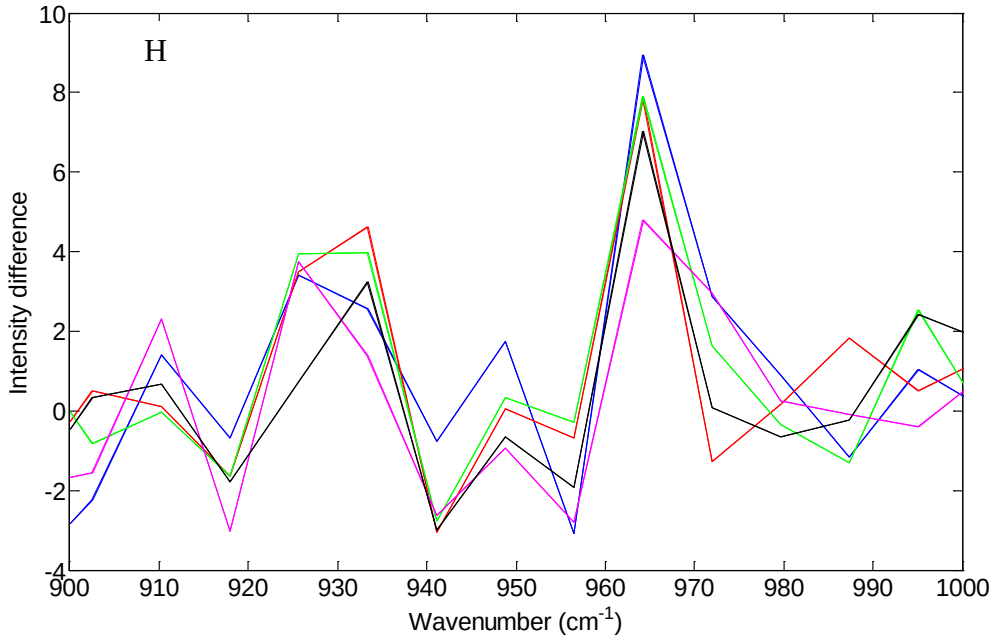
(Figure 4.8 continued)





(Figure 4.8 continued)





(Figure 4.8 continued)

signifies the analyte-active class, while a negative score specifies the analyte-inactive class. The classification results for the prediction sets based on the conventional classification rule are summarized in Table 4.5.

For prediction set *A*, the maximum values in the SVM score plot suggested that ammonia was present over the interferogram number range from 1000 to 1200 (Figure 4.8, A). Difference spectra were then calculated for several interfeorgrams in this range. The ammonia doublet signature was found in all of these difference spectra (Figure 4.8, B), which confirmed the finding from the SVM score plot.

Though prediction sets *B*, *C*, and *D* (Figure 4.8, C-H) had scores on a different scale, the overall pattern was the same as that observed for prediction set *A*. Whenever the scores rose to a maximum, implying the presence of ammonia, difference spectra verified the findings. A minor variation was observed in prediction set *C*, where the ammonia feature was found in the difference spectra with the baseline slightly shifted. This suggests difficulty in establishing a stable background spectrum for use in the calculation of the difference spectra.

One implementation issue observed with the prediction sets was that the conventional SVM classification rule did not necessarily produce the optimal classification results. Because the SVM classifier was developed with synthetic ammonia-active data and then applied to field data collected at a different time and location than the data used to train the classifier, a slight offset in the SVM scores of the prediction set was induced. The presence of this offset suggests that tuning the classification threshold for the detection of ammonia to a value derived from the prediction set itself may provide benefits in improving the accuracy of the classifications.

Table 4.5 Summary of classification results for prediction sets based on standard SVM classification

Model #	Kernel parameter (γ)	Regularization parameter (C)	No. of support vectors	Prediction set *A*		Prediction set *B*		Prediction set *C*		Prediction set *D*	
				miss rate (%)	false rate (%)	miss rate (%)	false rate (%)	miss rate (%)	false rate (%)	miss rate (%)	false rate (%)
1	10.24	3.0×10^{3}	2112	36.3	0.05	25.9	0.10	0.00	2.40	57.1	0.09
2	2.56	3.9×10^{4}	2119	37.3	0.05	25.9	0.10	0.00	2.46	61.9	0.09
3	1.28	8.7×10^{4}	2128	36.3	0.05	25.9	0.10	0.00	2.63	61.9	0.09
4	2.56	2.7×10^{4}	2123	35.3	0.05	25.9	0.16	0.00	2.52	57.1	0.18
5	2.56	1.5×10^{4}	2144	36.3	0.05	25.9	0.10	0.00	2.40	61.9	0.09

The nature of the airborne monitoring experiment produces a time-based profile of SVM output scores that is somewhat analogous to profiles encountered in statistical process control. Because the spectrometer has both a narrow FOV and a high scan rate, detections of analyte plumes being released from a ground source will necessarily be narrow in time. Stated differently, analyte detections will be rare events superimposed on a background trace in which the analyte signature is not present. This is analogous to a process control setting in which the process is typically within specification and the monitoring goal is to detect the occurrence of extreme events that signal deviation from the norm.

Control charts are a standard tool used in statistical process control to detect extreme events that signal when a process is out of specification.[87] In this work, control charts based on the SVM score profiles were developed for the purpose of automating the detection decisions and tuning the classification threshold to the current prediction set. For each prediction set, the mean SVM score, median SVM score, the average moving range, and the estimate of the standard deviation of the average moving range of the score profile were calculated. The level of the mean or median SVM score was viewed as the basis for separating the binary data classes. An upper control limit (UCL) and a lower control limit (LCL) were calculated as two times the standard deviation of the average moving range above or below the mean or median SVM score.

As noted above, the nature of the airborne passive FT-IR measurements dictates that for a single flight track, the vast majority of data will belong to the ammonia-inactive class, with at most a very small number of ammonia-active data points. Thereby it is legitimate to assume that, if correctly classified, a data point from the ammonia-inactive class will have an SVM score within the range between LCL and UCL, while data from the ammonia-active class will be expected to have an SVM score higher than the UCL. For a given classifier, if it generates an

SVM score lower than the UCL for a confirmed active data point, it is considered a case of missed detection. Similarly, if the classifier produces an SVM score higher than the UCL for a confirmed inactive data point, this case is counted as a false detection. In the case of an inactive data point with an SVM score lower than the LCL, it is considered an outlier and not subject to classification. Control chart settings, including the mean SVM score, LCL and UCL (computed relative to the mean) are overlaid on Figure 4.8, A, C, E, and G. All classification results are summarized in Table 4.6. Results are provided for both the use of the mean and median SVM score as the base response. Impressively, the missed detection rates for all four prediction sets were found to be 43 – 87 % lower than those obtained with the conventional SVM classification (Table 4.5) when the mean score was used as the baseline response. Missed detection rates were even lower when the median SVM score was used as the baseline response. False detection rates did increase relative to the conventional classification rule (e.g., 1.4 to 3.8 %), and further increased false detection rates were seen with the control limits computed on the basis of the median SVM score. Most of the false detections were single interferograms, however, and could be eliminated by a simple requirement that two consecutive positive detections are required for an alarm decision. The results subject to this rule are summarized in Table 4.7. Overall, however, the performance of the SVM classifiers with the external prediction sets was excellent and proved the robustness of the classification methodology based on synthesized data.

4.4. Conclusions

This research demonstrated that, with an appropriate interferogram simulation protocol and signal processing strategy, a classifier, developed from supervised pattern recognition algorithms with selected short interferogram segments as input patterns, was able to recognize the signature of ammonia in airborne passive infrared data with low rates of false detections.

Table 4.6 Summary of classification results for prediction sets based on process control adjusted SVM

Prediction sets	Mean SVM	Estimated standard deviation	LCL	UCL	Missed detection rate (%)	False detection rate (%)
A	-1.21	0.27	-1.76	-0.67	7.8	1.45
B	-1.15	0.25	-1.65	-0.66	14.8	1.69
C	-1.01	0.40	-1.81	-0.22	0.0	3.75
D	-1.21	0.24	-1.69	-0.72	19.0	3.82
Prediction sets	Median SVM	Estimated standard deviation	LCL	UCL	Missed detection rate (%)	False detection rate (%)
A	-1.35	0.27	-1.90	-0.80	4.9	2.79
B	-1.28	0.25	-1.77	-0.78	11.1	3.53
C	-1.32	0.40	-2.12	-0.53	0.0	6.54
D	-1.29	0.24	-1.78	-0.80	14.3	4.88

Table 4.7 Summary of classification results for prediction sets based on process control adjusted SVM subject to the rule of two-consecutive positive detections

Prediction sets	Mean SVM	Estimated standard deviation	LCL	UCL	Missed detection rate (%)	False detection rate (%)
A	-1.21	0.27	-1.76	-0.67	23.53	0.155
B	-1.15	0.25	-1.65	-0.66	27.78	0.00
C	-1.01	0.40	-1.81	-0.22	12.50	1.29
D	-1.21	0.24	-1.69	-0.72	28.57	0.799
Prediction sets	Median SVM	Estimated standard deviation	LCL	UCL	Missed detection rate (%)	False detection rate (%)
A	-1.35	0.27	-1.90	-0.80	21.57	0.155
B	-1.28	0.25	-1.77	-0.78	24.07	5.27×10^{-2}
C	-1.32	0.40	-2.12	-0.53	12.50	1.73
D	-1.29	0.24	-1.78	-0.80	19.05	1.07

The overall methodology could be transferred to an application scenario in which automated monitoring is required for continuously detecting the presence of a specific chemical species in the atmosphere. Direct analysis of interferograms to extract analyte features is specifically designed for data collected in passive remote sensing sessions in which computing spectra in absorbance/transmittance units is not realistic due to the lack of a reliable and constant spectral background.

The interferogram simulation protocol takes advantage of the availability of a large pool of IR background data and uses these data to synthesize as much analyte-active data as needed for the development of classification models. The use of a laboratory reference spectrum for the analyte in the data synthesis procedure allows classification models to be developed for any compound without the need for an analyte-specific data collection effort.

In this work, ANOVA studies were used to assess the importance of adjustable parameters related to interferogram segment selection and digital filter design. The ANOVA of a full factorial experimental design revealed that segment-related factors, including segment length, segment starting point, and the interaction between the two segment factors, played the most significant roles with respect to classification performance on the monitoring set. The optimal choice of a 120-point segment starting from point 88 was consistent with previous applications of the interferogram-based data analysis methodology.[75,76]

For the detection of ammonia, the design of the IIR filters seemed to have relatively little effect on the classification performance. This may be attributed to the unique narrow doublet signature of ammonia located in the mid-IR region which produces a unique interferogram signature and is thus not significantly affected by the presence in the interferogram of contributions from other spectral frequencies. Because other analytes with less distinct

bandshapes may be affected by contributions from frequencies outside the region of the band, the general strategy of studying the segment and filtering parameters together is still considered important.

The PLDA and SVM methods were the two supervised pattern recognition algorithms employed in this work. The PLDA method was used for the grid search of optimal conditions for the interferogram segment and digital filter. Good classification power achieved with relatively simple computations makes the PLDA method suitable for this task. With its capacity to deal with more complex pattern recognition situations but requiring extensive study of two configuration parameters, the SVM technique is not well suited to an experimental design in which a variety of other parameters are being studied. The combined use of PLDA and SVM was a workable compromise demonstrated through this research. Comparison of the classification performance of the final SVM and PLDA classifiers under the same set of segment and filter conditions on the monitoring set confirmed the better selectivity and sensitivity of SVM as a pattern recognition algorithm.

For the first time in passive remote sensing research, statistical process control methods were used in conjunction with the SVM classifier to implement classification rules for signaling an analyte detection. Control charts generated from the profile of SVM scores were used to define UCL and LCL bounds tuned to each data set. Excellent classification performance, resulting from comparison of SVM scores against the UCL or LCL calculated specifically for each data collection session proved the robustness of the ammonia classifier when applied to data outside of the training set.

For future work, to further improve classifier performance, such as the ability to detect ammonia with greater variation or the capacity to detect ammonia at lower concentrations, more

consideration related to the training set design and model selection process is needed. This is the focus of the work described in Chapter 5.

Chapter 5

CALIBRATION STRATEGIES FOR THE AUTOMATED DETECTION OF AMMONIA WITH LOWER LIMITS OF DETECTION BASED ON PASSIVE FOURIER TRANSFORM INFRARED SPECTROMETRY

5.1 Introduction

For several decades, open-path Fourier transform infrared (OP-FT-IR) spectrometry has proven to be a reliable remote sensing measurement technique by providing fast, real-time, simple-to-use identification or quantification of trace gaseous agents in open environments. [33,34,50,88-92] Different than normal laboratory FT-IR air monitoring where a physical sample cell is used to contain an air sample, in this approach, during a data measurement in the field, the incident light from the light source is passed through a path, typically > 100 m in length, that crosses the location of interest. If the target analyte appears in the optical path, its corresponding spectral signature will be registered in the radiance arriving at the IR detector.

Depending on whether an artificial light source is used, OP-FTIR measurement techniques can be divided into two categories: (1) the active measurement technique, where a manmade IR light source is used in conjunction with the spectrometer, [89,91,93] and the passive measurement approach, where naturally occurring ambient IR radiance is used as the light source. [35,84,94] Active OP-FTIR, quite similar to the classical setup of a laboratory FT-IR spectrometer except with a much longer optical path length, is able to provide data with quality close to laboratory data. Among OP-FT-IR measurements, this technique is most commonly applied, especially when low limits of detection (LOD) or high sensitivity are desired. By comparison, passive OP-FT-IR, conceptually more close to the core idea of remote sensing, has advantages over the active technique to provide mobile and fast responses for real-time *in situ* applications. Because there is no requirement to place the source or a retroreflector on the opposite side of the sampling path being monitored, the passive measurement is compatible with

mounting on a moving platform such as an aircraft. In such an implementation, the spectrometer would typically be mounted in a downward-looking orientation and the upwelling ground radiance would serve as the IR background.

One of the principal drawbacks of the passive measurement is the inherently unstable nature of the IR background radiance. This is especially true when the spectrometer is mounted on a moving platform. Given this continuous change in the ambient IR source radiance and the relatively small dynamic range of temperature contrast between the background and any atmospheric constituents within the field-of-view (FOV), passive OP-FT-IR is considered less sensitive and typically has higher detection limits than active OP-FT-IR. This intrinsic limitation cannot be overcome by any improvement in the physical sensor configuration but can be mitigated to some extent through application of sophisticated data handling methods.

In our laboratory we have developed a number of data analysis capabilities for use with passive OP-FT-IR remote sensing data. [50,82,95,96] Starting with passive FT-IR data, a variety of data mining tools, including signal processing, pattern recognition, and multivariate modeling techniques have been employed to develop automated classifiers for compound recognition or to provide quantitative information for estimating analyte amounts. The work presented here is a further refinement of these capabilities.

In Chapter 4, with ambient ammonia vapor as the target analyte, by employing laboratory pure-component based interferogram simulation, signal extraction directly from the interferogram, and a model selection process driven by PLDA, an automated SVM classifier was developed, validated, and tested successfully with passive IR data collected with an airborne spectrometer. Based upon this established ammonia classification methodology, the effort in the

follow-on work described here focuses on improving the detection limit of the SVM classifier through optimization of the training data sets used in its formulation.

With respect to supervised classification in which an iterative process is used to optimize a classification model on the basis of a set of training patterns, it was hypothesized that a key to lowering the LOD of the classifier is to optimize the composition of the training set. In this work, several training sets with simulated interferograms representing ammonia-active patterns with various concentrations were assembled and evaluated in terms of their fitting capacity. The selected training sets were subsequently used to develop SVM classification models. The detection limits of the final selected SVM classifiers were estimated by applying the computed classifiers to another group of simulated ammonia-active data. A variety of comparisons between the training sets designed in this work and the training set used in the previous work were applied in order to investigate the relationships between the distribution of the training data and the achieved detection limits.

Another emerging issue related to SVM classification is how to interpret the SVM scores, the output of the SVM discriminant function, for implementation in field applications. In the previous chapter, we explored the limitations of the traditional SVM classification rule (positive scores indicating one class and negative scores indicating the opposite class). Essentially, with this classification rule, the SVM scores can be deemed uncalibrated and thus have no probabilistic interpretation. This can lead to misleading or suspect classifications in real applications.

In Chapter 4, a control chart analysis was proposed and tested with SVM scores obtained from true external predication data. An in-depth comparison between the traditional SVM

classification rule and the control chart method is described here in combination with the assessment of training set composition.

5.2 Experimental

5.2.1 Instrumentation and data collection

As described in Chapter 4, all airborne interferogram data used in this work were collected by our research collaborators at the United States Environmental Protection Agency. A Bomem Model MR254/AB spectrometer (ABB Bomem, Quebec City, Canada) was employed to acquire passive FT-IR interferograms. The spectrometer was equipped with a liquid-nitrogen cooled Hg:Cd:Te (MCT) detector operating over the 800 – 1200 cm^{-1} (MWIR) region. The FOV of the spectrometer was restricted to 0.3^{o} with a Cassegrain telescope. The resulting raw interferograms were double-sided single-scans containing 512 points. The spectral bandwidth was 1975 cm^{-1}, corresponding to sampling at every eighth zero-crossing of the HeNe reference laser. Forman phase correction using 64 points on each side of the centerburst of the raw interferogram and normalization to a vector magnitude of unity were applied prior to further processing.

To collect airborne data, the spectrometer mentioned above was mounted in a downward-looking orientation on a twin-engine high-wing aircraft (Aero Commander 680 FL, Aero Commander, Culver City, CA). The aircraft flew at altitudes of 2000 – 3000 ft. with a cruising speed of 100 – 150 knots. Open-air passive-mode interferograms were acquired at an approximate rate of 80 interferograms per second while the aircraft was flying over ground targets of interest. Four airborne remote sensing experiments (*A*, *B*, *C* and *D*) were conducted at two different sites. Three data sets, *A*, *B*, and *D*, were collected in the Houston, Texas area at three different times during the aftermath of Hurricane Rita in September 2005. Session *C* was collected in February 2005 in Missouri where emission of an unknown concentration of

ammonia was reported. All of these airborne data sets served as external prediction sets for use in testing the SVM classifiers.

A database of >500,000 passive IR background interferograms acquired with the same instrument at a variety of times and locations was also used in this work to represent ammonia-inactive data. These interferograms were also used as part of the synthesis procedure to generate ammonia-active data.

5.2.2 Data analysis implementation

Digital filters were designed with the aid of the Filter Design and Analysis Tool (FDATool) provided with the Signal Processing Toolbox (Version 5) of Matlab Version 7.4 (The MathWorks, Natick, MA). The SVM classifiers were trained and tested with the public-domain package, SVM^{light} (Version 6.01, http://svmlight.joachims.org). The K-nearest neighbor (KNN) algorithm was implemented via in-house software. Digital filtering, SVM classification and all other data analysis tasks were performed on a Dell Precision 490 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Version 5.2, Red Hat, Inc., Raleigh, NC).

5.3 Results and Discussion

5.3.1 Overview of Methodology

In the work described in Chapter 4, a comprehensive set of spectral processing and pattern recognition tools was utilized in order to develop an automated classification model for airborne ammonia monitoring. The excellent classification results for four airborne test sets produced by the optimized classifier confirmed the feasibility and effectiveness of the entire methodology. However, as noted in the previous chapter, the absolute classification percentages for the ammonia-active patterns obtained with the training and monitoring data were relatively low (e.g., 30% missed detections). Because the ammonia signal levels were generated by a data

synthesis procedure, the signal strengths and distribution of signal strengths within the training set can be set to any level desired. The key question, however, is what signal levels and distribution in the training set produce the best-performing classifiers.

This situation can be illustrated by application of principal component analysis [97,98] (PCA) to the training set used in Chapter 4 and subsequent inspection of plots based on the principal component (PC) scores. As described in Chapter 3, PCA is a multivariate modeling method that attempts to remove redundant information from a data matrix and thereby model the data with fewer dimensions than originally used. In this way, high-dimensional data can often be visualized effectively in plots based on two or three dimensions.

As shown in Figure 5.1, PCA was implemented with either the entire raw interferogram (Figure 5.1, A) or with patterns based upon optimally preprocessed interferogram segments (Figure 5.1, B) that employed the preprocessing parameters selected in Chapter 4. In both cases, the data matrix was mean-centered prior to application of PCA. For plots A and B, 77.1 and 96.4 % of the data variance was explained by two PCs. The data submitted to the PCA calculation included all synthetic active patterns (green), 25 % of the airborne inactive patterns (black) uniformly chosen from the training set, confirmed airborne ammonia-active patterns from prediction sets *A*, *B*, *C*, and *D* (red), and 25 % of the airborne inactive patterns (blue) from prediction sets *A*, *B*, *C*, and *D*).

Comparing the two figures, it is clear that signal preprocessing, including interferogram segment selection and digital filtering, helps to remove irrelevant information in the raw

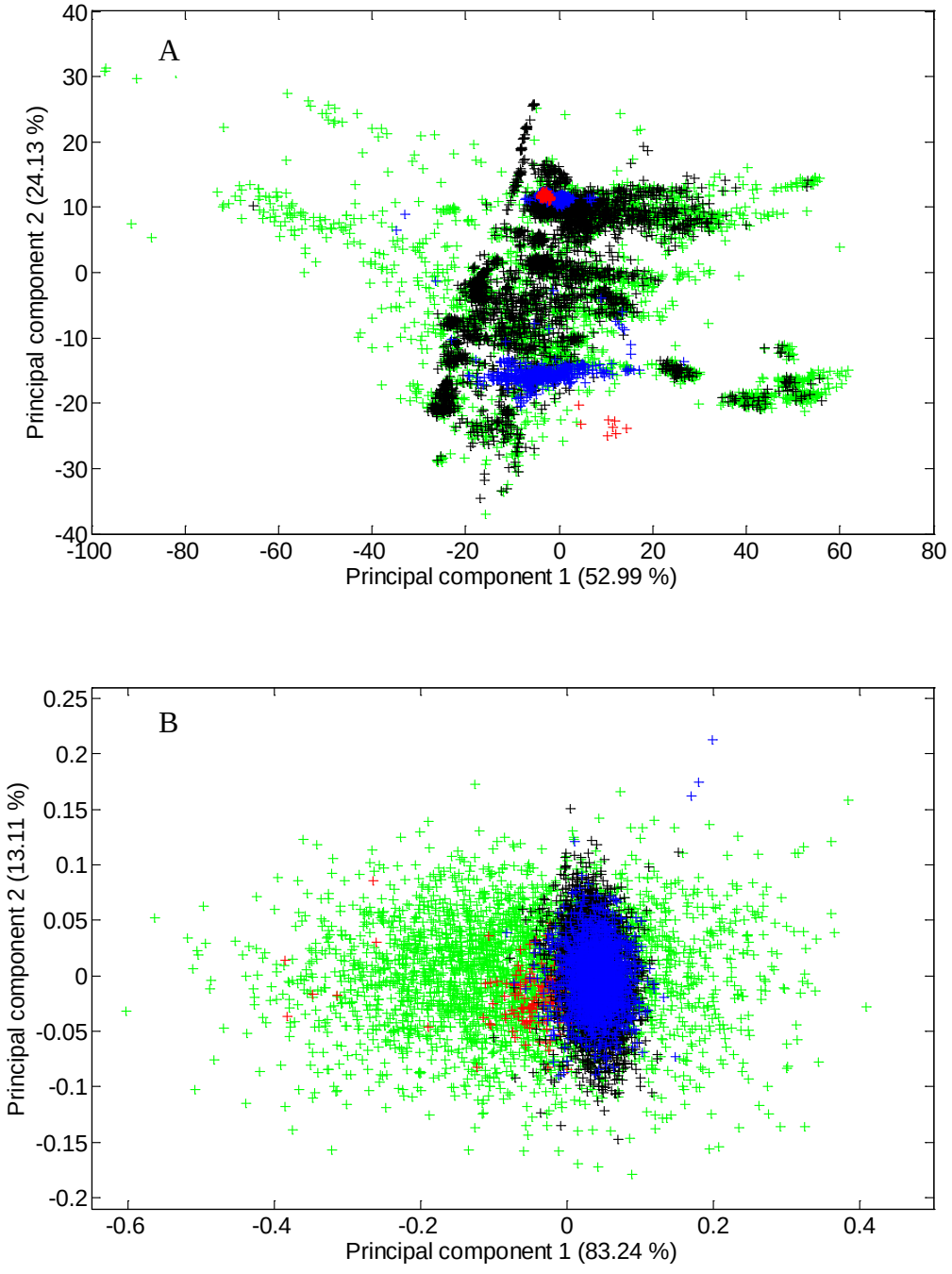


Figure 5.1 Principal component score plots illustrating the degree of clustering and separation among different pattern groups. Data plotted include airborne ammonia-active patterns in the combined prediction set (red), airborne ammonia-inactive patterns in the combined prediction set (blue), synthetic ammonia-active patterns in the training set used in Chapter 4 (green), and ammonia-inactive patterns in the same training set (black). The entire raw interferogram was used to generate plot A, while the optimal preprocessing parameters developed in Chapter 4 were used in plot B. A 120-point interferogram segment starting at point 88 (relative to the

centerburst) was used. The digital filter used was a Chebyshev Type II IIR filter centered at 964 cm^{-1}, with a passband FWHM of 28 cm^{-1} and a stopband attenuation of 40 dB.

interfergram (Figure 5.1, A) and extract the information most related to discrimination of the ammonia-active and ammonia-inactive data classes (Figure 5.1, B). As illustrated by the PC scores (Figure 5.1, B), within the combined prediction set, the active patterns (red) and inactive patterns (blue) were separable but very marginally. Across the data sets, the inactive patterns (black and blue) in both the training set and the combined prediction set overlapped closely, indicating the training patterns are globally representative.

The major issue here arises from the synthetic active patterns (green) comprising the training set. These patterns cover not only the airborne active patterns (red) in the combined prediction set, but also are highly overlapped with the inactive patterns from both the training and prediction sets. This illustrates the reason the class separation percentages were relatively low in the work reported in Chapter 4 and also establishes the basis for exploring the impact of the class distribution on the effectiveness of the trained classifiers. Stated differently, is vagueness in the boundary of the two data classes desirable or problematic for the SVM modeling procedure and its ability to produce a classifier sensitive to the analyte while being resistant to false detections?

In order to develop a robust classifier that achieves a low LOD, one strategy is to have active patterns representing concentrations as low as possible included in the training set, while attempting to control or refine the boundary between the analyte-active and analyte-inactive data classes. In this work, a new protocol to assemble the training set in a quantitatively controlled fashion was proposed by drawing the active patterns from a pool of synthetic spectral data representing a certain range of low concentration ammonia. Non-iterative pattern recognition methods, including PCA and KNN analysis,[99] were used to monitor the classification boundary and therefore to select optimal training sets.

In this work, a monitoring set was designed to include synthetic ammonia patterns of an array of concentrations reflecting the LOD expected from the developed classifier. Once the training sets were selected and the monitoring set was designed, SVM classifiers with varying kernel parameters (γ) and regularization parameters (C) were developed. The optimal SVM classifiers were selected by performance on the training and monitoring sets, and were subsequently applied to the four airborne prediction sets to evaluate performance with field data outside of the training set. The LOD for each classifier could be estimated through its classification results for the qualitatively designed active patterns in the monitoring set.

5.3.2 Design and selection of training sets

The procedure used to generate the synthetic ammonia-active spectra remained the same as previously described in Chapter 4, but here the scaling factors used to simulate spectra of various ammonia concentrations were drawn from a specifically designed pool of numbers. The whole pool consisted of 50 bins that corresponded to equal steps within the range of -6 to +4 ppm-m, where negative and positive concentrations specify emission and absorption signals, respectively. Each bin was composed of 60 normally distributed random numbers. The numbers in each bin had a standard deviation of 0.2, while the mean varied from -5.9 to +3.9 with an increment of 0.2.

To simulate airborne ammonia spectra with various concentrations, the scaling factors based upon this number pool were multiplied by the deresolved 1 ppm-m reference ammonia spectrum employed previously in Chapter 4. These scaled ammonia spectra were then superimposed onto randomly chosen airborne ammonia-inactive spectral profiles. As a result, 3000 synthetic ammonia spectra were generated simulating both emission and absorption spectra. These spectra covered the concentration range from 3.5×10^{-2} ppm-m to 6.0 ppm-m for emission and 3.6×10^{-2} ppm-m to 4.0 ppm-m for absorption.

The concentration distribution of the overall pool is illustrated in Figure 5.2. It should be noted that the signal strengths corresponding to these concentrations represent the maximum possible levels that could be obtained in a remote sensing measurement. Effectively, these signal levels correspond to a situation in which the temperature differential between the analyte and background is large enough to have no suppressing effect. Because emission signals necessarily arise from hot gases, the temperature differential is typically higher relative to the background than for the case of absorption data. For this reason, the emission signal levels in the synthetic data are more accurately correlated with their concentration labels than are the absorption signals.

The concentration distribution depicted in Figure 5.2 also reflects bias toward emission spectra. In remote sensing applications where hot gases are monitored, emission occurs more frequently than absorption. An example of this monitoring scenario would be the detection of stack emissions. To maximize performance for these applications, the population of spectra was skewed toward emission, including 1800 emission spectra and 1200 absorption spectra with various concentrations.

Once the overall pool of synthetic ammonia spectra was established, subsets with different concentration ranges were drawn from this pool. The intrinsic capacity for binary classification was estimated for each subset by the KNN method. In the KNN classifications, class memberships for each pattern were assigned on the basis of the majority classifications of its five nearest neighbors. Patterns submitted to the KNN calculations were based on the optimal segment and filtering parameters as determined in Chapter 4.

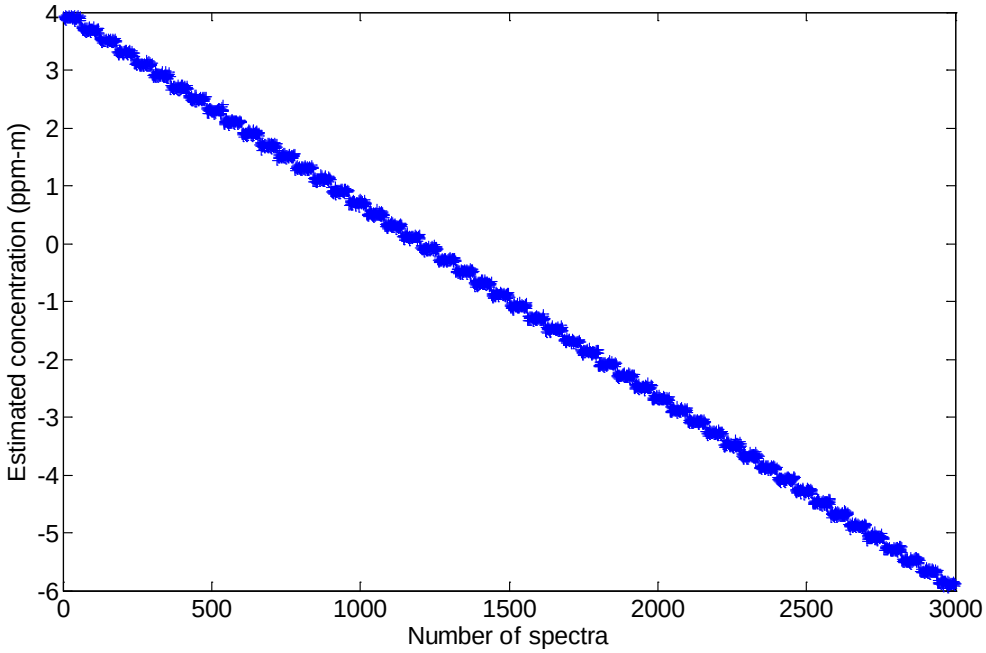


Figure 5.2 Concentration profile of synthetic ammonia-active interferograms designed for the training set.

Table 5.1 indicates the composition, the designed concentration limits, and the KNN results for six proposed training sets (*Trn_1*, *Trn_2*, …, *Trn_6*) assembled from the synthesized ammonia-active and field ammonia-inactive data. In addition, the table includes the corresponding results for the training set used in Chapter 4 (*Trn_0*) and the combined data (*Prd*) from prediction sets *A*, *B*, *C*, and *D*. The six new training sets included the same 12,000 airborne interferograms in the ammonia-inactive class used previously in Chapter 4, but contained different numbers of active interferograms drawn from the pool of 3000 synthetic ammonia interferograms described above.

Training set *Trn_1* included interferograms derived from all 3000 synthetic active spectra from the pool, corresponding to ammonia concentrations as low as 10^{-2} ppm-m for both emission and absorption cases. Training sets *Trn_2*, *Trn_3* and *Trn_4* contained 2400, 1800, and 1200 active patterns, corresponding to minimum ammonia concentrations of 1.0, 2.0, and 3.0 ppm-m for both emission and absorption cases. Training sets *Trn_5* and *Trn_6* both contained 900 active patterns, but the former had an emission lower limit of 4.0 ppm-m and an absorption lower limit of 3.0 ppm-m, while the latter had a lower limit of ~ 3.5 ppm-m for both emission and absorption cases.

The combined prediction set, *Prd*, totaled 6933 interferograms collected from four different airborne data collection sessions, and was classified visually to have 185 ammonia-active patterns and 6748 non-ammonia patterns. Interferograms with indeterminate classifications based on visual inspections were not used. The KNN results showed that 6720 out

Table 5.1 Data partitioning and composition

Data sets	Data composition			KNN results			
	No. inactive patterns	No. active patterns	Designed lowest NH_3 level [1] (ppm-m)	No. inactive patterns detected	Inactive detection rate (%)	No. active patterns detected	Active detection rate (%)
Prd [2]	6748	185	NA	6720	99.59	125	67.57
Trn_0 [3]	12,000	3000	NA	11,806	98.38	2232	74.40
Trn_1	12,000	3000	-3.49×10^{-2} (3.60×10^{-2})	11,571	96.43	1449	48.30
Trn_2	12,000	2400	-1.03 (1.03)	11,661	97.18	1331	55.46
Trn_3	12,000	1800	-2.03 (2.04)	11,773	98.11	1143	63.50
Trn_4	12,000	1200	-3.03 (3.05)	11,882	99.02	869	72.43
Trn_5	12,000	900	-4.04 (3.05)	11,928	99.40	716	79.56
Trn_6	12,000	900	-3.62 (3.44)	11,941	99.51	738	82.00

[1] Corresponds to the lowest concentration of the pure ammonia absorbance spectrum used for spectral simulation. The (+) sign in front of the concentration value indicates an absorption band is superimposed, while the (-) sign indicates an ammonia emission band.
[2] Corresponds to a combined prediction set including airborne data from the four individual prediction sets (*A*, *B*, *C*, *D*).
[3] *Trn 0* denotes the training set formed for the work described in Chapter 4, while *Trn_1* to *Trn_6* specify the six training sets designed in this chapter.

of 6748 patterns in the ammonia-inactive class could be correctly classified (99.6 %), while 125 out of 185 patterns in the ammonia-active class could be correctly classified (67.6 %). The KNN results for *Trn_0*, the training set used in the work described in Chapter 4, were correct classification of 11,806 out of 12,000 patterns in the ammonia-inactive class (98.4 %) and 2232 out of 3000 patterns in the ammonia-active class (74.4 %). These classification rates serve as a baseline for comparison to the results obtained with the new training sets investigated in this chapter.

Inspection of Table 5.1 reveals that, as expected, the ability to recognize both the ammonia-active and ammonia-inactive patterns improves as the ammonia signal strengths increase. The recognition of both classes improves because the increased ammonia signals make the classes more distinct and thus the nearest neighbors in the data space will more likely be from the same class. Even at the highest ammonia levels used here, however, there is still significant class overlap as evidenced by only ~80 % correct classifications for the ammonia-active patterns in *Trn*_5 and *Trn*_6.

Figure 5.3 displays score plots generated from the application of PCA to the preprocessed interferogram data from training sets *Trn_3* (A) and *Trn_4* (B). The first two principal components accounted for 88.0 and 87.8 % of the data variance in *Trn_3* and *Trn_4*, respectively, and were used in generating the score plots. As before, the data were mean-centered before application of PCA. The color scheme used in Figure 5.3 is the same as that employed in

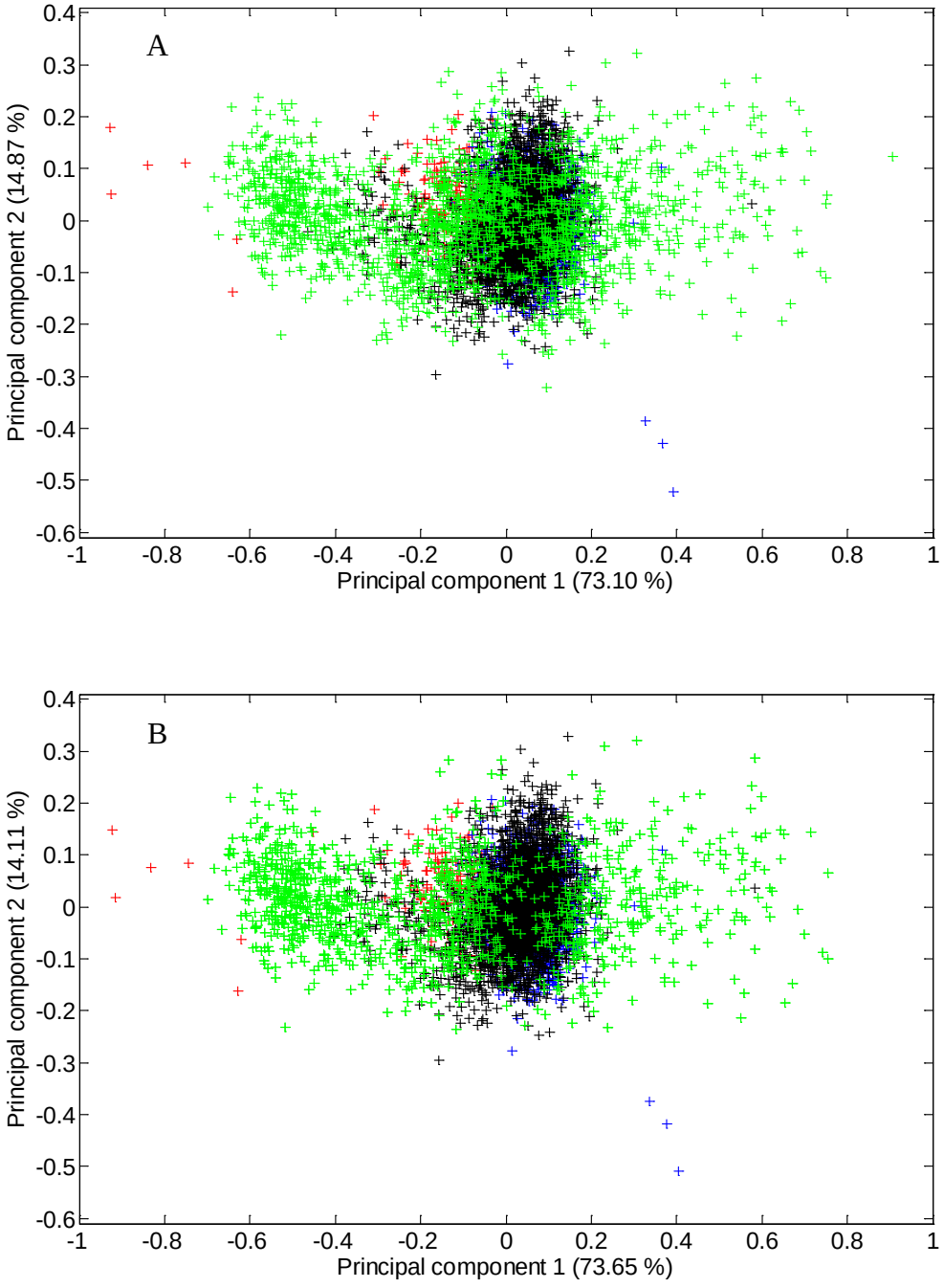


Figure 5.3 Principal component score plots illustrating the degree of clustering and separation among different pattern groups. Data plotted include airborne ammonia-active patterns in the combined prediction set (red), airborne ammonia-inactive patterns in the combined prediction set (blue), synthetic ammonia-active patterns in the training set (green), and ammonia-inactive patterns in the training set (black). Interferograms were preprocessed with the optimal segment and digital filtering parameters determined in Chapter 4. Plots A and B correspond to the data from training sets *Trn_3* and *Trn_4*, respectively.

Figure 5.1. Compared to *Trn_0* (Figure 5.1), the two training sets displayed in Figure 5.3 exhibit a slightly greater separation between the synthetic active patterns (green) and airborne inactive patterns (black in the training set and blue in the combined prediction set). Meanwhile, though widely generalized across the space, the synthetic active patterns (green) appear to be similar to the airborne active patterns (red).

As illustrated by Figures 5.1 and 5.3, one of the attributes of using synthetic data to represent the analyte-active class is the ability to tune the degree of overlap between the data

classes. This leads to a key question to be addressed in this work: does adjusting the degree of class separation produce classifiers with improved limits of detection? For example, will the more clear-cut grouping across the different pattern categories observed in the PCA score plots in Figure 5.3 lead to better-defined separation hyperplanes than achieved previously in Chapter 4 with training set *Trn_0*.

5.3.3 Design for the monitoring set

The monitoring set was made with a design scheme similar to that used for the training set. The concentration distribution of the overall pool is illustrated in Figure 5.4. An array of scaling factors was assembled consisting of 48 bins with 60 uniformly distributed random numbers in each bin. Concentrations across bins gradually changed, with each bin representing an equally-distanced sub-range of concentrations. For example, bin 25 represented a range of 9.5×10^{-3} to 0.30 ppm-m, and bin 26 represented a range of 0.30 to 0.60 ppm-m. Later, to evaluate the computed SVM classifiers, the ammonia missed detection rate was calculated separately for each bin as a way to estimate the detection limit of the classifier.

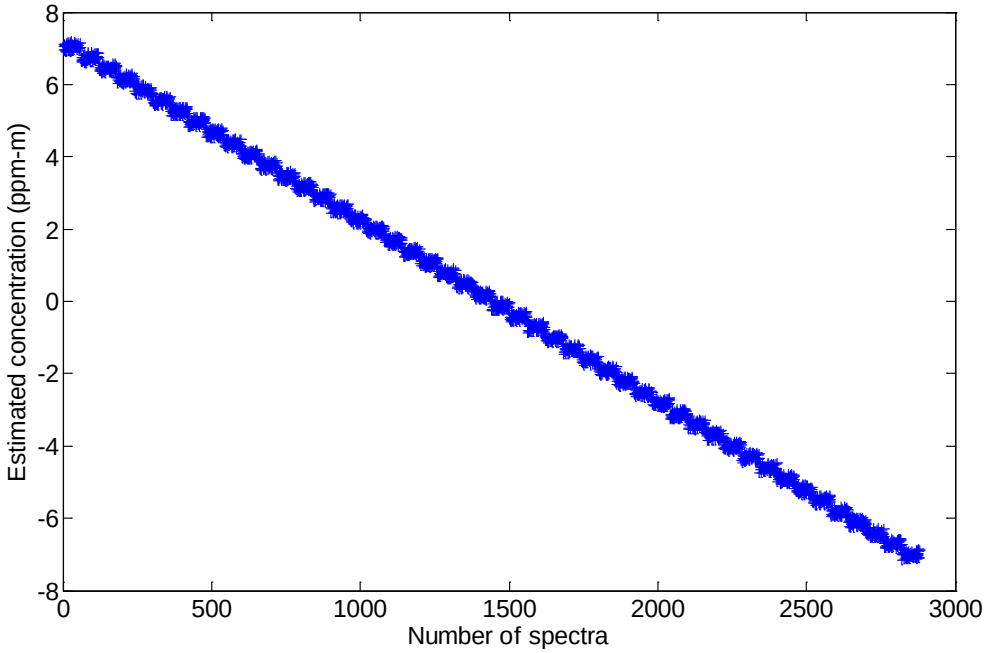


Figure 5.4 Concentration profile of synthetic ammonia-active interferograms designed for the monitoring set.

The concentrations of the resulting total of 2880 synthetic ammonia spectra spanned from 9.5×10^{-3} ppm-m to 7.2 ppm-m for emission, and from 9.9×10^{-3} ppm-m to 7.2 ppm-m for absorption. Here spectra were equally split between emission and absorption, i.e., 1440 spectra for each group. These concentration ranges, slightly extended from the designed concentration ranges of the training sets, were specifically designed and enabled the monitoring set to test the detection limit for each training set. Once the synthetic spectra were obtained, the inverse FT was used to produce the corresponding synthetic interferograms.

Together with the synthetic ammonia interferograms, 17,367 airborne ammonia-inactive interferograms were added to the monitoring set. This is the same group of ammonia-inactive interferograms used in the monitoring set employed in Chapter 4. None of these interferograms were used in the training sets. The false detection rate for the inactive patterns was also used as a criterion in assessing the performance of the SVM classifiers.

5.3.4 Calculation of SVM classifiers

The optimal preprocessing parameters for interferogram segment selection and digital filtering described previously in Chapter 4 were again used in the development of SVM classifiers with the computed training sets. These parameters included a segment starting point of 88, a segment length of 120, a Chebyshev Type II filter centered at 964 cm^{-1}, with a passband width of 28 cm^{-1} (FWHM) and a stopband attenuation of 40 dB. The same set of SVM configuration parameters as described in Chapter 4, including 10 levels of the kernel parameter, γ, ranging from 0.02 to 10.24, and 20 levels of the regularization parameter, C, ranging from 3.00×10^{3} to 2.31×10^{5}, were also used here to construct a pool of 200 SVM classifiers. All parameters used for signal processing and pattern recognition are listed in Table 5.2.

