Investigation of soil reflectances for detecting hydrocarbon pipeline leakages

Sharon Leigh Cundill
March, 2008
<table>
<thead>
<tr>
<th><strong>Course Title:</strong></th>
<th>Geo-Information Science and Earth Observation for Environmental Modelling and Management</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level:</strong></td>
<td>Master of Science (Msc)</td>
</tr>
<tr>
<td><strong>Course Duration:</strong></td>
<td>September 2006 - March 2008</td>
</tr>
</tbody>
</table>
| **Consortium partners:** | University of Southampton (UK)  
Lund University (Sweden)  
University of Warsaw (Poland)  
International Institute for Geo-Information Science and Earth Observation (ITC) (The Netherlands) |
| **GEM thesis number:** | 2006-28                                                                         |
Investigation of soil reflectances for detecting hydrocarbon pipeline leakages

by

Sharon Leigh Cundill

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation for Environmental Modelling and Management

Thesis Assessment Board

Chair

Prof. Dr. F.D. van der Meer

External Examiner

Prof. T.P. Dawson

Member

Dr. M. Schlerf

First Supervisor

Dr. M. van der Meijde

Second Supervisor

Dr. H.M.A. van der Werff

International Institute for Geo-Information Science and Earth Observation

Enschede, The Netherlands
Disclaimer

This document describes work undertaken as part of a programme of study at the International Institute for Geo-information Science and Earth Observation. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.
Abstract

There is a need for a non-destructive, cost- and time-effective technique for detecting leakages from hydrocarbon pipelines. This study investigated soil reflectances with regard to benzene condensate pollution which leaked from a subterranean gas pipeline in the northern part of the Netherlands. GER3700 visible/near infrared spectra, FTIR thermal infrared spectra and field gamma ray spectrometry values were measured for a limited number of soil samples from both polluted and non-polluted sites. Some auger soil samples were also included for depths of 0.5m and 0.9m. The data was assessed visually and statistically analysed with box plots and logistic regression. No distinct features were apparent in any of the datasets that consistently differentiated between the polluted and clean samples. Bands 1337.89nm and 1362.93nm could possibly aid in the differentiation of polluted soils from clean, but they require further investigation. The samples from depth showed possible gradient patterns in reflectance with an increase in depth, which include shifts in colour and albedo. The results require further investigation with larger datasets.

Keywords: hydrocarbon pollution, soil reflectance, GER 3700, FTIR, gamma ray spectrometry.
Acknowledgements

There are so many people who need to be thanked for their contribution to my completing this thesis. Firstly, to my supervisors Mark van der Meijde and Harald van der Werff, thank you for your continual support, constructive criticisms, making me think and making me laugh. It has been a privilege to work under your supervision. Thanks to Freek van der Meer, Chris Hecker, Boudewijn de Smeth, David Rossiter and others of the ITC staff for their assistance and advice during the course of my research. To the staff from the four institutions, thank you for sharing your knowledge, and to some, for becoming my friends.

To my fellow GEMs, who have become family as we travelled around Europe together, thank you for your friendship, companionship, support and laughter through the good times and bad. To my special GEMs: to Saili, my little Akka, with whom I shared everything; to Niladri (Nil), my travel advisor and sanity; to Ben (Bennykins), the one whose whistle I always listen for and who can deal with my bad moods and to Rob (Robster), where to start… lets just say, thank you for those long chats and for making me laugh.

A special thank you to Wiebke and Nicky for feeding, caring, distracting, hugging, encouraging, advising, making me laugh and keeping me sane in the last stages of thesis writing. You guys are my sunshine!! To Alison, thank you for always making me believe in myself. To all my dear friends back home and abroad, thank you for your friendship and random support during this time. To GeoTerraImage (Pty) Ltd, thank you for encouraging me to pursue further studies and for allowing me the time off work.

To my amazing family, this has been a tough time away from you, but I thank you for your continual love and support over the distance. To my parents, Issie and Pops, thank you for loving and supporting me in whatever I decide to do. I thank my Heavenly Father for giving me this ability and strength to achieve thus far. Without Him, I am nothing.
## Table of contents