Table 5.2 Parameters used for signal processing and pattern recognition

Parameters		Levels	Values
Chebyshev Type II filter	Passband center (cm^{-1})	1	964
	Passband width, FWHM (cm^{-1})	1	28
	Stopband attenuation (dB)	1	40
Interferogram segment	Starting point	1	88
	Segment length	4	60, 80, 100, 120
SVM configuration	Kernel parameter (γ)	10	0.02, 0.04, 0.08, 0.16, 0.32, 0.64, 1.28, 2.56, 5.12, 10.24
	Regularization parameter (C)	20	3.0×10^{3}, 1.5×10^{4}, 2.7×10^{4}, 3.9×10^{4}, 5.1×10^{4}, 6.3×10^{4}, 7.5×10^{4}, 8.7×10^{4}, 9.9×10^{4}, 1.11×10^{5}, 1.23×10^{5}, 1.35×10^{5}, 1.47×10^{5}, 1.59×10^{5}, 1.71×10^{5}, 1.83×10^{5}, 1.95×10^{5}, 2.07×10^{5}, 2.19×10^{5}, 2.31×10^{5}

The computed SVM classifiers were evaluated on the basis of three criteria: (1) whether they achieved a smaller training error than produced by *Trn_0* (i.e., 5.27 to 5.45 % of the overall training population for the top five classifiers developed in Chapter 4), (2) for those classifiers meeting criterion 1, the minimum training error was considered, and (3) the minimum difference between the missed detection and false detection rates for the monitoring set. Employing these criteria, the top five classifiers based on each training set were identified.

Table 5.3 lists the performance results across the top five classifiers computed with each training set except *Trn_1* and *Trn_2*. Both of these training sets exhibited poor convergence of the optimization algorithm and took a prohibitive amount of computational time to train. This was attributed to too much overlap between the data classes. On the basis of lower training error and fewer support vectors when compared to the previous results obtained with *Trn_0*, training sets *Trn_3* to *Trn_6* all reflected greater class separation. Classification results listed in the table were based on the conventional SVM classification rule in which scores greater than zero were assigned to the ammonia-active class, while scores less than or equal to zero were interpreted as inactive classifications.

However, when applied to the same monitoring set, the classifiers based on the training set from Chapter 4 (*Trn_0*) appeared less likely to miss active patterns than the classifiers developed from training sets *Trn_3* to *Trn_6* (e.g., missed detection rates of 12 to 20 % compared to 48 to 53 %). By contrast, each training set's capacity to recognize the inactive patterns, identical across the training sets, remained almost the same (e.g., false detection rates of 1 to 2 %).

Table 5.3 Summary of selected classifiers

Training set #	Model #	Kernel parameter (γ)	Regularization parameter (C)	Support vectors	Trn. error rate (%)	Trn. miss rate (%)	Trn. false rate (%)	Mon. missed rate (%)	Mon. false rate (%)
Trn_3	1	5.12	3.00×10^{3}	1309	2.6	16.6	0.54	46.0	2.15
	2	10.24	3.00×10^{3}	1277	2.1	13.7	0.41	47.0	2.47
	3	2.56	3.0×10^{3}	1394	3.2	19.9	0.62	48.1	1.96
	4	0.64	3.90×10^{4}	1411	3.3	20.8	0.62	48.6	1.96
	5	0.64	2.70×10^{4}	1439	3.4	21.3	0.68	48.7	1.94
Trn_4	1	5.12	3.00×10^{3}	701	1.2	11.4	0.22	47.8	1.40
	2	10.24	3.00×10^{3}	709	0.8	7.75	0.14	48.4	1.64
	3	0.64	6.30×10^{4}	743	1.6	14.1	0.31	49.6	1.27
	4	1.28	1.50×10^{4}	744	1.6	14.2	0.30	49.9	1.29
	5	0.32	2.31×10^{5}	737	1.6	14.7	0.30	49.9	1.20
Trn_5	1	5.12	1.11×10^{5}	671	0.8	4.0	0.53	52.6	2.00
	2	5.12	9.90×10^{4}	398	0.1	1.8	0.00	52.7	1.46
	3	5.12	8.70×10^{4}	405	0.1	1.8	0.01	52.7	1.43
	4	5.12	7.50×10^{4}	401	0.1	1.8	0.02	52.8	1.46
	5	5.12	6.30×10^{4}	406	0.1	1.9	0.01	53.0	1.42
Trn_6	1	10.24	2.70×10^{4}	427	0.02	0.3	0.00	52.3	1.40
	2	10.24	8.70×10^{4}	428	0	0.0	0.00	52.5	1.58
	3	10.24	9.90×10^{4}	427	0	0.0	0.00	52.5	1.58
	4	10.24	1.11×10^{5}	427	0	0.0	0.00	52.5	1.58
	5	10.24	1.23×10^{5}	427	0	0.0	0.00	52.5	1.58
Trn_0	1	10.24	3.00×10^{3}	2112	53	24.8	0.38	12.0	1.71
	2	2.56	3.90×10^{4}	2119	5.3	25.1	0.39	21.6	1.72
	3	1.28	8.70×10^{4}	2128	5.4	25.2	0.46	18.3	1.70
	4	2.56	2.70×10^{4}	2123	5.4	25.2	0.42	17.8	1.75
	5	2.56	1.50×10^{4}	2144	5.4	25.6	0.42	13.5	1.72

Consideration of the monitoring results in Table 5.3 suggested that the classifiers built with training sets *Trn_3* and *Trn_4* were the best for further testing. The missed detection rates slowly increased and the false detection rates decreased from *Trn_3* to *Trn_6* as the class overlap in the training set decreased. This can also be seen in the almost perfect training results for *Trn_5* and *Trn_6* (Table 5.3) and the KNN classification results discussed previously (Table 5.1). With KNN, the classification rates for the active class were 80 and 82 % for *Trn_5* and *Trn_6*, respectively, while the corresponding results for *Trn_3* and *Trn_4* were 64 and 72 %.

Comparing the results for *Trn_3* to those for *Trn_6*, an approximately 4 % increase in the missed detections was offset by only an approximately 0.5 % improvement in the rate of false detections. This suggests the tradeoff choice lies toward the classifiers built with *Trn_3* and *Trn_4*.

5.3.5 Performance of SVM classifiers with airborne prediction sets

The 10 top classifiers generated from training sets *Trn_3* and *Trn_4* were tested with prediction sets *A*, *B*, *C*, and *D*. The SVM output scores were plotted with respect to the interferogram sequence number for each prediction set. The score profiles generated from the different SVM classifiers had similar patterns, differing only in magnitude. A general trend was that classifiers computed from training set *Trn_3* gave rise to a larger separation of scores between the two data classes. Classifier Model #1, based on training set, *Trn_3*, with $\gamma = 5.12$ and $C = 3.0\times10^3$, produced SVM scores with the largest difference between the maximum and minimum SVM scores for the prediction sets. This classifier was selected for further examination and for illustrating the overall classifier performance.

The SVM score profiles produced by this classifier for the four prediction sets are shown as the red traces in Figure 5.5. The blue traces correspond to the results obtained with the

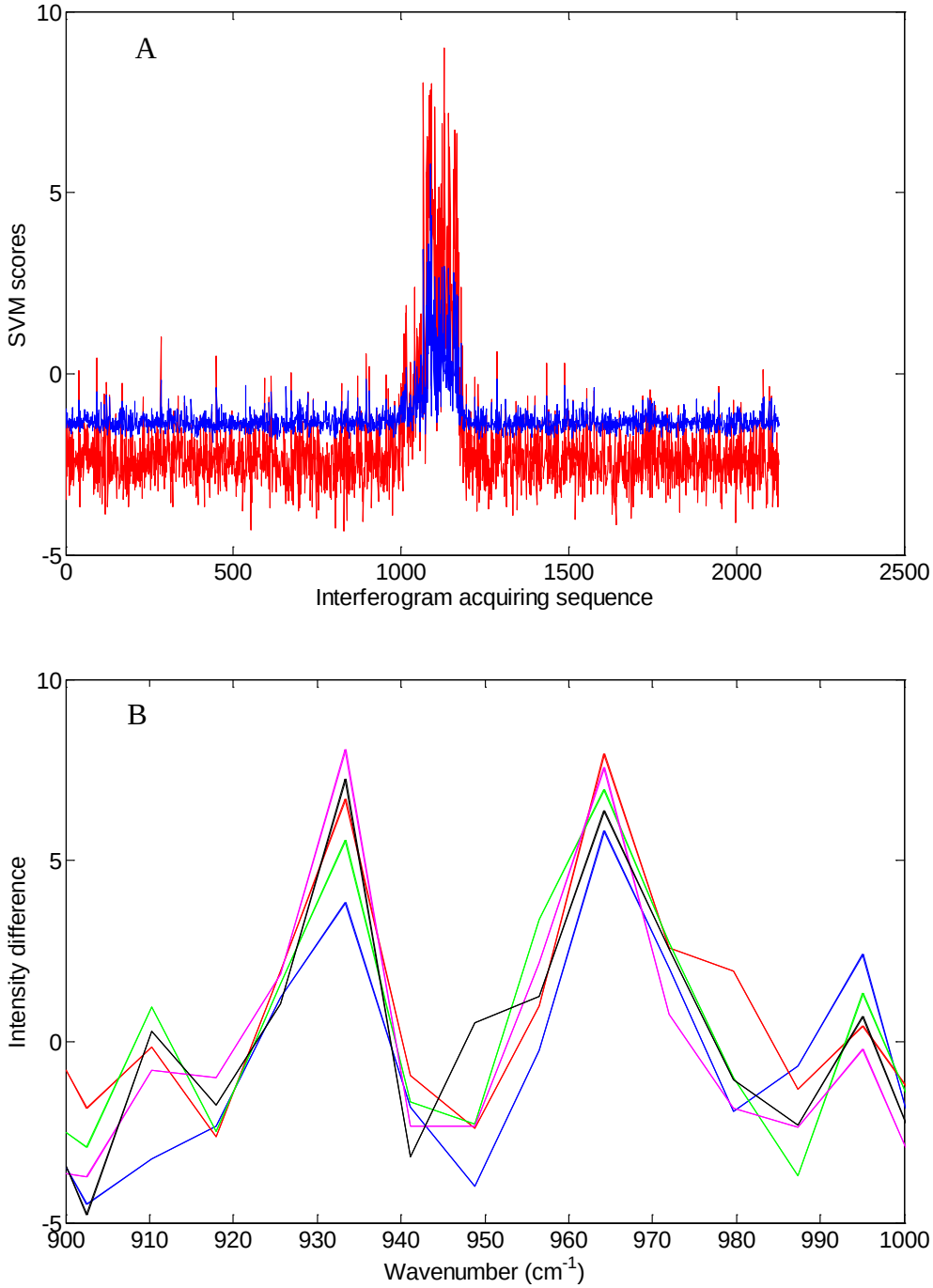
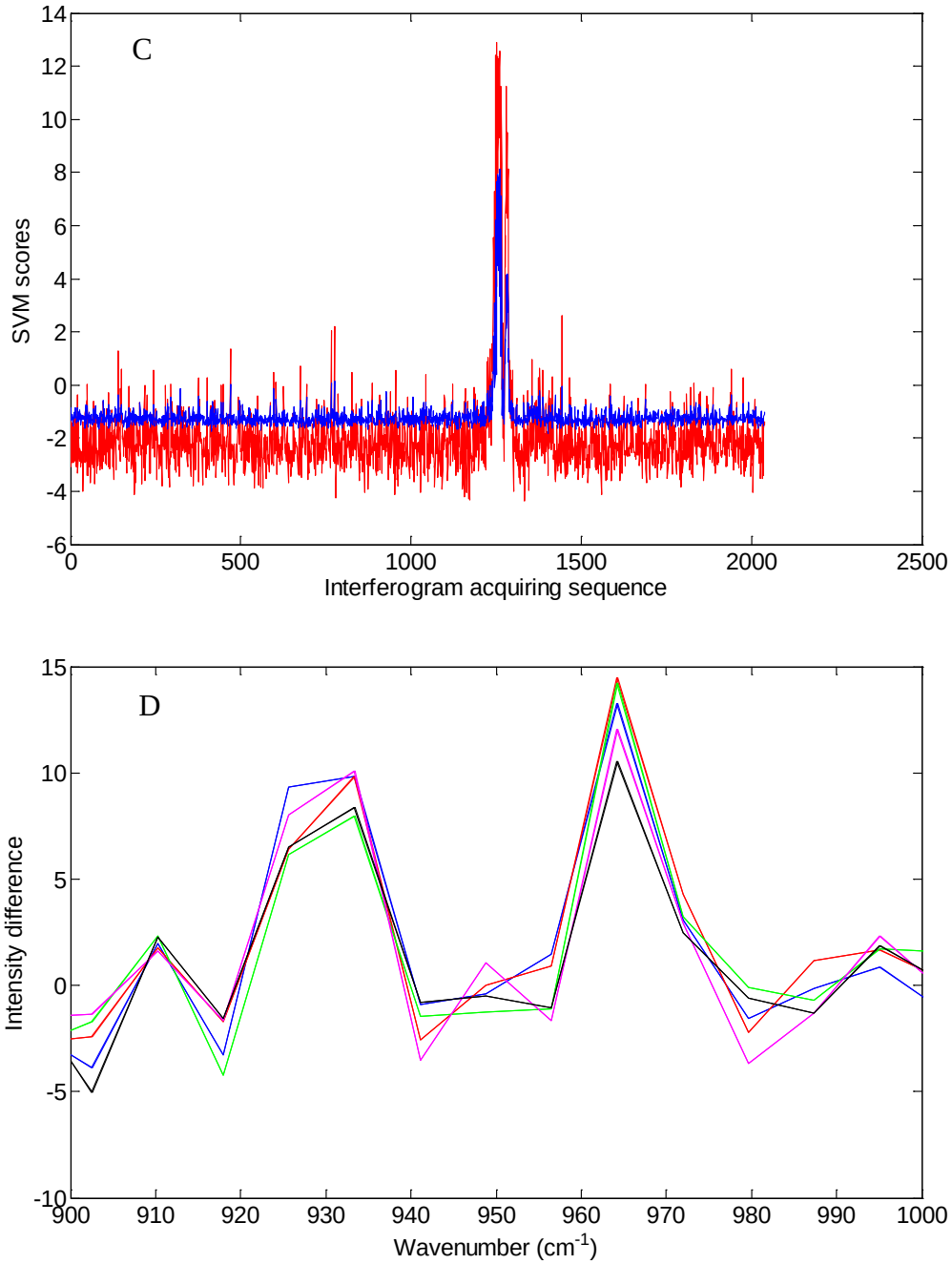
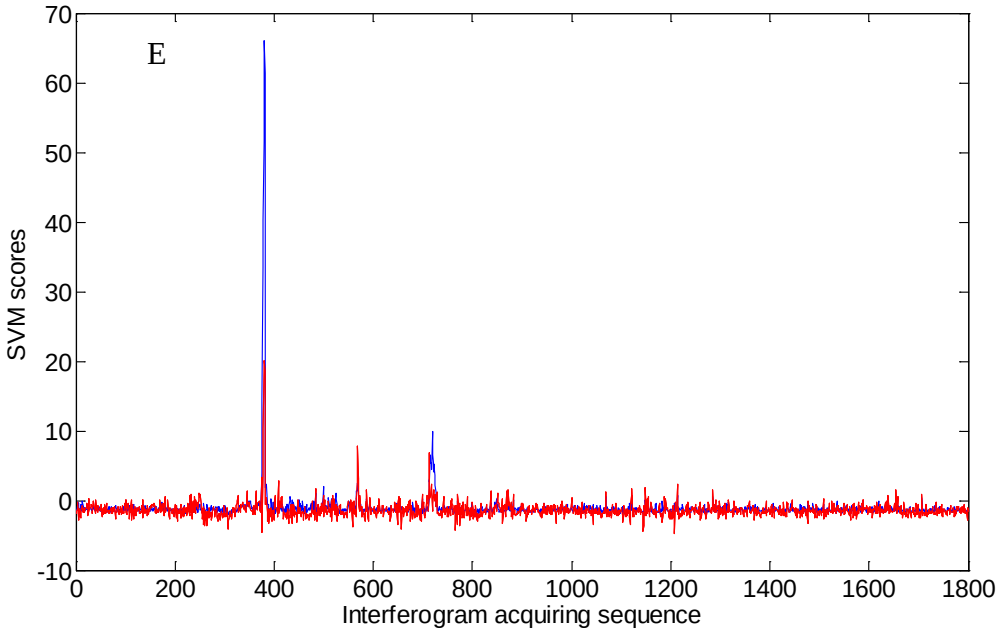


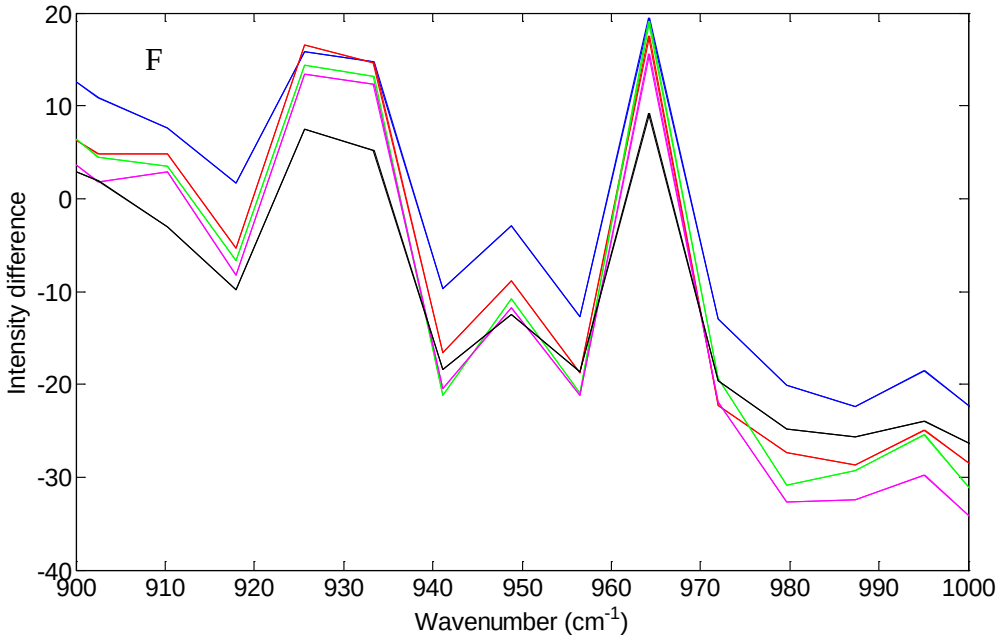
Figure 5.5 Profile of SVM scores with respect to acquisition sequence and corresponding difference spectra derived from the positive peaks in the score profiles for the four prediction sets. (A) Profile of SVM scores from prediction set *A*. (B) Difference spectra corresponding to interferograms #1067, #1077, #1081, #1089, and #1091 in prediction set *A*. (C) Profile of SVM scores from prediction set *B*.; (D) Difference spectra corresponding to interferograms #1253, #1256, #1258, #1260, and #1265 in prediction set *B*. (E) Profile of SVM scores from prediction set *C*. (F) Difference spectra corresponding to interferograms #378, #379, #380, #381, and #382

in prediction set *C*. (G) Difference spectra corresponding to interferograms #715, #717, #719, #721, and #723 in prediction set C. (H) Difference spectra corresponding to interferograms #566, #567, #568, #569, and #570 in prediction set C. (I) Profile of SVM scores from prediction set *D*. (J) Difference spectra corresponding to interferograms #375, #376, #378, #379, and #380 in prediction set *D*. For each time profile, plots A, C, E, and I, SVM scores generated from the current selected classifier (red, *Trn_3*) were compared with the scores produced by the selected classifier from Chapter 4 (blue, *Trn_0*).

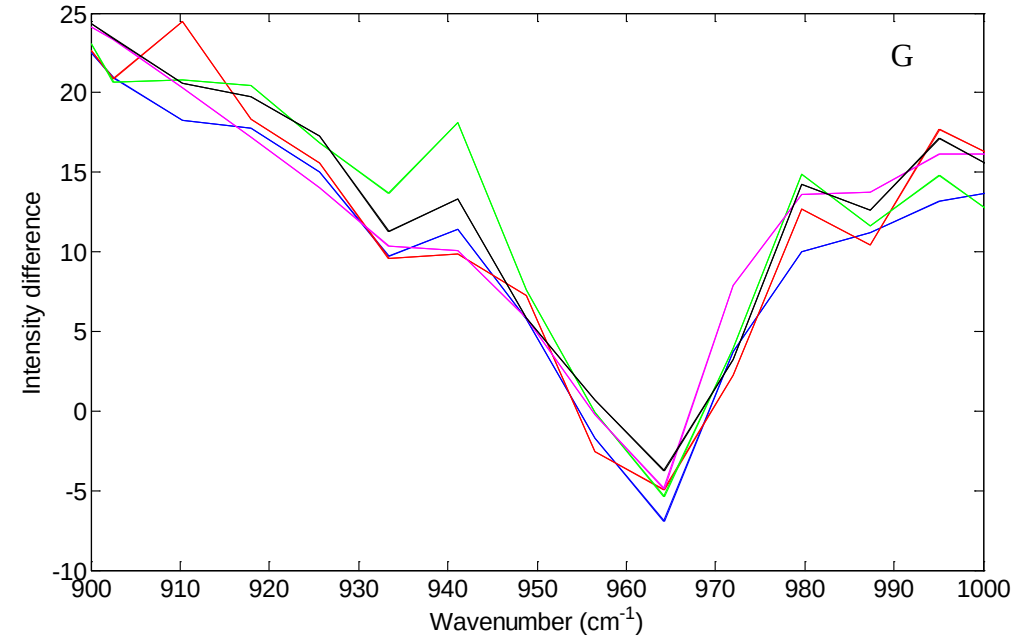


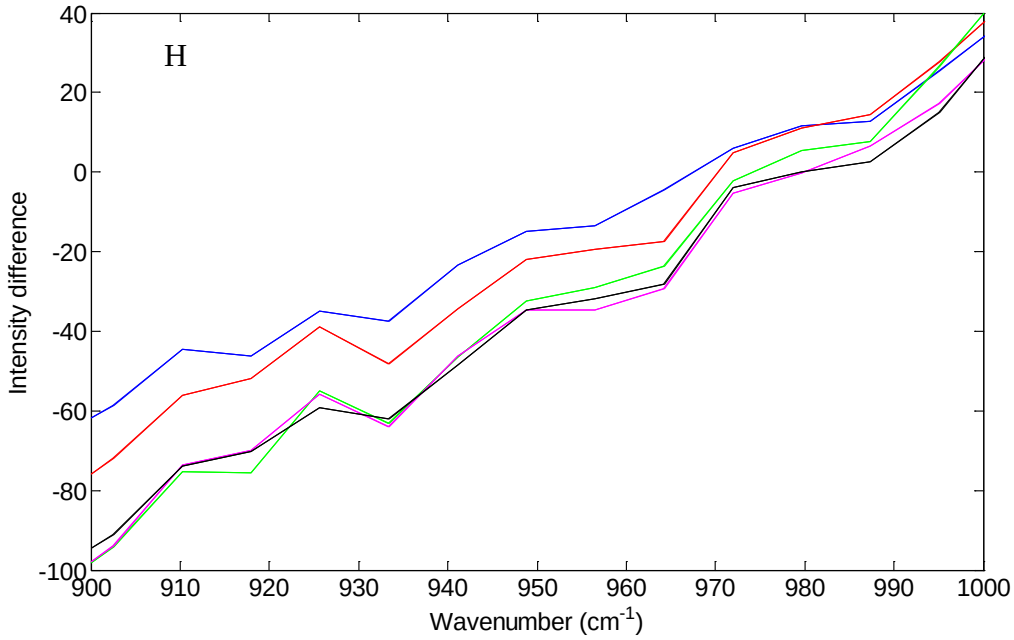
(Figure 5.5 continued)



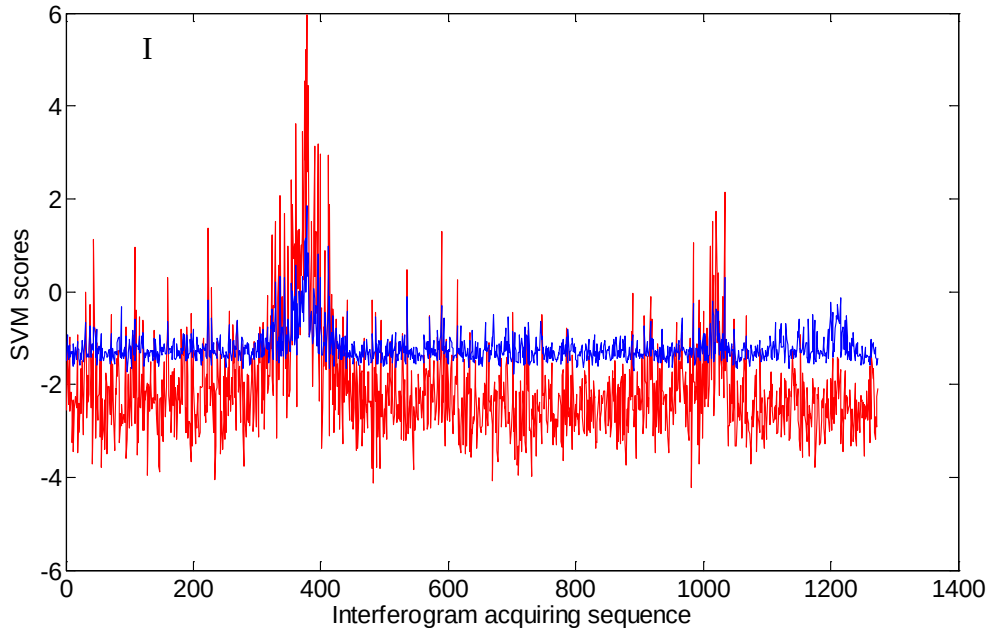


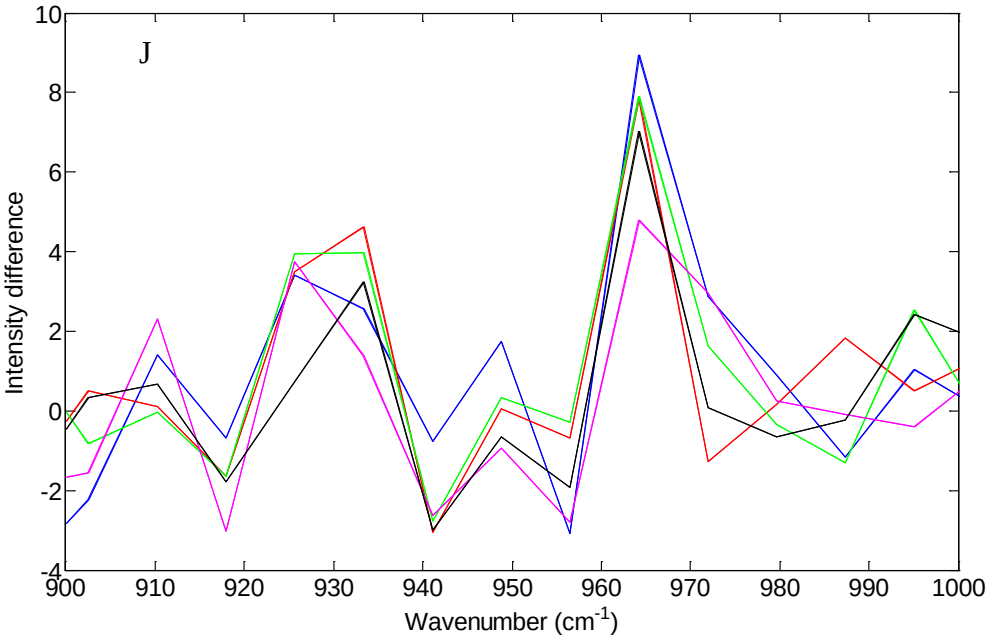
(Figure 5.5 continued)





(Figure 5.5 continued)





(Figure 5.5 continued)

classifier developed in Chapter 4 with training set *Trn_0*. In general, both classifiers shared the same patterns of SVM scores for all the prediction sets. The inspection of absorbance difference spectra with selected interferograms with positive scores previously conducted in Chapter 4 confirmed the accuracy of the classifications. One exception came from prediction set *C* (Figure 5.5, E-H). The SVM score profile for prediction set *C* produced by the classifier based on *Trn_0* showed that there were two positive score spikes, a weak one followed by a stronger one. Inspection of difference spectra for the interferograms with positive SVM scores revealed that both spikes corresponded to ammonia-active patterns, the major spike corresponding to ammonia emission spectra and the adjacent minor spike corresponding to ammonia absorption spectra. In the case of the selected classifier based on *Trn_3*, one additional spike was present along with the two spikes mentioned above. The additional spike was confirmed to be ammonia absorption by visual inspection of the corresponding difference spectra. Both classifiers detected ammonia absorption correctly.

The occurrence of both absorption and emission signatures within the same field data set is not uncommon. Emission signatures arise from hot gas plumes while absorption signals correspond to locations downwind from the emission source in which the gas temperature has equilibrated with the surrounding atmosphere. Because the effective ground temperature that defines the background radiance is typically hotter than that of the air, analyte signals downwind from the plume will be realized as absorptions.[yousuf p96] Difference spectra for these cases are displayed in Figure 5.5, G and H.

Besides the similarity in profiles across the prediction sets, there were noticeable differences between the results obtained from the classifiers based on the *Trn_0* and *Trn_3* training sets. As shown in Figure 5.5, A, C, E, and I, the predicted SVM profiles produced from

the classifier based on *Trn_0* (blue) tended to have a positive baseline drift when compared to the profile of the classifier based on *Trn_3* (red trace). The profile for the classifier developed from *Trn_3* exhibited greater variation over time.

Classification percentages for the four prediction sets are presented in Table 5.4. The classification results were obtained from a comparison of the SVM scores to the reference classifications obtained by visual inspections of difference spectra. The statistics presented were based on both the conventional SVM classification rule used previously with the monitoring set and the control chart methodology employed in Chapter 4.

The results obtained with the conventional classification rule are summarized in Table 5.5. The SVM output for prediction sets generated from training set *trn_3* tended to detect the active patterns more sensitively, (e.g., smaller missed detection rate) while the scores obtained from the classifier based on *trn_0* had a lower false detection rate.

These results can be contrasted to those obtained through the use of control charts, listed in Table 5.5 as well. The classification rule used here was the same as described in Chapter 4. For an individual SVM score profile generated for a prediction set, the UCL and LCL were calculated as two times the standard deviation of the moving average about the mean of the entire SVM trace. If a confirmed active pattern resulted in an SVM score lower than the UCL, this data point was considered missed. Similarly, if a confirmed inactive pattern had an SVM score higher than the UCL, it was considered a false alarm.

Comparison of the results obtained with the conventional classification rule showed that the classification based on control charts yielded better detection of active patterns at a cost of

Table 5.4 Summary of classification results for prediction sets

Training set	Prediction sets	Target	Estimated standard deviation	LCL	UCL	Missed detection rate (%)	False detection rate (%)
Trn_3	A	-2.02	0.77	-3.57	-0.48	2.94	1.65
	B	-1.97	0.82	-3.61	-0.32	5.56	2.42
	C	-1.35	0.73	-2.82	0.11	12.50	4.14
	D	-2.10	0.80	-3.70	-0.51	14.29	2.22
Trn_4	A	-1.95	0.59	-3.12	-0.77	3.92	1.91
	B	-1.87	0.60	-3.06	-0.68	5.56	2.05
	C	-1.59	0.62	-2.83	-0.34	50.00	4.25
	D	-1.99	0.57	-3.13	-0.84	14.29	1.95
Trn_0	A	-1.21	0.27	-1.76	-0.67	7.84	1.45
	B	-1.15	0.25	-1.65	-0.66	14.81	1.69
	C	-1.01	0.40	-1.81	-0.22	0.00	3.75
	D	-1.21	0.24	-1.69	-0.72	19.05	3.82

Table 5.5 Statistics of SVM scores for inactive patterns from different sources

Inactive pattern source	Identically configured top classifiers				Final selected classifiers			
	γ	C	Scores mean	Scores standard deviation	γ	C	Scores mean	Scores standard deviation
Trn_0	10.24	3.0×10^3	-1.31	0.28	1.28	8.7×10^4	-1.34	0.33
Trn_3	10.24	3.0×10^3	-2.40	1.61	5.12	3.0×10^3	-2.19	0.99
Trn_4	10.24	3.0×10^3	-2.72	1.13	5.12	3.0×10^3	-2.36	0.87
Mon (*trn_0*)	10.24	3.0×10^3	-1.26	0.48	1.28	3.0×10^3	-1.24	0.59
Mon (*trn3*)	10.24	3.0×10^3	-2.29	1.68	5.12	3.0×10^3	-2.13	1.22
Mon (*trn4*)	10.24	3.0×10^3	-2.67	1.32	5.12	3.0×10^3	-2.33	1.12

more false alarms for the actual inactive patterns. For example, for prediction set *A*, the model based on training set *Trn_3* (γ = 5.12 and $C = 3.0\times10^3$) resulted in missed detection rates of 7.8 % and 2.9 %, respectively, for the conventional classification rule and classification based on control charts. The corresponding false detection rates were 0.8 % and 1.6 %. Given that single false alarms can be removed easily by simply requiring two consecutive positive detections to signal an alarm, use of the control chart methodology is judged to be the best choice for implementing the classifier.

The statistics associated with the control charts can also be used to help compare the SVM profiles across the classifiers. The mean SVM score, computed across the aircraft pass, tended to be positively biased from the classifier based on *Trn_0* in comparison to the classifiers based on *Trn_3* or *Trn_4* (~ 40 % more positive). From the estimate of the standard deviation of the average moving range, the classifiers based on *Trn_3* and *Trn_4* produced SVM scores that had much greater variance than the classifier derived from *Trn_0* (1-2 times higher standard deviation).

These trends were manifest in the classification results and resulted in a slightly improved false detection rate and occasionally degraded missed detection performance with the classifiers based on training sets *Trn_3* and *Trn_4*. This was especially seen with prediction set *C* which featured significant variation in the spectra acquired across the aircraft pass. For this data set, one or four out of eight ammonia active spectra (12.5 % or 50.0 %) were not detected with the selected classifiers developed from the *Trn_3* and *Trn_4* training sets, respectively.

Bias-variance trade-off theory from the machine learning community [100,101] may help to explain the statistical differences observed here across the classifiers. It states that, for a given classifier, the overall error can be broken down into bias, describing the average error from the

true classifier, variance, describing how much an individual classifier varies about the classifier's mean, and noise, describing whatever error remains. In general, the more biased a classifier becomes, the less variance it has, and vice versa. Conceptually, this theory is analogous to the decomposition of the variance about a linear regression model into variance about the regression, variance due to the regression, and the residual variance. Note that the literal definitions of bias and variance vary across these research fields, however.

In this case, the classifier based on *Trn_0* was more positively biased, and therefore resulted in a smoother (i.e., having less variance) profile of SVM scores. The less biased classifiers built with training sets *Trn_3* and *Trn_4* produced SVM score profiles with more variation.

5.3.6 Evaluation of SVM classifiers with respect to training data composition

The classification performance with prediction sets *A*, *B*, *C*, and *D* revealed that the selected classifiers developed from training sets *Trn_0*, *Trn_3*, and *Trn_4* have more or less comparable predictive capacity in terms of the classification results based on control charts. They were all able to detect low levels of ambient ammonia while not being prone to false alarms. The classifier based on *Trn_0* performed better with respect to ammonia detection while the classifiers built with training sets *Trn_3* and *Trn_4* had better resistance to false detections.

It is useful to characterize the prediction performance of the classifiers specifically in terms of the distribution of the training data. To do so, from the top classifier list for each training set listed in Table 5.3, the same set of SVM configuration parameters, $\gamma = 10.24$ and $C = 3\times10^3$, was employed for training sets *Trn_0*, *Trn_3*, and *Trn_4*. Classifiers developed from each of these training sets using this set of SVM configuration parameters were used to generate SVM score profiles for both the training set and the monitoring set.

The resulting profiles are shown in Figure 5.6. Note that though the number of interferograms in *Trn_3* and *Trn_4* were 13,800 and 13,200, to make them graphically comparable to *Trn_0* (15,000 interferograms), zeroes were appended to the score profiles for *Trn_3* and *Trn_4* such that the same 12,000 ammonia-inactive interferograms used in all three training sets were aligned on the right side of the plot.

For both the training sets and the monitoring set, the classifier based on *Trn_0* (blue trace) results in higher score values for the active patterns (i.e., interferogram numbers 1-3000 for *Trn_0*, 1-1800 for *Trn_3*, 1-1200 for *Trn_4*, and 1-2880 for the monitoring set). This trend made the classifier developed from *Trn_0* more favorable to the detection of ammonia-active patterns relative to the other two classifiers. However, the classifier developed with *Trn_0* also generated higher score values for the inactive patterns (i.e., the last 12,000 interferograms for all training sets and interferogram numbers 2881-20,247 for the monitoring set). Higher scores for the ammonia-inactive patterns can lead to a higher probability of false detections. These findings were consistent with the classification results summarized previously in Table 5.3.

As seen in the left section of Table 5.6, statistics computed from the scores plotted in Figure 5.6 for the ammonia-inactive patterns showed that the classifiers based on *Trn_3* and *Trn_4* produced very consistent scores for the training and monitoring set. The classifier computed from *Trn_3* produced mean SVM scores of -2.40 and -2.29 when applied to the inactive patterns in the training and monitoring set, respectively. The corresponding standard deviations were 1.61 and 1.68. Similarly, for the classifier based on *Trn_4*, the mean scores were -2.72 and -2.67, and the corresponding standard deviations were 1.13 and 1.32 for the training and monitoring set, respectively.

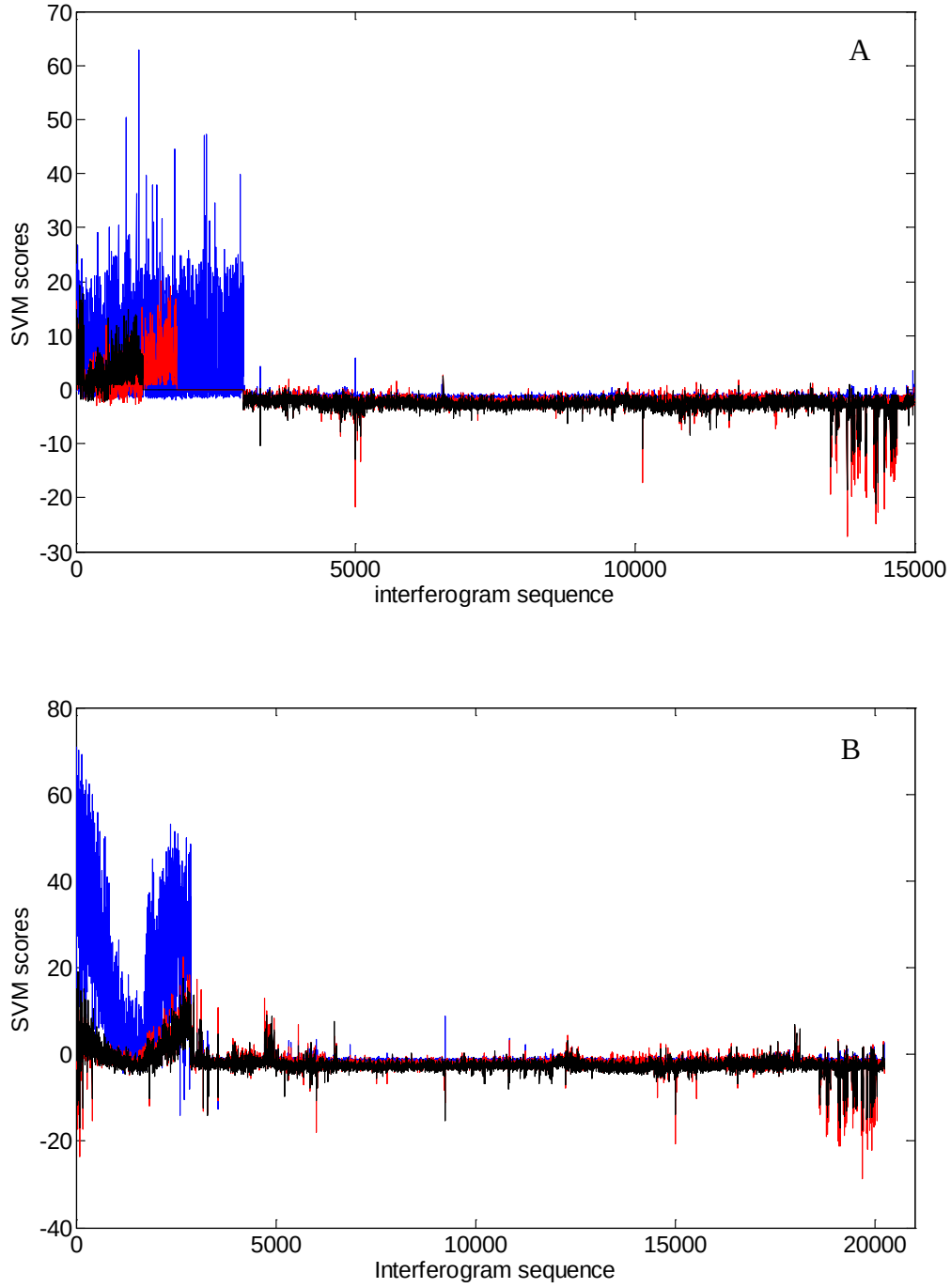


Figure 5.6 Profiles of SVM scores obtained from the three classifiers when applied to their corresponding training sets (A) and the monitoring set (B). Blue, red, and black traces correspond to training sets *Trn_0*, *Trn_3*, and *Trn_4*, respectively. The three SVM classifiers were all configured with γ =10.24 and C = 3000.

Table 5.6 Classification performance evaluated from different methods

Data sets		Control chart classification		SVM conventional classification	
Training sets	Prediction sets	Missed detection rate (%)	False detection rate (%)	Missed detection rate (%)	False detection rate (%)
Trn_3	A	2.94	1.65	7.84	0.78
	B	5.56	2.42	9.26	1.48
	C	12.50	4.14	12.50	4.92
	D	14.29	2.22	14.29	0.89
Trn_4	A	3.92	1.91	11.77	0.26
	B	5.56	2.05	20.37	0.63
	C	50.00	4.25	50.00	2.96
	D	14.29	1.95	33.33	0.44
Trn_0	A	7.84	1.45	35.29	0.05
	B	14.81	1.69	25.93	0.16
	C	0.00	3.75	0.00	2.52
	D	19.05	3.82	57.14	0.18

Given that the inactive patterns in the training and monitoring sets were both selected from a pool of >500,000 interferograms, these sets should approximate a reasonable parent distribution for the inactive data class. The consistency in response between the training and monitoring sets for the classifiers computed from *Trn_3* and *Trn_4* lends confidence that these classifiers have trained successfully with good generalizability to data outside of the training set.

For the classifier computed with *Trn_0,* the mean SVM scores were quite different between the training and monitoring sets. The score means did not show much drift (-1.31 to -1.26), but the score standard deviations dramatically increased from 0.28 to 0.48 (70 % increase) for the training and monitoring sets, respectively. Such inconsistent results for patterns presumably of similar distributions indicated that the classifier based on *Trn_0* was not optimized well with respect to its generalizability outside of the training set. Stated differently, there is evidence the SVM model was overfit to the training data.

The right section of Table 5.6 displays similar statistics for the case in which the final selected classifiers based on the three training sets (i.e., those classifiers with various SVM configuration parameters) were applied to the ammonia-inactive patterns in the training and monitoring sets. The classifiers based on *Trn_3* and *Trn_4* produced scores with very little change in the mean values and slightly improved (i.e., smaller) standard deviations when compared to the classifiers that had the same SVM configuration parameters. However, the classifier based on *Trn_0* again showed significant variation between the SVM scores for the training and monitoring sets. An even higher standard deviation (6.91 vs. 5.79) of the scores in the monitoring set was observed for this classifier relative to the one discussed above. This suggests an issue with the composition of the training set itself rather than an artifact of the configuration parameters.

According to the conventional SVM classification rule, a separation hyperplane is set where the discriminant function is equal to zero. As for prediction, a pattern with a positive score is classified into one class, while a pattern with a negative score is placed into the other class. It has been realized that such a rigid classification definition might produce problems when the classifier is applied to data outside of the training set. Other researchers have tried various methods to make corrections to the conventional classification rule.[102,103]

The classification methods developed in Chapter 4 based on control charts take into account the trend of the pattern of SVM scores with time and assume the overall score profile begins with and is dominated by patterns from one class. This is the case for monitoring of discrete emission sources by overflights of the aircraft. As mentioned earlier, the nature of the passive FT-IR remote sensing data acquisition is that, if occurring, only a small number of ammonia-active interferograms (e.g., < 1 %) would be acquired among a majority of inactive interfrograms during one aircraft pass over a suspected emission source. Furthermore, the aircraft pass and associated data acquisition would likely begin prior to the area of the target. As a consequence of these characteristics of the specific application under development for this research, it can be argued that the trend of the trace of SVM scores with time is more significant than the arbitrary classification threshold of 0.0.

Table 5.5 provides a comparison of the classification results obtained for the four prediction sets with both the conventional SVM classification rule and the use of control charts. The classification results based on control charts were superior to those obtained with the conventional classification rule. Both approaches yielded similar rates of false detections, but the classifications obtained with the control charts featured fewer missed detections. The “trend-

focused" method employing control charts was superior to the "value-focused" conventional classification rule.

The classifier computed with training set *Trn_0* realized the greatest benefit from the use of control charts. By employing control charts, the bias in the SVM scores noted previously is mitigated. Though not the best performer when control charts were used, the classifier based on *Trn_3* outperformed the other two classifiers when the conventional classification rule was employed.

5.3.7 Estimate of detection limits for selected classifiers

In order to estimate the detection limits corresponding to the selected classifiers, the missed detection rate determined with the conventional SVM classification rule was computed separately for each concentration bin in the active class of the monitoring set. These missed detection rates were then plotted in Figure 5.7 with respect to the corresponding mean concentration of the bin.

The conventional definition of the LOD is the concentration corresponding to an analyte signal that is statistically distinct from the background at approximately the 99% level. To apply this definition to the current application, the SVM score was considered the analyte signal and a missed detection rate of 1 % was established as the performance level required to meet the definition of the LOD.

To estimate the concentration corresponding to the signal at the LOD, each half of the curve in Figure 5.7 was fit to a 4^{th} order polynomial function by multiple linear regression and the concentration corresponding to the intersection of the resulting fitted equation with 1 % missed detection was taken as the LOD.

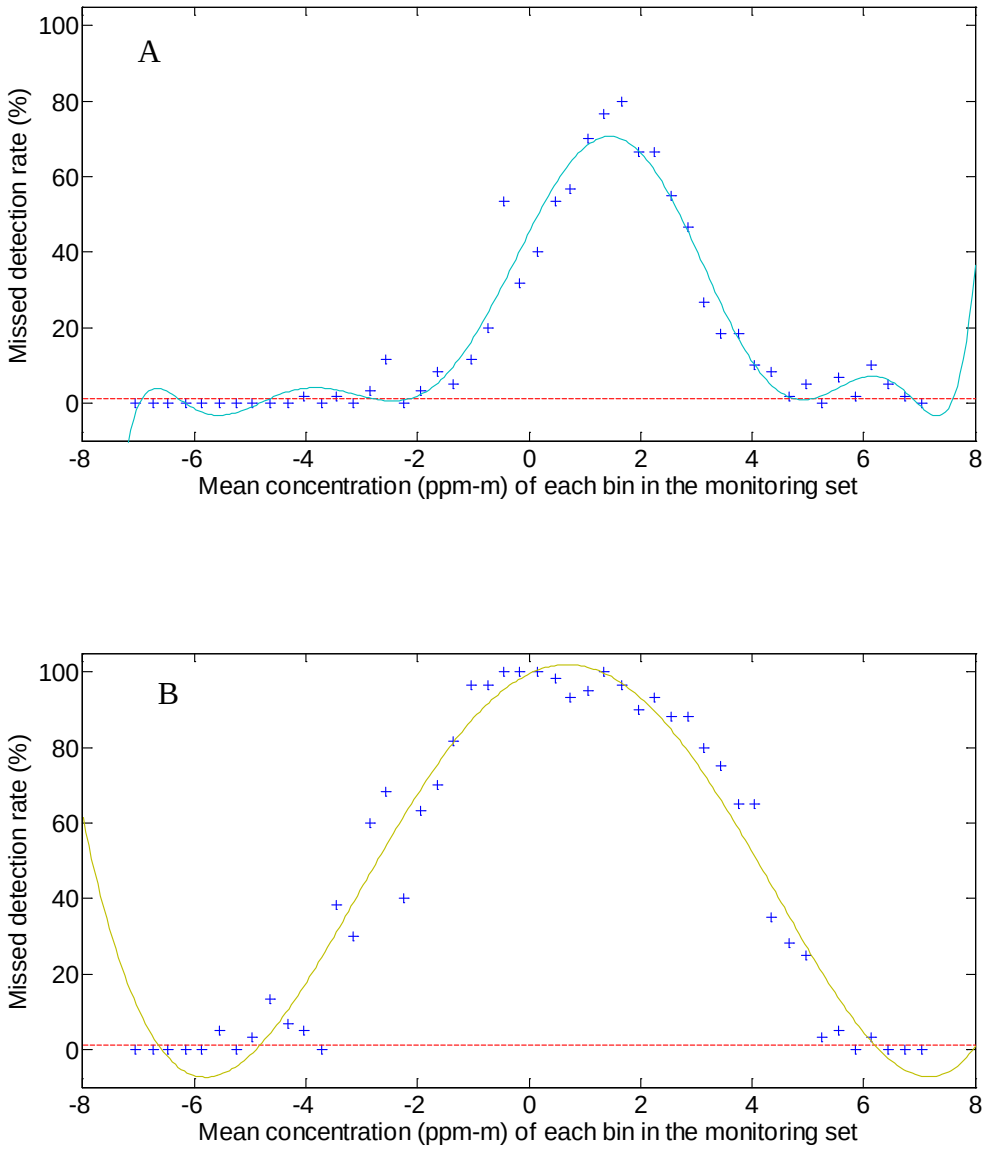
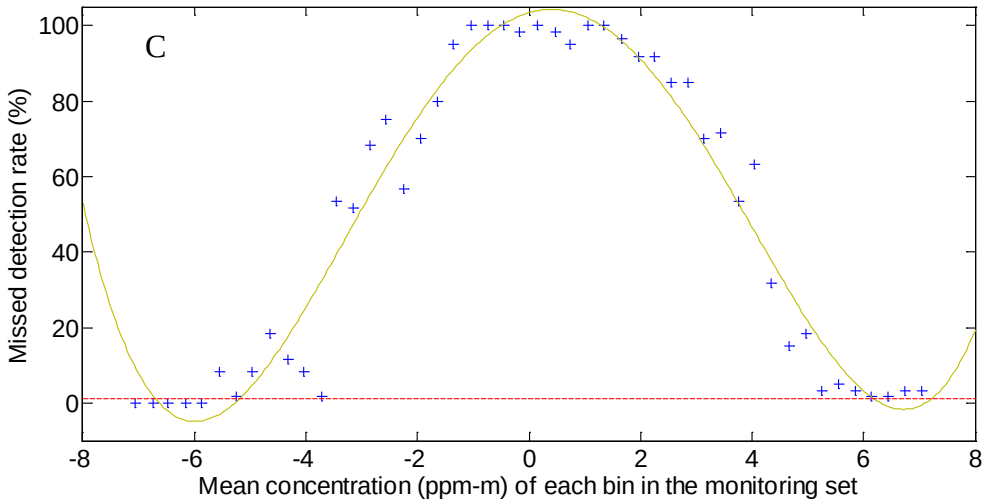


Figure 5.7 Missed detection rates as a function of mean ammonia concentration (ppm-m) for each concentration level within the monitoring set. Panels A, B, and C correspond to the classifiers developed with training sets *Trn_0*, *Trn_3*, and *Trn_4*, respectively. Positive and negative concentration values correspond to absorption and emission signals, respectively.



(Figure 5.7 continued)

According to this rule, the LOD for the classifier based on *Trn_0* was 2 ppm-m for the emission case and 5 ppm-m for the absorption case. The analogous values for emission and absorption for the classifier built with *Trn_3* were 5 and 6 ppm-m, respectively. For the classifier based on *Trn_4*, the values of the LOD were also 5 and 6 ppm-m for emission and absorption, respectively. These results are consistent with values reported with active-mode open-path FT-IR (e.g., 0.3-1.5 ppm-m depending on instrumental resolution).[30] As stated previously, however, our estimates of the LOD correspond to the best-case conditions in which the temperature differential between the sample and background is large enough to have no effect of suppressing the analyte signal (i.e., analogous to an active-mode measurement).

5.4 Conclusions

The work described previously in Chapter 4 produced a classifier development procedure incorporating an interferogram simulation protocol and signal processing strategy, supervised pattern recognition algorithms, and a novel classification calculation based on control charts. A selected classifier at optimal conditions was able to recognize the signature of ammonia in airborne passive infrared data, with both low missed detections and false detections. In this chapter, the relationships between the data distribution of the training set and the resulting performance of the SVM classifiers were examined more fully. A by-product of this work was the establishment of a protocol for estimating the detection limit of the classifier.

With parameters related to the interferogram segment and preprocessing digital filter fixed at their optimal conditions established in Chapter 4, this work focused on designing various training sets with different analyte concentration distributions and investigating the impact of the concentration distribution on classification performance. Active patterns were drawn in a stepwise fashion from a parent simulated active pattern pool with known ammonia concentration ranges. The top training sets were selected by a combination of classification results from the

KNN method and visual examinations of PCA score plots. The selection of the training sets was guided by comparing the KNN classification statistics to those of the SVM classifier developed in Chapter 4.

When applied to the airborne prediction sets, the SVM classifiers developed in this work resulted in overall comparable classification performance to the classifier developed in Chapter 4. The classifiers developed here with training sets *Trn_3* and *Trn_4* were better than the classifier from Chapter 4 based on training set *Trn_0* if the conventional SVM classification rule was used. However, the classifier from Chapter 4 derived greater benefit from the use of control charts in performing the classification due to a bias in its background SVM score profile.

For future work, incorporating the impact of temperature differential on the determination of the LOD is of paramount importance. In addition, adding more airborne prediction data sets with broader variation in background conditions would be useful. Also of interest is the assessment of whether the classification methodology based on control charts can overcome instrumental variation with time such as the effects of instrument repair or upgrades.

Chapter 6

AUTOMATED DETECTION OF CESIUM-137 BY AIRBORNE GAMMA-RAY SPECTROMETRY

6.1 Introduction

The energy of gamma-ray radiation produced by a particular nuclide due to radioactive decay is characteristic. When these emissions are captured and analyzed by a gamma-ray spectrometer, a gamma-ray energy spectrum can be produced. An analysis of the gamma-ray spectrum is typically used to determine the identity and quantity of radionuclides present in a sample.

Airborne gamma-ray spectrometry survey (AGRS) is a remote sensing technique that places a gamma-ray spectrometer in a downward-looking orientation on board a flying aircraft. Emissions of gamma-rays from manmade or naturally-occurring ground sources can thus be detected in a remote manner for the purpose of radioisotope identification or quantification. This technique has found applications in areas such as isotope exploration, [104,105] geographical mineralization mapping, [106,107] and environmental monitoring. [108,109]

The radioisotope of interest in this work is cesium-137 (^{137}Cs), the most common radioactive form of cesium and a significant environmental contaminant. Cesium-137 is produced when uranium and plutonium absorb neutrons and undergo fission. Its presence in the environment today mainly comes from fallout of nuclear weapons testing, nuclear reactor waste or accidental releases such as in the Chernobyl and Fukushima reactor accidents. [108, 110] The very first step to cleanup of ^{137}Cs is to detect it in the environment. Cesium-137 undergoes radioactive decay with emission of beta particles and relatively strong gamma-ray radiation, which makes it detectable with airborne gamma-ray spectroscopy.

It was desired to develop a rapid and accurate automated detection of ^{137}Cs from gamma-ray spectra collected during aerial surveys. Typical survey conditions are aircraft passes at 300 – 500 ft altitudes (~ 100 – 150 m) above the survey site with spectral acquisitions on the order of one spectrum per second. The goal for the automated classifier is to be able to produce a yes/no decision regarding the presence of the signatures of one or more target isotopes in near real time. In this way, rapid pinpointing of ground locations can be accomplished and feedback is provided to the operator regarding the need to change the aircraft track.

The implementation of automated methodology for identifying radioisotopic signatures from gamma-ray spectra faces several challenges. First, even though every radioisotope has its characteristic gamma-ray signature, airborne gamma-ray spectra of a single target isotope can vary, highly dependent on the altitude at which the spectra are acquired, the process of radioactive decay, and the degree to which Compton scattering has affected the overall spectrum. When combined with the relatively low signal-to-noise ratios observed in the airborne spectra, it becomes quite challenging to recognize relatively weak target features with changing shapes and intensities arising from a varying background.

Several signal preprocessing techniques have been proposed to extract target features from the complex background observed with gamma-ray spectra. Methods include factor analysis,[111] principal component analysis,[112,113] wavelet analysis,[114,115] and target-based calibration methods.[116] Once the target feature is extracted, it is reasonable to develop classification methods for target isotopes based on the preprocessed spectra. Pattern recognition methods such as artificial neural networks [117,118] and decision trees [119-121] have been reported to develop classifiers based on airborne gamma-ray spectra.

In Chapters 4 and 5, methods incorporating spectral simulation, digital filtering and supervised pattern recognition were developed for the automated detection of ammonia vapor from passive FT-IR spectra collected from the air. Despite fundamental differences in the phenomenon being monitored and the instrumentation used to record the data, the issues inherent in the gamma-ray detection problem are remarkably similar to those described previously for the passive FT-IR work.

Both methods suffer from weak analyte signals superimposed on varying backgrounds, thus necessitating a feature extraction step. The gamma-ray spectra are more challenging in this regard because the shape of the analyte signature changes with altitude because of the effects of Compton scattering. With both techniques, the collection of analyte-active field data for use in developing classifiers is difficult and costly. The practical challenges of releasing toxic chemicals into the atmosphere for the collection of FT-IR data is more than matched by the resistance to placing radioactive sources in the open air. With both methods, however, the collection of airborne background data is easily accomplished as a part of normal flight operations. Given that laboratory infrared and gamma-ray reference spectra are available for the analytes of interest, the use of data synthesis methods for simulating analyte-active data is possible with both types of data.

Because of these similarities, the same methodology developed in Chapters 4 and 5 was adopted to develop an automated detection of ^{137}Cs from airborne gamma-ray spectroscopic data. Spectral simulation was used to provide representative spectra of ^{137}Cs for assembling training and monitoring sets for the development of SVM classifiers. Digital filtering was again used to help extract features associated with the target isotope from the underlying background. Once

developed and optimized, the SVM classifier was applied to airborne test sets in order to evaluate the performance of the classifier in an actual airborne monitoring scenario.

6.2 Experimental

6.2.1 Instrumentation and data collection

Gamma-ray spectroscopic data used in this work were collected by our research collaborators at the United States Environmental Protection Agency. A RSX-4 airborne gamma spectrometer (Radiation Solutions Inc, Mississauga, Ontario, Canada) specifically designed for airborne detection and measurement of gamma radiation from both naturally occurring and manmade sources was employed. The spectrometer was equipped with two detector units. Each unit consisted of four 2''×4''×16'' thallium-activated sodium iodide (NaI[Tl]) scintillation crystals. An automatic multi-peak gain stabilization detector setup converted the analog signals to 1024 digital channels, with a spectral resolution of 2.98 keV.

To collect airborne data, the spectrometer mentioned above was mounted in a downward-looking orientation on a twin-engine aircraft (Aero Commander 680 FL, Aero Commander, Culver City, CA). The aircraft flew at altitudes ranging from ~ 100 to 3000 ft with a cruising speed of 100 – 110 knots. Survey passes over targeted ground locations were typically flown at altitudes in the range of 100 to 1000 ft. The data acquisition rate was approximately one spectrum per second. A radar altimeter (TRA 3000, FreeFlight Systems, South Waco, TX) was employed to estimate the height measurment above the ground level (AGL). Typically, each spectral point represented an area of 10 acres (40,469 m^2), with an effective field-of-view of 6 acres (24,281 m^2).

During the course of the target surveys, attempts were made to collect background data in areas away from the suspected target areas. These collections were performed at altitudes of 3000 ft and at lower altitudes. The data collected at 3000 feet contained no ground emissions

because of the long atmospheric path and the high probability of loss of all ground signals by scattering. For the work described here, none of these backgrounds were used as part of the data processing because it was desired to develop methodology that did not require background data at the measurement site.

Three types of data were employed in this work: (1) a reference pure-component spectrum of ^{137}Cs acquired with the aircraft parked over a cesium point source, (2) a database of 19,939 background spectra acquired across a range of altitudes and locations, all containing no ^{137}Cs signatures, and (3) three test sets acquired at three locations in which it was known that ^{137}Cs was present. The reference spectrum and background data were used in the construction of the automated classifiers while the test sets were used to evaluate the performance of the classifiers in actual field survey applications.

The three test sets, designated *Sedan*, *Smallboy* and *Desert Rock*, were conducted at three different sites located with the U. S. Department of Energy's Nevada Test Site. The *Sedan* and *Smallboy* data sets were overflights of sites of underground nuclear tests in the 1960's. As such, there was no reference information available as to the exact locations of ^{137}Cs contamination. The *Desert Rock* data set was collected at an airport site in which manmade ^{137}Cs and cobalt-60 (^{60}Co) point sources were placed at either end of an airport runway.

For the three data sets, the aircraft flew a series of passes over the survey area in a grid pattern. Separate grids were typically flown for different altitudes. For the *Sedan* data set, grids were flown at 500 ft (150 m) and 1000 ft (300 m). With the *Smallboy* data set, the altitudes investigated were 300 ft (90 m), 500 ft (150 m), and 1000 ft (300 m). The *Desert Rock* data contained aircraft passes at approximately 80, 150, 350, 500, 650, 800, and 1000 ft. The corresponding altitudes in meters were approximately 25, 50, 100, 150, 200, 250, and 300 m. For

the *Sedan* and *Smallboy* prediction sets, data collected at each altitude were treated as separate data sets so that classification results could be studied as a function of altitude. Because of the simpler nature of the data (i.e., all of the active patterns concentrated in a single known location) and the larger number of altitudes present, it was simplest to treat all of the altitudes together in the *Desert Rock* data set.

Spectra in the three test sets were visually examined in an effort to assign reference classifications (i.e., ^{137}Cs active or inactive) so that performance statistics could be computed for the SVM classifiers. This was done by a blind peer-review process in which two independent reviewers inspected the spectra and classified them into active, inactive, or uncertain categories. Those spectra for which the judgments were in agreement were assigned the corresponding reference classifications. If there was disagreement, the corresponding spectrum was put in the uncertain category and was not used subsequently in the calculation of performance statistics.

Visual inspections were judged to be reliable for the *Sedan* and *Desert Rock* data sets. Figures 6.1, 6.2, and 6.3 are images generated from the results of the visual review of the *Desert Rock* data (prediction set *A*), *Sedan* data acquired at 500 ft altitude (prediction set *B*), and *Sedan* data acquired at 1000 ft altitude (prediction set *C*), respectively. Images were generated with the latitude and longitude recorded by the aircraft corresponding to its position when each gamma-ray spectrum was collected. The symbols plotted for each location are color-coded to indicate the reference classifications (i.e., red = active, blue = uncertain, violet = inactive). Spectra in the uncertain class are generally adjacent to the active spectra.

Cesium-137 signals were very weak in the *Smallboy* data, and reference classifications were judged unreliable. For this data set, the interpretation of the classification results will be qualitative.

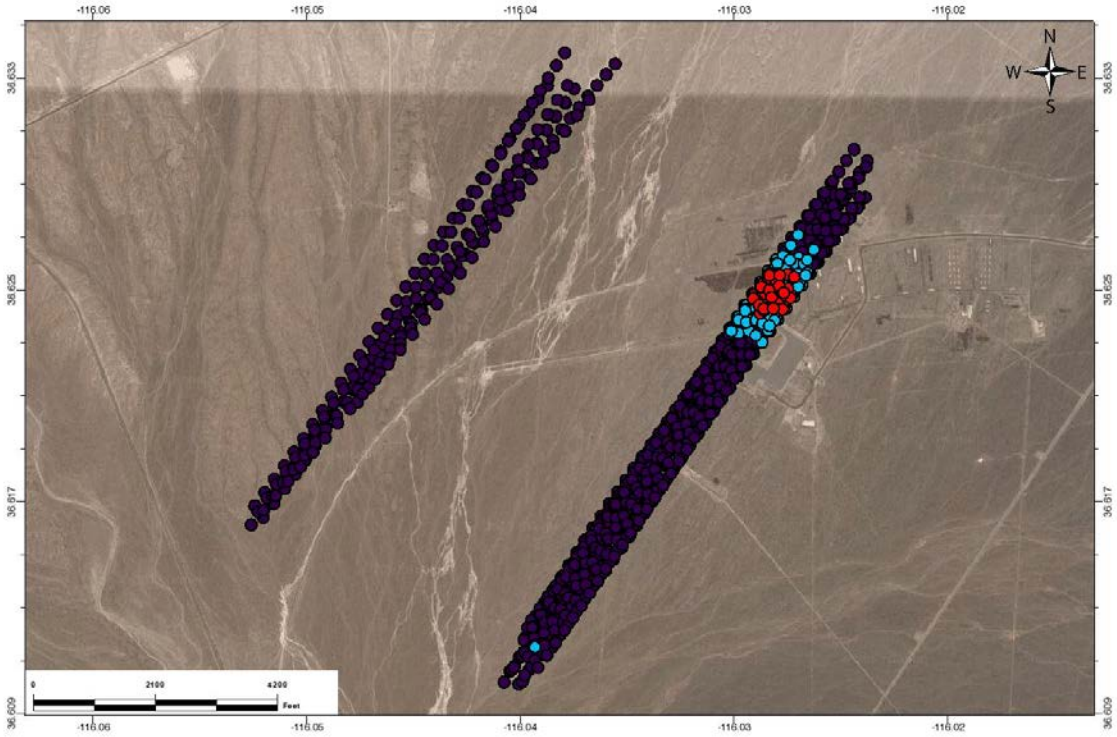


Figure 6.1 Image generated from the reference classifications of the *Desert Rock* data set (prediction set *A*). Points plotted in red, blue, and violet correspond to the active, uncertain, and inactive data classes, respectively, as determined by visual inspections of the gamma-ray spectra.

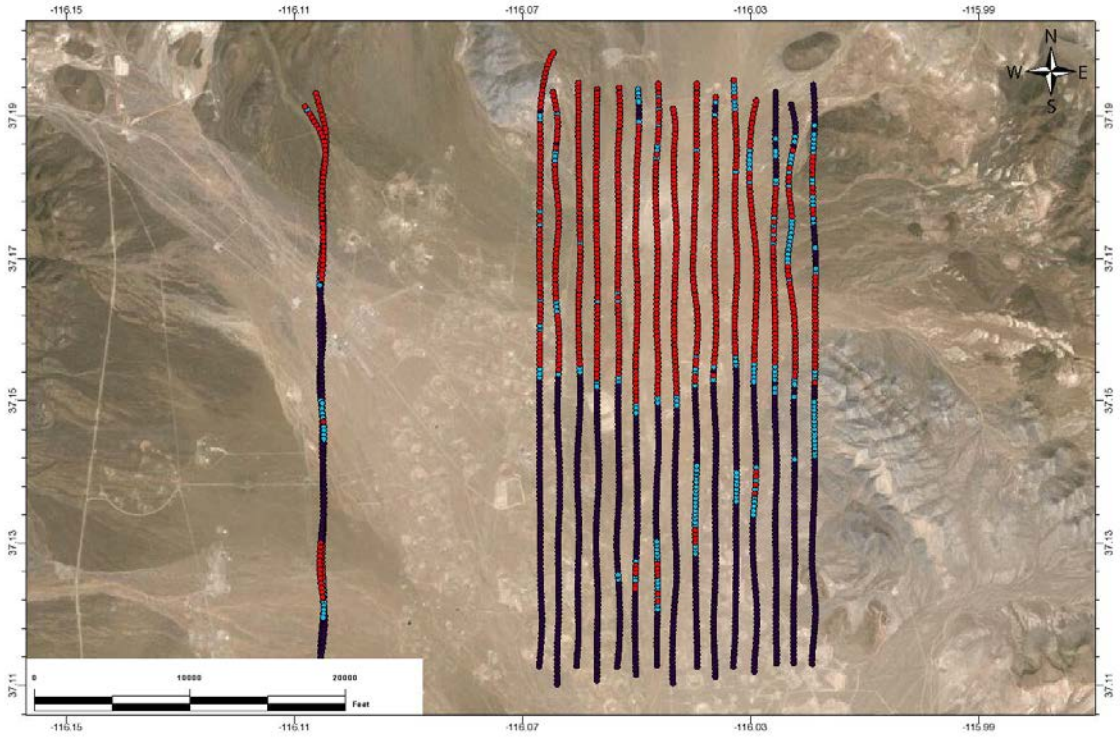


Figure 6.2 Image generated from the reference classifications of the spectra in the *Sedan* data set collected at 500 ft altitude (prediction set *B*). Points plotted in red, blue, and violet correspond to the active, uncertain, and inactive data classes, respectively, as determined by visual inspections of the gamma-ray spectra.

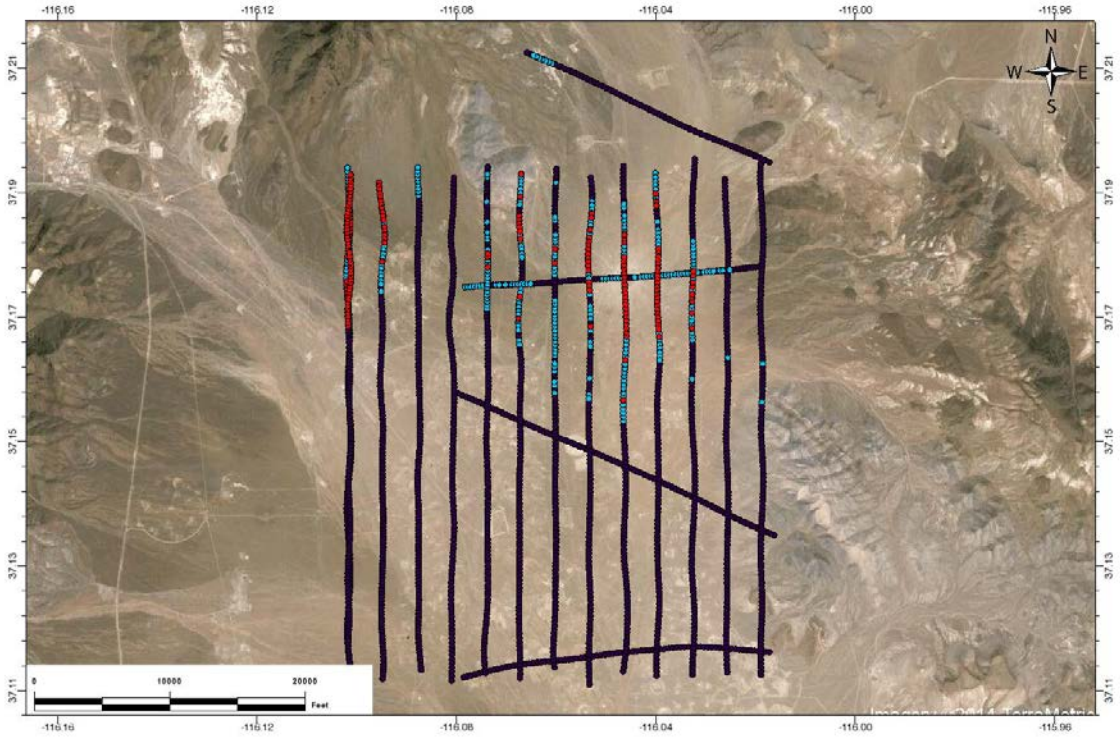


Figure 6.3 Image generated from the reference classifications of the spectra in the *Sedan* data set collected at 1000 ft altitude (prediction set *C*). Points plotted in red, blue, and violet correspond to the active, uncertain, and inactive data classes, respectively, as determined by visual inspections of the gamma-ray spectra. When compared to Figure 6.2, many fewer detections are made at 1000 ft.

6.2.2 Data analysis implementation

Digital filters were designed with the aid of the Filter Design and Analysis Tool (FDATool) provided with the Signal Processing Toolbox (version 5) of Matlab Version 7.4 (The MathWorks, Natick, MA). The SVM classifiers were trained and tested with the public-domain package, SVM^{light} (Version 6.01, http://svmlight.joachims.org). Digital filtering, SVM classification and all other data analysis tasks were performed on a Dell Precision 490 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Version 5.2, Red Hat, Inc., Raleigh, NC). Displays of SVM scores as a function of latitude and longitude were computed with the ENVI software system (Version 4.8, Excelis, Inc., Boulder, CO).

6.3 Results and Discussion

6.3.1 Overview of Methodology

As the target analyte to be detected in this work, ^{137}Cs has a characteristic photopeak at 662 keV, corresponding to photon emission while unstable Barium-137m (512 keV, 95 %), the major product of beta decay of ^{137}Cs , relaxes to stable Barium-137 (173 keV, 5%). In the airborne ^{137}Cs gamma-ray spectrum, the characteristic peak of ^{137}Cs is present along with a broad Compton scattering band at lower energies. As shown in Figure 6.4, the peak at 662 keV is the principal feature used to detect the presence of ^{137}Cs.

As emitted gamma-rays travel upward through the atmosphere, they lose energy as a consequence of collisions with atmospheric constituents such as nitrogen. This causes the gamma-ray spectrum to be altitude-dependent and suggests the need for altitude correction. In this work, measured gamma-ray spectral intensities at each spectral point i, y_i, were corrected for altitude as

$$y_i' = y_i \cdot e^{0.00524A} \qquad (6\text{-}1)$$

where A is the altitude in meters above ground level. The correction factor assumes a single exponential dependence on the loss of signal intensity with altitude. The attenuation coefficient in air, 0.00524 m^{-1}, is energy-dependent, but the value used here is typical of that employed in airborne gamma-ray spectrometry.

Figure 6.5 displays airborne background gamma-ray spectra corresponding to altitudes of 25, 100, and 300 m above ground level from the *Desert Rock* data set. Spectra are shown before and after application of Eq. 6-1. The spectra plotted are averages of 29 to 37 scans along the flight track. All of the spectra displayed were collected from the same geographical area. The altitude correction improves the similarity in intensities across the spectra from the three altitudes but some deviations in intensity remain, especially at low energies. Overall, the altitude correction appears effective in the region of the 662 keV peak of ^{137}Cs. For the data plotted in Figure 6.5, the percent relative standard deviation in intensity is 57 % at 662 keV in the uncorrected data. This value decreases to 13 % after applying the correction.

Previously, as described in Chapters 4 and 5, automated classifiers were developed for use in identifying VOCs from airborne passive FT-IR interferogam data. The same methodology was adopted here for the automated detection of ^{137}Cs from an analysis of airborne gamma-ray spectra. The SVM was again employed as the pattern recognition method. A training set composed of patterns representing the ^{137}Cs-active and ^{137}Cs-inactive data classes was assembled to develop the SVM models. Once developed, the SVM models were tested with the monitoring

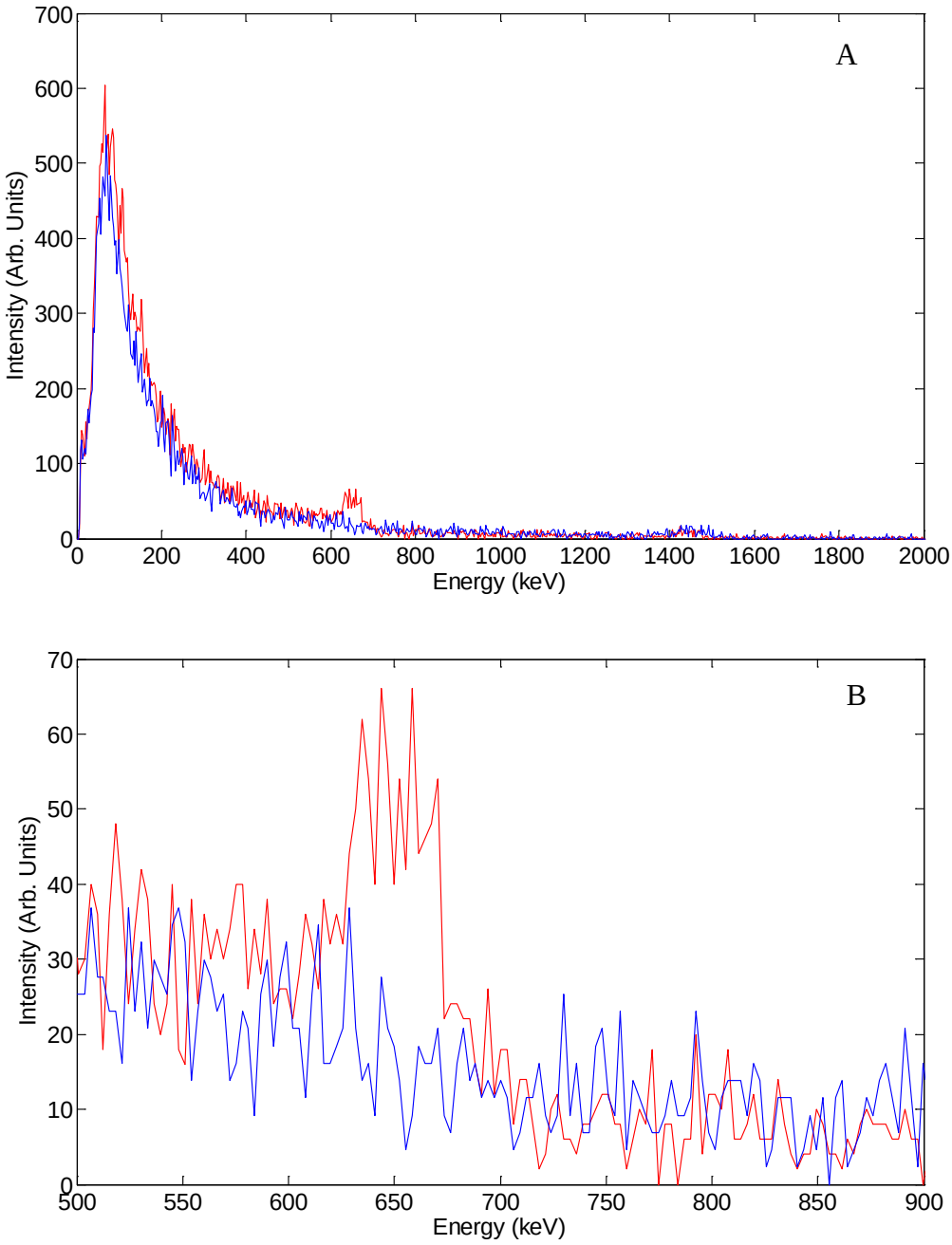


Figure 6.4 A. Airborne gamma-ray spectrum with ^{137}Cs (red) in comparison with a background spectrum (blue) in which no ^{137}Cs is present. B. Expanded view of the region of 500 to 900 keV.

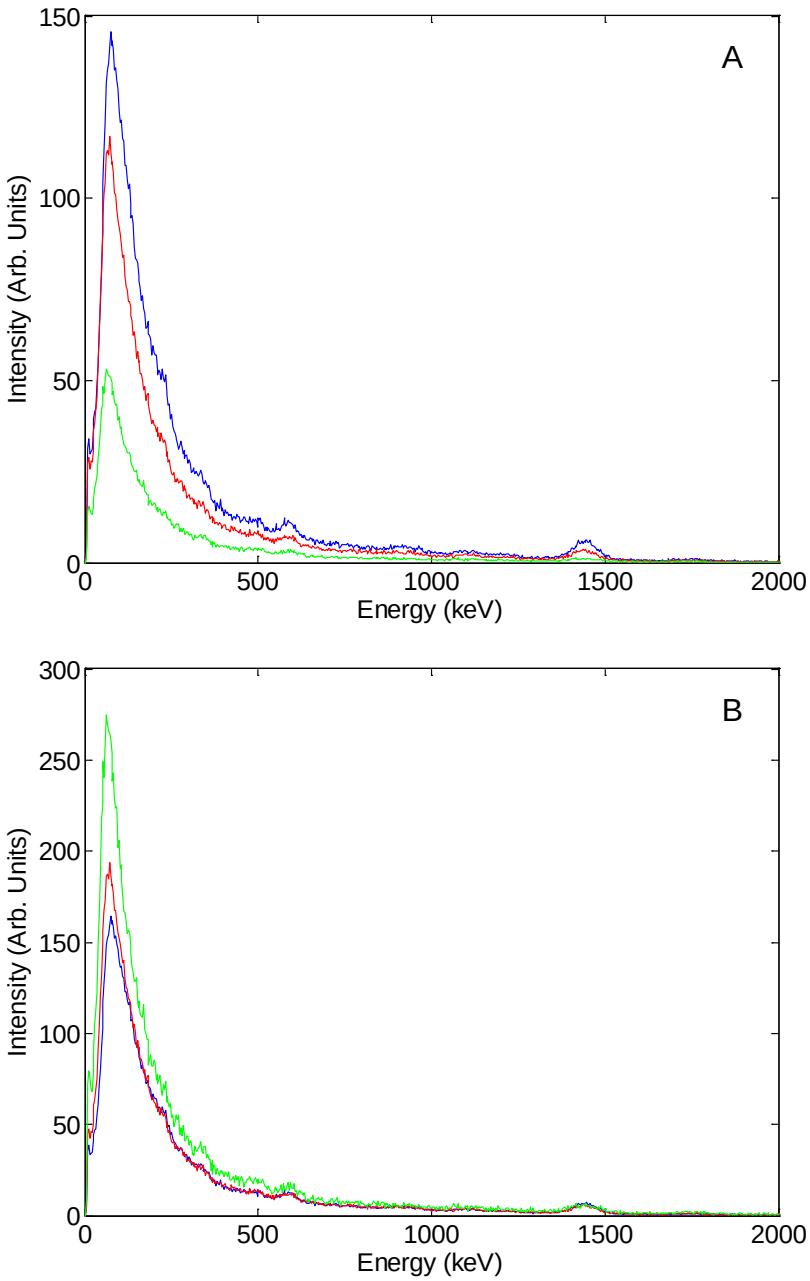


Figure 6.5 Average gamma-ray background spectra from the *Desert Rock* data set corresponding to altitudes of approximately 25 (blue), 100 (red), and 300 (green) m. Plots A and B, respectively, correspond to the original data and data after altitude correction with Eq. 6-1.

set, which was also composed of both ^{137}Cs-active and ^{137}Cs-inactive patterns. Based upon the classification results on the monitoring set, an optimal SVM model was chosen to be applied to the three test sets described previously. The model's performance was evaluated based upon the classifications for the tests sets as compared to the reference classifications and the geographical locations of the active classifications.

To provide patterns representing the ^{137}Cs-active class in the training and monitoring sets, mathematically synthesized gamma-ray spectra were generated. As illustrated in Figure 6.6, an array of ^{137}Cs signatures representing varying signal levels was produced by randomly scaling the "pure" ^{137}Cs gamma-ray spectrum collected from the calibration source. Synthetic spectra were then generated by adding the scaled ^{137}Cs spectra to randomly chosen background spectra collected from the air. The backgrounds used were a subset of 5000 selected from the 14,877 available backgrounds that had corresponding altitudes < 3000 ft. The subset selection procedure developed by Carpenter and Small [78] was used to select the 5000 background spectra.

Figure 6.7 displays five of the synthetic active spectra, with the region of the ^{137}Cs signature at 662 keV expanded in the inset. The signal levels displayed are typical of those encountered during field measurements. Figure 6.8 compares one of the synthetic spectra with scan 2345 in the *Sedan* test set, a spectrum confirmed to have the ^{137}Cs signature and which possesses a similar height of the peak at 662 keV. Both spectra were altitude corrected through the use of Eq. 6-1. From a comparison of the synthetic and field spectra, the match is good in the region above 400 keV, but the traces deviate at lower energy. This illustrates that the current data synthesis procedure does not incorporate the effects of Compton scattering.

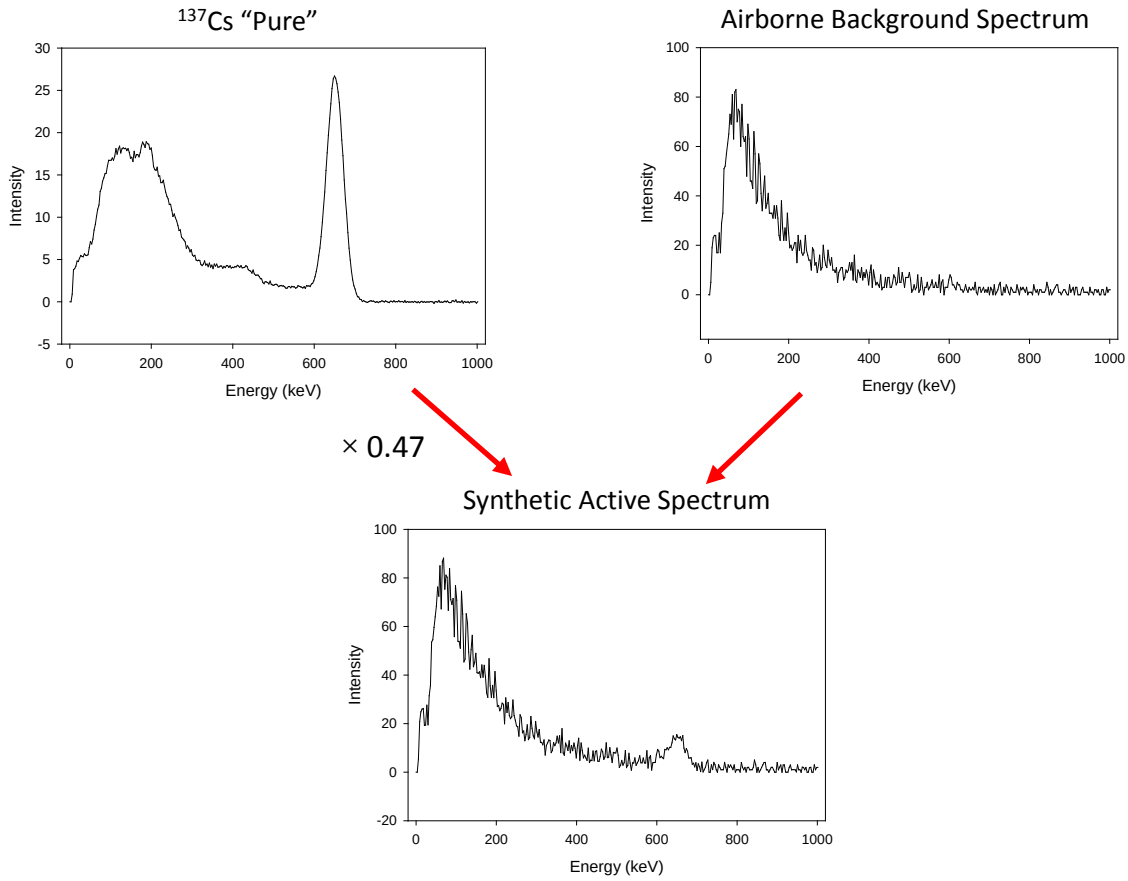


Figure 6.6 Schematic diagram of the calculation of synthetic spectra. A spectrum of ^{137}Cs collected from a calibration source is scaled by a randomly selected constant (e.g., 0.47 in the pictured example) and added to a randomly selected background spectrum collected from the air. The resulting synthetic active spectrum can be used in the training or monitoring set for use in developing and optimizing the SVM classifiers.

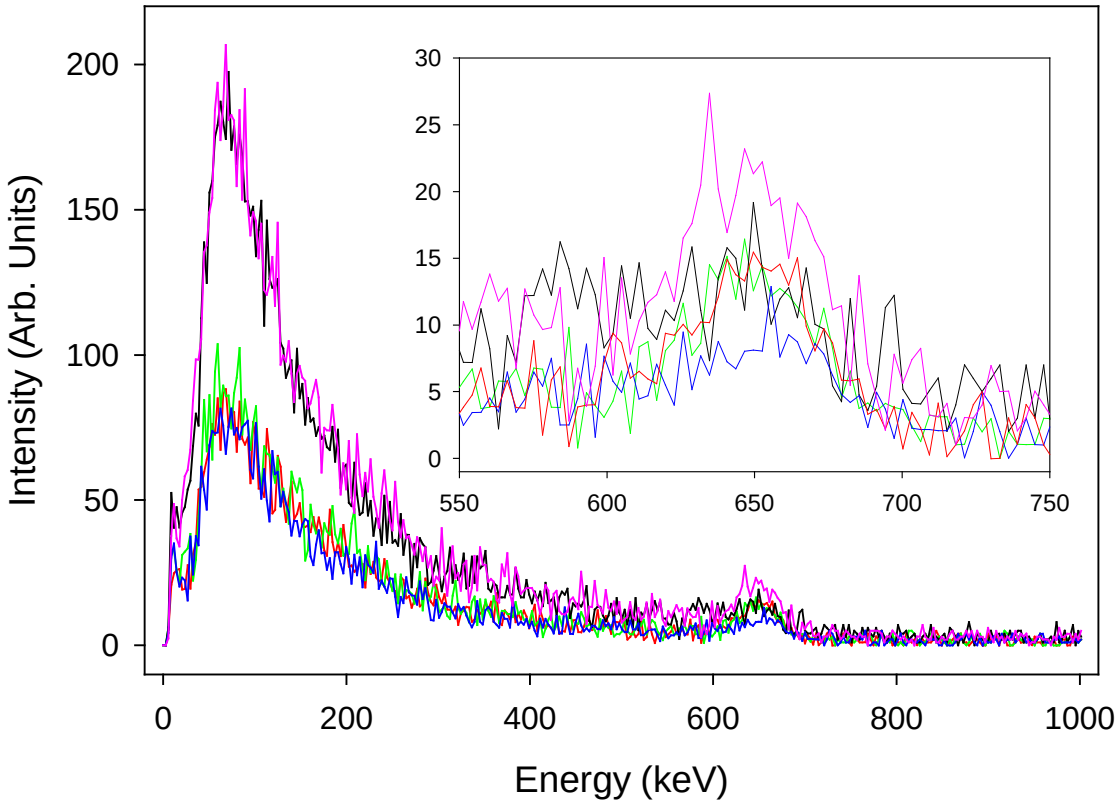


Figure 6.7 Five example spectra produced through the data synthesis procedure. The inset is an expanded view of the region of the characteristic signal of ^{137}Cs at 662 keV.