1. **Introduction** .................................................................................................................. 1  
   1.1. Background and significance ............................................................................. 1  
   1.2. Research Problem .............................................................................................. 3  
   1.3. Research Objective ............................................................................................ 4  
   1.4. Research Questions ........................................................................................... 4  
   1.5. Hypothesis .......................................................................................................... 4  
   1.6. Research Approach ............................................................................................ 4  
2. **Materials** .................................................................................................................... 5  
   2.1. Study area ............................................................................................................ 5  
   2.2. Data available ...................................................................................................... 6  
   2.3. Data collected ...................................................................................................... 7  
   2.3.1. August 2007 field campaign ......................................................................... 7  
   2.3.2. October 2007 field campaign ...................................................................... 10  
3. **Methods** .................................................................................................................... 15  
   3.1. Data preparation ................................................................................................. 15  
   3.2. Description of datasets used .............................................................................. 16  
   3.2.1. GER subsets .................................................................................................. 17  
   3.2.2. FTIR subsets .................................................................................................. 19  
   3.2.3. Gamma ray subsets ....................................................................................... 20  
   3.3. Visual assessment ............................................................................................... 20  
   3.4. Logistic regression .............................................................................................. 21  
   3.5. Prediction ............................................................................................................ 22  
   3.6. CIE 1964 colour matching ................................................................................. 23  
4. **Results** ....................................................................................................................... 24  
   4.1. GER data ............................................................................................................. 24  
   4.1.1. Reflectance spectra ......................................................................................... 24  
   4.1.2. Continuum removed spectra ........................................................................... 28  
   4.1.3. First derivatives ............................................................................................... 33  
   4.1.4. Noise ............................................................................................................... 37  
   4.1.5. Prediction ........................................................................................................ 38  
   4.1.6. CIE 1964 colour matching ............................................................................. 40  
   4.1.7. Depth .............................................................................................................. 42  
   4.2. FTIR data ............................................................................................................ 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1.</td>
<td>Reflectance spectra</td>
<td>46</td>
</tr>
<tr>
<td>4.2.2.</td>
<td>Continuum removed spectra</td>
<td>48</td>
</tr>
<tr>
<td>4.2.3.</td>
<td>First derivatives</td>
<td>51</td>
</tr>
<tr>
<td>4.3.</td>
<td>Gamma ray spectrometer data</td>
<td>53</td>
</tr>
<tr>
<td>5.</td>
<td>Discussion</td>
<td>55</td>
</tr>
<tr>
<td>5.1.</td>
<td>GER spectrometer data</td>
<td>55</td>
</tr>
<tr>
<td>5.1.1.</td>
<td>Surface samples</td>
<td>55</td>
</tr>
<tr>
<td>5.1.2.</td>
<td>Depth samples</td>
<td>57</td>
</tr>
<tr>
<td>5.2.</td>
<td>FTIR spectrometer data</td>
<td>58</td>
</tr>
<tr>
<td>5.3.</td>
<td>Gamma ray spectrometer data</td>
<td>59</td>
</tr>
<tr>
<td>6.</td>
<td>Conclusions and recommendations</td>
<td>60</td>
</tr>
<tr>
<td>6.1.</td>
<td>Conclusions</td>
<td>60</td>
</tr>
<tr>
<td>6.2.</td>
<td>Recommendations</td>
<td>61</td>
</tr>
<tr>
<td>7.</td>
<td>REFERENCES</td>
<td>63</td>
</tr>
<tr>
<td>8.</td>
<td>APPENDIX</td>
<td>67</td>
</tr>
<tr>
<td>8.1.</td>
<td>CIE chromaticity colour matching</td>
<td>67</td>
</tr>
</tbody>
</table>
List of figures

Figure 1: Location of study area in northern Netherlands.......................... 5
Figure 2: Sampling grid for Field 1....................................................... 8
Figure 3: Sampling grid for Field 9.......................................................... 9
Figure 4: Sampling grid for Field 4b West............................................... 10
Figure 5: Sampling grid for Field 4b East............................................... 11
Figure 6: Sampling grid for Field 13....................................................... 12
Figure 7: GER reflectance spectra for polluted and clean soil samples ...... 25
Figure 8: GER reflectance spectra at wavelength 1730nm. ....................... 25
Figure 9: GER reflectance spectra at wavelength 2310nm. ....................... 26
Figure 10: Box plots of GER reflectance spectra.................................... 27
Figure 11: GER continuum removed spectra......................................... 29
Figure 12: GER continuum removed smoothed spectra........................... 29
Figure 13: GER continuum removed smoothed spectra at wavelength 2310nm................................................................. 30
Figure 14: Box plots of GER continuum removed spectra...................... 31
Figure 15: First derivatives of GER reflectance spectra.......................... 33
Figure 16: First derivatives of GER reflectance spectra at 1730nm............. 34
Figure 17: First derivatives of GER reflectance spectra at 2310nm............. 34
Figure 18: Box plots of the first derivative of GER reflectance spectra..... 35
Figure 19: Comparison of original and smoothed spectra........................ 38
Figure 20: Prediction of polluted and clean sites for Field 4b West.......... 39
Figure 21: Prediction of polluted and clean sites for Field 4b East.......... 40
Figure 22: Prediction of polluted and clean sites for the Field 13............. 41
Figure 23: CIE 1964 colour matching function x and y coordinates plot... 42
Figure 24: Comparison of continuum removed spectra of depth samples,. 43
Figure 25: Box plots of the continuum removed spectra for the depth samples................................................................. 44
Figure 26: CIE 1964 colour matching function x and y coordinates plot comparing the 0.15m, 0.5m and the 0.9m samples. ......................... 45
Figure 27: FTIR reflectance spectra for polluted and clean soil samples... 46
Figure 28: Box plots of FTIR reflectance spectra................................. 47
Figure 29: FTIR continuum removed spectra....................................... 49
Figure 30: Box plots of FTIR continuum removed spectra..................... 49
Figure 31: First derivative of FTIR reflectance spectra. .............................. 51
Figure 32: Box plots of the first derivative of FTIR reflectance spectra...... 52
Figure 33: Box plots of the gamma ray spectrometer measurements........ 54
Figure 34: The CIE chromaticity diagram.............................................. 67
List of tables