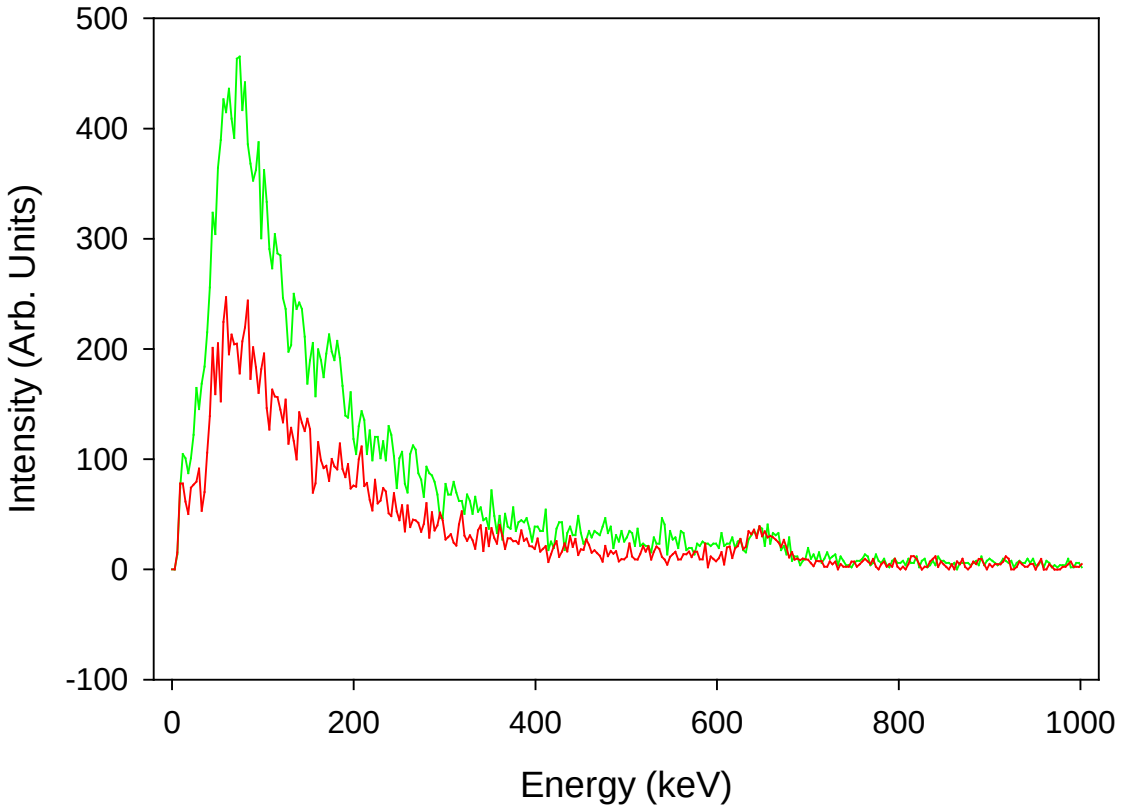


Figure 6.8 Comparison of synthetic (red) and active field data (green) in which ^{137}Cs is present. The field data corresponds to scan 2345 in the *Sedan* data set. Altitude scaling (Eq. 6-1) was applied to both spectra.

In order to enhance the selectivity and sensitivity of the SVM classifiers, data preprocessing steps were explored as a means to retain features most relevant to the ^{137}Cs target, while suppressing noise and unwanted spectral features. It was also hoped that data preprocessing could help to minimize differences between the synthetic and field data. In this work, bandpass digital filtering was chosen as the major data preprocessing method.

When applied to spectra such as those displayed in Figure 6.8, a digital filter considers the spectrum to be a time trace composed of a mixture of underlying harmonic frequencies. The filter attenuates or passes these underlying frequencies as specified by its frequency response function. Broad spectral features such as slowly varying background components are modeled by low frequencies, while narrow or rapidly changing features such as random noise are modeled by high frequencies. Thus, it is possible for a single bandpass filter to suppress both background and noise features.

For example, if the frequency response of the filter is optimized, spectral features of the width of the ^{137}Cs peak at 662 keV could be largely retained while wider features such as the Compton scattering background or narrower features such as random noise could be suppressed. As in the work described in Chapters 4 and 5, IIR digital filters are investigated here for their utility in improving the SVM classifiers.

6.3.2 Data partitioning

As mentioned earlier, to develop classifiers based on supervised pattern recognition algorithms, the spectral data are partitioned into training, monitoring and prediction sets. The data partitioning scheme is summarized in Table 6.1. In this work, the training set and monitoring set were formed in a similar manner. The training set contained 3000 ^{137}Cs-active patterns and 12,000 ^{137}Cs-inactive patterns, while the monitoring set consisted of 5000 ^{137}Cs-active patterns and 7939 inactive patterns.

Active patterns for both training and monitoring sets were from synthetic data, while the inactive patterns were selected from the database of background data collected during field surveys conducted by the aircraft. Both the actives in the training and monitoring sets made use of the same set of 5000 backgrounds in the data synthesis, but different random combinations of background spectrum and scaling factor were employed.

For the 12,000 inactive patterns in the training set, 5000 were the same pool of backgrounds used in the data synthesis, 5000 were another subset selected from the 14,877 backgrounds with altitudes < 3000 ft, and 2000 from picked from the 5062 backgrounds collected in the region of 3000 ft. The inactives in the monitoring set contained the remaining 3062 backgrounds collected at 3000 ft and the 4877 background spectra not used previously from the original group of 14,877. Subset selections were again performed by use of the method of Carpenter and Small. [78]

As described previously, the prediction sets used in this work were derived from the *Sedan*, *Smallboy*, and *Desert Rock* data sets. The *Sedan* and *Desert Rock* data had assigned reference classifications based on visual inspections, while the *Smallboy* data was assessed qualitatively. The *Sedan* and *Smallboy* prediction sets were divided into two and three subsets, respectively, on the basis of altitude. Table 6.1 lists the prediction sets and provides details regarding the number of active and inactive spectra where available.

6.3.3 Characterization of Synthetic Data

The strengths of the ^{137}Cs signals in all the data subsets were estimated on the basis of the spectral signal-to-noise (S/N) ratio as described previously.[83] [84] A difference spectrum was calculated as the residual obtained after performing a linear regression of the ^{137}Cs active spectrum under study onto a representative inactive spectrum. The difference between the maximum and minimum in the residual spectrum over the range of 540 to 780 keV was

calculated to estimate the ^{137}Cs signal intensity. The noise level in the difference spectrum, s_d, was computed as the standard deviation of the spectrum over the range of 2502 – 3000 keV, where no substantial peaks were observed. Assuming the noise levels in the active and inactive spectra are equal, and there are no errors in the regression coefficients, b_0, and b_1, the noise level in the original spectrum, s_s, can be approximated as:

$$s_s = \sqrt{\frac{{s_d}^2}{1+ {b_1}^2}} \qquad (6\text{-}2)$$

Then, the S/N ratio was computed as the calculated ^{137}Cs signal intensity over the estimated noise level.

The computed S/N ratios for the data sets are displayed as box plots in Figure 6.9. The horizontal lines associated with each box denote the median S/N ratio and the box limits define the quartiles of the distribution. Whiskers are drawn to the most extreme data value that lies within 1.5 times the interquartile range (difference between first and third quartiles). The whiskers span the nominal expected range of the data values. Points lying outside this range are plotted with individual symbols. The box plots show that the median S/N ratios for the training and monitoring sets, both composed of synthetic spectra, were 2-4 times higher than the corresponding values for the prediction sets, composed of field data in which the ^{137}Cs signal was present on the basis of visual inspection. Ideally, the strength of the signals across the different data sets should be at a similar level, in order to align the SVM model with the data to which it will ultimately be applied. The occurrence of higher S/N ratios in the active patterns in the training and monitoring may have a negative impact on the ability of the SVM model to detect low concentrations.

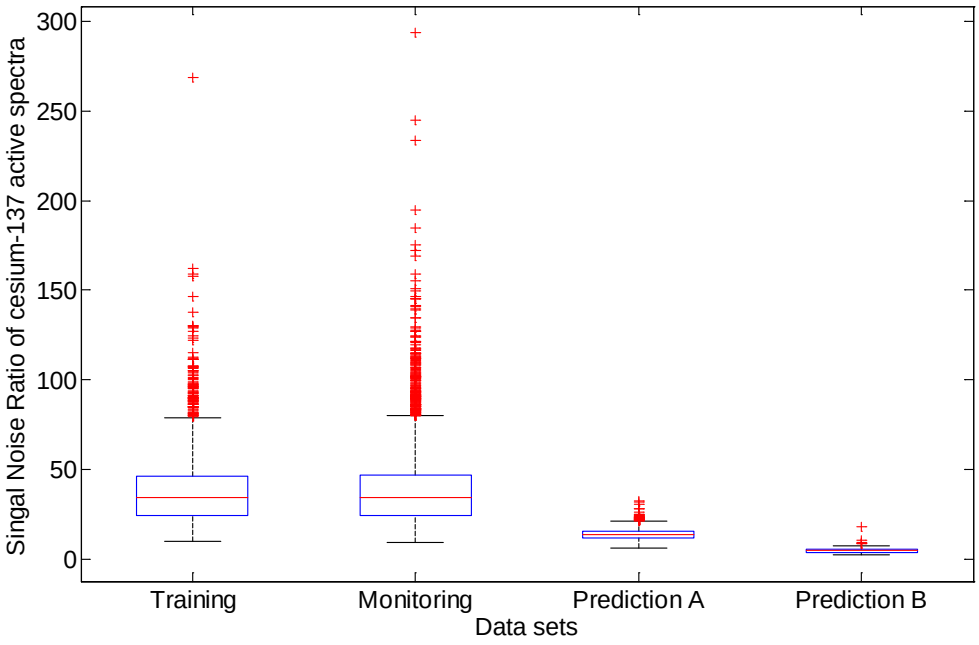


Figure 6.9 Box plots computed from S/N ratios of ^{137}Cs-active patterns across different data sets. Prediction sets A, B, and C correspond, respectively, to the *Sedan* data set at 500 ft altitude, the *Sedan* data set at 1000 ft altitude, and the *Desert Rock* data set across all altitudes.

6.3.4 Effects of digital filtering

The Chebyshev Type II infinite impulse response (IIR) filter was employed as the design method for the bandpass digital filters used in this work as preprocessing tools. Nine filters with varying passband lower limits, 0.01, 0.02, and 0.03 and upper limits, 0.05, 0.06, and 0.07, resulting in varying passband widths from 0.02 to 0.06, were designed and tested. The range of upper limits was established in preliminary work, where a low-pass filter with a cutoff about 0.6 was used successfully. The range of lower limits was added to obtain passbands that suppressed low-frequency background components while being stable in the Chebyshev Type II filter design. All filters had a stopband attenuation of 60 dB. The dB scale is logarithmic, with 20, 40, and 60 dB representing 10, 100, and 1000 times reduction of the input signal intensity. The specifications of the computed filters can be found in Table 6.2.

As an illustration, the frequency response of the filter designed with a passband of 0.03 to 0.05 (normalized frequency scale of 0 to 1) is shown in Figure 6.10. The action of this filter is illustrated in Figure 6.11 with 10 spectra from the *Sedan* data set. The filter was applied with a single pass through the data, starting 14 points before the first point plotted. Scans 2347-2351 (red) were confirmed by visual inspection to have the ^{137}Cs signature at 662 keV. Similarly, scans 4649-4653 (green) were confirmed to have no evidence of ^{137}Cs. Figure 6.11 A shows the spectra over the range of 400 to 900 keV after application of altitude correction. The weak ^{137}Cs feature can be observed in the spectra plotted in red. Figure 6.11 B shows the corresponding spectra after application of the filter whose frequency response is shown in Figure 6.10. The filter suppresses both random noise and baseline components, producing a derivative-like result. The active (red) and inactive (green) traces are distinct over the region of approximately 550 to 850 keV.

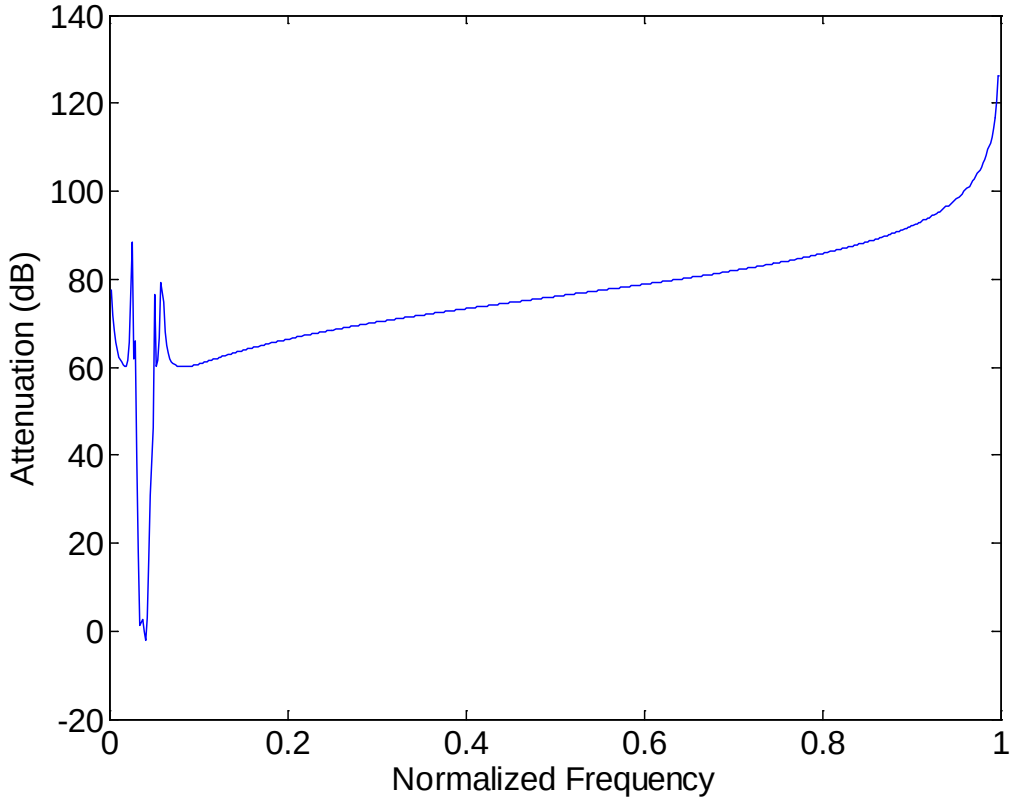


Figure 6.10 Frequency response of Chebyshev Type II filter with passband starting and stopping limits of 0.03 to 0.05 on the normalized frequency scale of 0 to 1. The stopband attenuation is 60 dB. The dB scale is logarithmic, with 20, 40, and 60 dB corresponding to factors of 10, 100, and 1000 by which the input signal is suppressed. This filter has the narrowest passband generated.

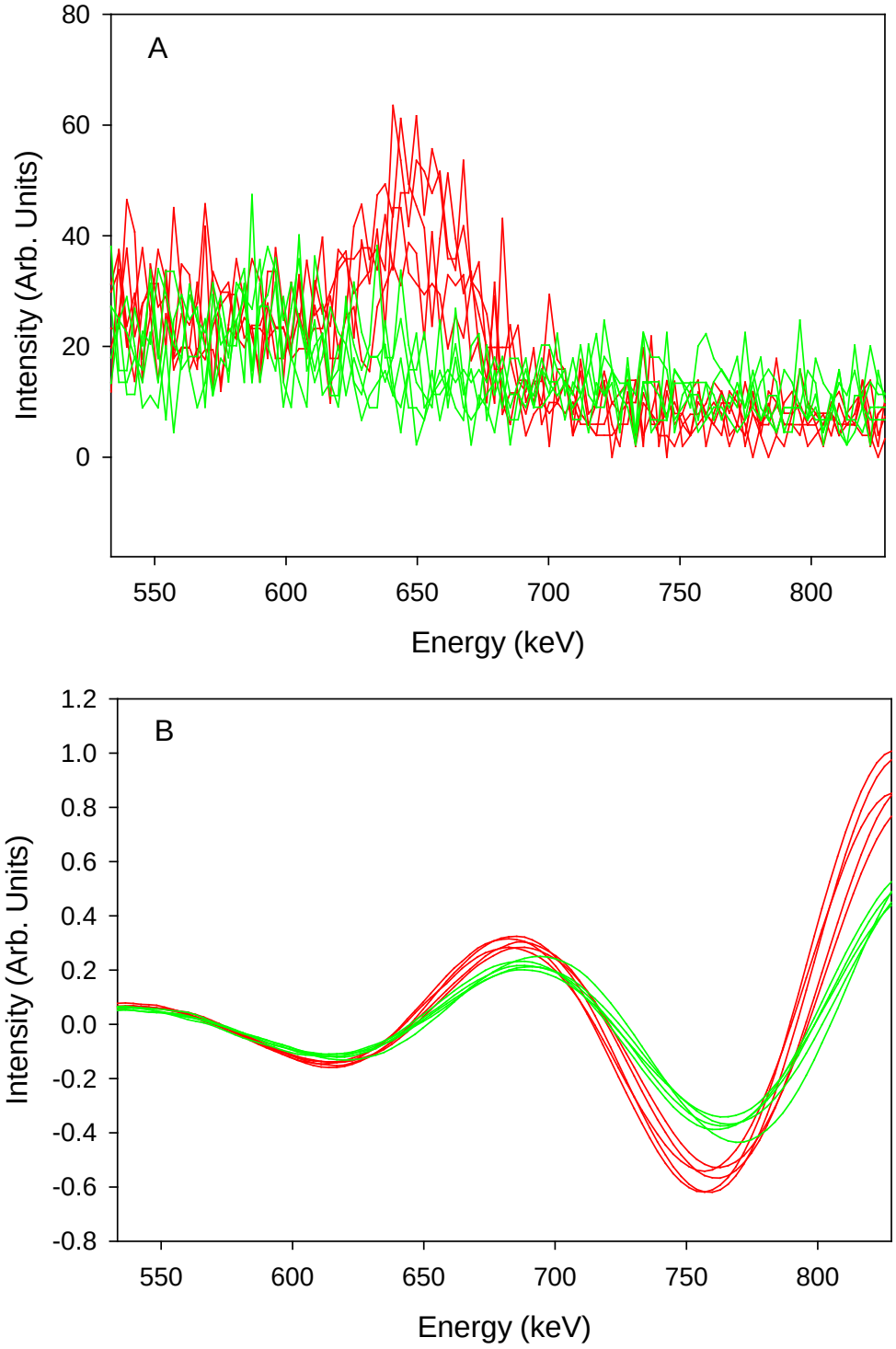


Figure 6.11 Spectra from the *Sedan* data set corresponding to scans 2347-2351 (red) and 4649-4653 (green). A. Spectra after altitude correction. The spectra plotted in red have the characteristic peak of ^{137}Cs at 662 keV. B. Spectra after application of the filter whose frequency response is plotted in Figure 6.10.

Figure 6.12 compares a subset of the spectra in Figure 6.7 (2347, 2349, 2351, 4649, 4651, 4653) to three of the synthetic ^{137}Cs-active spectra from the training set. Panel A plots the altitude-corrected spectra and panel B plots the spectra after application of the filter whose frequency response is displayed in Figure 6.10. Two of the synthetic spectra from the training set (blue traces) are higher in intensity than the plotted ^{137}Cs-active spectra from the *Sedan* data set. As in Figure 6.11, the filtering produces derivative-like traces. The active and inactive spectra are distinct in the filtered spectra, with the increasing ^{137}Cs signal strength producing a consistent pattern.

6.3.5 SVM model selection

A full factorial experimental design, as detailed in Table 6.2, was developed to include five factors at different levels. Three levels of filter passband lower limit, three levels of filter passband upper limit, 12 levels of spectral segment starting point, 12 levels of the SVM kernel parameter (γ), and 12 levels of the SVM regularization parameter (C) were explored. The stopband attenuation was set at 60 dB, the gap between the segment starting point and filter starting point was set to 14 points (i.e., the filter began operating 14 points before the first point used in constructing the patterns for classification), and the segment length was fixed at 100 points. The resulting combinations with varying conditions were applied to the training set to generate a pool of 15,552 SVM classifiers.

Once developed, all SVM classifiers were tested with the monitoring set using the same processing steps as applied to the training set. In calculating classification performance, the conventional SVM classification rule was used (i.e., positive SVM scores denote the active class,

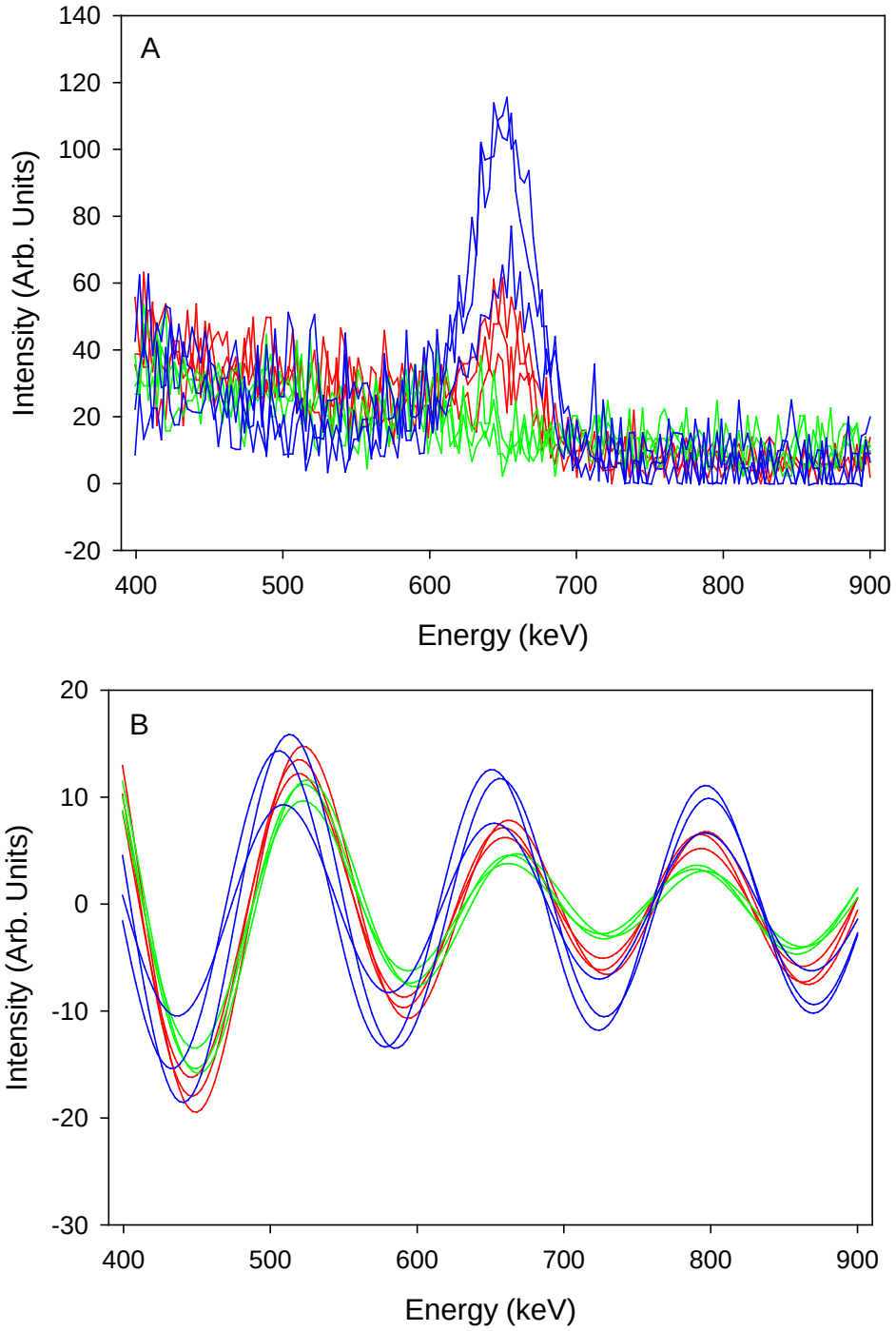


Figure 6.12 Spectra from the *Sedan* data set corresponding to scans 2347, 2349, 2351 (red) and 4649, 4651, 4653 (green) and three representative synthetic ^{137}Cs-active spectra from the training set (blue, sequence numbers 50, 53, 59). A. Spectra after altitude correction. The spectra plotted in red and blue have the characteristic peak of ^{137}Cs at 662 keV. B. Spectra after application of the filter whose frequency response is plotted in Figure 6.6. The active and inactive spectra are distinct from approximately 550 to 850 keV.

while scores ≤ 0 denote the inactive class). The classification performance obtained with the monitoring set, balanced by the results produced by the training set, was used to evaluate every classifier, and eventually to determine the optimal classifier and its corresponding segment and digital filter specifications.

Four criteria were used in combination to evaluate all classifiers developed, including (1) the number of support vectors, (2) the error count in the training step (i.e., the count of patterns incorrectly classified, including both missed and false detections), (3) the monitoring missed detection rate (i.e., the percentage describing the fraction of active patterns classified into the inactive class), and (4) the monitoring false detection rate (i.e., the percentage describing the fraction of inactive patterns classified as active patterns). The top classifiers chosen were those with the monitoring set false detection rate less than 1 %, the monitoring set missed detection rate less than 15 %, the training set error counts as small as possible, and the number of support vectors as large as possible.

According to these criteria, 10 top-performing classifiers were chosen. Their performance is summarized in Table 6.3. Variation in performance across the top classifiers was not dramatic as indicated by the sum of the missed and false detection percentages (last column in Table 6.3). A trend was found that models processed by the filter with passband starting and stopping specifications of 0.03 and 0.05, respectively, (i.e., models 5-8, and 10 in Table 6.3) resulted in better missed detection rates with the monitoring set, although with slightly higher rates of false detections. This filter was the narrowest investigated and its frequency response function is plotted in Figure 6.6. The spectra displayed in Figures 6.7 B and 6.8 B were produced by use of this filter.

6.3.6 Classification results with prediction sets

Each classifier listed in Table 6.3 was applied to all the prediction sets. Output scores from the SVM models were plotted for the prediction sets to facilitate comparisons. The score profiles generated from the different SVM classifiers were similar in terms of overall shape, differing only in the magnitude of the scores. The results obtained with Models 1 and 5 (Table 6.3) are presented here as representative. Model 1 produced the lowest rate of false alarms (0.24 %) with the monitoring set, while Model 5 had the fewest missed alarms (5.48 %) with the monitoring data and had the fewest misclassified patterns in the training set (103).

For prediction sets *A*, *B*, and *C*, the availability of reference classifications allowed classification percentages to be computed. Both the conventional SVM classification rule (i.e., SVM scores greater than zero denote the active class) and an approach based on the use of the SVM scores of the monitoring set as an external reference distribution were employed.

In Chapters 4 and 5, control charts were used in an approach that considered the inactive classification to be the norm and thereby treated active patterns as disruptions of the normal "process" being monitored. In the previous work, control limits for a run were computed based on the average and standard deviation of the SVM scores of that run. This approach assumed that the VOCs being detected arose from sources such as leaks or stack emissions and therefore represented a small fraction of the scans within the run.

This approach does not apply well to the use of gamma-ray spectroscopy in airborne surveys. While the source of gamma-rays could be a point source (e.g., a lost or stolen manmade radioactive source), it is also possible that radioactive material could have been spread over an extended area. For example, the *Sedan* and *Smallboy* prediction sets corresponded to locations where there is widespread presence of ^{137}Cs in the soil. For this reason, assignment of control limits for the classification must be done with either a fixed set of external reference data with

known class identities or through the collection of known “clean” backgrounds at the start of a data run.

An attempt at the latter approach was made during the collection of the *Sedan* data set. Two closely spaced aircraft passes were made away from the main survey area in an effort to acquire clean backgrounds. These scans can be seen at the far left of Figure 6.2. However, as indicated in the figure, visual inspections of the resulting spectra revealed clear evidence of ^{137}Cs. On the basis of this result, any attempt to define control limits on the basis of assumed clean backgrounds appeared highly problematic.

The approach adopted here made use of the SVM scores of the monitoring data to define a fixed set of control limits for a given classification model. Examination of histograms of the SVM scores suggested that the distribution was non-Gaussian and thus a simple representation based on the mean and standard deviation was judged to be inaccurate. An alternate approach was to define the set of SVM scores for the monitoring set as an external reference distribution and assign control limits on the basis of classification percentage as a function of SVM score.

For both Models 1 and 5 in Table 6.3, the SVM scores for the monitoring set were sorted and divided into bins spaced at increments of 0.5. This bin size reflected a tradeoff between good resolution of the SVM scores while also having a high percentage of occupied bins. The percentage of correctly classified patterns falling within each bin was tallied, and classification percentage was plotted as a function of bin center. Figures 6.13 A and B are the resulting plots for Models 1 and 5, respectively. The *x*-axis in both figures was truncated to the region of [-10, 10] for clarity.

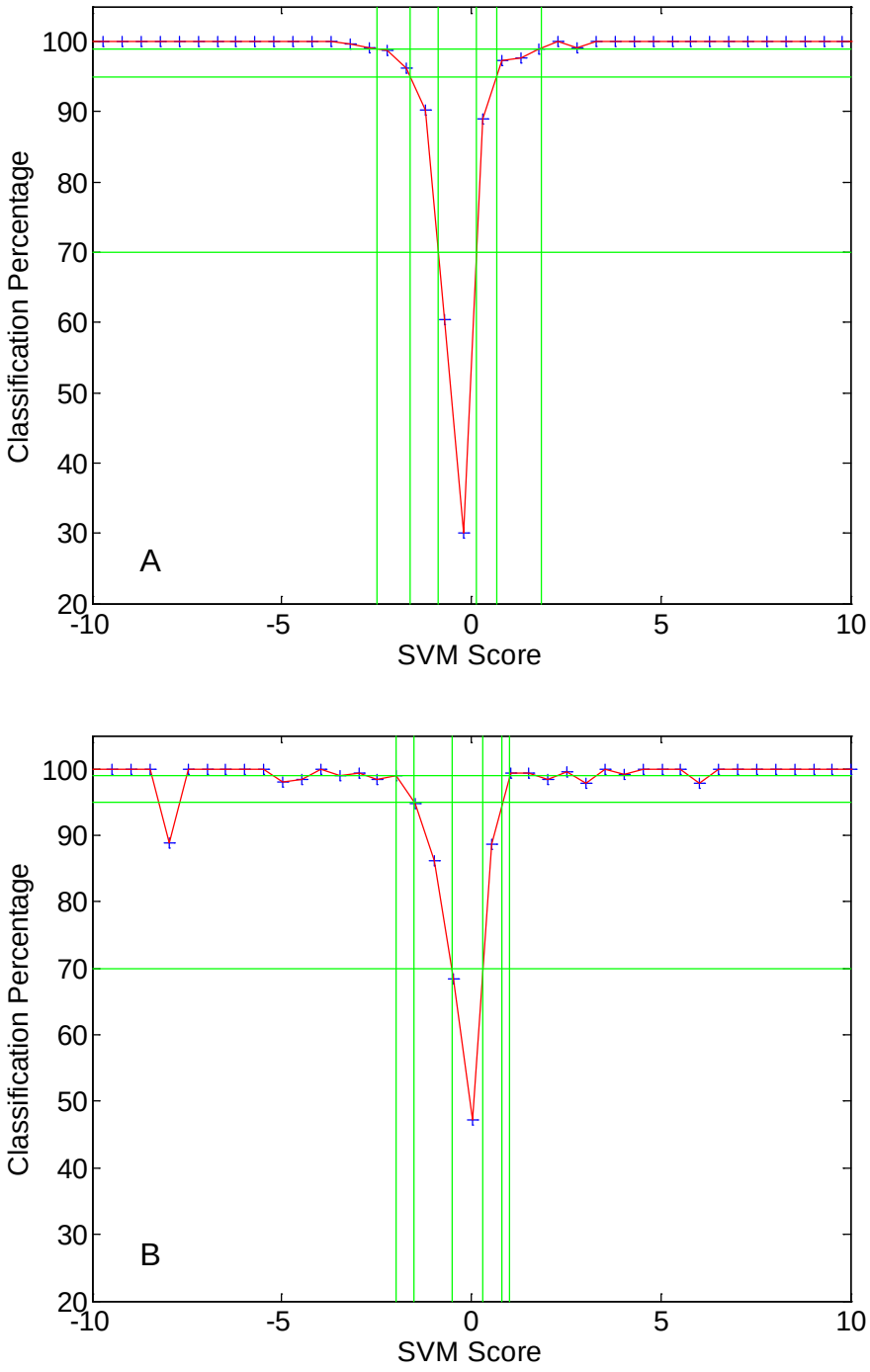


Figure 6.13 Classification percentage as a function of SVM scores for the data in the monitoring set. Plots A and B were derived from the scores for Models 1 and 5 in Table 6.3, respectively. Horizontal grid lines are drawn at 70, 95, and 99 % correct classification. The vertical grids specify the intersection of the horizontal grids with the corresponding value of SVM score. The *x*-axis was truncated to [-10, 10] for clarity.

Inspection of these figures reveals that at the positive and negative extremes of the SVM scores, 100 % classification is achieved. The classification percentage drops as the scores approach the conventional decision threshold of 0.0. Control limits were selected at 70, 95, and 99 % correct classification on each side of 0.0. For Model 1, this produced values of -2.5, -1.6, and -0.86 for 99, 95, and 70 % correct classification of the ^{137}Cs-inactive class. The corresponding values for the ^{137}Cs-active class were +1.8, +0.66, and +0.14 for 99, 95, and 70 % correct classification. The corresponding values for Model 5 were -2.0, -1.5, -0.52, +0.30, +0.82, and +1.0.

For prediction sets *A*, *B*, and *C*, Table 6.4 lists the classification percentages obtained by use of both the conventional classification rule and the approach based on control limits assigned from the distribution of SVM scores for the monitoring set. For this tabulation, the 70 % positive cutoff (i.e., 0.14 for Model 1 and 0.30 for Model 5) was used as the upper control limit (UCL). For patterns in the active class, SVM scores less than this cutoff signaled a missed detection. Patterns from the inactive class that exceeded the UCL were counted as false detections.

Results for the *Desert Rock* data set (prediction set *A*) were uniformly excellent. Model 1 slightly outperformed Model 5 in terms of minimizing both missed and false detections, but both models exhibited negligible errors when the results were viewed in absolute terms. Furthermore, the two classification rules yielded very similar results in terms of the patterns assigned to each class.

Figure 6.14 presents classification images resulting from Model 1 (A) and Model 5 (B). The presentation is analogous to Figure 6.2, but here colors are assigned on the basis of the SVM scores and derived control limits for each model. For patterns placed in the inactive class, black,

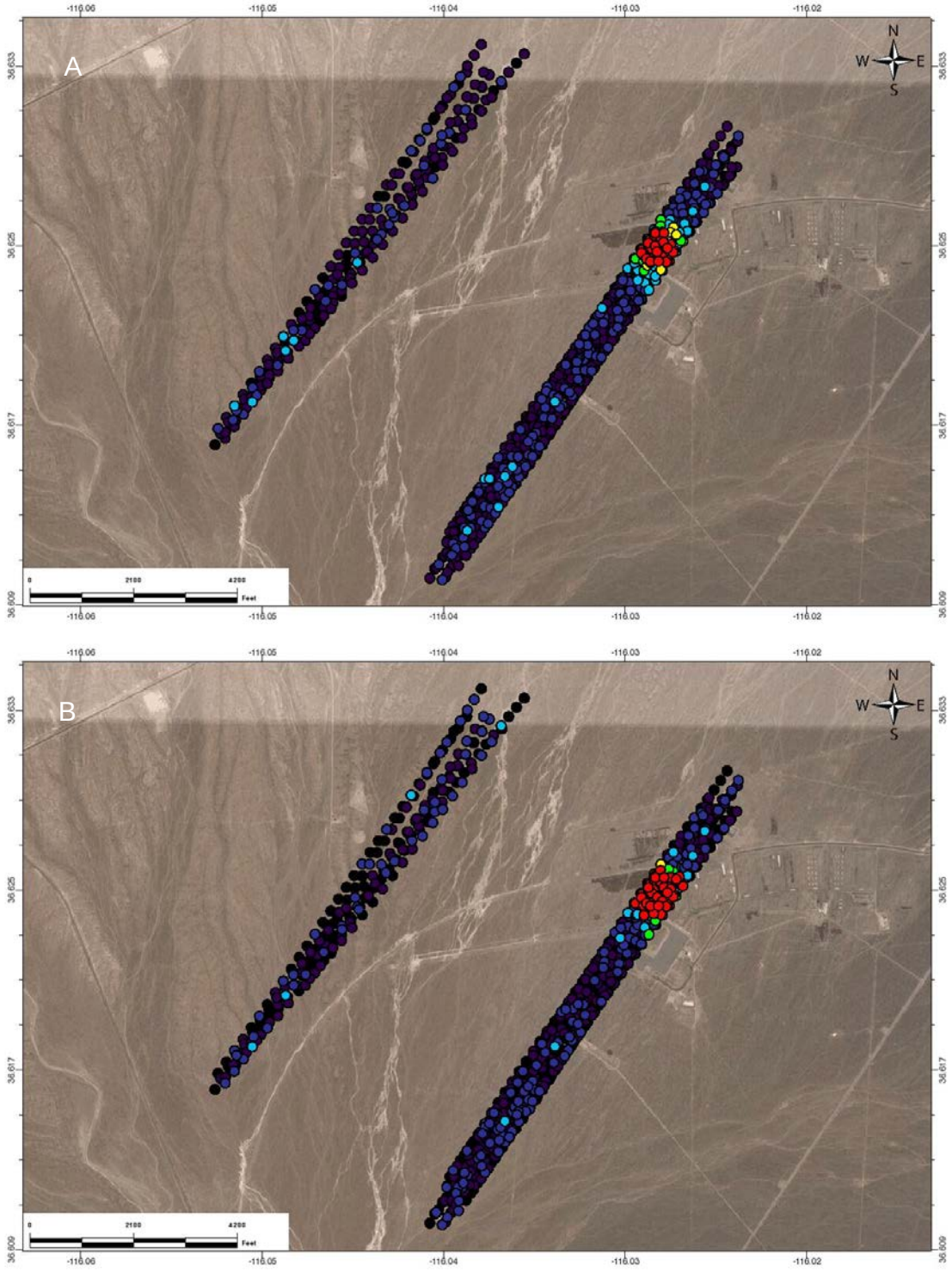


Figure 6.14 Classification images for the Desert Rock data set generated from the SVM scores for Models 1 (A) and 5 (B) in Table 6.3. Black, violet, and dark blue symbols specify the 99, 95, and 70 % levels for the inactive class, while red, yellow, and green symbols identify the corresponding levels for the active class. Light blue symbols denote uncertain classification.

violet, and dark blue symbols correspond to the 99, 95, and 70 % levels. The corresponding colors for patterns identified as ^{137}Cs-active are red, yellow, and green for the 99, 95, and 70 % levels, respectively. Thus, red and black symbols denote 99% or greater confidence that the corresponding pattern is active or inactive, respectively. Points plotted in light blue specify an uncertain classification as these spectra did not meet the 70 % cutoff for either class. A comparison of Figure 6.14 with Figure 6.1 confirms the excellent classification performance of both SVM models. The presence of more red symbols in Figure 6.14 B indicates that Model 5 has greater sensitivity than Model 1 as more spectra have been classified as active with a higher level of confidence.

The results listed in Table 6.4 for the 500 and 1000 ft altitudes of the *Sedan* data set (prediction sets *B* and *C*, respectively) are significantly worse in terms of classification percentages. False detections are better with Model 1 but still exceed 4 % in the best case. Missed detections approach 50 % for the data collected at 500 ft and 95 % for the data acquired at 1000 ft.

Figures 6.15 and 6.16 display the classification images obtained for prediction sets *B* and *C*, respectively. The format is identical to Figure 6.14. These images can be compared to Figures 6.2 and 6.3 to identify matching and mismatching locations relative to the classification results based on visual inspections of the spectra.

Inspection of Figure 6.15 reveals that the ^{137}Cs detections are overall in the correct locations in the upper half of the image. Both SVM models produce similar images, with Model 1 appearing to exhibit stronger detections (i.e., more red symbols specifying the 99% level). The

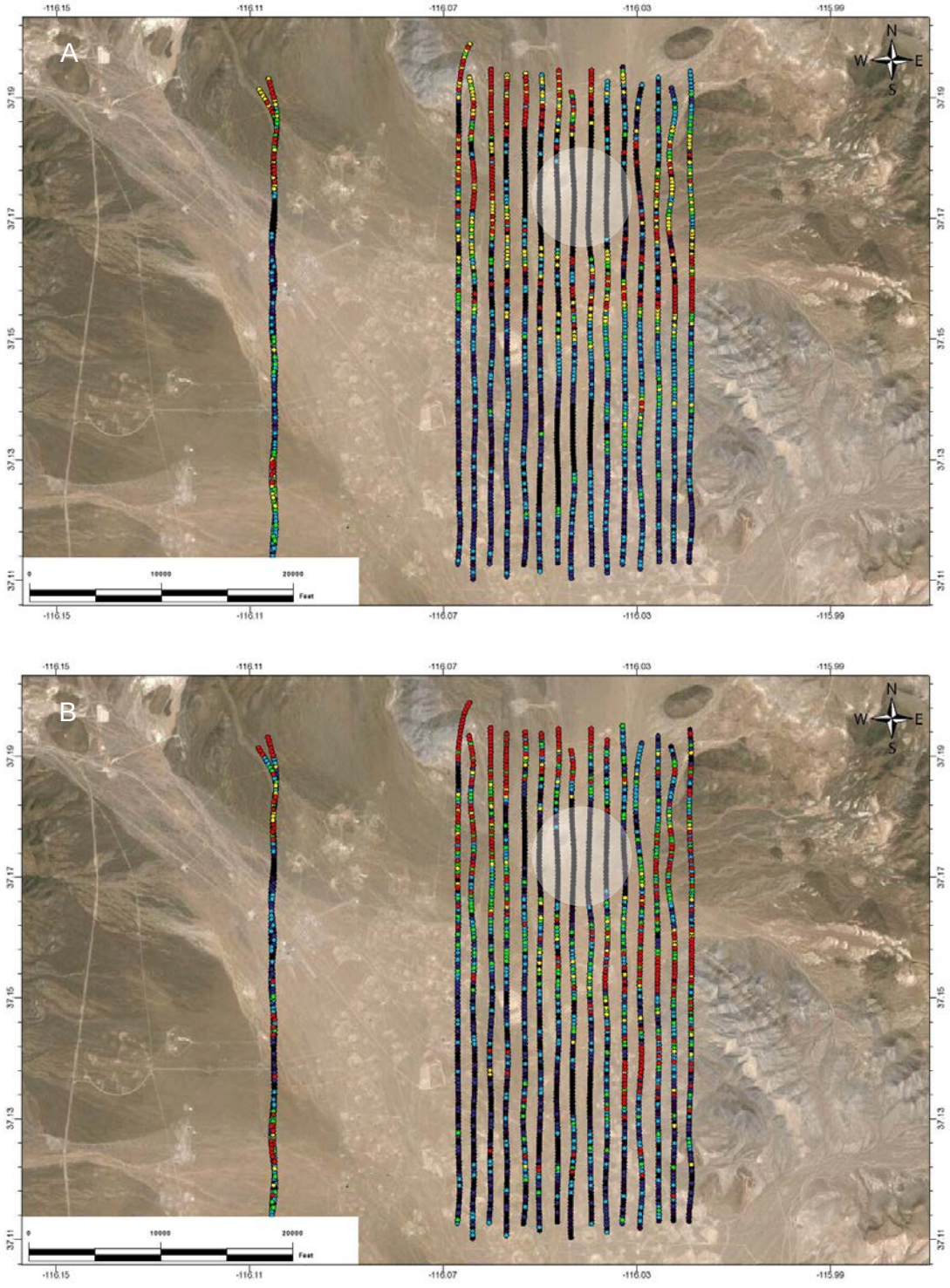


Figure 6.15 Classification images for the *Sedan* data set (500 ft altitude, prediction set *B*) generated from the SVM scores for Models 1 (A) and 5 (B) in Table 6.3. The presentation format is the same as that used in Figure 6.14. The circles in both plots highlight a region of strong ^{137}Cs signals that went undetected.

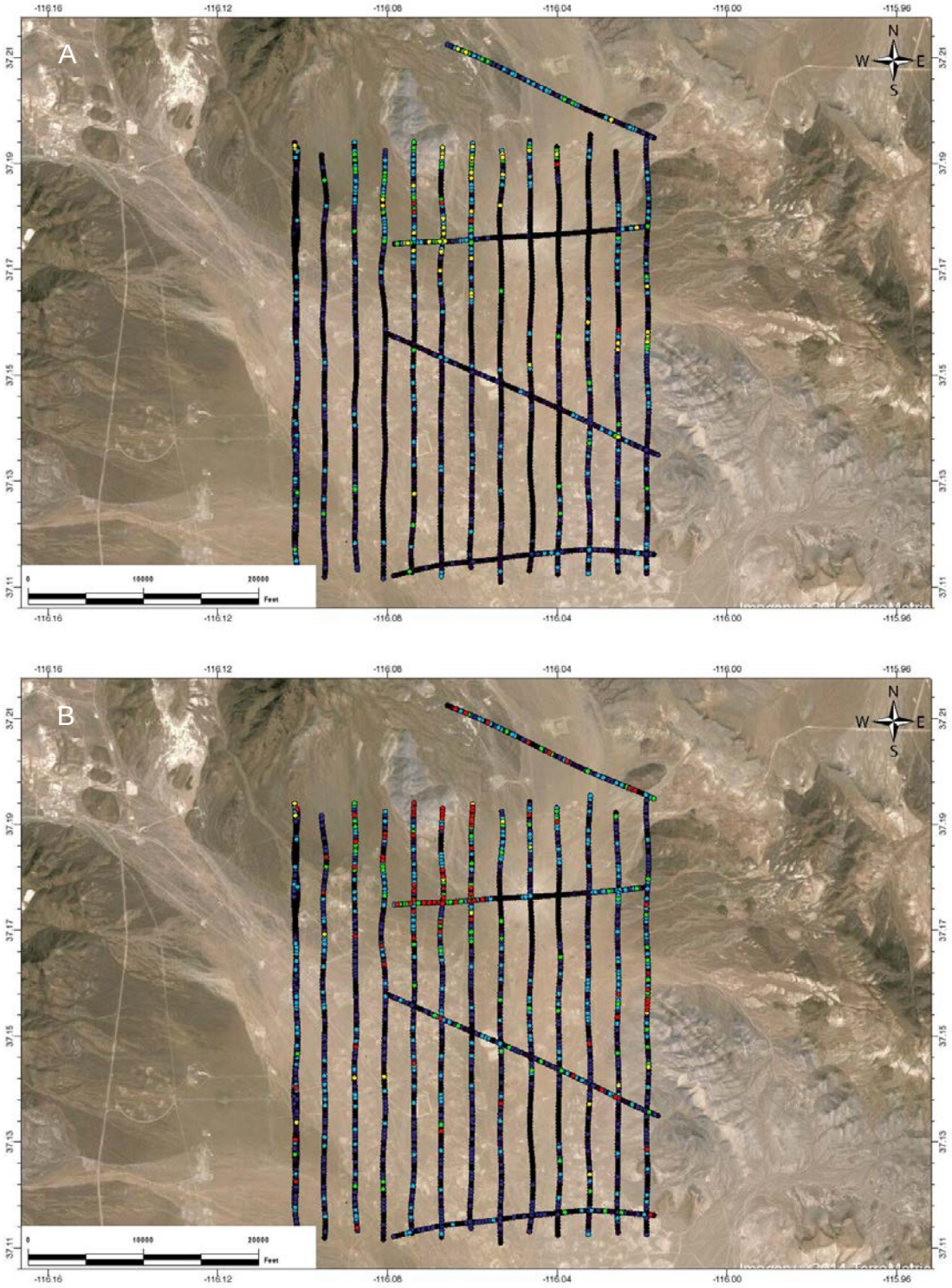


Figure 6.16 Classification images for the *Sedan* data set (1000 ft altitude, prediction set *C*) generated from the SVM scores for Models 1 (A) and 5 (B) in Table 6.3. The presentation format is the same as that used in Figure 6.14. In the upper middle of the plot, horizontal and vertical flight tracks cross at the Sedan Crater.

most interesting aspect of both plots, however, is a large contiguous circular region of missed detections. This region is highlighted by the white circle superimposed on each plot. The symbol color throughout this region is primarily black, indicating that these patterns are being identified at the 99 % level of the *inactive* data class.

The indicated region in Figure 6.15 is the area in which the nuclear detonation occurred 194 m underground in 1962. The crater made by this underground explosion can be seen in Figure 6.16 in the upper middle of both images near the point at which a horizontal and a vertical flight track cross. Visual inspections of the spectra corresponding to these locations reveal very strong ^{137}Cs signals.

To investigate these results further, PCA was performed on the patterns input to Model 5 (Table 6.3) derived from the spectra in the *Sedan* data set collected at 500 ft altitude. Three PCs were computed from the mean-centered patterns, accounting for 99.96 % of the data variance. Figures 6.17 and 6.18 are PC score plots in which the 2375 patterns are color-coded in different ways.

Figure 6.17 codes the patterns on the basis of the visual evaluation of the corresponding spectra for the presence of the ^{137}Cs signature at 662 keV. Red, black, and green symbols denote confirmed ^{137}Cs active patterns, confirmed inactive patterns, and patterns with uncertain classification. As expected, the red and black patterns occupy distinct regions in the data space with the green patterns largely being positioned at the interface between the active and inactive classes. The diffuse region of red patterns on the right side of the score plot correspond to the strongly active spectra collected from the Sedan crater area.

Figure 6.18 codes the patterns according to their classification by SVM Model 5. Red and black symbols specify the correctly classified active and inactive patterns, respectively. Cyan

symbols are false positives, while blue symbols denote missed detections. The false detections lie at the interface between the two data classes and primarily consist of those patterns judged to be of uncertain classification in the visual inspection of spectra. Many of these spectra may in fact have weak ^{137}Cs signatures. The interesting result revealed in Figure 6.18 is the large group of missed detections on the right side of the plot. These are the strongly active spectra collected near the Sedan crater. This result reveals that the nonlinear SVM classifier has effectively modeled the interface between the inactive and weakly active patterns, but has failed to group the strong ^{137}Cs signatures with the weak ones.

In general, a result of this type signals that the training set used to generate the SVM classification model has failed to capture the variety of signatures that will be encountered in the active class when the model is applied outside of the training set. Initial attempts to increase the population of strong actives in the synthetic training data did not correct the problem, however. This remains an issue under current study.

Classification images generated from the application of SVM Models 1 and 5 (Table 6.3) to the *Smallboy* prediction set are presented in Figure 6.19 for the data collected at 300 ft altitude and in Figure 6.20 for the data acquired at all altitudes. The same color scheme used in generating the classification images presented previously is employed. As discussed previously, the very weak ^{137}Cs signatures present in this data set prevented an accurate assessment of classification percentages. Interpretation of these results will be done in a qualitative manner.

Both SVM models detected a broad contiguous area of ^{137}Cs active patterns at 300 ft and a scattering of detections at higher altitudes. Model 5 judged the detections to be higher in probability, with a large area at >99%. To help confirm the classification performance with the 300 ft data, four groups of four consecutive spectra were chosen for visual analysis. The

corresponding ground locations for the groups of spectra are marked by the circles numbers 1-4 in Figures 6.19 A and B. Locations 1 and 2 specify areas where both SVM models classified the spectra as inactive for the presence of ^{137}Cs, while locations 3 and 4 were judged active by both models.

For locations 1-4, respectively, Figures 6.21 to 6.24 plot the four individual spectra in each group. Figure 6.25 overlays the mean spectra from each of the four groups. From visual analysis of the spectra in Figures 6.21 and 6.22, there is no evidence of the ^{137}Cs peak at 662 keV at locations 1 and 2. From inspection of Figures 6.23 and 6.24, very weak ^{137}Cs signatures are observed at locations 3 and 4. These observations are confirmed in Figure 6.25, where the signal averaging has reduced the noise level by a factor of two. Locations 3 and 4 do appear to contain ^{137}Cs.

While we cannot quote a reliable classification percentage, visual inspection of the results from the *Smallboy* prediction set suggest both classification models are performing well, with Model 5 exhibiting somewhat greater sensitivity in the form of higher detection probabilities.

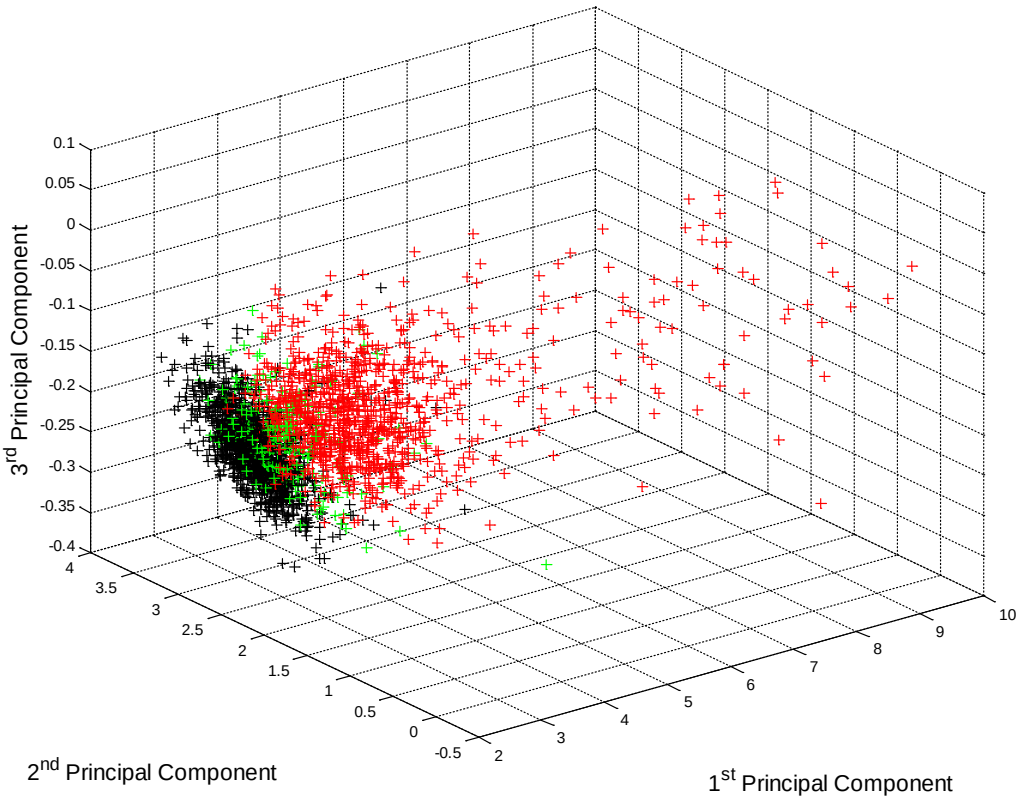


Figure 6.17 Score plots based on the first three principal components computed from the input patterns for SVM Model 5 (Table 6.3) corresponding to the spectra collected at 500 ft altitude in the *Sedan* data set. Red, black, and green points correspond to active, inactive, and uncertain classifications, respectively, for the presence of ^{137}Cs on the basis of the visual review of spectra. The scattered red points on the right of the plot were collected during overflights of the *Sedan* crater area.

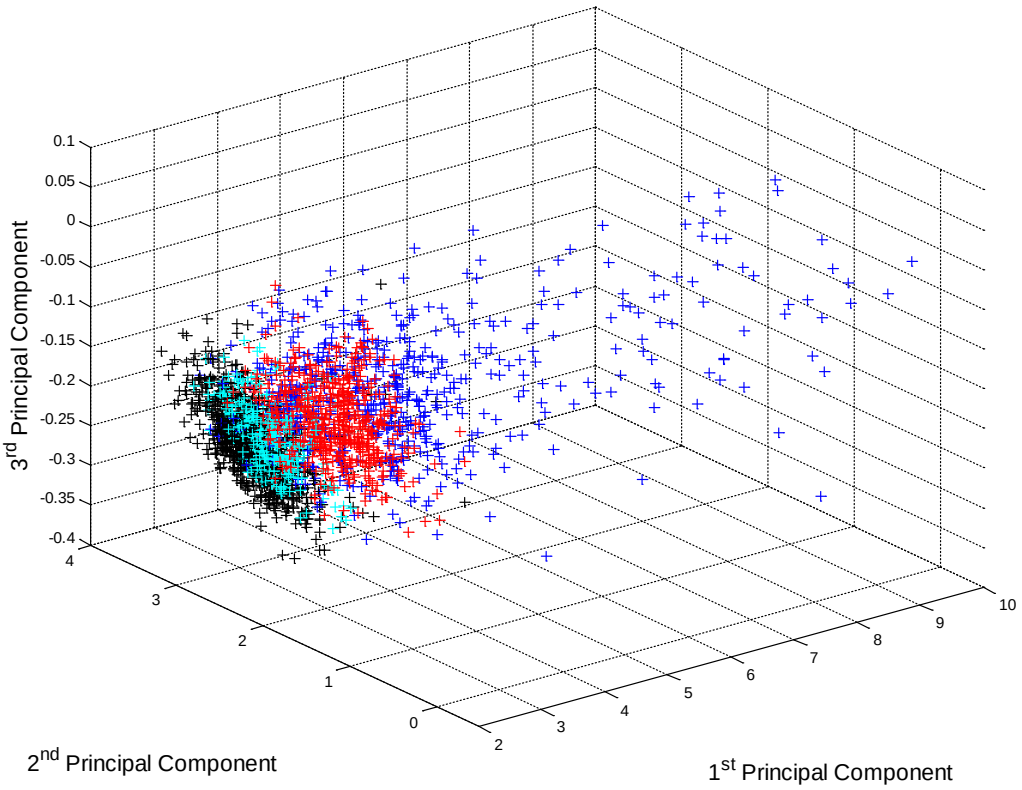


Figure 6.18 Score plots based on the first three principal components computed from the input patterns for SVM Model 5 (Table 6.3) corresponding to the spectra collected at 500 ft altitude in the *Sedan* data set. Red, black, cyan, and blue points correspond to correctly classified active, correctly classified inactive, misclassified inactive (i.e., false positives), and misclassified active (i.e., missed detections) patterns, respectively. The missed detections are predominantly located in *Sedan* crater area and represent very strong ^{137}Cs signals.

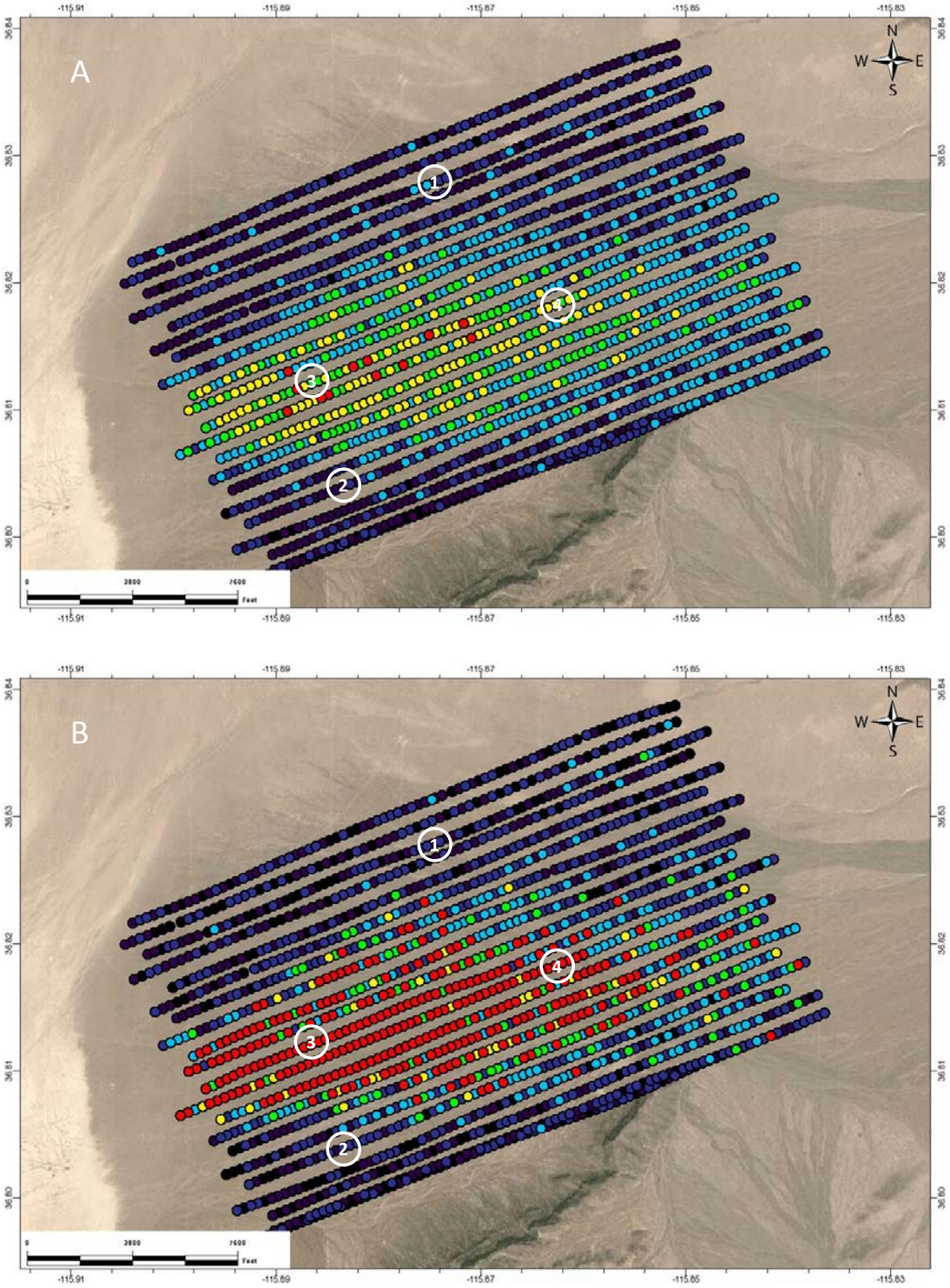


Figure 6.19 Classification images for the *Smallboy* data set (300 ft altitude, prediction set *D*) generated from the SVM scores for Models 1 (A) and 5 (B) in Table 6.3. The presentation format is the same as that used in Figure 6.14. The circles numbered 1-4 identify the locations of scans 1318-1321, 2444-2447, 1947-1950, and 2040-2043, respectively.

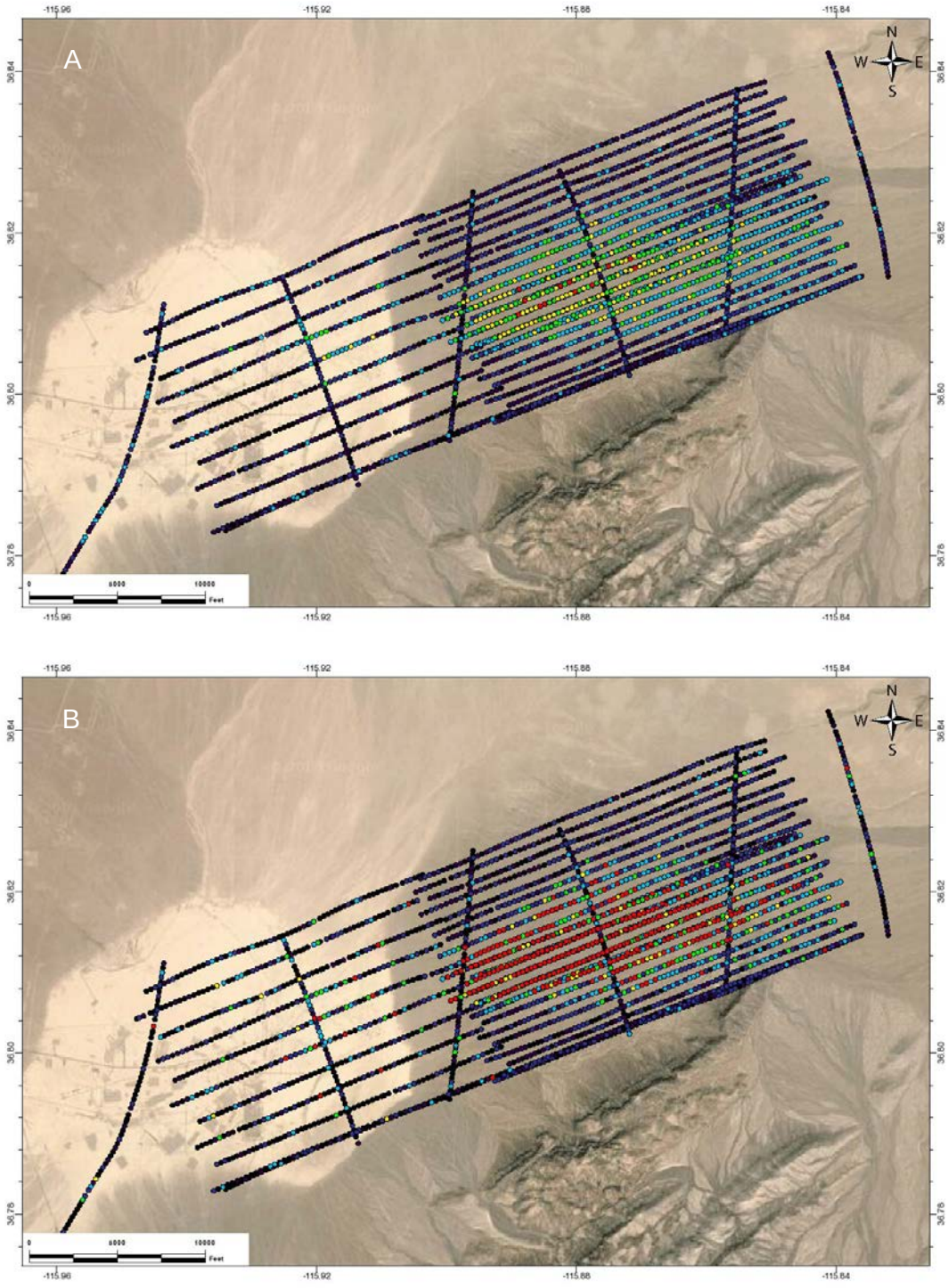


Figure 6.20 Classification images for the *Smallboy* data set at all altitudes (300, 500, 1000 ft, prediction sets *D*, *E*, *F*) generated from the SVM scores for Models 1 (A) and 5 (B) in Table 6.3. The presentation format is the same as that used in Figure 6.14.

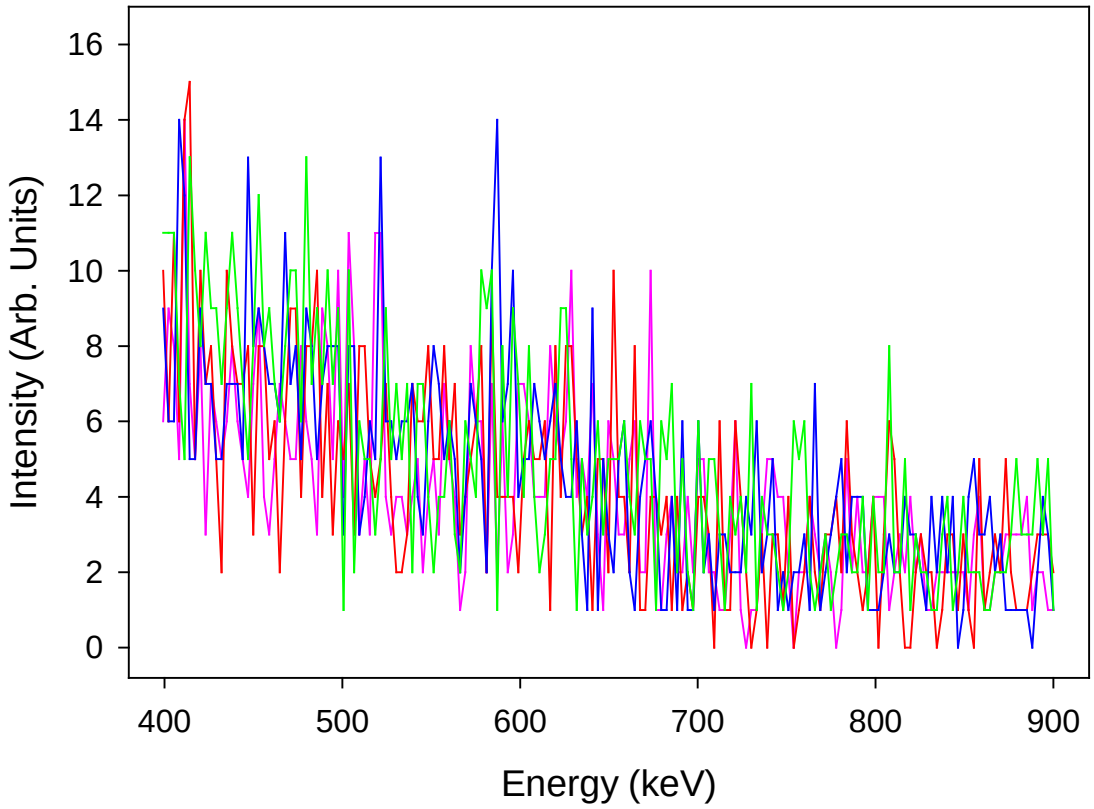


Figure 6.21 Single-scan spectra from the *Smallboy* data set collected at 300 ft altitude. The pink, red, blue, and green lines correspond to spectral scans 1318, 1319, 1320, and 1321, respectively. These spectra correspond to the circle labeled "1" in Figure 6.18. The SVM classifiers based on Models 1 and 5 (Table 6.3) classified these spectra into the inactive class. No evidence of the characteristic peak of ^{137}Cs at 662 keV can be seen in these spectra.

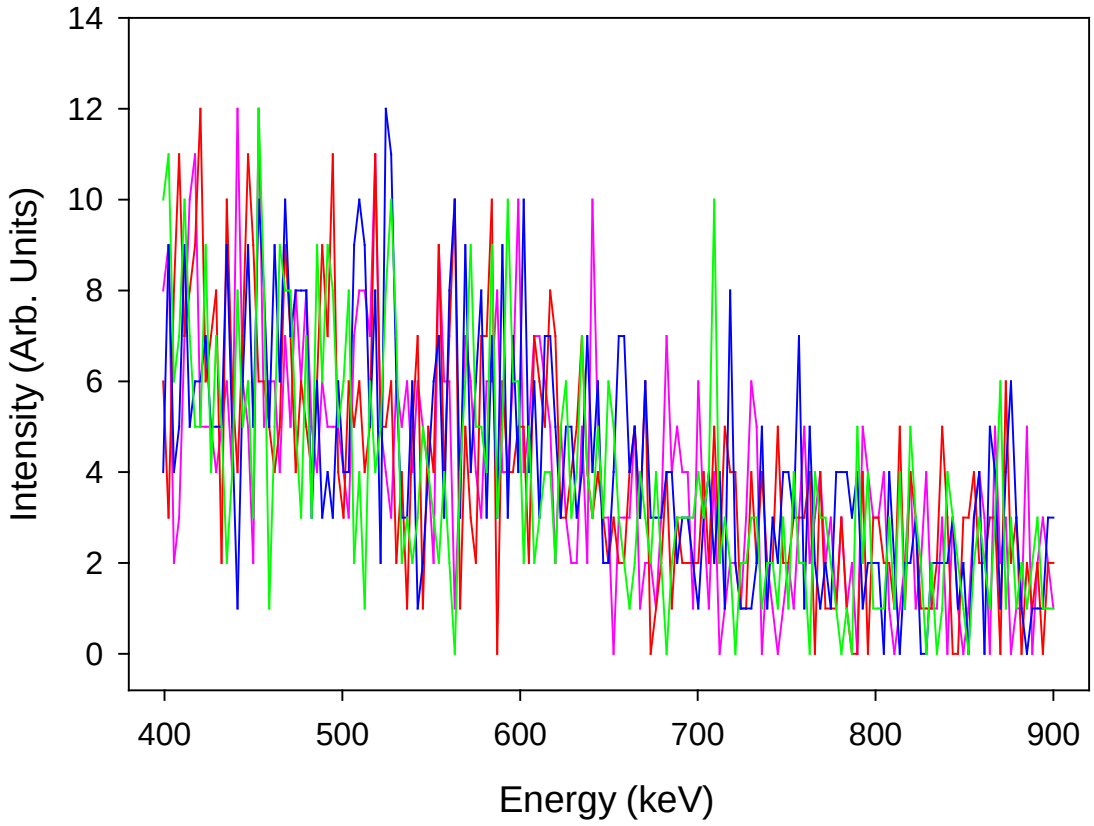


Figure 6.22 Single-scan spectra from the *Smallboy* data set collected at 300 ft altitude. The pink, red, blue, and green lines correspond to spectral scans 2444, 2445, 2446, and 2447, respectively. These spectra correspond to the circle labeled "2" in Figure 6.18. The SVM classifiers based on Models 1 and 5 (Table 6.3) classified these spectra into the inactive class. No evidence of the characteristic peak of ^{137}Cs at 662 keV can be seen in these spectra.

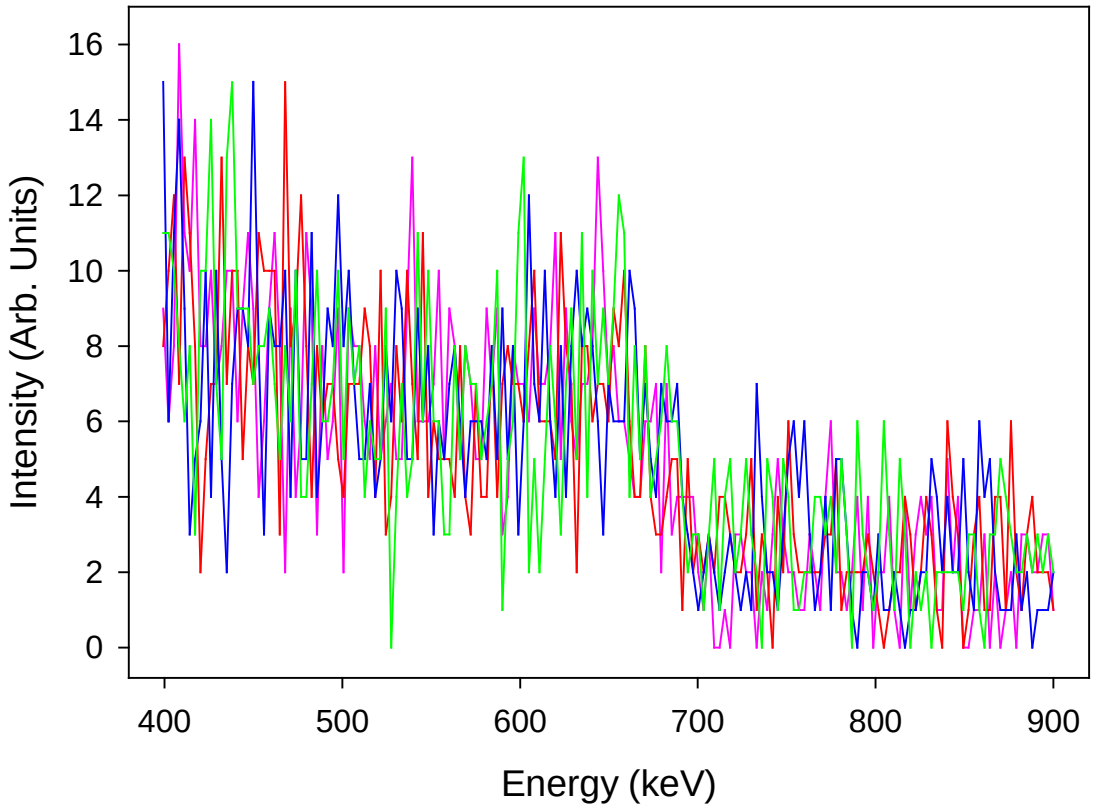


Figure 6.23 Single-scan spectra from the *Smallboy* data set collected at 300 ft altitude. The pink, red, blue, and green lines correspond to spectral scans 1947, 1948, 1949, and 1950, respectively. These spectra correspond to the circle labeled "3" in Figure 6.18. The SVM classifiers based on Models 1 and 5 (Table 6.3) classified these spectra into the active class. The characteristic peak of ^{137}Cs at 662 keV can be seen in the spectra, although the signal is very weak.

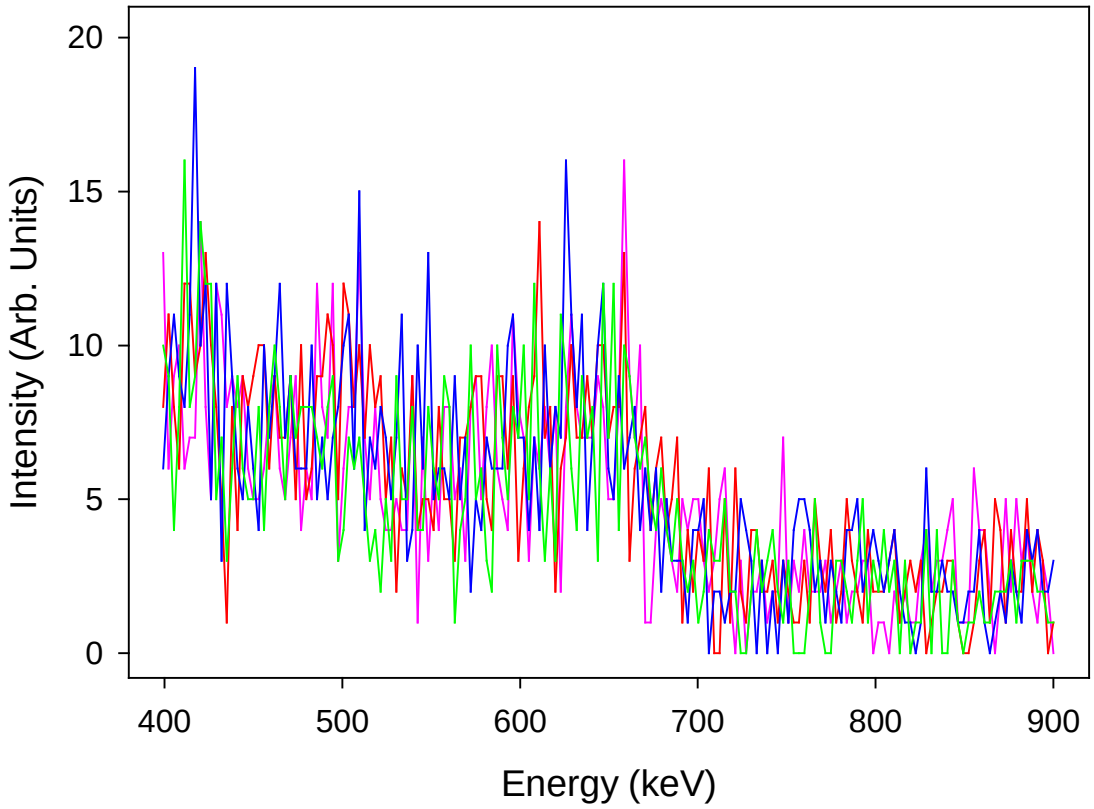


Figure 6.24 Single-scan spectra from the *Smallboy* data set collected at 300 ft altitude. The pink, red, blue, and green lines correspond to spectral scans 2040, 2041, 2042, and 2043, respectively. These spectra correspond to the circle labeled "4" in Figure 6.18. The SVM classifiers based on Models 1 and 5 (Table 6.3) classified these spectra into the active class. The characteristic peak of ^{137}Cs at 662 keV can be seen in the spectra, although the signal is very weak.

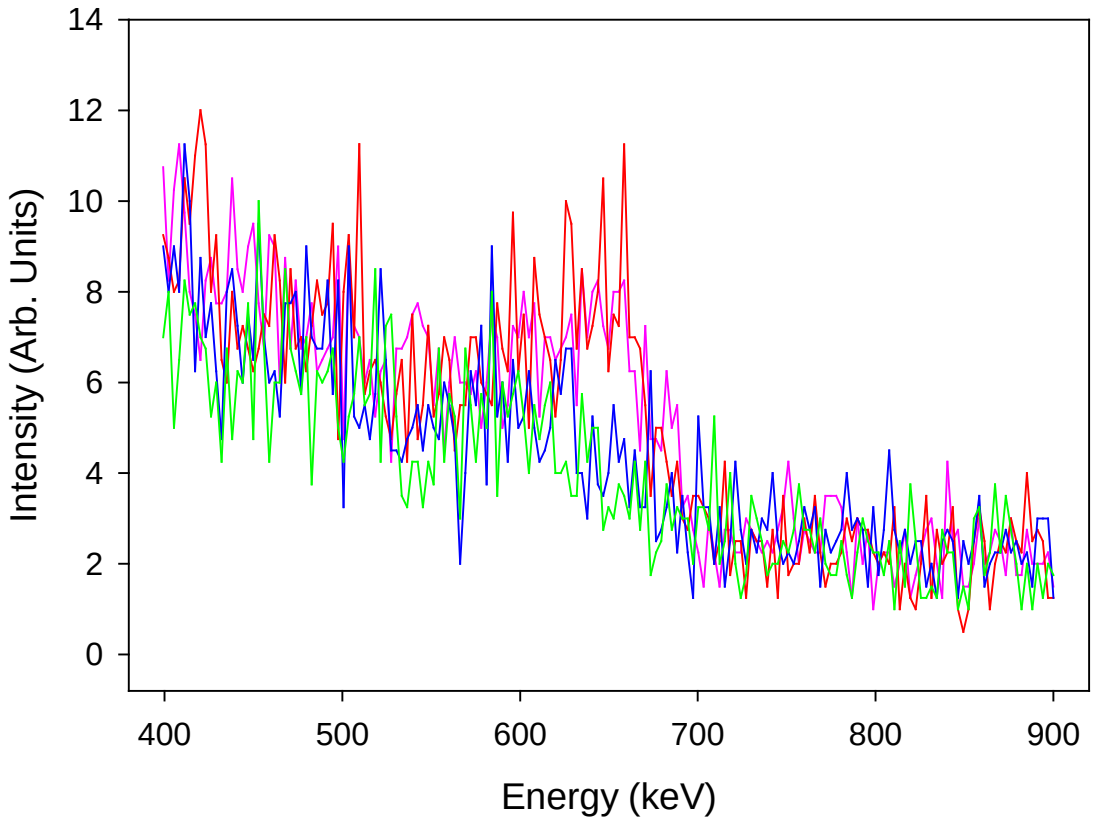


Figure 6.25 Signal-averaged spectra from the *Smallboy* data set collected at 300 ft altitude based on averaging four consecutive scans. The pink, red, blue, and green lines correspond to spectral scans 1947-1950, 2040-2043, 1318-1321, and 2444-2447, respectively. The characteristic peak of ^{137}Cs at 662 keV can be seen in the red and pink traces.

6.4 Conclusions

In this work, an automated identification of the radisotope, ^{137}Cs, from analysis of airborne gamma-ray spectra has been developed. This preliminary work demonstrated that, with an appropriate spectral simulation protocol and signal processing strategy, a classifier, developed from a supervised pattern recognition algorithm is able to recognize the signature of ^{137}Cs in airborne gamma-ray data.

Application of bandpass digital filtering and spectral segment selection were employed to extract weak signatures of ^{137}Cs out of the airborne gamma-ray spectra. This work was complicated by the effects of Compton scattering and changes of altitude on the spectral peak shapes and intensities. A full factorial experimental design was implemented to find the optimal digital filter and spectral segment, which extracted the ^{137}Cs feature only, and suppressed the low-frequency Compton scattering and high-frequency noise. The chosen spectral segment, 540-840 keV, reflected the characteristic peak of ^{137}Cs at 662 keV, and the selected narrow passband filter helped to keep the ^{137}Cs active spectra distinguishable from the inactive ones, while suppressing the background signals and noise.

Spectral simulation was applied to generate synthetic active patterns for use in the training and monitoring sets. While not a perfect match to the actual field-collected spectra, the synthetic spectra provided a practical alternative to requiring the collection of field-active spectra for use in computing classification models. Improvement of the accuracy of the spectral simulation process is an important area of future work.

When the chosen classifiers (Models 1 and 5 in Table 6.3) were applied to the prediction sets collected from three field surveys, the resulting classification results were good for discriminating weak ^{137}Cs signatures. Use of the monitoring data to derive probability-based classification thresholds allowed confidence levels to be assigned to the predictions.

The key remaining issue is the failure of the SVM classifiers to detect the strong ^{137}Cs signatures in the *Sedan* prediction set. Options to address this problem include further study of the population of training patterns used to define the active class, examination of the effect of the digital filter bandpass on this problem, and the architecture of the SVM classifiers.

Table 6.1 Description of data sets

Data subsets	No. ^{137}Cs-active patterns	No. ^{137}Cs-inactive patterns	Total No. patterns in data set[a]
Training set	3000	12,000	15,000
Monitoring set	5000	7939	12,979
Prediction set *A* (*Desert Rock*)	40	933	1038
Prediction set *B* (*Sedan*, 500 ft)	990	1160	2375
Prediction set *C* (*Sedan*, 1000 ft)	183	1767	2219
Prediction set *D* (*Smallboy*, 300 ft)	N/A	N/A	1723
Prediction set *E* (*Smallboy*, 500 ft)	N/A	N/A	486
Prediction set *F* (*Smallboy*, 1000 ft)	N/A	N/A	1044

[a]For cases in which the total number of patterns does not equal the sum of the active and inactive patterns, the missing patterns correspond to those whose classification was deemed uncertain in the visual review of the data.

Table 6.2 Parameters used for signal processing and pattern recognition

Parameters		Levels	Values
Chebyshev Type II filter	Passband lower limit (w_{stop1})	3	0.01 0.02 0.03
	Passband upper limit (w_{stop2})	3	0.05 0.06 0.07
	Stopband attenuation (dB)	1	60
Spectral segment	Starting point (keV)	12	56.6 116.2 175.8 235.4 295.0 354.6 414.2 473.8 533.4 593.0 652.6 712.2
	Segment length	1	100
SVM configuration	Kernel parameter (γ)	12	0.001 0.01 0.05 0.1 0.5 1 5 10 50 100 500 1000
	Regularization parameter (C)	12	1 30 70 100 1000 4000 1×10^4 2×10^4 4×10^4 1×10^5 1×10^6 1×10^7

Table 6.3 Top classifiers derived from training and monitoring results

Model	Passband lower limit	Passband upper limit	Segment starting point (keV)	SVM γ parameter	SVM *C* parameter	No. of support vectors	Trn. error counts	Mon. miss rate (%)	Mon. false rate (%)	Sum of mon. false/miss (%)
1	0.03	0.07	473.8	0.01	4000	669	200	6.6	0.24	6.8
2	0.01	0.06	533.4	0.001	1000	557	195	6.0	0.30	6.3
3	0.01	0.06	533.4	0.001	4000	548	189	6.0	0.33	6.3
4	0.01	0.05	533.4	0.001	20000	558	197	6.1	0.33	6.4
5	0.03	0.05	533.4	10	1000	993	103	5.5	0.92	6.4
6	0.03	0.05	533.4	5	1000	810	128	5.5	0.86	6.4
7	0.03	0.05	533.4	1	4000	608	153	5.5	0.63	6.1
8	0.03	0.05	593.0	0.05	400000	557	172	5.5	0.58	6.1
9	0.02	0.05	473.8	0.01	40000	561	167	5.6	0.72	6.3
10	0.03	0.05	593.0	0.1	400000	558	163	5.6	0.66	6.3

Table 6.4 Summary of classification results for prediction sets *A, B, C*

	Model 1[a]				Model 5			
	Conventional[b]		Control limits[c]		Conventional		Control limits	
Prediction set	% Missed	% False	% Missed	% False	% Missed	% False	% Missed	% False
Prediction set *A* (*Desert Rock*)	0.0	0.1	0.00	0.1	2.5	0.6	2.5	0.3
Prediction set *B* (*Sedan*, 500 ft)	55.1	7.0	56.5	4.4	47.8	18.8	56.8	13.3
Prediction set *C* (*Sedan*, 1000 ft)	95.1	4.5	95.1	3.7	94.0	10.9	95.6	7.2

[a]Model specifications correspond to Table 6.3.

[b]Conventional classification results were based on a threshold of 0.0 as the class boundary.

[c]Classifications based on control limits used the 70 % level (threshold values of 0.14 and 0.30 for Models 1 and 5, respectively.

Chapter 7

PERFORMANCE DIAGNOSTICS FOR LONG-TERM MONITORING OF GLUCOSE BY NEAR INFRARED SPECTROSCOPY BASED ON MULTIVARIATE CALIBRATION AND PATTERN RECOGNITION METHODS

7.1 Introduction

Near-infrared (near-IR) spectroscopy has proven to be a promising data collection technique for use in quantitative determinations of a variety of analytes in complex matrixes.[11,122-126] Analytes with physiological significance, such as glucose, have characteristic vibrational spectral signatures in the near-IR region, whereas common background constituents such as water exhibit reduced absorbance relative to the mid-IR. This allows the quantitative determination of species, directly with little or no sample preparation, in aqueous solutions with optical path lengths in the millimeter range. These characteristics give near-IR analyses promise for use in direct determinations of analytes in biological samples such as blood or plasma.[127] It also boosts the hope for *in vivo* measurements of analytes, directly from tissue in a noninvasive fashion or for the purpose of long-term monitoring.[128,129]

To develop quantitative analyses in these application scenarios, multivariate modeling techniques such as partial least-squares (PLS) regression are employed to build calibration models that relate spectral intensities to analyte concentrations. These calculations require a set of spectra of calibration samples with known reference concentrations.[130,131]

Multivariate modeling methods such as PLS provide an efficient means to extract concentration information of the target analyte from spectra taken from complex matrixes. However, one of the main problems with multivariate calibration is the tendency for the performance of the calibration model to degrade over time. This degradation can be caused by numerous sources of variation, such as changes in the instrumental response, changes in the environmental conditions associated with the data measurement, or sample-related spectral

changes with time. In effect, the computed model is keyed to the calibration spectra used in its generation. Any changes in the instrument, experimental conditions, or sample composition that occur with time dictate that the spectra submitted to the model will be different than those originally used in its formulation. As this mismatch grows between the input spectra and the original spectra upon which the model is based, the ability of the model to produce accurate concentration predictions degrades.

Data preprocessing methods can be used to suppress some sources of spectral variation before the data are submitted to the calibration model. However, the fact that this spectral variation cannot be completely removed and that the incorporation of this variation into the calibration model will tend to worsen its prediction performance demands periodic calibration updating.

Numerous calibration updating and standardization methods have been reported that attempt to minimize the performance deterioration of calibration models over time. They include direct standardization,[132] piecewise direct standardization (PDS),[133] and prediction augmented classical least-squares (PA-CLS).[40] Signal processing techniques such as orthogonal signal correction (OSC)[134] and the wavelet transform[135] can also be employed to the same effect.

However, all of these techniques require time and expense for collecting data for updated calibration samples. In such scenarios, to avoid unnecessary cost, it is important to develop a calibration updating strategy that incorporates an evaluation of the performance of the calibration model itself, thereby allowing a forecast to be made regarding the timing as to when a calibration update is needed.

For each test sample, the multivariate calibration model will calculate a predicted concentration by extracting targeted information from the input spectrum and weighting that

information appropriately in the estimation of the analyte concentration. If the true reference concentration of a sample is known, the corresponding residual concentration (i.e., the difference between the known and estimated concentrations) can be obtained. In formulating the model, it is a common practice to use the residual concentrations as a diagnostic of model performance. However, when the model is applied to a sample with unknown concentration, a residual concentration is not available.

A by-product of a prediction is the residual spectrum, the remaining spectral information that is not extracted and submitted to the calibration model. In principle, for the model to perform well, the residual spectrum associated with the prediction sample must lie within the span of the residual spectra associated with the formulation of the model (i.e., the residual spectra associated with the calibration data). The work presented in this chapter explores ways to use the residual spectra to develop a calibration diagnostic for use in determining when a model is failing.

In this work, a residual modeling strategy is proposed, aiming to estimate residual concentrations from the residual spectra left from the application of the calibration model. In order to develop efficient residual models, two prominent modeling techniques, PLS[122] and support vector regression (SVR),[136] and a hybrid modeling technique, PLS-aided SVR,[137,138] are used. Also, for the first time, an amplified PLS-aided SVR technique is proposed and investigated.

In this work, residual models are built with single-beam spectra. A set of classification tests is proposed and implemented with the residual models in order to evaluate the capacity of these models to provide performance diagnostics for the calibration model. On the basis of the

performance diagnostics generated from the selected residual model, a calibration updating strategy is developed.