Table 1: Summary of datasets used during the analysis process.............. 16
Table 2: Samples selected for the GER primary dataset .................... 17
Table 3: Samples selected for the FTIR primary dataset ..................... 19
Table 4: Samples selected for the gamma ray primary dataset .............. 20
Table 5: Characteristic hydrocarbon absorption wavelengths in the
visible/near infrared part of the spectrum.............................................. 24
Table 6: Summary of the bands selected for the regression equations for the
different GER datasets................................................................. 36
Table 7: Bands included in the multivariate linear regression equation for the
depth dataset...................................................................................... 44
Table 8: Characteristic absorption regions in the mid-infrared spectrum for
aromatic compounds............................................................................. 46
Table 9: Summary of the bands selected for the regression equations for the
different FTIR datasets................................................................. 52
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII</td>
<td>American Standard Code for Information Interchange</td>
</tr>
<tr>
<td>ASD</td>
<td>Analytical Spectral Devices, Inc.</td>
</tr>
<tr>
<td>CIE</td>
<td>Commision Internationale de l’Eclairage</td>
</tr>
<tr>
<td>DLR</td>
<td>Deutsches Zentrum für Luft- und Raumfahrt e.V (German Aerospace Center)</td>
</tr>
<tr>
<td>DRIFT</td>
<td>diffuse reflectance infra-red Fourier transform</td>
</tr>
<tr>
<td>FTIR</td>
<td>Fourier Transform Infrared (spectrometer)</td>
</tr>
<tr>
<td>GPS</td>
<td>global positioning system</td>
</tr>
<tr>
<td>ITC</td>
<td>International Institute for Geo-Information Science and Earth Observation</td>
</tr>
<tr>
<td>LDS</td>
<td>leak detection system</td>
</tr>
</tbody>
</table>
1. Introduction

1.1. Background and significance

Pipelines are used extensively to transport gas and oil over land (Degermenci, 2001; Thomas and Dawe, 2003) as this is often the most economical and relatively safe method when compared to other transportation methods. However, pipelines do develop leaks. These leaks are difficult to detect since many pipelines are subterranean. Due to the hazardous and toxic nature of many of the hydrocarbons transported, the leaks result in both environmental and economic damages.

Traditional methods of gas leak detection often require extensive periods in the field taking measurements. These methods are costly, time-consuming, require trained personnel and are often destructive (e.g. drilling) (Nanni and Demattê, 2006; Odlare et al., 2004; Zirnig et al., 2002). New 'state-of-the-art' detection methods are usually installed at the time of pipeline construction. These are typically either external based leak detection systems (like acoustic emission detectors, fibre optical sensing cable, vapour sensing cable, and liquid sensing cable based systems) or internal leak detection systems (like balancing systems, Real Time Transient Model LDS, pressure/flow monitoring and statistical analysis LDS; Geiger, 2006). Obviously these systems are not useful for pipelines which are already installed. However, these leak detection systems may fail. In addition, external organisations or governments may wish supplementary means to monitor the pipelines for leakages for policing purposes.