In the research presented here, a data set collected from aqueous samples containing physiological levels of glucose (analyte) and five other interfering components (lactate, urea, ascorbate, alanine, triacetin) is used. This sample matrix represents the types of components and degree of spectral overlap encountered in biological samples such as blood plasma or serum. Initially, based upon a calibration data set, an optimal PLS model for glucose concentration is built and optimized. The robustness of this model is investigated by evaluating its external prediction performance over a period of 416 days. Residual modeling is applied, performance diagnostics are developed, and calibration updating is applied to these data as described above.

7.2 Experimental

7.2.1 Sample preparation

The preparation of samples used in this work has been described in a previous study.[131] A brief summary of the experimental procedures is presented here. In each sample, six components, glucose, sodium lactate, urea, sodium ascorbate, alanine, and triacetin, were present in a background matrix of phosphate buffer. A uniform experimental design[139] was used to generate an initial set of 2129 mixture combinations. Employing a combination of the subset selection method first developed by Carpenter and Small[78] and the Kennard-Stone algorithm,[140] 64 calibration samples and 25 prediction samples were picked from the pool of 2129 combinations. The concentration ranges for glucose, lactate, urea, ascorbate, alanine, and triacetin were 1.6-30, 0.8-32, 1-30, 1.2-19.8, 1.5-28.5, and 1.4-30 mM, respectively. The concentration values for each component per sample were assigned to minimize correlations among the constituents.

Seven reagents were used in this work, including glucose (ACS reagent, Fisher Scientific, Fair Lawn, NJ), sodium L-lactate (98%, Sigma, St. Louis, MO), urea (ACS reagent, Fisher),

sodium ascorbate (ACS reagent, Aldrich, Milwaukee, WI), alanine (Aldrich), and triacetin (Sigma). The 0.1 M, pH 7.4 phosphate buffer solution that served as the solvent contained 5 g/L of sodium benzoate as a preservative. All samples were prepared by volumetric dilution of pure-component stock solutions prepared in the phosphate buffer. The concentrations of glucose for all the samples were confirmed with an YSI model 2300 STAT Plus glucose analyzer (YSI, Inc., Yellow Springs, OH).

7.2.2 Instrumentation

All spectra were collected with a Nicolet Nexus 6700 Fourier transform (FT) spectrometer (Thermo Fisher Scientific, Inc., Madison, WI) equipped with a calcium fluoride beam splitter and liquid nitrogen-cooled indium antimonide detector. A tungsten-halogen lamp was used as the light source, while a K-band optical interference filter (Barr Associates, Westford, MA) was applied to restrict the light throughput reaching the detector to the range of 5000 – 4000 cm^{-1}. To ensure detector linearity, the incident light was further attenuated by use of a 63 % transmittance thin-film type neutral density filter (Rolyn Optics, Covina, CA).

7.2.3 Data Collection Protocol

The entire data collection spanned 416 days. All 64 samples in the calibration set were measured in the first 4 days. This calibration data set is termed, C_{full}. For the rest of the period, starting 5 days after the completion of the calibration data collection, the 25 samples in the prediction set were measured repeatedly on roughly a biweekly basis. This produced a series of 29 prediction sets, termed *P1* to *P29*. During the course of each measurement session, triplicate measurements of each sample were made without sample replacement or reloading. Meanwhile, to assess instrument performance and to allow the calculation of spectra in absorbance units (AU), background spectra of the phosphate buffer were collected periodically. A total of 22

buffer spectra were acquired daily, with 16, 3, and 3 buffer spectra collected at the beginning, middle, and end of each data collection day, respectively.

Table 7.1 summarizes the data collection protocol used in this work. Note that, in practice, the experimental run orders for all samples were randomized to minimize the correlation of component concentrations and time. At the end of each measurement session, all buffer and samples were stored in the freezer. Prior to the next session, they were taken from the freezer and thawed. To confirm the stability of the samples through the freeze/thaw cycles, the glucose concentrations of all samples were tested periodically with the YSI analyzer.

The raw data consisted of 256 co-added double-sided interferograms containing 8192 points collected at every zero crossing of the He-Ne reference laser (spectral resolution of 8 cm^{-1}, bandwidth of 15,803 cm^{-1}).

7.2.4 Data analysis implementation

All interferograms were Fourier processed to single-beam spectra with software implemented on a Silicon Graphics Origin 200 Computer (Silicon Graphics, Inc., Mountain View, CA) operating under Irix (version 6.5, Silicon Graphics, Inc.). Triangular apodization and Mertz phase correction were applied during the Fourier processing step. All other computational tasks were implemented under MATLAB (Version 7.4, The MathWorks, Natick, MA). The SVM classifiers and regression models were trained and tested with the public-domain package, SVM^{light} (Version 6.01). The SVM, PLS, and all other data analysis steps were performed on a Dell Precision 490 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Version 5.2, Red Hat, Inc., Raleigh, NC).

Table 7.1 Summary of Data Collection

Days	Data Type	Data Set ID	Number of Samples	Number of Sample Spectraa	Number of Buffer Spectrab
1-4	Calibration	C2	64	192	88
9-10	Prediction	P1	25	75	44
23-24	Prediction	P2	25	75	44
37-39	Prediction	P3	25	75	44
51-52	Prediction	P4	25	75	44
65-66	Prediction	P5	25	75	44
79-80	Prediction	P6	25	75	44
100-101	Prediction	P7	25	75	44
114-115	Prediction	P8	25	75	44
128-129	Prediction	P9	25	75	44
142-143	Prediction	P10	25	75	44
156-157	Prediction	P11	25	75	44
177-178	Prediction	P12	25	75	44
191-192	Prediction	P13	25	75	44
205-206	Prediction	P14	25	75	44
219-220	Prediction	P15	25	75	44
233-234	Prediction	P16	25	75	44
248-249	Prediction	P17	25	75	44
261-262	Prediction	P18	25	75	44
275-276	Prediction	P19	25	75	44
289-290	Prediction	P20	25	75	44
303-304	Prediction	P21	25	75	44
317-318	Prediction	P22	25	75	44
331-332	Prediction	P23	25	75	44
345-346	Prediction	P24	25	75	44
359-360	Prediction	P25	25	75	44
373-374	Prediction	P26	25	75	44
387-388	Prediction	P27	25	75	44
401-402	Prediction	P28	25	75	44
415-416	Prediction	P29	25	75	44

[a]Three consecutive replicate spectra were collected for each sample.
[b]Spectra of phosphate buffer were collected each day for diagnostic purposes.

7.3 Results and Discussion

7.3.1 Overview of Methodology

The pure-component absorbance spectra[141] of the seven chemical components present in the mixture samples are shown in Figure 7.1, A. As the target analyte, glucose features an O-H combination band at 4700 cm^{-1}, and two C-H combination bands at 4300 and 4400 cm^{-1}, respectively. The other components contain varying spectral signatures within the range of 5000 – 4000 cm^{-1}. There is clearly substantial spectral overlap among the components.

Figure 7.1, B and C plot single-beam and absorbance spectra, respectively, of one of the mixture samples. No features of the chemical components can be seen visually in the single-beam spectrum because of the large dynamic range of the spectral background. The absorbance spectrum has several features arising from the contributions of the mixture components. In order to quantify glucose levels in the mixture spectra, multivariate calibration methods, such as PLS, are used to extract the analyte information from the background contributions of the sample matrix.

As one of the most widely used multivariate calibration techniques, PLS regression features a latent-variable based linear regression method which incorporates the analyte concentration into the computation of a series of latent variables extracted from the spectral data matrix. The PLS scores are projections of spectra onto each latent variable and serve as the independent variables in the construction of a multiple linear regression model for predicting analyte concentrations. Use of the PLS technique to extract information from the data matrix

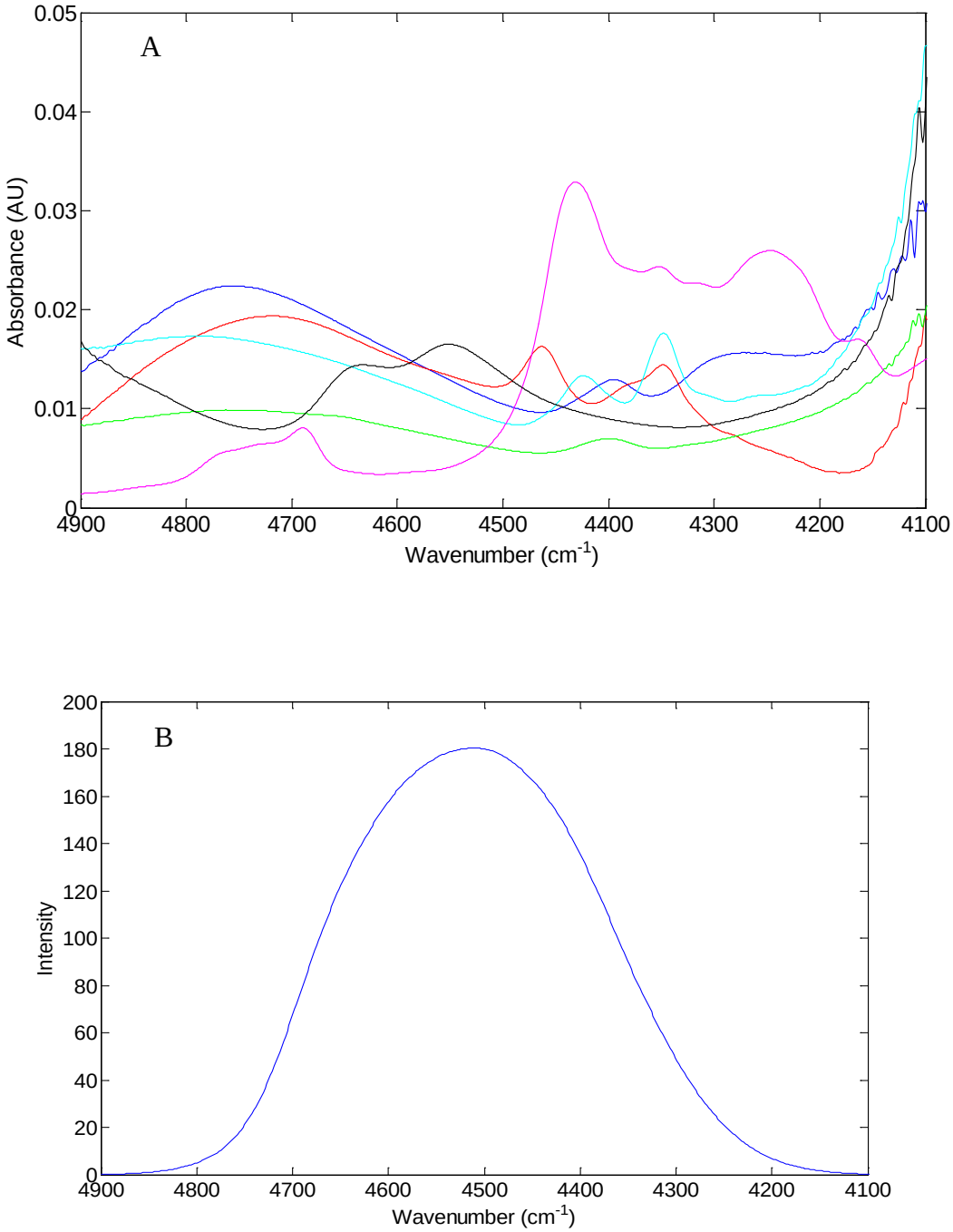
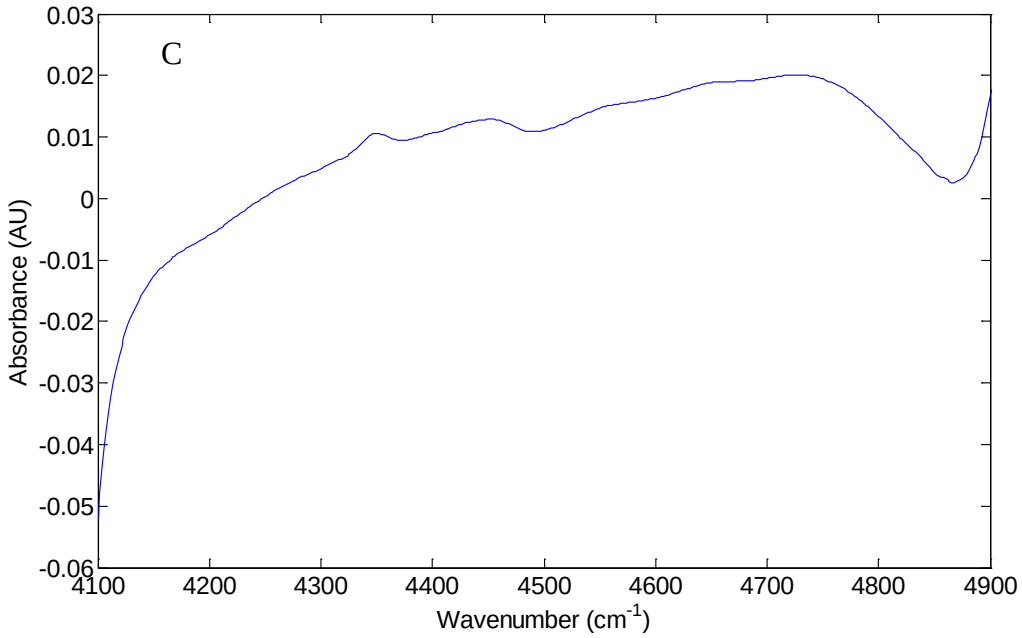


Figure 7.1 A. Pure-component absorbance spectra of 100 mM glucose (blue), 100 mM alanine (red), 50 mM ascorbate (green), 150 mM lactate (cyan), 100 mM triacetin (magenta), and 90 mM urea (black). A spectrum of phosphate buffer was used as the background for calculation of all absorbance spectra. B. Near-infrared spectra of a six-component mixture solution in single-beam mode. The glucose concentration was 29.6 mM. C. Near-infrared spectra of a six-component mixture solution in absorbance units relative to a background of phosphate buffer. The glucose concentration was 29.6 mM.



(Figure 7.1 continued)

before the calculation of the predictive model allows the model to be based on fewer terms than if raw spectral intensities were employed as the independent variables.

In this work, calibration models for glucose based on both single-beam and absorbance near-IR spectra were developed by implementing the PLS algorithm. For the absorbance calculations, the background was taken as the mean of the 22 buffer single-beam spectra acquired during the same day as the sample.

Use of single-beam spectra directly in the construction of calibration models based on aqueous near-IR spectra of dilute solutions is motivated by the extreme temperature sensitivity of the background absorbance of water. Because a matrix-matched background cannot be collected at the exact temperature of the samples, spectral artifacts (e.g., baseline curvature) are introduced into absorbance spectra by the shifts in the water absorption bands with changing temperature. Thus, the absorbance calculation can make the spectra more complex rather than simpler. In previous work, our laboratory has demonstrated successful glucose calibration models based on near-IR single-beam spectra. [142]

The calibration spectra were mean-centered before submission to the PLS algorithm. A pool of models with varying wavenumber ranges and numbers of latent variables (i.e., PLS scores) were generated, and a leave-one-out cross-validation (LOOCV) procedure for each sample (three replicate spectra) was utilized for model selection. For each model, the cross-validation standard error of prediction (CV-SEP) was computed as an assessment of the model's generalized precision. The CV-SEP is a pooled error in predicted concentration accumulated across each spectrum withheld from the calibration calculations. An optimal model was chosen as that corresponding to the minimum CV-SEP.

Subsequently, the optimal calibration models were applied to the external test sets. To obtain a predicted concentration, the PLS latent variables previously computed with the calibration data are used to obtain the PLS scores for the prediction samples. As a part of this calculation, a residual spectrum can be computed for each prediction sample corresponding to the remaining spectral information that did not project onto the PLS latent variables. The goal of this research was to develop methodology to interrogate the residual spectra as a way to diagnose model performance and to correct for deficiencies in the predicted concentrations.

The prediction performance was evaluated, and the resulting residual spectra were later used to develop models that sought to relate information in the residual spectra to the corresponding concentration residuals (i.e., the difference between the predicted and actual concentrations). Several regression techniques were used for building the residual models, including PLS, support vector regression (SVR), PLS-aided SVR, and a newly proposed amplified PLS-aided SVR. The SVR procedure features a supervised nonlinear regression which is achieved by mapping low-dimensional data in the original input data space to a high-dimensional feature space through the use of a kernel function. Details regarding the specific procedures used will be provided in a subsequent section.

The robustness of the residual models and their prediction performance was studied and will be described below. A set of classification tests was designed and implemented to evaluate and select the optimal residual models. Selected residual models retrieved residual concentrations directly or indirectly from the residual spectra, and provided a set of threshold ranges for classification purposes. Based upon those predicted residual concentrations and the determined thresholds, performance diagnostics of the calibration model were established, followed by a calibration updating procedure.

7.3.2 Characterization of Spectral Data

The quality of the measured data was assessed by evaluating the noise distribution within replicate spectra for each sample across all data sets. For each sample, ratios of two out of the triplicate spectra in all possible combinations were taken to obtain 100 % lines, which were subsequently converted to absorbance units (AU). Root-mean-square (RMS) noise levels were calculated for the 100 % lines in μAU over the spectral range of 4500 – 4300 cm^{-1} about a fitted second order polynomial function. For each data set, the mean and standard deviation of the RMS noise levels for all samples within replicates were calculated as a characteristic of the measured data. The detailed noise distribution across all data sets was captured in Figure 7.2. The mean ± standard deviation of the RMS noise estimates across all data sets was 5.6 ± 4.1 μAU.

7.3.3 Calibration with PLS

Details regarding the calibration models, including data partitioning, model configurations, and results are summarized in Table 7.2. To build calibration models for glucose concentrations, 96 full-range (518-point) single-beam spectra collected from the first 32 samples of the original calibration set, C_{full}, formed a new calibration set, $C_{½}$, and were used to build glucose calibration models. The remaining 96 spectra from the original calibration set served as an external test set, termed test set *B*, while all the prediction data sets from *P1* to *P29* served as a second external test set, termed test set *A*. Both test sets were later used to evaluate the performance of the established calibration model.

In building the calibration model, a wavenumber selection protocol was utilized. Within the spectral window of 5000 - 4000 cm^{-1}, ranges with widths varying from 600 to 50 cm^{-1} in steps of 25 cm^{-1} moved across the spectral window in steps of 25 cm^{-1}. This procedure generated

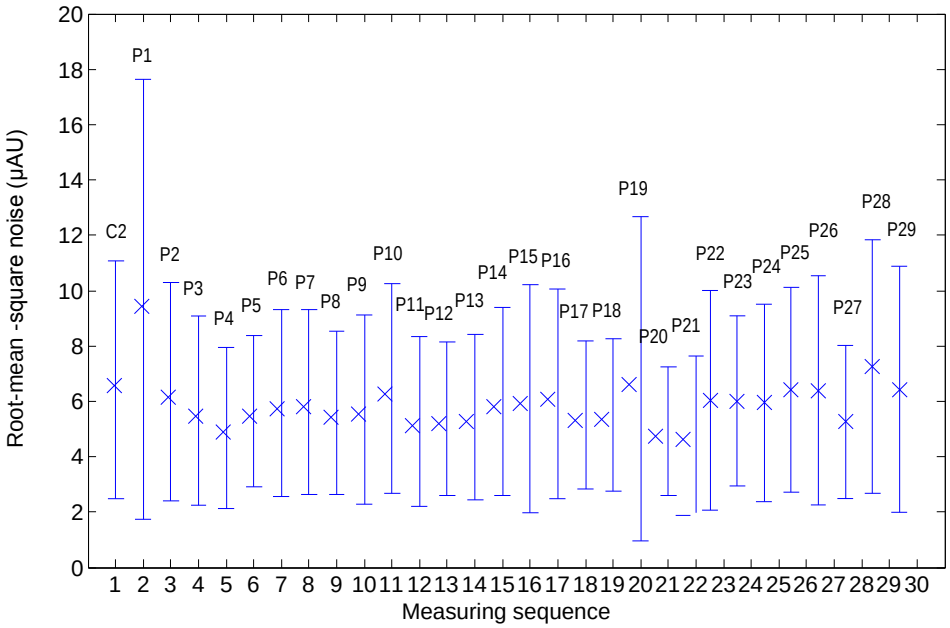


Figure 7.2 Characterization of RMS noise levels over 4500-4300 cm^{-1} in the measured sample spectra. The "X" symbols denote the mean values and error bars are drawn as one standard deviation about the mean. The label, C2, denotes the calibration set, while the labels, P1 to P29, specify the individual prediction sets.

Table 7.2 Summary of Selected Calibration Models

Calibration results	Single-beam	Absorbance
Data source of calibration set	1st half of C_{full}	1st half of C_{full}
Spectra Number in calibration set	96	96
Data source of test set *A*	[P1:P29]	[P1:P29]
Number of spectra in test set *A*	2175	2175
Data source of test set *B*	2nd half of C_{full}	2nd half of C_{full}
Number of spectra in test set *B*	96	96
Spectral range (cm^{-1})	4609.8 – 4111.3	4490.2 – 4290.8
Number of latent variables	14	9
CV-SEP (mM)	0.51	0.45
R^2	0.9991	0.9986
SEP for test set *A* (mM)	1.88	1.49
SEP for test set *B* (mM)	0.45	0.53

459 ranges. For each range, PLS models based upon 1 – 21 latent variables were built. The performance of each candidate model was assessed by computing its CV-SEP as described previously. The optimal spectral range was determined from the sorted CV-SEP values. The corresponding number of PLS factors to use was established via the *F*-test statistic at the 95% confidence level. At the optimal wavenumber range, the number of latent variables that produced a model with a value of CV-SEP not statistically different from the minimum CV-SEP was chosen as optimal. The final selected model was then applied to the prediction sets.

For the calibration model based on single-beam spectra, the optimal spectral range was 4609.8 – 4111.3 cm^{-1}. This model was based on 14 latent variables. The resulting CV-SEP was 0.51 mM, and the coefficient of multiple determination (R^2) was 0.9991 (Table 7.2). Similar results were found for the model based on absorbance spectra. A final chosen model had a spectral range of 4490.2 – 4290.8 cm^{-1} and 9 latent variables, resulting in a CV-SEP of 0.45 mM and an R^2 value of 0.9991. The fewer latent variables associated with the model based on absorbance data reflects the smaller spectral range used, the removal of background components by the absorbance calculation, and the linear relationship between absorbance and concentration afforded by the Beer-Lambert law. The low CV-SEP values and strong correlations between predicted and prepared concentrations, randomly distributed residuals about zero, as illustrated in Figure 7.3, suggest both models were well-behaved.

The selected models were then applied to prediction sets *A* and *B*. None of these data were used during any of the calibration calculations. Table 7.2 lists the values of the standard error of prediction (SEP) obtained with prediction sets *A* and *B*. The SEP is analogous to CV-SEP in that it represents a pooled error in the predicted concentrations across the given set of spectra. Both calibration models worked well with prediction set *B*, producing SEP values in the

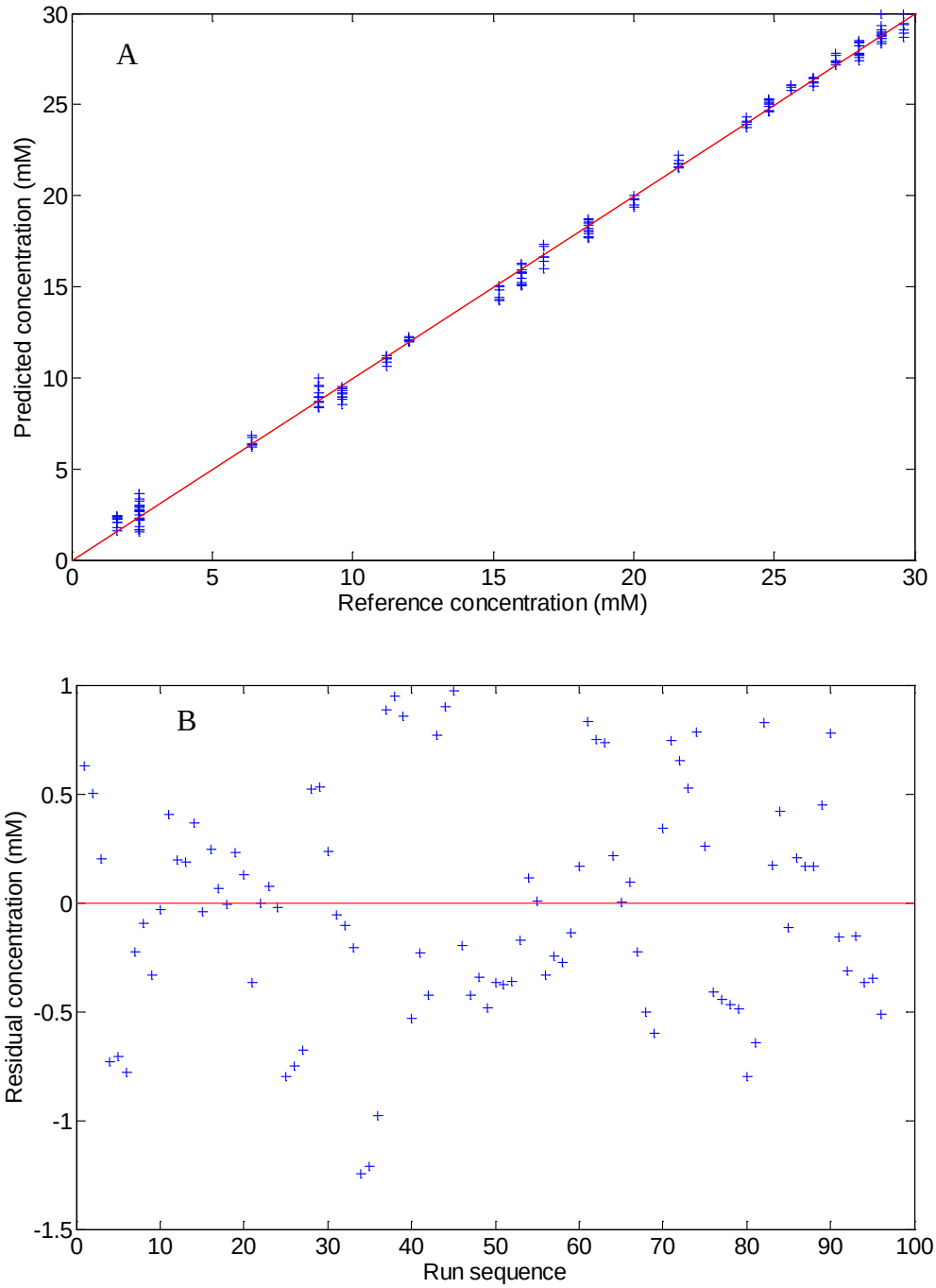
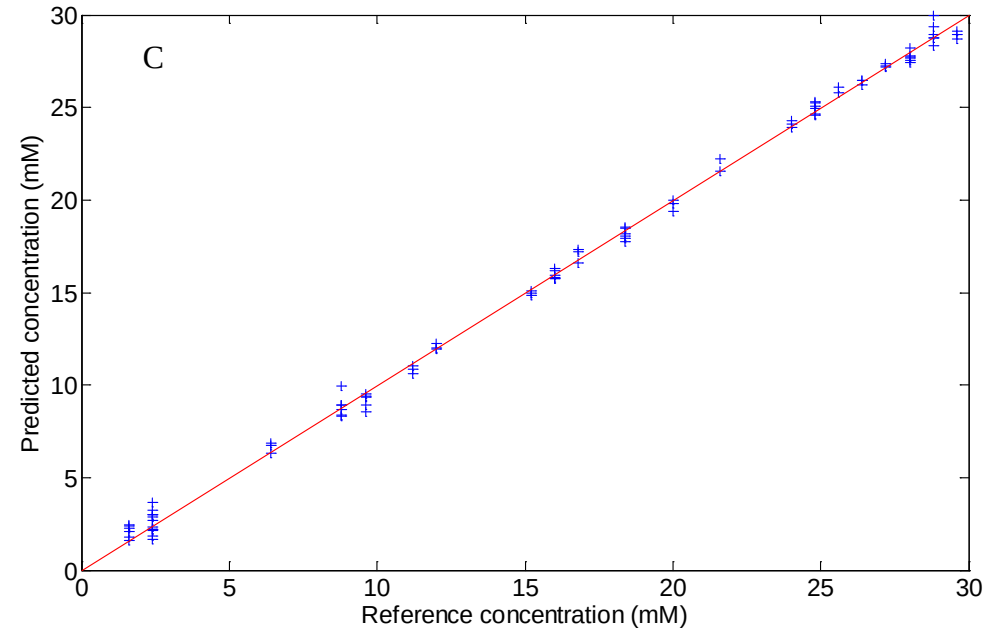
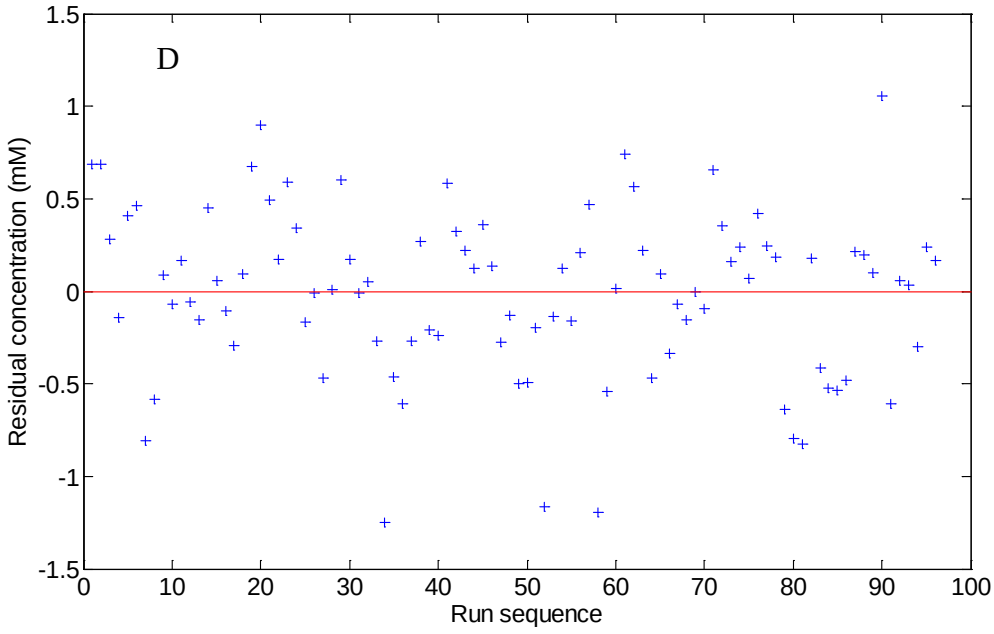


Figure 7.3 Characteristics of the calibration models for the 96 spectra in calibration set $C_{½}$. (A) Correlation plot of predicted vs. reference glucose concentrations for the model based on single-beam spectra. The red line denotes perfect correlation. (B) Plot of concentration residuals vs. sequence number for the model based on single-beam spectra. The red line specifies residuals of 0.0 mM. (C) Correlation plot for model based on absorbance data. (D) Plot of residuals for model based on absorbance data.





(Figure 7.3 continued)

range of 0.5 mM. This excellent prediction performance reflects the minimal separation in time between the calibration and prediction data. With prediction set *A*, however, the prediction performance degraded with time for both models and was worse with the model based on single-beam data (overall SEP of 1.88 vs. 1.49 mM). As noted above, the absorbance calculation has some ability to correct for changes in the instrumental background with time because of the use of the background buffer spectrum collected on the same day.

Figure 7.4 summarizes the external prediction results (SEP values) for individual prediction sets *P1* to *P29*. Plots A and C correspond to the calibration models based on single-beam and absorbance spectra, respectively. Plots B and D display the individual concentration residuals for the spectra in prediction sets *A* (red) and *B* (green) produced by the calibration models based on single-beam and absorbance data, respectively. Results for both models show a clear deterioration with time. For the calibration model based on single-beam spectra, SEP values for the individual prediction sets (*P1* to *P29*) rose from 0.5 mM to more than 12 mM over the 416-day period, although the trend was not a simple ramp (Figure 7.4, A). A similar trend was observed with the model based on absorbance data (Figure 7.4, C), with the SEP values ranging over time from 0.5 to about 3 mM. The residuals are smaller when absorbance data are employed, although the same time-based dependence is observed in each plot.

The degradation in prediction performance with time indicates the spectra increasingly contain characteristics that are not present in the calibration data. Glucose concentrations were confirmed by external reference measurements periodically throughout the data collection and were not found to change significantly. The variation observed most likely arises from instrumental drift with time or non-glucose sample changes caused by the freeze-thaw processes involved in keeping the samples for the 416-day study. These results illustrate the need for

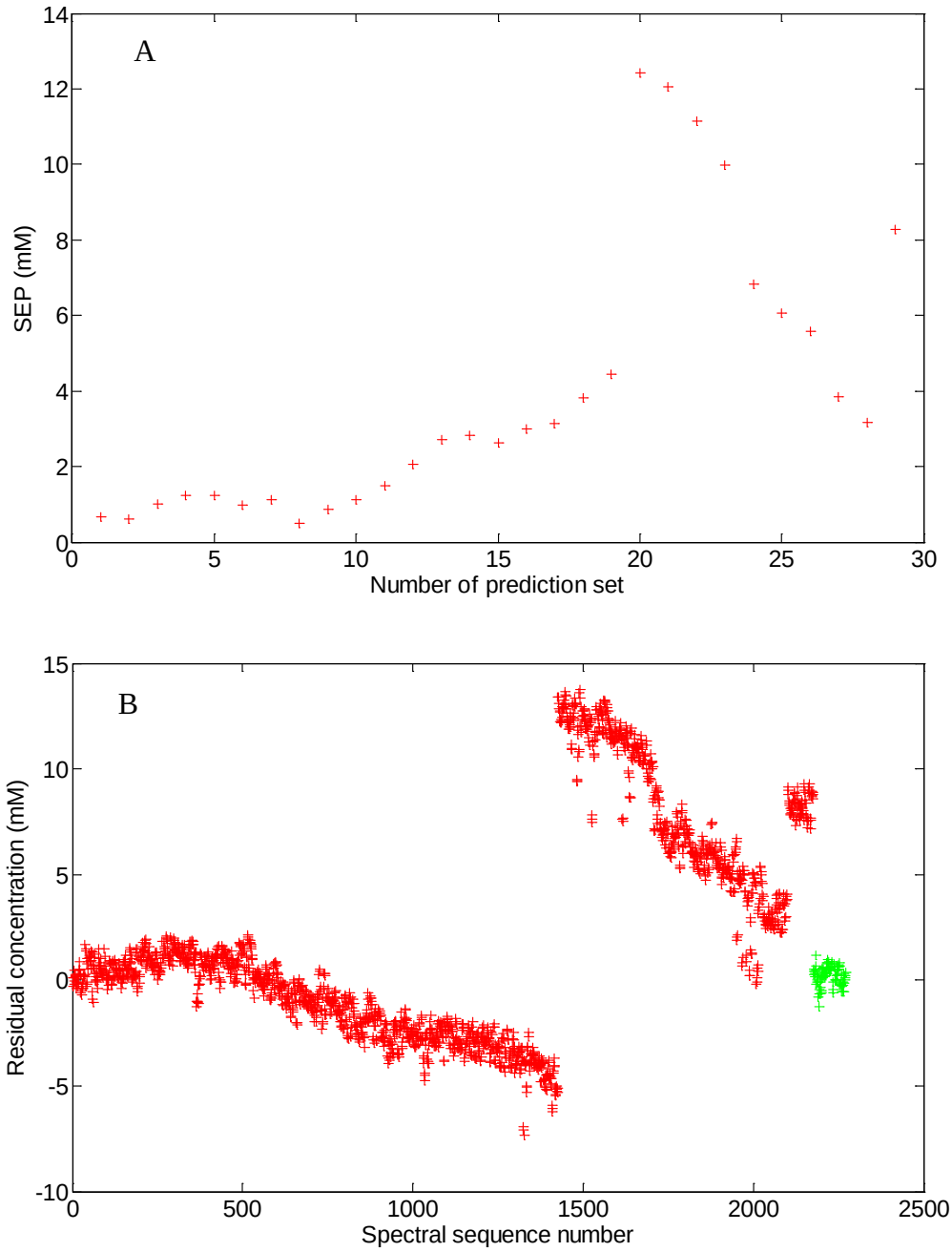
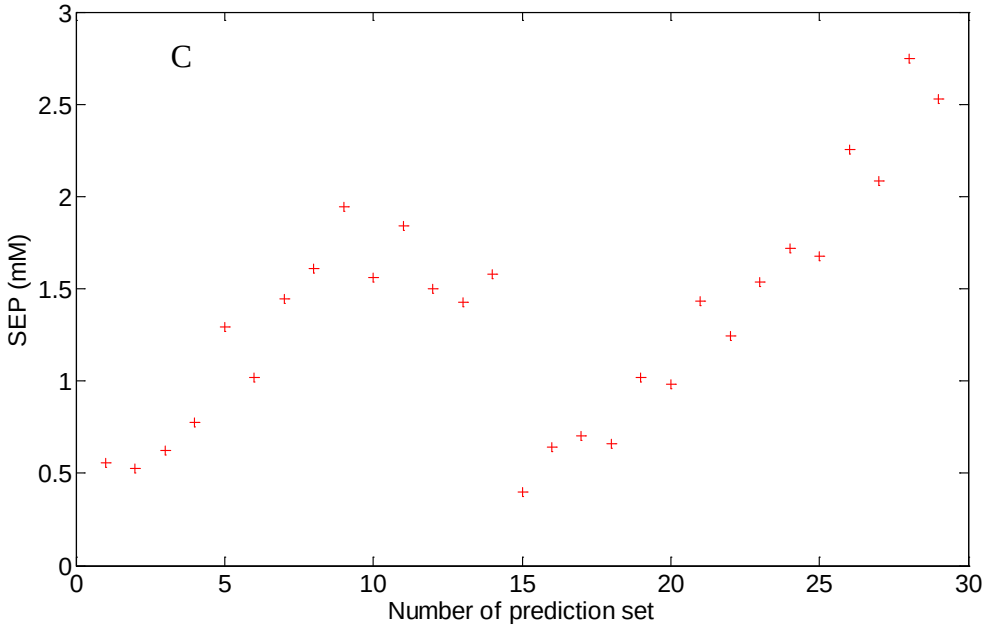
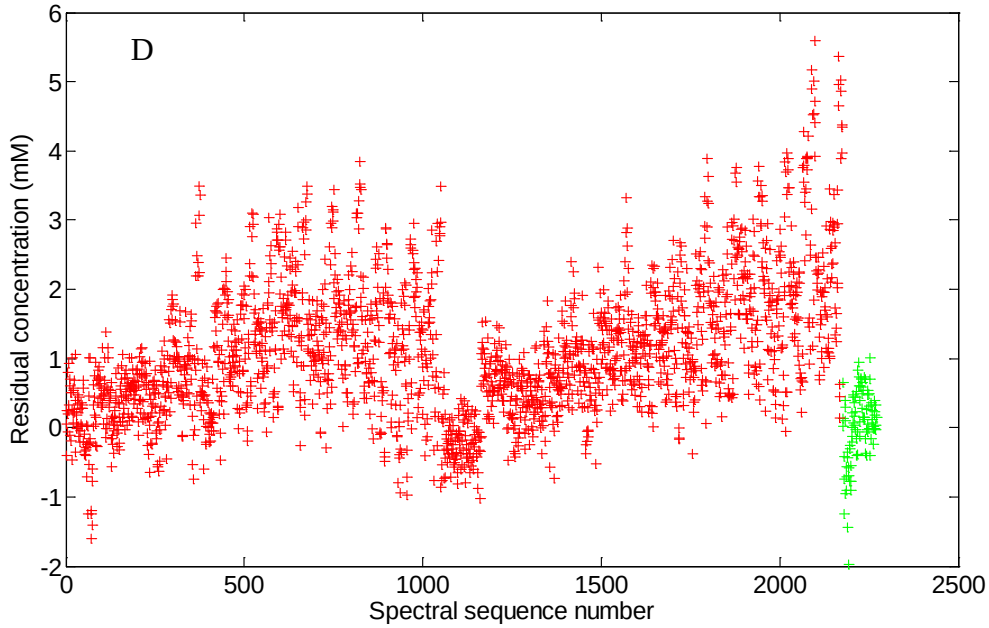


Figure 7.4 External prediction results (SEP values) for individual prediction sets *P1* to *P29*. Plots A and C correspond to the calibration models based on single-beam and absorbance spectra, respectively. Plots B and D display the individual concentration residuals for the spectra in prediction sets *A* (red) and *B* (green) produced by the calibration models based on single-beam and absorbance data, respectively. Results for both models show a clear degradation in prediction performance with time, although the model based on absorbance data exhibits smaller residuals overall than the model derived from single-beam spectra.





(Figure 7.4 continued)

methods to discern when an established calibration model is no longer able to perform adequately and therefore needs to be updated. Given the cost and effort involved in establishing a working calibration, such a diagnostic would be extremely valuable in maximizing the usable lifetime of a developed model.

7.3.4 Residual Modeling

Preliminary work indicated an evident correlation between the residual spectra and residual concentrations after the calibration models were applied. Efforts were then made to model the residual concentrations as a function of the residual spectra. Models were built with the residuals arising from the application of both calibration models. As noted previously, several model forms were applied (SVR, PLS, and PLS-aided SVR).

For each method used, 750 residual spectral segments from the odd-numbered prediction sets ranging from *P1* to *P20*, i.e., *P1*, *P3*, *P5*, …, *P19*, comprised the training set for the residual model. The rest of the 750 residual spectral segments from the corresponding even-numbered prediction sets, i.e., *P2*, *P4*, *P6*, …, *P20*, comprised a monitoring set of spectra for use in evaluating the residual models. The remaining 771 residual spectral segments from prediction sets *P21* to *P29* were used to form test set *C*. The 96 residual spectral segments from the second half of the calibration set, C_{full}, formed test set *D*. This data partitioning is summarized in Table 7.3.

Table 7.3 Data Partitioning for Residual Modeling

	Training set	Monitoring set	Test set *C*	Test set *D*
Data source	[*P1*, *P3*, …, *P19*]	[*P2*, *P4*, …, *P20*]	[P21, P22, …, P29]	2nd half of C_{full}
Number of observations	750	750	675	96

7.3.4.1 Residual Models Based on SVR

As introduced earlier, the residual models used the set of concentration residuals, ε_i, obtained from the application of the PLS calibration model to the spectra in the training set as the dependent variable. The values of ε_i are defined as

$$\epsilon_i = y_i - \hat{y}_i \qquad (7\text{-}1)$$

where y_i is the reference glucose concentration for sample i and $\hat{y}_i$ is the concentration predicted by the model. The corresponding intensities in the residual spectral segments (e.g., the range of 4609.8 to 4111.3 cm^{-1} for the model based on single-beam spectra) were used as independent variables in building the residual models.

For the initial models based on SVR, the SVM architecture used previously in Chapters 4, 5, and 6 was employed, with the difference being that instead of labeling each data object with a numerical class label (e.g., 1 or -1), the corresponding actual residual concentration was used. As with the previous work, radial basis functions were chosen as the kernel function mapping features in the original space to the higher-dimensional feature space. Two parameters related to the architecture of the SVM, the kernel parameter, γ, and regularization parameter, C, were investigated. A grid search was employed by varying γ from 0.16 to 15.52 over 25 levels and C from 32 to 40000 over 40 levels. This generated 1000 different SVM configurations.

For each architecture, an SVR model was developed by use of the 750 residual spectral segments in the training set. The computed model was subsequently applied to the residual spectra in the monitoring set. For each input spectral residual, the output of the SVR model was a predicted residual concentration. The standard errors in the predicted residual concentrations for the training set (termed the standard error of calibration or SEC) and the SEP for the monitoring data were calculated. The optimal model was chosen based on three criteria: (1) the minimum

SEP for the monitoring set, (2) the minimum SEC for the training set, and (3) the largest number of support vectors. The selected model was then applied to test sets *C* and *D* and the overall performance of the model was evaluated. The configuration parameters and the performance results of the selected models are listed in Table 7.4.

For the model based on single-beam spectra, the residual spectral segments obtained after application of the calibration model spanned a wavenumber range of 4609.8 - 4111.3 cm^{-1} and totaled 259 data points. The values of these 259 independent variables served as the 259-dimensional input pattern for SVR. The optimal SVR residual model was constructed with a γ parameter of 5.28 and a *C* parameter of 40,000 (Table 7.4). Figure 7.5 A is a correlation plot of predicted vs. true residual concentrations for this model.

As shown in Figure 7.5 B, the original residual concentrations under investigation spanned from -10 to +15 mM. Over this broad range, the selected SVR model was able to predict the residual concentrations accurately for all data subsets, including the training set, the monitoring set and test sets *C* and *D* (Figure 7.5 A). The residual concentrations remaining after this SVR modeling were plotted with respect to the spectral time sequence as shown in Figure 7.5, B, for a total of 2271 observations. This included the spectra measured from the second half of the calibration set (test set *D*) and the spectra collected from all prediction sets for a continuous 416-day period.

After application of the SVR residual model, the initial residual concentration range of -10 to +15 mM was reduced to a much lower and more randomly distributed range. Values of

Table 7.4 Summary of Residual Models

Residual models	Single-beam models			Absorbance models			
	SVR	PLS	PLS-aided SVR	SVR	PLS	PLS-aided SVR	PLS-aided SVR +
Spectral range (cm^{-1})	N/A	4524.5 – 4126.5	4524.5 – 4126.5	N/A	4475.6 – 4296.9	4475.6 – 4296.9	4475.6 – 4296.9
Number of latent variables	N/A	7	7	N/A	6	6	6
γ parameter	5.28	N/A	5.92	15.52	N/A	15.52	11.68
C parameter	40,000	N/A	32,768	27,304	N/A	40,000	32
SEC (mM)	0.42	0.50	0.51	0.78	0.48	0.73	0.48
SEP (monitoring) (mM)	0.54	0.87	0.55	0.66	0.50	0.62	0.50
SEP (Test set *C*) (mM)	1.13	1.65	0.91	1.15	0.54	0.99	0.51
SEP (Test set *D*) (mM)	0.47	0.51	0.52	0.86	0.65	0.93	0.64

SEP for the residual concentrations ranged from 0.47 mM for test set *D* to 1.13 mM for test set *C*. These values compared well to the standard errors in the residual concentrations for the training and monitoring sets (0.42 and 0.54 mM, respectively). The larger value for test set *C* reflects increasing time relative to both the original calibration model and the residual model. Overall, however, both the correlation and residual plots in Figure 7.5 A and B suggest that, for a broad residual concentration range resulting from long-term monitoring with a static calibration model, the selected SVR model worked well to retrieve the residual concentrations from the residual spectra remaining after application of the calibration model.

The same approach was also applied to the residual spectra obtained from the calibration model based on absorbance data. For this model, the residual spectra spanned the relatively narrow range of 4490.2 - 4290.8 cm^{-1} (104 spectral points). These spectral residuals were used as the input data for SVR modeling. A pool of residual models was developed, optimized and tested in the same manner described above. The configuration and evaluation results for the final selected model are listed in Table 7.4.

Figure 7.5 C is a correlation plot derived from the residual modeling of the absorbance data. The presentation format is the same as used in Figure 7.5 A. There is virtually no correlation evident, meaning that the SVR model failed to retrieve significant information about the residual concentrations directly from the residual spectra. This is also evident in the residual vs. spectral sequence plot in Figure 7.5 D. The profile of the remaining residual concentrations is very similar to the profile of the original concentration residuals in Figure 7.5 D.

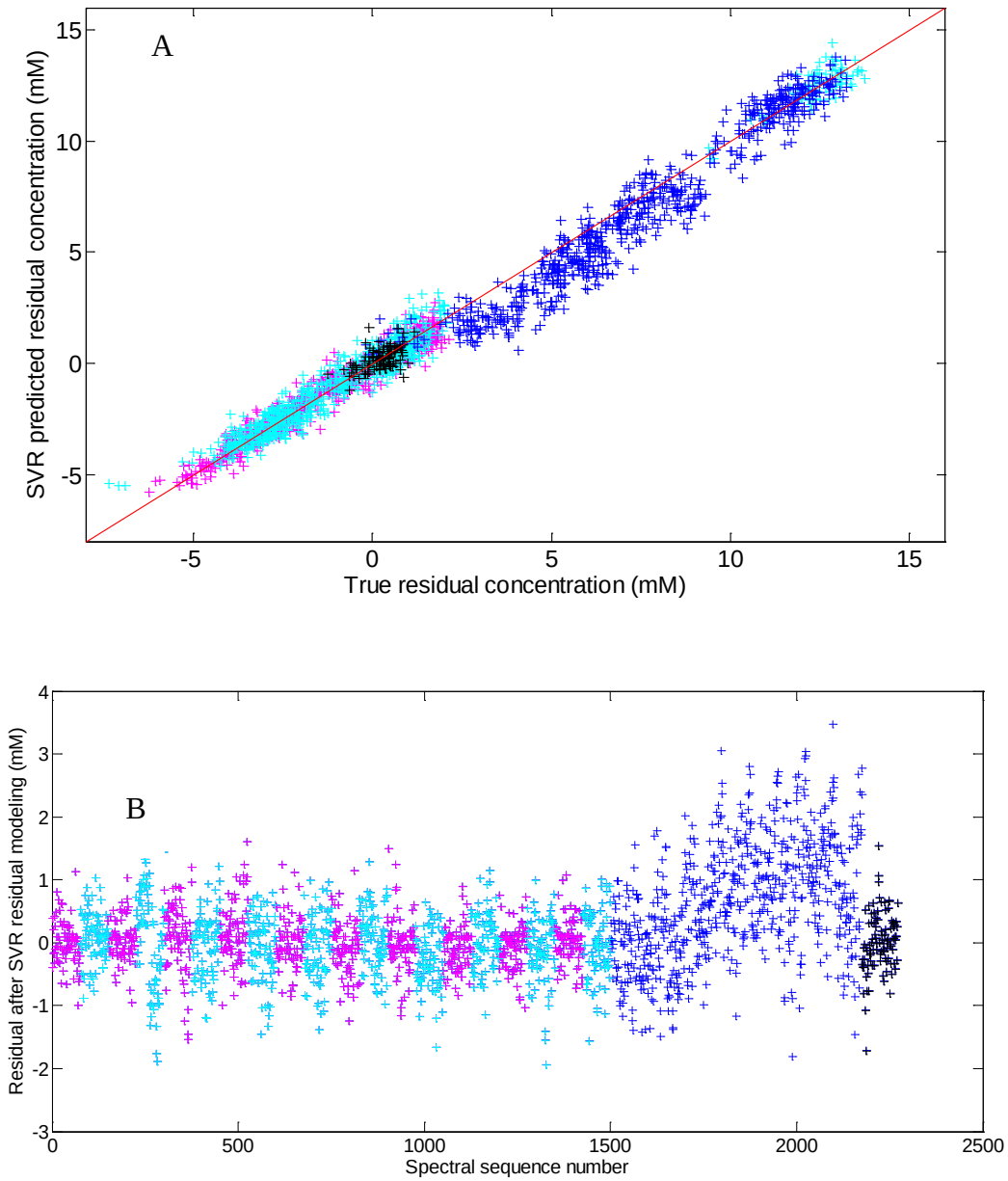
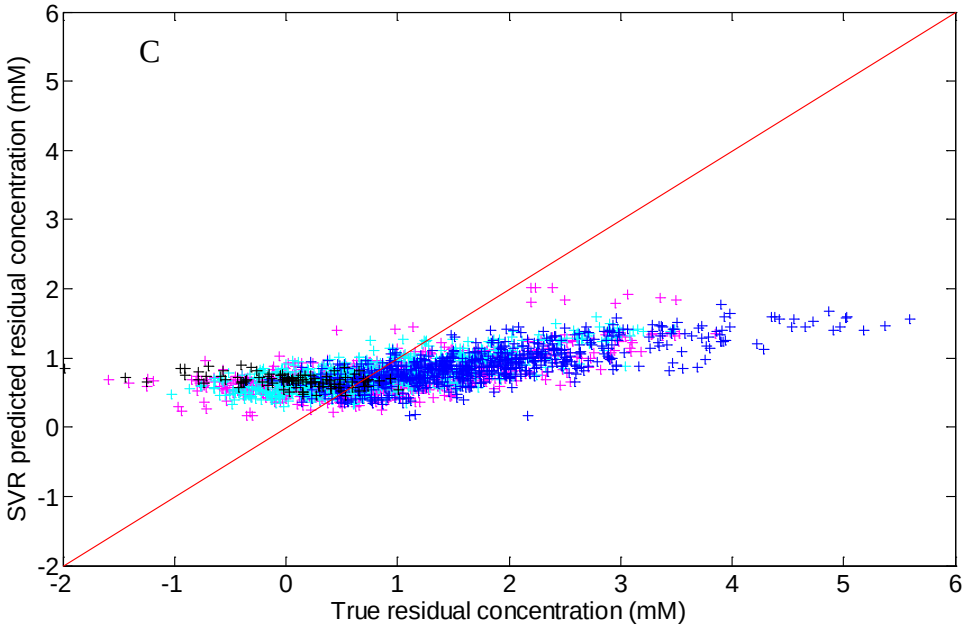
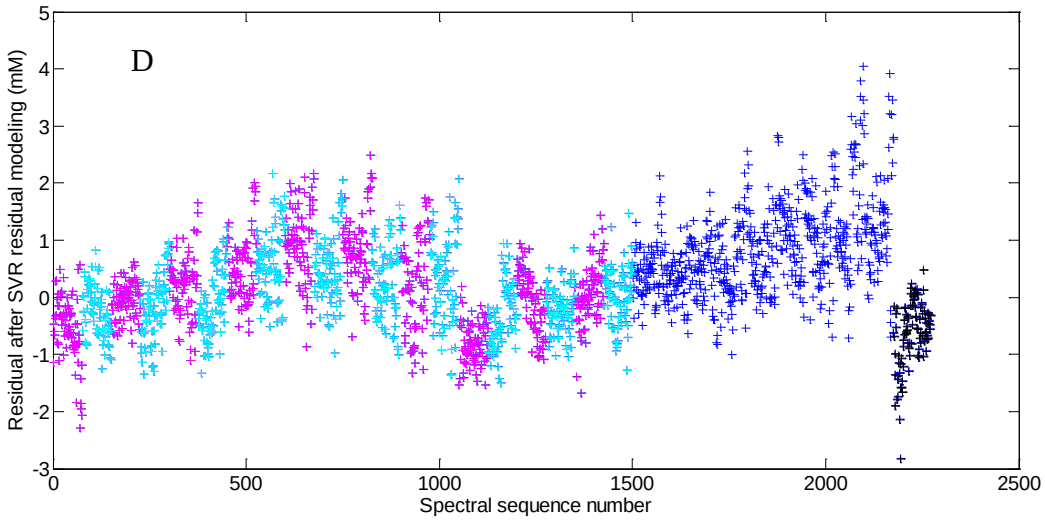


Figure 7.5 Characterization of SVR residual models. (A) Correlation plot of predicted residual concentrations vs. true residual concentrations derived from the calibration model based on single-beam spectra. Different colors are used to identify the data subsets: magenta – training set, cyan – monitoring set, blue – test set *C*, black – test set *D*. The correlation coefficient is 0.9894. (B) Residual trace (mM) vs. spectral sequence number application of the SVR residual model based on single-beam data. The blue and black traces correspond to test sets *C* and *D*, respectively. (C) Correlation plot of predicted vs. true residual concentrations derived from the model based on absorbance data. The color scheme is the same as in plot A. The correlation coefficient is 0.7175. (D) Residual trace (mM) vs. spectral sequence number after application of the SVR residual model based on absorbance data. The color scheme is the same as in plot B.





(Figure 7.5 continued)

This result indicates that either there is little useful information in the residual spectra associated with the calibration model based on absorbance spectra or that the SVR modeling protocol was unable to extract that information. To address this problem further, other modeling procedures were evaluated.

7.3.4.2 Residual Models Based on PLS

The PLS method was next used to build analogous residual models to those described above based on SVR. The concentration residuals in the training set served as the dependent variable in the PLS calculations, and the residual spectra obtained from the application of the original calibration model served as the independent variables. Within the spectral range of the residual spectra, the same wavenumber selection strategy and range of latent variables studied previously in the optimization of the original calibration model were employed. This produced a pool of possible PLS residual models. The SEP for the monitoring set was calculated for each model, and the optimal model was chosen as that which produced the lowest SEP for the monitoring set. The selected model was then applied to the two test sets, *C* and *D*, and the overall model performance was evaluated. The configuration parameters and performance results for the models based on both single-beam and absorbance data can be found in Table 7.4.

For the PLS residual model based on single-beam spectra, the optimal residual PLS model had a spectral range of 4524.5 – 4126.5 cm^{-1}, with 7 latent variables. Figure 7.6 A and B are correlation and residual sequence plots for this model analogous to those shown previously in Figure 7.5 A and B for the SVR model. Assessment of the statistics in Table 7.4 and visual comparisons of the corresponding plots in Figure 7.4 and 7.5 reveal that the PLS residual model is somewhat less effective than the model based on SVR. Information is clearly being extracted from the residual spectra by the PLS model, but the overall performance is not as good.

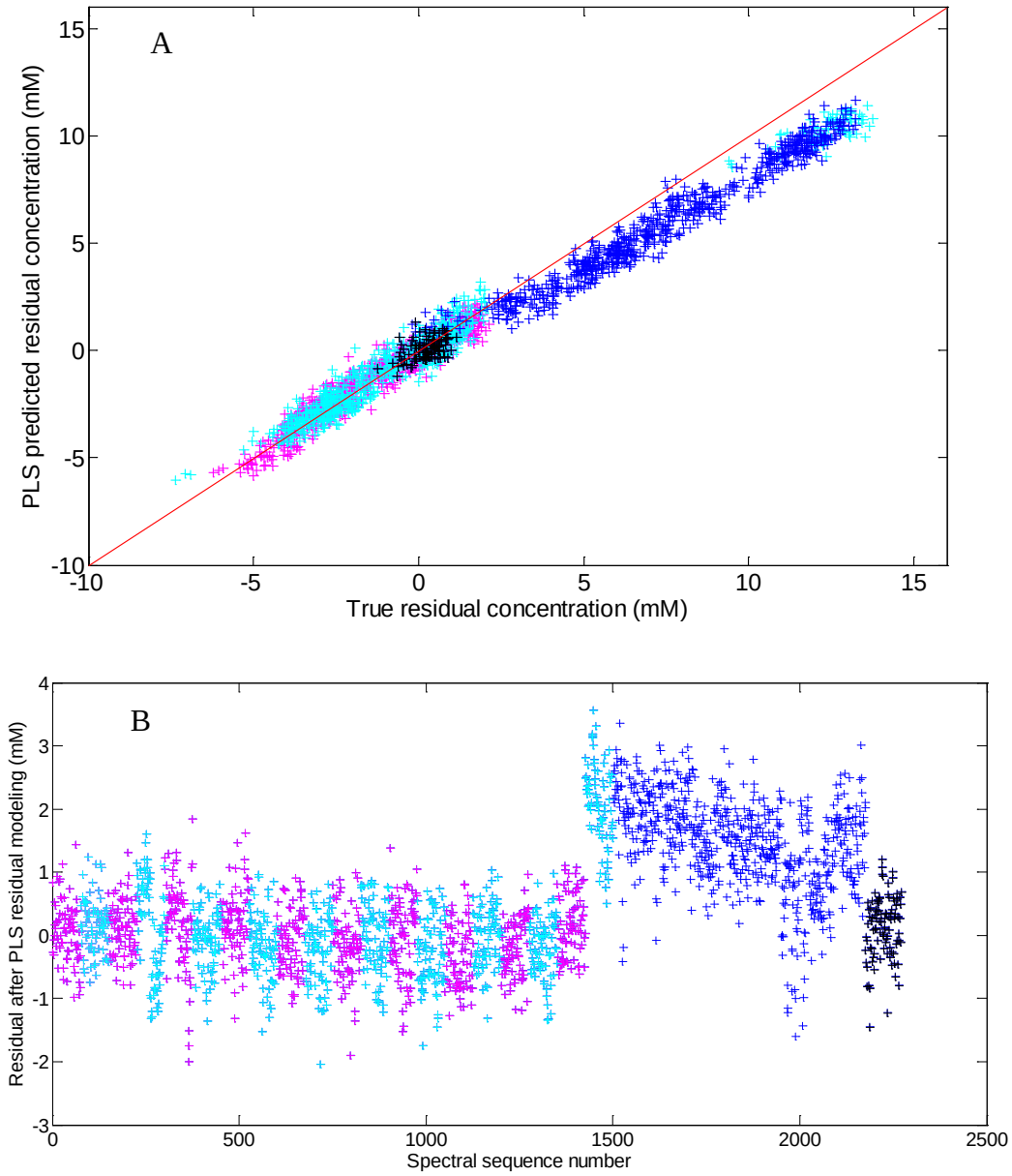
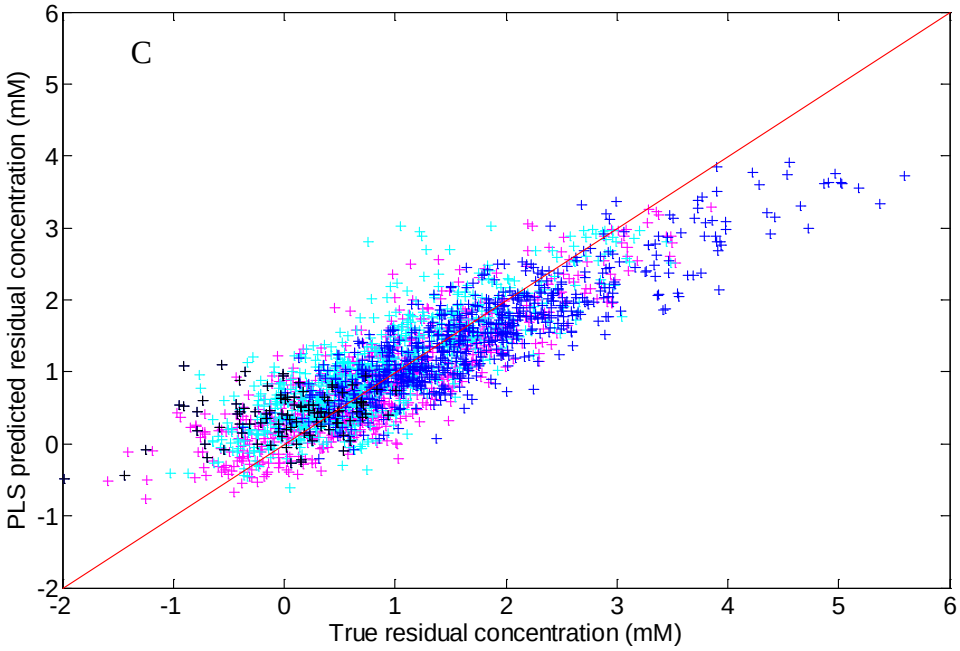
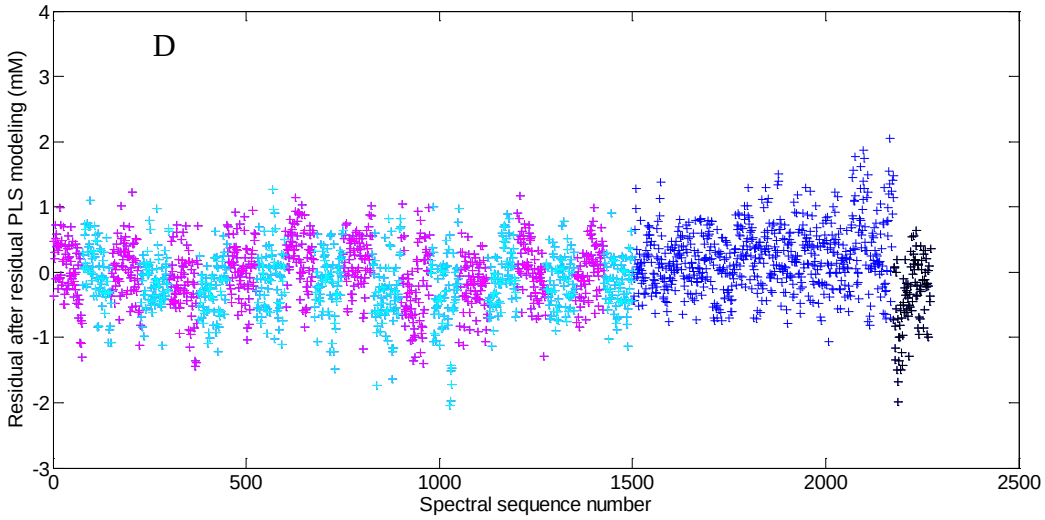


Figure 7.6 Characterization of PLS residual models. (A) Correlation plot of predicted residual concentrations vs. true residual concentrations derived from the calibration model based on single-beam spectra. Different colors were used to identify the data subsets: magenta – training set, cyan – monitoring set, blue – test set *C*, black – test set *D*. The correlation coefficient is 0.9912. (B) Residual trace (mM) vs. spectral sequence number application of the PLS residual model based on single-beam data. The blue and black traces correspond to test sets *C* and *D*, respectively. (C) Correlation plot of predicted vs. true residual concentrations derived from the model based on absorbance data. The color scheme is the same as in plot A. The correlation coefficient is 0.8601. (D) Residual trace (mM) vs. spectral sequence number after application of the PLS residual model based on absorbance data. The color scheme is the same as in plot B.





(Figure 7.6 continued)

For the residuals derived from the calibration model based on absorbance spectra, the optimal residual PLS model was based on 6 latent variables and a spectral range of 4475.6 – 4296.9 cm^{-1}. Statistics for this model are presented in Table 7.4, and Figure 7.5 C and D are the corresponding correlation and residual sequence plots. Compared to the analogous model based on SVR, the PLS model was much more successful in retrieving information regarding the residual concentrations from the residual absorbance spectra. This conclusion is supported by the clear correlation between the predicted and true residual concentrations in Figure 7.6 C and the random structure of the residual sequence plot in Figure 7.6 D. The fact that the SEP values for the residual concentrations are approaching the CV-SEP of approximately 0.5 mM associated with the original calibration model suggests that the useful information in the residual spectra has been extracted and that the remaining information is noise.

7.3.4.3 Residual Models Based on PLS-aided SVR

The next study was designed to gain insight into why the SVR residual model failed with the absorbance data while outperforming the PLS model for the single-beam data. A key difference between the two approaches was the much greater dimensionality of the input into the SVR model than the PLS model. For example, in the case of the residuals derived from the absorbance data, the SVR procedure operated with 104-dimensional inputs. By contrast, the PLS procedure embodied a feature extraction step that also included a study of wavenumber subsets within the 104-dimensional input. The final PLS model reduced the 104-dimensional input to 6 latent variables that formed the basis for the model.

To evaluate the impact of input dimensionality on the SVR procedure, PLS scores derived from the residual spectra were used as inputs into the SVR model. In principle, this approach combines the feature extraction capability of PLS with the nonlinear modeling ability

of SVR. The hypothesis tested was whether simplifying the SVR input would add greater robustness to the resulting model.

This combined approach was called PLS-aided SVR. Here, instead of residual spectral segments as input patterns, the PLS scores of the residual PLS model were fed into SVR as the input data. The remainder of the SVR procedure was the same as that employed in the original work.

For the model based on single-beam spectra, the seven PLS scores computed from the 4524.5 – 4126.5 cm^{-1} range of the residual spectra served as the input data for SVR. Results for the resulting SVR model are summarized in Table 7.4, and correlation and residual plots are provided in Figure 7.7 A and B, respectively. The optimal model was constructed with a γ parameter of 5.92 and a *C* parameter of 32,768. For the training set, monitoring set, and test set *D*, the correlation plot and residual plot both indicated similarly good prediction performance to that obtained with the original SVR residual model and better performance than the corresponding PLS residual model. Values of SEC for the training set and SEP for the monitoring set and test set D were slightly worse for the PLS-aided SVR model when compared to the original SVR residual model. For test set *C*, however, the SEP was reduced to 0.91 mM, an improvement of 19.5 % relative to the original SVR model. This improvement can be seen clearly by comparing the residual sequence trace in Figure 7.7 B to the corresponding plot in Figure 7.5 B for the original SVR model. The improvement is primarily seen in the latter portion of test set C after sequence number 1500.

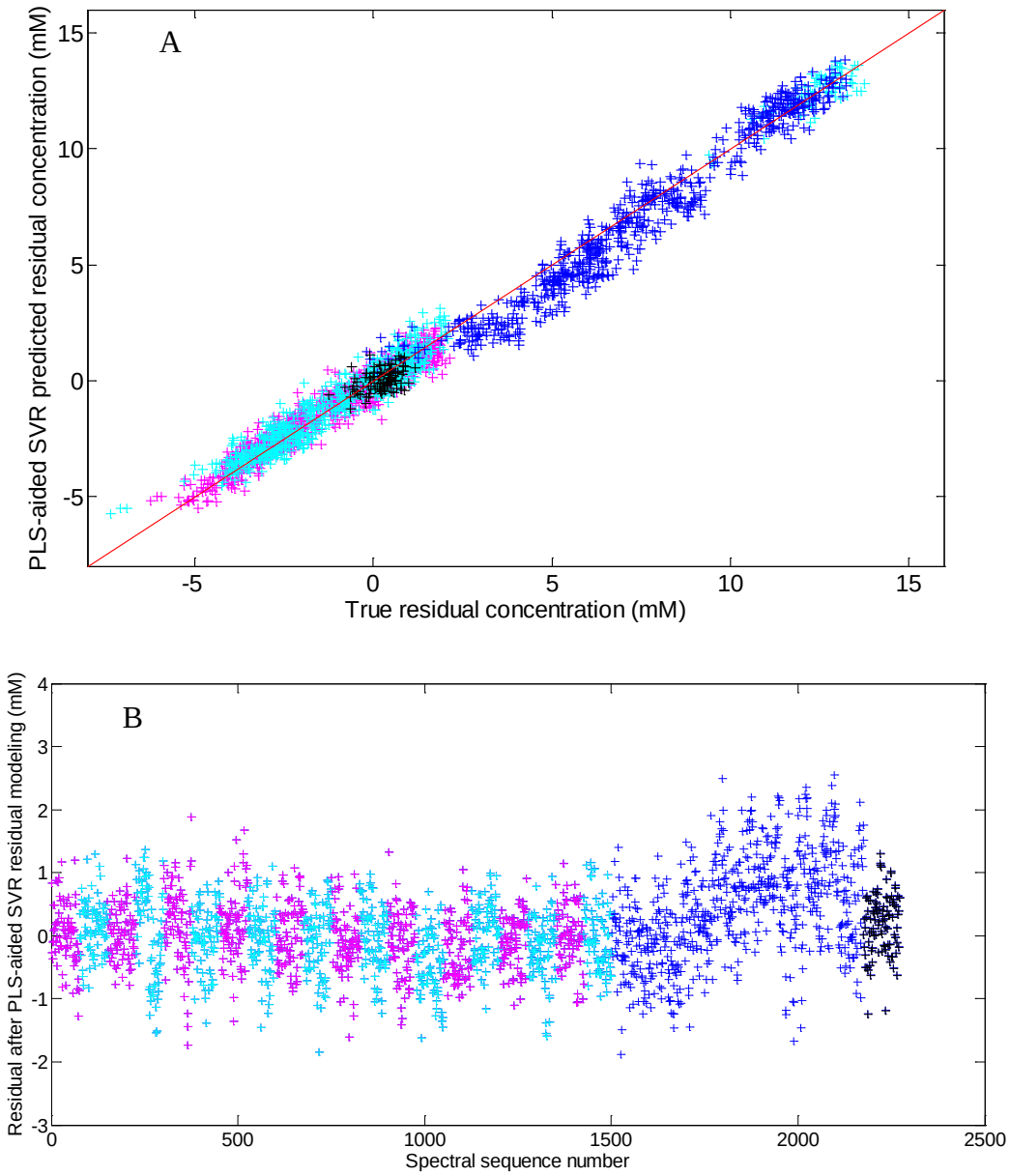
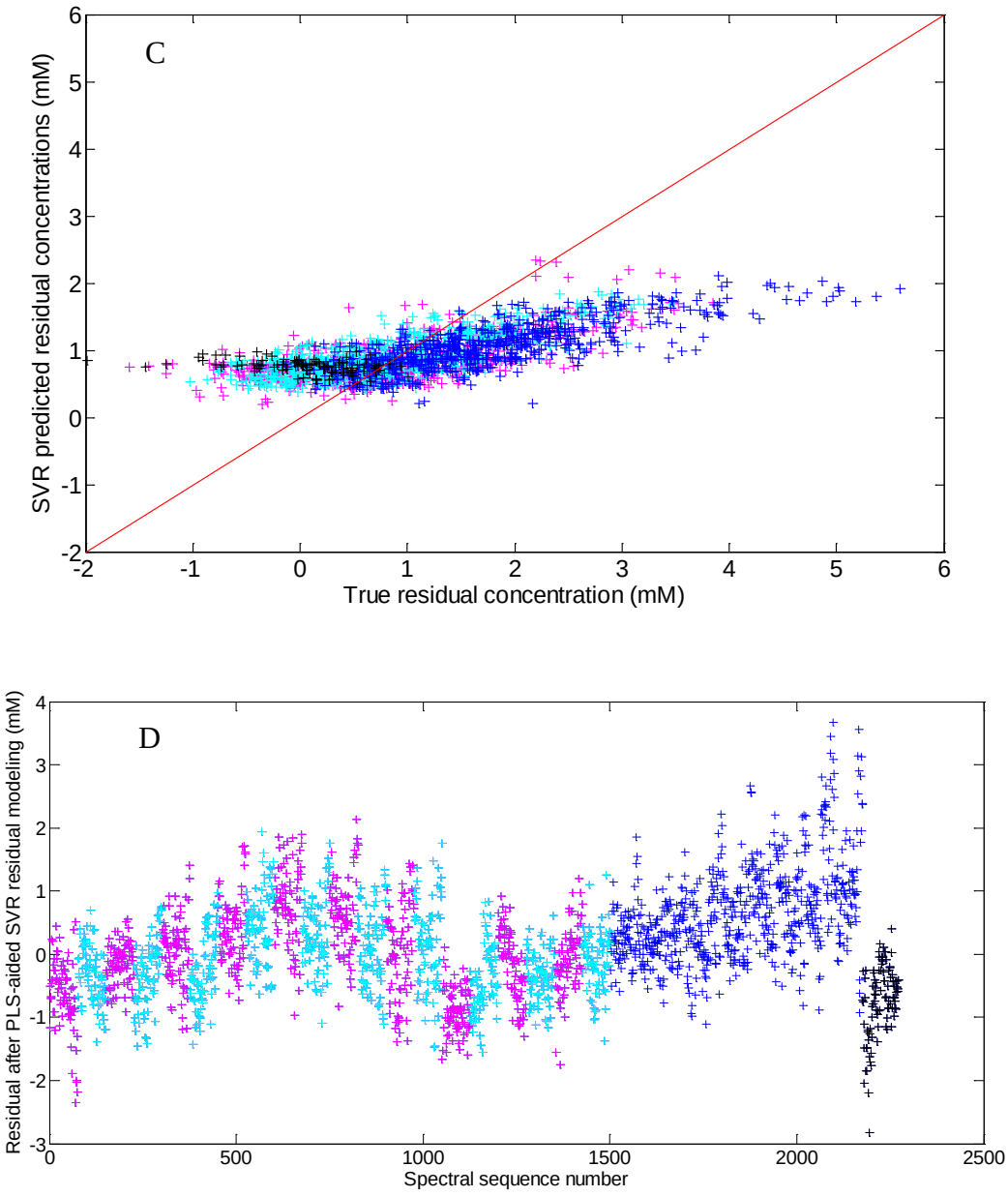


Figure 7.7 Characterization of PLS-aided SVR residual models. (A) Correlation plot of predicted residual concentrations vs. true residual concentrations derived from the calibration model based on single-beam spectra. Different colors were used to identify the data subsets: magenta – training set, cyan – monitoring set, blue – test set *C*, black – test set *D*. The correlation coefficient is 0.9909. (B) Residual trace (mM) vs. spectral sequence number application of the PLS-aided SVR residual model based on single-beam data. The blue and black traces correspond to test sets *C* and *D*, respectively. (C) Correlation plot of predicted vs. true residual concentrations derived from the model based on absorbance data. The color scheme is the same as in plot A. The correlation coefficient is 0.7392. (D) Residual trace (mM) vs. spectral sequence number after application of the PLS-aided SVR residual model based on absorbance data. The color scheme is the same as in plot B.



(Figure 7.7 continued)

The same approach was applied to the model based on absorbance data. Here, the six PLS scores computed from the 4475.6 – 4296.9 cm^{-1} range served as inputs to the SVR model. The same SVR modeling procedure was used as described previously. The characteristics of the resulting optimal model are presented in Table 7.4. Figure 7.7 C and D are the corresponding correlation and residual sequence plots, respectively.

When Figure 7.7 C is compared to the correlation plot for the original SVR model in Figure 7.4 C, only very slight improvement is noted. The largely horizontal shape of the correlation plot confirms that little useful information is being extracted from the residual spectra for use in modeling the residual concentrations. The residual sequence trace in Figure 7.7 D is again only slightly better than that obtained with the original SVR model (Figure 7.5 D). Similarly small improvements are observed in the SEP values.

7.3.4.4 Residual Models Based on Amplified PLS-aided SVR

The results presented in the previous section did not resolve the issue of why the SVR-based models did not perform well with absorbance data. In investigating the results, it was noted that the scale of the PLS scores may play a role. A survey of the PLS scores revealed that PLS scores based on single-beam data were on the order of 10^{-1}, while the corresponding scores based on absorbance data were in the range of 10^{-4}. This result was reasonable given that the single-beam spectrum is plotted in arbitrary units of light intensity, while the absorbance spectrum follows a logarithmic scale based on the fixed 0 to 1 range of light transmittance. This can be observed in the spectra plotted in Figure 7.1 B and C.

To evaluate whether scaling played a role in the modeling, a scaling factor of 600 was multiplied by the original six sets of PLS scores derived from the residual absorbance spectra (computed from the previously optimized range of 4475.6 – 4296.9 cm^{-1}). This value was chosen so that the amplified scores were brought to approximately the same magnitude as the scores

derived from the residual single-beam spectra. The scaled scores then served as the input to the SVR model. The rest of the procedure was the same as that used previously (optimization of SVM parameters, etc.).

The results obtained are summarized in the last column of Table 7.4. The values of SEC for the training set and SEP for the monitoring set and test sets *C* and *D* are much improved relative to those obtained with the original PLS-aided SVR method. Compared to the linear PLS model, the values of SEC and SEP for the monitoring set are identical, while there is a slight improvement in the SEP values for test sets *C* and *D*. Figure 7.8 A and B are correlation and residual sequence plots, respectively, that are displayed in the same format used previously in Figures 7.5, 7.6, and 7.7. These plots demonstrate that with the scaling modification, the SVR model is now able to extract useful information from the residual spectra. When compared to the linear PLS model (Figure 7.6 C and D), however, there is not a clear benefit to the SVR model, particularly given the complication of having to optimize the γ and C configuration parameters.

7.3.5 Effect of residual correction

It is clear that the residual models presented above, based upon either single-beam or absorbance residual spectra, can predict residual concentrations fairly accurately. Incorporation of those residual concentrations derived from selected residual models into the calibration models, allows quantitative information, once left behind by the established calibration models, to be retrieved. As a result, an improvement in the prediction performance of the calibration model can be realized.

For all the prediction sets, from *P1* to *P29*, SEP values were calculated based upon concentrations predicted by the calibration models only, and concentrations predicted by the calibration models and corrected by use of the selected residual models. The PLS-aided SVR

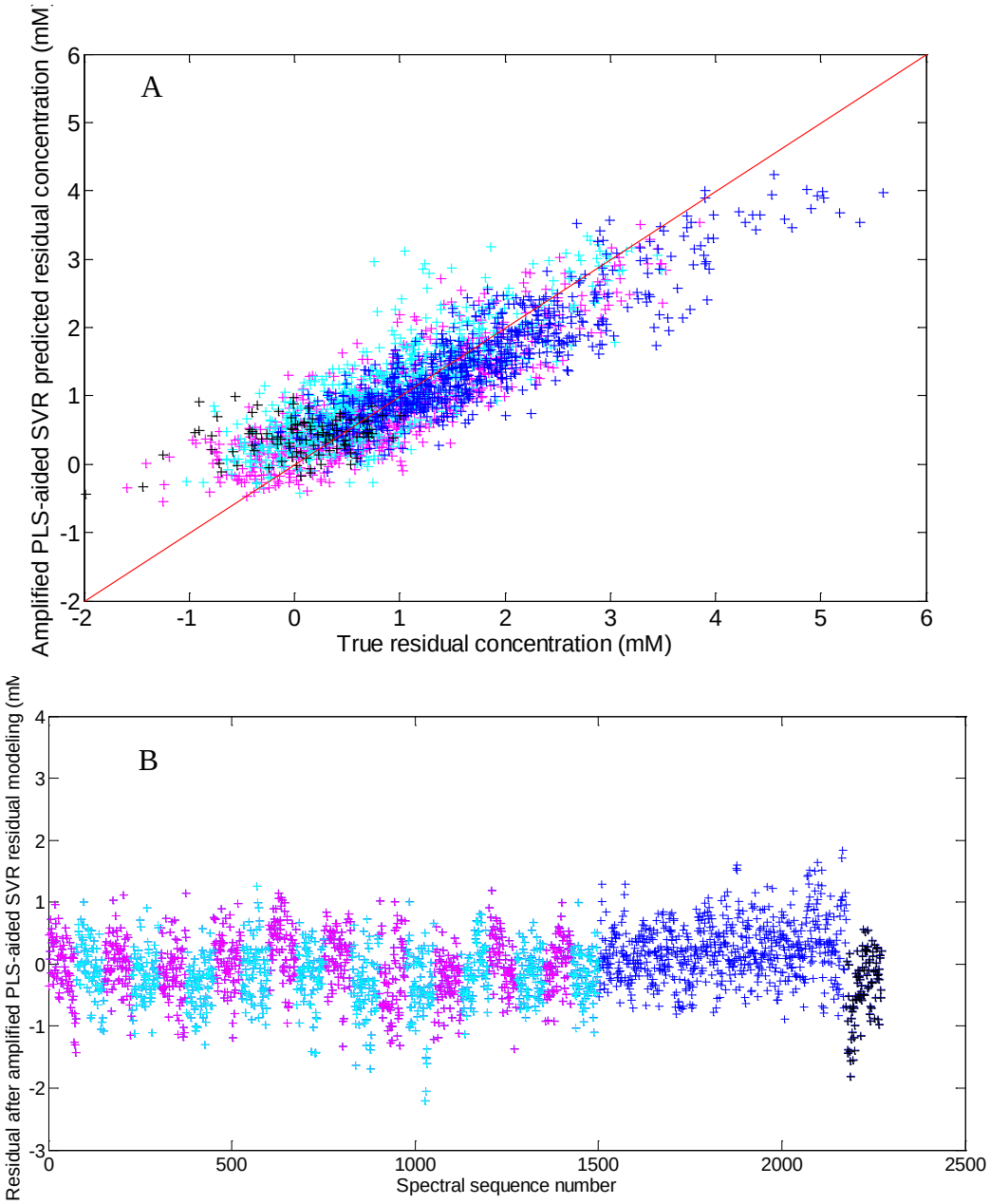


Figure 7.8 Characterization of amplified PLS-aided SVR residual models. (A) Correlation plot of predicted residual concentrations vs. true residual concentrations derived from the calibration model based on absorbance spectra. Different colors were used to identify the data subsets: magenta – training set, cyan – monitoring set, blue – test set *C*, black – test set *D*. The correlation coefficient is 0.8645. (B) Residual trace (mM) vs. spectral sequence number after application of the amplified PLS-aided SVR residual model based on absorbance data. The blue and black traces correspond to test sets *C* and *D*, respectively.

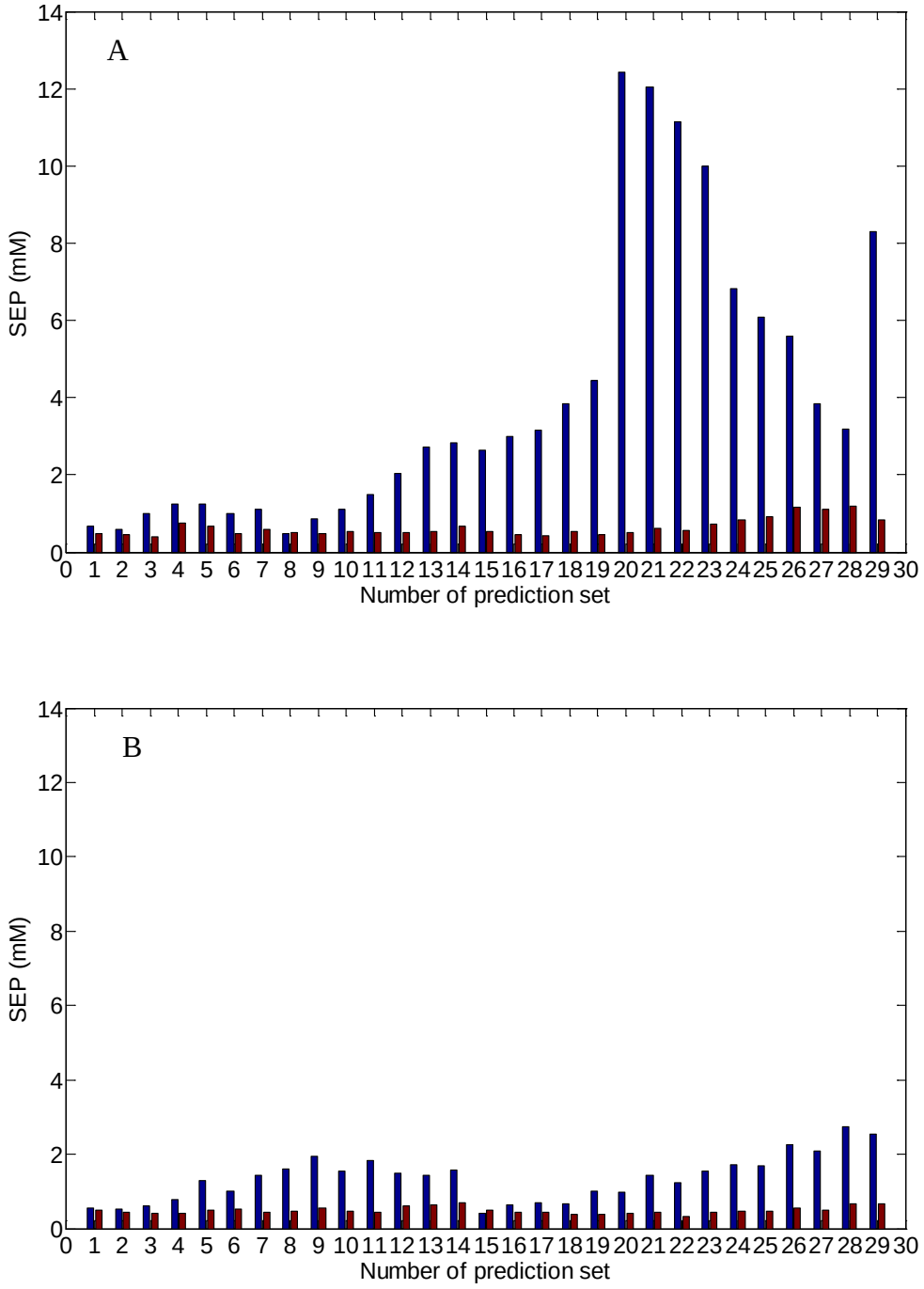


Figure 7.9 Effect of residual modeling on the prediction performance of the calibration model. Values of SEP (mM) for individual prediction sets from *P1* to *P29* are shown for the following cases: blue – SEP values without any residual correction, red – SEP values reflecting correction with residual concentrations (mM) retrieved from residual models. (A) PLS-aided SVR was selected as the residual model for the calibration model based on single-beam spectra. (B) Amplified PLS-aided SVR was selected as the residual model for the calibration model based on absorbance spectra.

method produced the selected residual model for the single-beam spectra, and the amplified PLS-aided SVR method produced the residual model used with the absorbance data. A side-by-side comparison of SEP values, without or with the residual correction, is shown in Figure 7.9.

With residual correction added, the prediction performance of the calibration model improved dramatically and its degradation over time appeared to be corrected as well. In the case of the model based on single-beam spectra, SEP values decreased from a little more than 12 mM at maximum down to less than 1.2 mM at maximum. The average SEP was 0.64 mM over the entire course of more than 412 days of data collection for all prediction sets. For the model based on absorbance spectra, the maximum SEP dropped from 2.7 mM to 0.70 mM, with an average of 0.49 mM over the time span of the data collection. The SEP values corrected by use of the residual models were comparable to the CV-SEP values, 0.51 mM for the single-beam calibration, and 0.45 mM for the absorbance calibration. This confirms the effectiveness of residual correction as a remedy to maintain the predictive ability of a calibration model over time.

7.3.6 Classification Performance of Residual Models

The work presented previously clearly demonstrates that information exists in the residual spectra that correlate with the residual concentrations. Next, studies were performed to assess the ability to use the residual models as a diagnostic tool to predict whether or not a given predicted concentration is reliable.

Each spectrum in all the data sets was assigned a binary class membership, either the alarm class or the non-alarm class, referenced against a pair of cutoffs corresponding to the residual concentrations. This terminology is analogous to that used in the previous chapters in discussions related to process monitoring. Membership in the alarm class signals a case in which the residual concentration is outside the distribution associated with the calibration model. This

case can be considered a failed prediction, necessitating an alarm condition in which the operator of a process would be notified. The non-alarm class would indicate a predicted concentration within the distribution associated with the calibration model. The predicted concentration would be considered valid in this case.

The SEP value, calculated from operating the calibration model on the first prediction set, *P1*, was utilized to set up the reference cutoff concentrations. Two sets of cutoff concentrations, either two times or three times the SEP value plus or minus about 0 mM, were implemented respectively. If a true residual concentration fell within the range spanned by the reference cutoffs, it would be assigned to the non-alarm class. Otherwise, if the true residual concentration fell beyond the upper reference cutoff or lower than the lower reference cutoff, it would be assigned to the alarm class, thus signaling a failed prediction.

Subsequently, the residual models generated predicted residual concentrations for each data point. A standard error, associated with the predicted residual concentrations, was calculated as the root sum of squares of the SEP value for the prediction set, *P1*, derived from the calibration model, and the SEC value from the training set of the selected residual models (i.e., $\sqrt{SEP^2 + SEC^2}$). This composite error value attempted to describe the error propagating through the calibration model and the residual model.

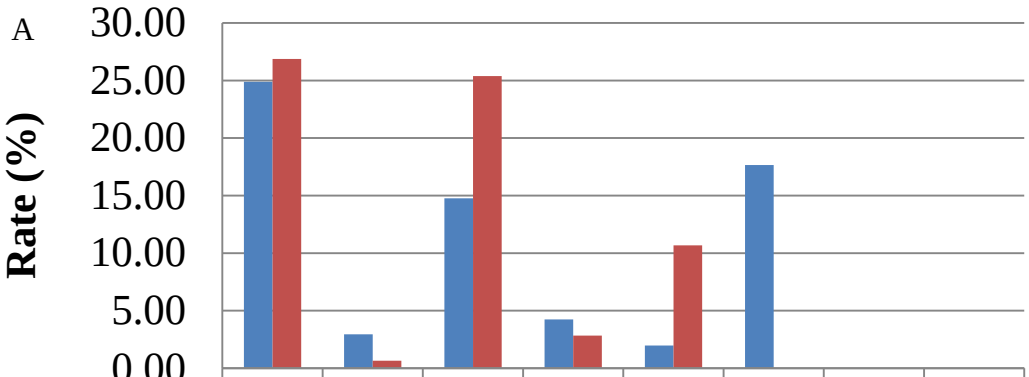
New sets of reference cutoff concentrations, calculated as two or three times this new standard error plus or minus about 0 mM, were then established for the predicted residual concentrations. For a data point belonging to the alarm class determined against the reference cutoff concentrations associated with the calibration model, if its predicted residual concentration was actually within the range of the pair of cutoff concentrations associated with residual model, this particular data point was considered misclassified by the residual model. Similarly, if a data

point belonging to the non-alarm class determined against the reference cutoff concentration associated with the calibration model produced a predicted residual concentration higher than the upper cutoff concentration or lower than the lower cutoff concentration associated with residual model, then this particular data point was considered a false detection by the residual model. The missed detection rate was the percentage of the count of misclassified data points relative to the total number of data points in the alarm class, and the false detection rate was the percentage of misclassified data points over the total number of data points in the non-alarm class.

Two selected residual models, the PLS-aided SVR model computed from single-beam spectra and the amplified PLS-aided SVR model based on absorbance data, were employed in the classification study based on the rules described above. The corresponding results are presented in Table 7.5. Note that the prefixes 2× or 3× in the table mean two times or three time the standard errors were used in establishing the reference cutoff concentrations (for both the calibration and residual models). The data partitioning here was the same as mentioned in Table 7.3.

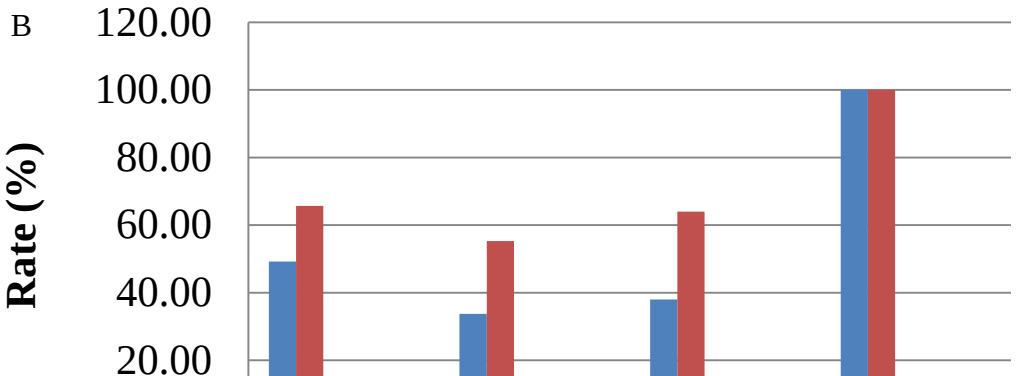
As illustrated in Table 7.5, in both single-beam mode and absorbance mode, the broader cutoff limits, 3×, helped to have more data points in the non-alarm class and decrease the false detection rate, while the narrower cutoffs, 2×, had more alarm data points and a lower missed detection rate. Data measured during the time closer to the calibration set, in other words, with smaller true residual concentrations, such as test set *D*, resulted in a higher population in the non-alarm class and a 0.00 % false detection rate. However, in this case, there were few or no data points falling into the alarm class, leading to missed detection percentages that were artificially high.

A



	trn miss	trn false	mon miss	mon false	test set C miss	test set C false	test set D miss (NA)	test set D false
2X PLS-SVR	24.88	2.94	14.76	4.24	1.98	17.65	0.00	0.00
3X PLS-SVR	26.87	0.66	25.38	2.84	10.67	0.00	0.00	0.00

B



	trn miss	trn false	mon miss	mon false	test set C miss	test set C false	test set D miss	test set D false
2X PLS-SVR +	49.20	2.40	33.74	4.56	37.97	2.49	100.00	0.00
3X PLS-SVR +	65.67	0.65	55.26	2.20	64.00	0.80	100.00	0.00

Table 7.1 (A) Classification results for PLS-aided SVR (PLS-SVR) residual model based on single-beam spectra. Note that the missed detection rate for test set *D* was not applicable, as there were no data points in the alarm class; (B) Classification results for the amplified PLS-SVR (PLS-SVR +) residual model computed from absorbance spectra. As described in the text, cutoff concentrations of 2× and 3× were investigated. In the table, the labels "trn", "mon", and "prd" denote the training, monitoring, and combined prediction data, respectively.

In comparison, data measured at much longer times relative to the calibration set, such as test set *C*, the monitoring set and the training set, produced higher populations in the alarm class, with fewer data points being misclassified (1.98 to 65.67 %). Meanwhile the false detection rate in the non-alarm class slightly increased (0.65 % to 17.65 %). In general, across all data sets, the residual model based on single-beam spectra resulted in more data points in the alarm class and lower missed detection rates, while the residual model based on absorbance spectra had more data points classified as non-alarm and lower false detection rates.

One exception was that there were no false detections out of 19 data points (false detection rate of 0.00%) in the non-alarm class of test set *C* from the residual model based on single-beam spectra compared to 3 cases of false detections out of 375 data points (false detection rate of 0.80 %) when the residual model based on absorbance spectra was used. The trends mentioned above imply that a tradeoff decision has to be made between improving the missed detection and false detection rates. Generally, with narrower reference limits and data with higher true residual concentrations, the residual model based on single-beam spectra produced fewer missed detections. For example, the lowest missed detection rate (1.98 %) was found from test set *C* using the single-beam residual model with 2× standard errors as reference limits. By comparison, broader reference limits and data with lower true residual concentrations tended to produce better false classification performance. However, the effect of the input spectral format, single-beam or absorbance, on the false detection rates was not as clear as that found for the missed detection rates. In this case, the lowest false detection rate (0.00 %) was found from test set *D* using any model with 2× or 3× standard errors as reference limits.

7.3.7 Performance Diagnostics of Calibration Model

Overall good classification results confirmed the reliability of the residual model as a tool to predict residual concentrations after calibration. From the perspective of performance

diagnostics in a practical scenario, a PLS calibration model would be applied to each spectrum with unknown concentration. As a result, a predicted concentration and a residual spectral segment would be generated for each unknown spectrum. In single-beam mode, if the PLS-aided SVR residual model was used, an estimated residual concentration would then be obtained. If the predicted residual concentration was within the previously established threshold range (e.g., three times the composite root sum squared error, [-2.52 mM, +2.52 mM]) the predicted concentration obtained from the calibration model would be judged acceptable. If the predicted residual concentration was outside the acceptable range, an alarm would be initiated that the calibration model might need an update. An operational cutoff would have to be established regarding the frequency of alarms that would be tolerated vs. a judgment that the calibration model could no longer be used.

7.4 Conclusions

As a part of a general calibration maintenance strategy, a protocol of performance diagnostics for a calibration model built for long-term near-IR-based glucose monitoring was proposed and investigated in this work. By building models from residual spectra remaining after application of the calibration model, residual concentrations were retrieved successfully from the residual models. A residual concentration threshold range could then be derived from the chosen residual models. Based upon the output of whether the retrieved residual concentrations were within or beyond a derived residual threshold range, a decision of whether or not a calibration update was needed could be made.

Residual models were built based on both single-beam spectra and absorbance spectra. Different modeling techniques, including PLS, SVR, PLS-aided SVR, and the newly proposed amplified PLS-aided SVR, were developed and optimized for residual modeling. For the residual model based on single-beam spectra, the PLS-aided SVR method outperformed the other models

and produced the lowest SEP (0.91 mM) for test set *C*. This modeling approach combined the latent variable advantage from PLS and nonlinear regression advantage from SVR. For the residual model based on absorbance spectra, the original hybrid model did not work unless an amplification of the PLS scores was made before submission of the data to the SVR procedure. The resulting amplified PLS-aided SVR model achieved slightly better prediction performance than the absorbance PLS residual model, with a SEP for test set *C* of 0.51 mM (vs. 0.54 mM).

Results from classification tests demonstrated that the residual modeling can provide efficient performance diagnostics for the established calibration model. Two levels of cutoffs, varying about zero mM with three or two times the propagated standard error of the calibration and residual models, were investigated as classification cutoffs. For the selected single-beam residual model, PLS-aided SVR, both the 3× and 2× cutoffs worked, resulting in missed detection rates of 10.68 % and 1.98 %, and false detection rates of 0.00 % and 17.65 %, respectively, for the test set most separated in time from the calibration data. For the best absorbance residual model, amplified PLS-aided SVR, the 2× cutoffs worked best, with a missed detection rate of 37.97 % and a false detection rate of 2.49 % for the same test set.

The absorbance-based residual model produced residual concentrations which provided slightly better correction to the original predicted concentrations (Figure 7.9), but this better SEP correction did not produce better performance diagnostics of the calibration model. Scaling might be a factor at play in this result. Due to the nature of the spectra, single-beam spectra carry much more variation in the spectra associated with the residual concentrations with much higher levels of fluctuation. The residual model based on single-beam data therefore appears to be derived from amplified data, which later might translate to higher sensitivity when applied to the diagnostics related to classification.

The amplified PLS-aided SVR method proposed in this work took advantage of the scalability of the PLS scores to address problems encountered in modeling the absorbance spectral residuals. From a broader point of view, it can also be viewed as a simple version of spectral data simulation. Different from what has been seen as simulating “synthetic” data from raw spectra, this work showed that it is also possible to manipulate the multivariate spectral information imbedded in the PLS scores. This opens a door towards a new type of spectral simulation for quantitative analysis. For the goal of this project, to extend the capacity of absorbance residual modeling, models built based on mathematically synthetic spectral data, based on raw spectra or spectral multivariate factors, deliberately designed to carry stronger concentration-spectrum correlation, might be considered in the future.

Chapter 8

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

This dissertation focused on the development of multivariate data analysis methods for automated detection and quantification with applications to passive infrared remote sensing, airborne gamma-ray spectroscopy, and near-infrared spectroscopy of aqueous solutions. Through this work, a set of novel methodology was developed that incorporated elements of signal processing, supervised pattern recognition, multivariate modeling, and statistical process control.

In Chapter 4, an automated detection of ambient ammonia vapor from airborne passive FT-IR spectrometry was developed. This application is complicated by the lack of stable infrared backgrounds for use in spectral processing and the difficulty encountered in assembling training data in which desired analytes are present. To overcome these challenges, direct analysis of interferogram data was used as a convenient way to separate the variable infrared background signature from the analyte-specific components of the data. Spectral simulation techniques produced mathematically synthesized interferograms containing the ammonia signature for use in training classification models. These data simulated the interferograms that would be collected in field remote sensing measurements made from an aircraft platform.

Development of an automated ammonia classifier combined windowing the interferogram to isolate an optimal segment with bandpass digital filtering to isolate an optimal set of spectral frequencies. Joint optimization of the digital filtering and segment parameters helped to extract the features most distinguishable in terms of binary classification (i.e., ammonia-active vs. ammonia-inactive). The PLDA method was used to define the optimal segment and filtering parameters, and final classification models were developed with an SVM classifier. Formal design of experiments methodology and ANOVA techniques were used to identify that a 120-point interferogram segment starting from point 88 was optimal. These

analyses also established that segment selection played a bigger role in feature extraction than digital filtering.

The optimized SVM classifier was tested by applying it to four prediction sets collected during emergency response missions. Techniques from statistical process control were used to develop an alternate implementation of the SVM classifier in which analyte detections were treated analogously to disruptions of a process in which the SVM score was the process variable being monitored. A classification rule was developed using control charts combined with a requirement that two consecutive alarms must occur before an alarm condition is established. Using this rule, results from the four prediction sets ranged from 13 to 27 % missed detections and 0.16 to 1.8 % false detections. These results validated the overall analysis approach and established a general methodology for the development of automated classifiers for the detection of VOCs from passive infrared data collected from the air.

In Chapter 5, the methodology developed in Chapter 4 was extended to investigate the relationship between the concentration profile of the synthetic patterns in the training set and the resulting performance of the computed SVM classifiers. Training sets were evaluated for the degree of discrimination between the active and inactive data classes by use of PCA and KNN classification. The most promising training sets were subsequently used to develop SVM classifiers, and the resulting classifiers were applied to the same four prediction sets used previously in the work described in Chapter 4.

Comparisons of classification performance over the prediction sets revealed that the classifiers developed in Chapter 5 were more sensitive to the identification of the ammonia-active patterns, while also somewhat more prone to false detections. Estimation of the limit of detection for the selected classifiers indicated that ammonia vapor of concentrations as low as 5

ppm-m in emission and 6 ppm-m in absorption could be identified. This was about 2-3 ppm-m above the lowest ammonia concentration found in the training set.

In Chapter 6, the same concepts incorporating spectral simulation, signal processing and pattern recognition developed in Chapter 4 were adopted to develop automated classifiers for the detection of the radioisotope ^{137}Cs from gamma-ray spectra collected from the same aircraft platform used in the passive FT-IR measurements.

The particular challenge in the detection of radioisotopes from airborne gamma-ray spectra is that the spectral intensities and bandshape change as a consequence of inelastic scattering between gamma rays and atmospheric species. This effect is altitude-based, and thus complicates the recognition of the spectral signature of a target species.

Just as in the FT-IR remote sensing case, the acquisition of training data with analyte signatures is very difficult. For this reason, a spectral simulation strategy was also employed in this work to synthesize training data for use in computing SVM classifiers. Bandpass digital filtering and spectral segment selection were used as in the FT-IR remote sensing work to extract the features most relevant to the identification of the ^{137}Cs signature. The developed classifiers were applied to the identification of ^{137}Cs in three predication data sets acquired during field surveys. Excellent classification performance was obtained from two out of the three prediction sets, particularly in light of the overall low signal-to-noise ratios of the collected gamma-ray spectra. The sub-optimal results obtained with the third data set are not completely understood and required further investigation.

In Chapter 7, pattern recognition methods were again employed, but here the application was to aid in the quantitative determination of glucose from near-infrared spectra of a simulated biological matrix. The specific application of interest was the need to assess and maintain the

performance of a quantitative calibration model for predicting glucose concentrations from near-infrared spectra. Once developed through the use of a set of calibration data, such models tend to drift or go out of specification over time due to instrumental variations, changes in sample conditions, or other unknown sources of variance. Detecting when the calibration model has gone out of specification is not straightforward, however.

The methodology developed in this work was based upon the assumption that the glucose information present in the spectrum corresponding to a given sample is a constant, and the function of the calibration model is to extract that information from whatever background components are present. If the calibration model is no longer functioning properly, a greater than expected amount of information will remain in the spectrum after processing by the calibration model (i.e., the residual spectrum will be larger in magnitude than expected). One of the goals of the research described in Chapter 7 was to assess whether the remaining concentration information could be extracted from the residual spectrum and thereby used to correct the concentration prediction obtained with the original model.

Several modeling techniques, such as PLS, SVR, PLS score fed SVR and amplified PLS fed SVR, were employed to retrieve residual concentrations of glucose from residual spectra. Residual modeling could predict residual concentrations accurately with SEP values as low as 0.5 mM. These retrieved residual concentrations were used as a means to evaluate the performance of a calibration model over time. Control charts, based upon retrieved residual concentrations, either with two or three times the standard deviation about the mean as the cutoffs, provided a controlled updating approach for the established calibration model.

A common theme in all of the projects described in this dissertation was the use of SVMs. The SVM is a potent and versatile supervised pattern recognition algorithm. Its utility in

multidimensional pattern recognition and nonlinear regression has been verified in this dissertation. In Chapter 4 through Chapter 6, SVMs were applied to resolve complex binary classification problems from remote sensing data, based upon either FTIR interferograms or gamma-ray spectra.

One common challenge across these projects was to assemble the training set with patterns representing each class, especially the analyte-active class. It is difficult to obtain a sufficient number of analyte-active patterns from actual field measurements. Spectral simulation, essentially superimposing pure-component signatures of a target anlayte onto a background spectral profile, was shown to provide a supply of patterns sufficient for representing the features and variation of the analyte-active class. Meanwhile, signal processing techniques, including digital filtering and segment selection, suppressed irrelevant information contained in the patterns, thereby facilitating the extraction of the information most relevant to the target analyte.

The SVM is not limited to classification problems. Quantitative information can be obtained through SVM regression. In principle, nonlinear SVM regression outperforms the linear regression methods, such as PLS, because of its greater flexibility in fitting the observations comprising the dependent variable. The complexity involved in multidimensional nonlinear regression can be reduced when PLS scores, representing patterns effectively with fewer variables, replace raw spectral segments as input patterns to SVM regression. This combination of PLS and SVM was demonstrated in Chapter 7.

Though potent, SVMs have shortcomings. High computational capacity is required to conduct SVM modeling because of the need to optimize two parameters associated with the SVM architecture. If other parameters associated with spectral processing (e.g., digital filtering or segment selection) require optimization, the overall computational burden can be extreme. In

the dissertation work, other supervised pattern recognition methods such as PLDA and KNN, as well as unsupervised pattern recognition methods such as PCA were used prior to employing SVMs in order to reduce the computational load.

Along with the use of SVMs, the technique of control charts was implemented across all the work in this dissertation to increase the performance of the SVM models when applied outside of the training data. Because of inadequate representation in the training data of all signatures that might be ultimately be encountered, biased SVM models were produced that resulted in too many false detections when applied to prospective data. Implementing a classification rule using control charts allowed alarm decisions to be tuned to the specific data set being classified and thereby improved the overall performance of the model.

REFERENCES

1. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeeke, J., *Handbook of Chemometrics and Qualimetrics: Part A.* Elsevier: Amsterdam, 1997.
2. Kowalski, B. R., Chemometrics - view and propositions. *Journal of Chemical Information and Computer Sciences* **1975,** *15* (4), 201-203.
3. Brereton, R. G., Chemometrics: Application of Mathematics and Statistics to Laboratory Systems. Ellis Horwood: New York, 1990.
4. Hopke, P. K., The evolution of chemometrics. *Analytica Chimica Acta* **2003,** *500* (1-2), 365-377.
5. Brown, S. D.; Sum, S. T.; Despagne, F.; Lavine, B. K., Chemometrics. *Analytical Chemistry* **1996,** *68* (12), R21-R61.
6. Wold, S.; Sjostrom, M., Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems* **1998,** *44* (1-2), 3-14.
7. Lavine, B.; Workman, J. J., Chemometrics. *Analytical Chemistry* **2004,** *76* (12), 3365-3371.
8. De Braekeleer, K.; de Juan, A.; Massart, D. L., Purity assessment and resolution of tetracycline hydrochloride samples analysed using high-performance liquid chromatography with diode array detection. *Journal of Chromatography A* **1999,** *832* (1-2), 67-86.
9. Abdollahi, H.; Tauler, R., Uniqueness and rotation ambiguities in Multivariate Curve Resolution methods. *Chemometrics and Intelligent Laboratory Systems* **2011,** *108* (2), 100-111.
10. Lopes, J. A.; Costa, P. F.; Alves, T. P.; Menezes, J. C., Chemometrics in bioprocess engineering: process analytical technology (PAT) applications. *Chemometrics and Intelligent Laboratory Systems* **2004,** *74* (2), 269-275.
11. Reich, G., Near-infrared spectroscopy and imaging: Basic principles and pharmaceutical applications. *Advanced Drug Delivery Reviews* **2005,** *57* (8), 1109-1143.
12. Undey, C.; Low, D.; Menezes, J. C.; Koch, M., PAT Applied in Biopharmaceutical Process Development And Manufacturing: An Enabling Tool for Quality-by-Design CRC Press: 2011.
13. Mariey, L.; Signolle, J. P.; Amiel, C.; Travert, J., Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics. *Vibrational Spectroscopy* **2001,** *26* (2), 151-159.
14. Lopez-Diez, E. C.; Goodacre, R., Characterization of microorganisms using UV resonance Raman spectroscopy and chemometrics. *Analytical Chemistry* **2004,** *76* (3), 585-591.
15. Nielsen, N. P. V.; Carstensen, J. M.; Smedsgaard, J., Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **1998,** *805* (1-2), 17-35.
16. Lenz, E. M.; Bright, J.; Knight, R.; Wilson, I. D.; Major, H., Cyclosporin A-induced changes in endogenous meta-bolites in rat urine: a metabonomic investigation using high field H-1 NMR spectroscopy, HPLC-TOF/MS and chemometrics. *Journal of Pharmaceutical and Biomedical Analysis* **2004,** *35* (3), 599-608.
17. Workman, J.; Koch, M.; Veltkamp, D., Process analytical chemistry. *Analytical Chemistry* **2007,** *79* (12), 4345-4363.
18. Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J., Visualization of GC/TOF-MS-based metabolomics data for

identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry* **2008,** *80* (1), 115-122.
19. Hall, J. W.; McNeil, B.; Rollins, M. J.; Draper, I.; Thompson, B. G.; Macaloney, G., Near-infrared spectroscopic determination of acetate, ammonium, biomass, and glycerol in an industrial Escherichia coli fermentation. *Applied Spectroscopy* **1996,** *50* (1), 102-108.
20. Dyrby, M.; Engelsen, S. B.; Norgaard, L.; Bruhn, M.; Lundsberg-Nielsen, L., Chemometric quantitation of the active substance (containing C N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra. *Applied Spectroscopy* **2002,** *56* (5), 579-585.
21. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics* **2002,** *16* (3), 119-128.
22. Thissen, U.; Pepers, M.; Ustun, B.; Melssen, W. J.; Buydens, L. M. C., Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems* **2004,** *73* (2), 169-179.
23. Rantanen, J.; Wikstrom, H.; Turner, R.; Taylor, L. S., Use of in-line near-infrared spectroscopy in combination with chemometrics for improved understanding of pharmaceutical processes. *Analytical Chemistry* **2005,** *77* (2), 556-563.
24. Skoog, D. A.; Holler, J. F.; Crouch, S. R., *Principles of Instrumental Analysis*. 6 ed.; Thomson Brooks: 2007.
25. A., B. D.; Ciurczak, E. W., *Handbook of Near-Infrared Analysis*. Marcel Dekker: New York, 2001.
26. Griffiths, P. R.; de Haseth, J. A., *Fourier Transform Infrared Spectrometry*. 2 ed.; John Wiley & Son: Hoboken, New Jersey, 2007.
27. Tsoulfanidis, N., *Measurement and Detection of Radiation*. 2 ed.; Taylor & Francis: Washington, DC, 1995.
28. Novello, D. P.; Martineau, R. J., *The Clear Air Act Handbook*. 2 ed.; American Bar Association: New York, 2005.
29. Communities, C. o. t. E., Communications from the Commission to th Council and the Eauropean Parliament - Thematic Strategy on Air Pollution. Brussels, 2005.
30. Carter, R. E.; Thomas, M. J.; Marotz, G. A.; Lane, D. D.; Hudson, J. L., Compound detection and concentration estimation by open-path Fourier-transform infrared spectrometry and canisters under controlled field conditions. *Environmental Science & Technology* **1992,** *26* (11), 2175-2181.
31. Lieberzeit, P. A.; Dickert, F. L., Chemosensors in environmental monitoring: challenges in ruggedness and selectivity. *Analytical and Bioanalytical Chemistry* **2009,** *393* (2), 467-472.
32. El-Roz, M.; Kus, M.; Cool, P.; Thibault-Starzyk, F., New Operando IR Technique to Study the Photocatalytic Activity and Selectivity of TiO2 Nanotubes in Air Purification: Influence of Temperature, UV Intensity, and VOC Concentration. *Journal of Physical Chemistry C* **2012,** *116* (24), 13252-13263.
33. Beer, R., *Remote Sensing by Fourier Transform Spectrometry*. John Wiley and Sons: New York, 1992; Vol. 120.
34. Beil, A. R. D. R. H. G. M., Remote sensing of atmospheric pollution by passive FTIR spectrometry. EUROPTO '98 (Barcelona): Spectroscopic Atmospheric Environmental Monitoring Techniques. Proceedings of SPIE - The International Society for Optical Engineering **1998,** 3493, 12.

35. Ben-David, A., Remote detection of biological aerosols at a distance of 3 km with a passive Fourier transform infrared (FTIR) sensor. *Optics Express* **2003,** *11* (5), 418-429.
36. Robinson, M. R.; Eaton, R. P.; Haaland, D. M.; Koepp, G. W.; Thomas, E. V.; Stallard, B. R.; Robinson, P. L., Noninvasive glucose monitoring in diabetic patients - a preliminary evaluation. *Clinical Chemistry* **1992,** *38* (9), 1618-1622.
37. Hazen, K. H.; Arnold, M. A.; Small, G. W., Temperature-insensitive near-infrared spectroscopic measurement of glucose in aqueous solution. *Applied Spectroscopy* **1994,** *48* (4), 477-483.
38. Hazen, K. H.; Arnold, M. A.; Small, G. W., Measurement of glucose and other analytes in undiluted human serum with near-infrared transmission spectroscopy. *Analytica Chimica Acta* **1998,** *371* (2-3), 255-267.
39. Mittermayr, C. R.; Tan, H. W.; Brown, S. D., Robust calibration with respect to background variation. *Applied Spectroscopy* **2001,** *55* (7), 827-833.
40. Wehlburg, C. M.; Haaland, D. M.; Melgaard, D. K.; Martin, L. E., New hybrid algorithm for maintaining multivariate quantitative calibrations of a near-infrared spectrometer. *Applied Spectroscopy* **2002,** *56* (5), 605-614.
41. Kramer, K. E.; Small, G. W., Blank augmentation protocol for improving the robustness of multivariate calibrations. *Applied Spectroscopy* **2007,** *61* (5), 497-506.
42. Ingle, J. D.; Crouch, S. R., *Spectrochemical Analysis*. Prentice-Hall: New Jersy, 1988.
43. Stuart, B. H., *Infrared Spectroscopy: Fundamentals and Applications*. John Wiley & Sons: Chichester, 2004.
44. Smith, B. C., *Fundamentals of Fourier Transform Infrared Spectroscopy*. 2 ed.; CRC Press: Boca Raton, 2011.
45. Shaffer, R. E.; Combs, R. J. *NRL Memorandum Report ECBC-TR-084*; 2000.
46. Tarumi, T. Data Analysis Stretegies for Airborne Remote Sensing of Volatile Organic Compounds Using Passive Fourier Transform Infrared Spectrometry. Ph.D. Thesis. 2003.
47. Minty, B. R. S., Fundamentals of Airborne Gamma-ray Spectrometry. *AGSO Journal of Australian Geology & Geophysics* **1997,** *17* (2), 12.
48. Cardarelli, J. I.; Thomas, M.; Curry, T., Environmental Protection Agency (EPA) Airborne Gamma Spectrometry System for Environmental and Emergency Response Surveys. *Proc. of SPIE* **2010,** *7812* (781205-10).
49. Small, G. W.; Harms, A. C.; Kroutil, R. T.; Ditillo, J. T.; Loerop, W. R., Design of optimized finite impulse response digital filters for use with passive Fourier transform infrared interferograms. *Anal. Chem.* **1990,** *62* (17), 1768-1777.
50. Mattu, M. J.; Small, G. W., Quantitative analysis of bandpass filtered Fourier transform infrared interferogram. *Anal. Chem.* **1995,** *67* (13), 2269-2278.
51. Cingo, N. A.; Small, G. W.; Arnold, M. A., Determination of glucose in a synthetic biological matrix with decimated time-domain filtered near-infrared interferogram data. *Vib. Spectrosc.* **2000,** *23* (1), 103-117.
52. Sulub, Y.; Small, G. W., Spectral simulation methodology for calibration transfer of near-infrared spectra. *Appl. Spectrosc.* **2007,** *61* (4), 406-413.
53. Hamming, R. W., *Digital filters*. 3 ed.; Prentice-Hall: 1995.
54. Chen, C.-T., Digital signal processing: spectral somputatoin and filter design. Oxford University Press: 2000.
55. Mathworks, Signal Processing Toolbox for Use with Matlab: User's guide. 5 ed.; MathWorks: Natick, 2000.

56. Williams, A.; Taylor, F., *Electronic Filter Design Handbook*. McGraw-Hill Professional: 2006.
57. Oppenheim, A.; Shafer, R.; Buck, J., *Discrete-Time Signal Processing*. Prentice-Hall Press: 1989.
58. Chen, N.; Lu, W.; Yang, J.; Li, G., *Support vector machine in chemistry*. World Scientific: 2004.
59. Bishop, C. M., Pattern recognition and machine learning. Springer: 2006.
60. Wehrens, R., *Chemometrics with R*. Springer: 2011.
61. Cristianini, N.; sHAWE-Taylor, J., An introduction to support vector machines and other kernel-based leaning methods. Cambridge University Press: 2000.
62. Tou, J. T.; Gonzalez, R. C., *Pattern recogntion principles*. Addison-Wesley: Reading, MA, 1974.
63. Delgoda, R.; Pulfer, J. D., Application of pattern recognition techniques to mass spectrometric data for sequencing c-terminal peptide residue series. *J. Chem. Inf. Comput. Sci.* **1993,** *33* (3), 332-337.
64. Decell, H. P.; Odell, P. L.; Coberly, W. A., *Pattern recogntion*. 1981.
65. Martens, H.; Martens, M., *Multivariate analysis of quality: an introduction*. John Siley & Sons: Chichester, 2001.
66. Wold, S.; Sjostrom, M.; Eriksson, L., PLS-regression: a basic tool of chemometrics. *Chemometrics Intell. Lab. Syst.* **2001,** *58* (2), 109-130.
67. Kowalski, B. R.; Bender, C. F., Pattern recogntion - powerful approach to interpreting chemical data. *Journal of the American Chemical Society* **1972,** *94* (16), 5632-&.
68. Derde, M. P.; Buydens, L.; Guns, C.; Massart, D. L.; Hopke, P. K., Comparison of rule-building expert systems with pattern recognition for the classification of analytical data. *Anal. Chem.* **1987,** *59* (14), 1868-1871.
69. Esbensen, K. H.; Guyot, D.; Westad, F.; Houmøller, L. P., *Multivariate data analysis - in practice*. 5 ed.; CAMO software: 2010.
70. Bangalore, A. S.; Small, G. W.; Combs, R. J.; Knapp, R. B.; Kroutil, R. T.; Traynor, C. A.; Ko, J. D., Automated detection of trichloroethylene by Fourier transform infrared remote sensing measurements. *Analytical Chemistry* **1997,** *69* (2), 118-129.
71. Chaffin, C. T.; Marshall, T. L.; Chaffin, N. C., Passive FTIR remote sensing of smokestack emissions. *Field Analytical Chemistry and Technology* **1999,** *3* (2), 111-115.
72. Goff, F.; Love, S. P.; Warren, R. G.; Counce, D.; Obenholzner, J.; Siebe, C.; Schmidt, S. C., Passive infrared remote sensing evidence for large, intermittent CO2 emissions at Popocatepetl volcano, Mexico. *Chemical Geology* **2001,** *177* (1-2), 133-156.
73. Hammer, C. L.; Small, G. W.; Combs, R. J.; Knapp, R. B.; Kroutil, R. T., Artificial neural networks for the automated detection of trichloroethylene by passive Fourier transform infrared spectrometry. *Analytical Chemistry* **2000,** *72* (7), 1680-1689.
74. Tarumi, T.; Small, G. W.; Combs, R. J.; Kroutil, R. T., Remote detection of heated ethanol plumes by airborne passive Fourier transform infrared spectrometry. *Applied Spectroscopy* **2003,** *57* (11), 1432-1441.
75. Wabomba, M. J.; Small, G. W., Robust classifier for the automated detection of ammonia in heated plumes by passive Fourier transform infrared spectrometry. *Analytical Chemistry* **2003,** *75* (9), 2018-2026.
76. Wan, B.; Small, G. W., Airborne passive Fourier transform infrared remote sensing of methanol vapor from industrial emissions. *Analyst* **2008,** *133* (12), 1776-1784.

77. Zhang, L.; Small, G. W., Automated detection of chemical vapors by pattern recognition analysis of passive multispectral infrared remote sensing imaging data. *Applied Spectroscopy* **2002,** *56* (8), 1082-1093.
78. Carpenter, S. E.; Small, G. W., Selection of optimum training sets for use in pattern recognition analysis of chemical data. *Analytica Chimica Acta* **1991,** *249* (2), 305-321.
79. Herzberg, G., Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules. Van Nostrand Reinhold: New York, 1945.
80. Belousov, A. I.; Verzakov, S. A.; von Frese, J., A flexible classification approach with optimal generalisation performance: support vector machines. *Chemometrics and Intelligent Laboratory Systems* **2002,** *64* (1), 15-25.
81. Sharpe, S. W.; Johnson, T. J.; Sams, R. L.; Chu, P. M.; Rhoderick, G. C.; Johnson, P. A., Gas-phase databases for quantitative infrared spectroscopy. *Applied Spectroscopy* **2004,** *58* (12), 1452-1461.
82. Wan, B.; Small, G. W., Synthetic training sets for the development of discriminant functions for the detection of volatile organic compounds from passive infrared remote sensing data. *Analyst* **2011,** *136* (2), 309-316.
83. Wabomba, M. J.; Small, G. W., Design protocols for time-dependent finite impulse response digital filters based on regression analysis of Fourier transform infrared interferograms. *Chemometrics and Intelligent Laboratory Systems* **2003,** *69* (1-2), 103-121.
84. Harig, R., Passive remote sensing of pollutant clouds by Fourier-transform infrared spectrometry: signal-to-noise ratio as a function of spectral resolution. *Applied Optics* **2004,** *43* (23), 4603-4610.
85. Szostek, B.; Trojanowicz, M., Real-time digital filters for signal processing in flow injection analysis: general considerations and simulation study. *Analytica Chimica Acta* **1992,** *261* (1-2), 509-519.
86. Tarumi, T.; Small, G. W.; Combs, R. J.; Kroutil, R. T., Infinite impulse response filters for direct analysis of interferogram data from airborne passive Fourier transform infrared spectrometry. *Vibrational Spectroscopy* **2005,** *37* (1), 39-52.
87. NIST/SEMATECH, e-Handbook of Statistical Methods. http://www.itl.nist.gov/div898/handbook/, 2012.
88. Sawyer, G. M.; Oppenheimer, C.; Tsanev, V. I.; Yirgu, G., Magmatic degassing at Erta 'Ale volcano, Ethiopia. *J. Volcanol. Geotherm. Res.* **2008,** *178* (4), 837-846.
89. Childers, J. W.; Thompson, E. L.; Harris, D. B.; Kirchgessner, D. A.; Clayton, M.; Natschke, D. F.; Phillips, W. J., Multi-pollutant concentration measurements around a concentrated swine production facility using open-path FTIR spectrometry. *Atmos. Environ.* **2001,** *35* (11), 1923-1936.
90. Harig, R.; Matz, G., Toxic cloud imaging by infrared spectrometry: A scanning FTIR system for identification and visualization. *Field Anal. Chem. Technol.* **2001,** *5* (1-2), 75-90.
91. Shao, L. M.; Liu, B. X.; Griffiths, P. R.; Leytem, A. B., Using Multiple Calibration Sets to Improve the Quantitative Accuracy of Partial Least Squares (PLS) Regression on Open-Path Fourier Transform Infrared (OP/FT-IR) Spectra of Ammonia over Wide Concentration Ranges. *Applied Spectroscopy* **2011,** *65* (7), 820-824.
92. Ballard, J.; Remedios, J. J.; Roscoe, H. K., The Effect of Sample Emission on Measurements of Spectral Parameters Using a Fourier-Transform Absorption Spectrometer. *Journal of Quantitative Spectroscopy & Radiative Transfer* **1992,** *48* (5-6), 733-741.

93. Todd, L. A.; Ramanathan, M.; Mottus, K.; Katz, R.; Dodson, A.; Mihlan, G., Measuring chemical emissions using open-path Fourier transform infrared (OP-FTIR) spectroscopy and computer-assisted tomography. *Atmos. Environ.* **2001,** *35* (11), 1937-1947.
94. Holland, S. K.; Krauss, R. H.; Laufer, G., Effect of temperature on passive remote sensing of chemicals by differential absorption radiometry. *Optical Engineering* **2005,** *44* (10).
95. Shaffer, R. E.; Combs, R. J., Comparison of spectral and interferogram processing methods using simulated passive Fourier transform infrared remote sensing data. *Applied Spectroscopy* **2001,** *55* (10), 1404-1413.
96. Sulub, Y.; Small, G. W., Quantitative determination of ethanol in heated plumes by passive Fourier transform infrared remote sensing measurements. *Analyst* **2007,** *132* (4), 330-337.
97. Martens, H.; Naes, T., *Multivariate Calibration*. Wiley & Sons: 1992.
98. Brereton, R. G., Chemometrics: Data Analysis for th Laboratory and Chemical Plant. Wiley & Sons: Chichester, 2003.
99. Burgard, D. R.; Kuznicki, J. T., *Chemometrics: Chemical and Sensory Data*. CRC Press: 1990.
100. Friedman, J. H., On bias, variance, 0/1 - Loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.* **1997,** *1* (1), 55-77.
101. Valentini, G.; Dietterich, T. G., Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. *J. Mach. Learn. Res.* **2004,** *5*, 725-775.
102. Platt, J. C., Probabilities for SV Machines. In *Advances in Large Margin Classifiers*, Smola, A.; Bartlett, P.; Schölkopf, B.; Schuurmans, D., Eds. MIT Press: 1999; pp 61-74.
103. Sollich, P., Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Mach. Learn.* **2002,** *46* (1-3), 21-52.
104. Bristow, Q., Airborne gamma-ray spectrometry in uranium exploration: principles and current practice. *International Journal of Applied Radiation and Isotopes* **1983,** *34* (1), 199-&.
105. Darnley, A. G., The development of airborne gamma-ray spectrometry: case study in technological innovation and acceptance. *Nuclear Geophysics* **1991,** *5* (4), 377-402.
106. Beamish, D.; White, J. C., A radiometric airborne geophysical survey of the Isle of Wight. *Proceedings of the Geologists Association* **2011,** *122* (5), 787-799.
107. Wilford, J., A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. *Geoderma* **2012,** *183*, 124-142.
108. Aoyama, M.; Hirose, K., The temporal and spatial variation of Cs-137 concentration in the western north pacific and its marginal seas during the period from 1979 to 1988. *Journal of Environmental Radioactivity* **1995,** *29* (1), 57-74.
109. Tyler, A. N., High accuracy in situ radiometric mapping. *Journal of Environmental Radioactivity* **2004,** *72* (1-2), 195-202.
110. Watanabe, T.; Tsuchiya, N.; Oura, Y.; Ebihara, M.; Inoue, C.; Hirano, N.; Yamada, R.; Yamasaki, S.; Okamoto, A.; Nara, F. W.; Nunohara, K., Distribution of artificial radionuclides (Ag-110m, Te-129m, Cs-134, Cs-137) in surface soils from Miyagi Prefecture, northeast Japan, following the 2011 Fukushima Dai-ichi nuclear power plant accident. *Geochemical Journal* **2012,** *46* (4), 279-285.
111. Duval, J. S., High sensitivity gamma-ray spectrometry - state of art and trial application of factor factor analysis. *Geophysics* **1977,** *42* (3), 549-559.

112. Ranjbar, H.; Hassanzadeh, H.; Torabi, M.; Ilaghi, O., Integration and analysis of airborne geophysical data of the Darrehzar area, Kerman Province, Iran, using principal component analysis. *Journal of Applied Geophysics* **2001,** *48* (1), 33-41.
113. Dragovic, S.; Onjia, A., Classification of soil samples according to their geographic origin using gamma-ray spectrometry and principal component analysis. *Journal of Environmental Radioactivity* **2006,** *89* (2), 150-158.
114. Schulze, G.; Jirasek, A.; Yu, M. M. L.; Lim, A.; Turner, R. F. B.; Blades, M. W., Investigation of selected baseline removal techniques as candidates for automated implementation. *Applied Spectroscopy* **2005,** *59* (5), 545-574.
115. Sullivan, C. J.; Martinez, M. E.; Garner, S. E., Wavelet analysis of sodium iodide spectra. *Ieee Transactions on Nuclear Science* **2006,** *53* (5), 2916-2922.
116. Tyler, A. N., Monitoring anthropogenic radioactivity in salt marsh environments through in situ gamma-ray spectrometry. *Journal of Environmental Radioactivity* **1999,** *45* (3), 235-252.
117. Brown, W. M.; Gedeon, T. D.; Groves, D. I.; Barnes, R. G., Artificial neural networks: a new method for mineral prospectivity mapping. *Australian Journal of Earth Sciences* **2000,** *47* (4), 757-770.
118. Gan, T. Y.; Kalinga, O.; Singh, P., Comparison of snow water equivalent retrieved from SSM/I passive microwave data using artificial neural network, projection pursuit and nonlinear regressions. *Remote Sensing of Environment* **2009,** *113* (5), 919-927.
119. Ohm, S.; van Eldik, C.; Egberts, K., gamma/hadron separation in very-high-energy gamma-ray astronomy using a multivariate analysis method. *Astroparticle Physics* **2009,** *31* (5), 383-391.
120. Lacoste, M.; Lemercier, B.; Walter, C., Regional mapping of soil parent material by machine learning based on point data. *Geomorphology* **2011,** *133* (1-2), 90-99.
121. Lemercier, B.; Lacoste, M.; Loum, M.; Walter, C., Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. *Geoderma* **2012,** *171*, 75-84.
122. Andersson, M.; Folestad, S.; Gottfries, J.; Johansson, M. O.; Josefson, M.; Wahlund, K. G., Quantitative analysis of film coating in a fluidized bed process by in line NIR spectrometry and multivariate batch calibration. *Analytical Chemistry* **2000,** *72* (9), 2099-2108.
123. Larrechi, M. S.; Callao, M. P., Strategy for introducing NIR spectroscopy and multivariate calibration techniques in industry. *Trac-Trends in Analytical Chemistry* **2003,** *22* (10), 634-640.
124. Lafrance, D.; Lands, L. C.; Burns, D. H., In vivo lactate measurement in human tissue by near-infrared diffuse reflectance spectroscopy. *Vibrational Spectroscopy* **2004,** *36* (2), 195-202.
125. Brearley, A. M., Applications of Near-Infrared Spectroscopy (NIR) in the Chemical Industry. 2005; p 392-423.
126. Blanco, M.; Villarroya, I., NIR spectroscopy: a rapid-response analytical tool. *Trac-Trends in Analytical Chemistry* **2002,** *21* (4), 240-250.
127. Amerov, A. K.; Chen, J.; Small, G. W.; Arnold, M. A., Scattering and absorption effects in the determination of glucose in whole blood by near-infrared spectroscopy. *Analytical Chemistry* **2005,** *77* (14), 4587-4594.
128. Tromberg, B. J.; Shah, N.; Lanning, R.; Cerussi, A.; Espinoza, J.; Pham, T.; Svaasand, L.; Butler, J., Non-invasive in vivo characterization of breast tumors using photon migration spectroscopy. *Neoplasia* **2000,** *2* (1-2), 26-40.

129. Maruo, K.; Tsurugi, M.; Tamura, M.; Ozaki, Y., In vivo noninvasive measurement of blood glucose by near-infrared diffuse-reflectance spectroscopy. *Applied Spectroscopy* **2003,** *57* (10), 1236-1244.
130. Blanco, M.; Peguero, A., Analysis of pharmaceuticals by NIR spectroscopy without a reference method. *Trac-Trends in Analytical Chemistry* **2010,** *29* (10), 1127-1136.
131. Sulub, Y.; Small, G. W., Spectral Simulation Protocol for Extending the Lifetime of Near-Infrared Multivariate Calibrations. *Analytical Chemistry* **2009,** *81* (3), 1208-1216.
132. Bouveresse, E.; Hartmann, C.; Massart, D. L.; Last, I. R.; Prebble, K. A., Standardization of near-infrared spectrometric instruments. *Analytical Chemistry* **1996,** *68* (6), 982-990.
133. Wang, Y. D.; Lysaght, M. J.; Kowalski, B. R., Improvement of multivariate calibration through instrument standardization. *Analytical Chemistry* **1992,** *64* (5), 562-564.
134. Woody, N. A.; Feudale, R. N.; Myles, A. J.; Brown, S. D., Transfer of multivariate calibrations between four near-infrared spectrometers using orthogonal signal correction. *Analytical Chemistry* **2004,** *76* (9), 2595-2600.
135. Greensill, C. V.; Wolfs, P. J.; Spiegelman, C. H.; Walsh, K. B., Calibration transfer between PDA-based NIR spectrometers in the NIR assessment of melon soluble solids content. *Applied Spectroscopy* **2001,** *55* (5), 647-653.
136. Barman, I.; Kong, C.-R.; Dingari, N. C.; Dasari, R. R.; Feld, M. S., Development of Robust Calibration Models Using Support Vector Machines for Spectroscopic Monitoring of Blood Glucose. *Analytical Chemistry* **2010,** *82* (23), 9719-9726.
137. Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V., Application of LS-SVM to non-linear phenomena in NIR spectroscopy: development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems* **2004,** *71* (2), 141-150.
138. Barman, I.; Dingari, N. C.; Singh, G. P.; Soares, J. S.; Dasari, R. R.; Smulko, J. M., Investigation of Noise-Induced Instabilities in Quantitative Biological Spectroscopy and Its Implications for Noninvasive Glucose Monitoring. *Analytical Chemistry* **2012,** *84* (19), 8149-8156.
139. Fang, K. T., *Uniform Design and Uniform Design Tables*. Science Press: Beijing, 1994.
140. Kennard, R. W.; Stone, L. A., Computer aided design of experiments. *Technometrics* **1969,** *11* (1), 137-&.
141. Amerov, A. K.; Chen, J.; Arnold, M. A., Molar absorptivities of glucose and other biological molecules in aqueous solutions over the first overtone and combination regions of the near-infrared spectrum. *Applied Spectroscopy* **2004,** *58* (10), 1195-1204.
142. Kramer, K. E.; Small, G. W., Digital Filtering and Model Updating Methods for Improving the Robustness of Near-Infrared Multivariate Calibrations. *Applied Spectroscopy* **2009,** *63* (2), 246-255.