A non-destructive, cost- and time-effective technique is already under investigation as a viable tool for pipeline leak detection, namely imaging spectrometry. Imaging spectrometry is a field of remote sensing where the acquisition of images is in a large number of contiguous spectral bands. It is equally referred to as imaging spectroscopy or hyperspectral remote sensing (van der Meer and de Jong, 2001, p XXI). Current studies predominately focus on the effects of the leaked hydrocarbons on vegetation and the resultant reflectance responses. The majority of these studies initially use field spectrometry data spectrometry (Noomen et al., 2003; Smith et al., 2004; van der Meijde et al., 2005) before perhaps scaling up the investigation to imaging (van der Meijde et al., 2005).
The land surface above pipelines is not however always covered by vegetation. Of additional interest is that the effects witnessed in the vegetation are ascribed to the response of the plants to the presence of the pollutants in the soil. Thus there is a need to also investigate the effect of hydrocarbons on the reflectance of other land cover types, and in particular bare soil. However, there are not many studies which have reported the effect of pipeline leakages on the reflectance of soil surfaces.

There are a number of reasons why it is expected that imaging spectrometry and the investigation of the spectral response of the polluted soils may be viable for the detection of leakages. Many studies have been conducted which focus on natural macro- and micro-seepages of hydrocarbons. These are driven by the petroleum exploration industry and include the analysis of hydrocarbon spectral signatures. Hydrocarbons have common absorption features at 1730nm and 2310nm (Cloutis, 1989). Furthermore hydrocarbons, and in particular, aromatic compounds (like benzene) have a number of distinct absorption features in the thermal range of the electromagnetic spectrum (see Table 8 in section 4.2.1, Stuart, 2004).

Studies of seepages also explore the effects of seepages on vegetation, soil and rock mineral alterations, and how these alter the reflectance spectra of the land cover. Mineral alterations have been identified that are associated with seepages, and most of these can be recognized by their specific spectral response (van der Meer et al., 2001a). In addition, some mineral alterations cause changes in the natural radioactivity of the soil (International Atomic Energy Agency, 2003). These changes in radioactivity can be measured using a gamma ray spectrometer and used for the identification of areas of mineral alteration and thus possible seepages. However, natural seepages occur over the long-term (van der Meer et al., 2002). Leakages from pipelines are relatively short-term and thus many of the mineral alterations may not yet be evident.

In a study related to hydrocarbon pollution, Horig et al. (2001) conducted an experiment which showed that hydrocarbon-bearing substances, including soil contaminated with low concentrations of oil, could be detected with airborne hyperspectral imagery. There are many studies which have shown that soil composition can be determined using spectral analysis but which do not focus on hydrocarbon pollution (Chabrillat et al., 2002; Nanni and Demattè, 2006; Selige et al., 2006). Kooistra et al. (2001) had positive results when using field spectroscopy to detect cadmium and zinc contaminated soil. In a laboratory experiment, Winkelmann (2005) investigated reflectance spectra of various hydrocarbon
pollutants at different concentrations with a number of soil types to determine pollution detection levels.

The migration of pollutants, and in particular hydrocarbons, within the subsurface is complex with many influencing factors. These include location of the water table, water table fluctuation, capillary pressure, water flow and geology (Kamon et al., 2006; Kim and Corapcioglu, 2003). It is possible that with water table fluctuations, residual pollutants remain in the pores between or attached to the soil particles. In addition, many hydrocarbons are volatile and thus may be found in the soil gas as vapour (Kamon et al., 2006; Kim and Corapcioglu, 2003). In soils with fine particle sizes, like clays, capillary action may play a role in bringing pollution to the surface. Moreover, studies have found that vegetation anomalies can be used for the indirect detection of hydrocarbons in the subsurface (van der Meer et al., 2001a). Some of the plants used in these anomaly studies have relatively shallow root systems (Smith, 2004; van der Meijde et al., 2005) implying that the pollution has migrated upwards to within the root system.

As pipelines are not always covered by vegetated surfaces, this study proposed to investigate the spectral responses of soil with regard to pollution by contaminants from pipeline leakage, specifically benzene condensates. The effects on the spectra may be as a direct detection of the benzene condensates or due to alterations that have resulted in the soil composition. It was expected that spectral features would enable the detection of benzene condensate pollution, using non-destructive field spectrometry measurements. This would be a useful compliment to the current subsurface investigation methods. It is anticipated that this study will lead to further studies on scaling up to operational level, using imaging spectrometry to survey large areas. Additionally, since hydrocarbons share similar absorption features, namely 1730nm and 2310nm (Cloutis, 1989), some of information gained from the study may be transferable to other hydrocarbon pipeline leakages.

1.2. Research Problem

It was not known whether benzene condensate leakages from underground gas pipelines display distinct, detectable spectral features at the soil surface.
1.3. Research Objective

The objective of this study was to determine, with the use of field spectroscopy and other geophysical data, if areas of soil surfaces which may be contaminated/polluted due to benzene condensate leakages from subterranean gas pipelines could be detected using reflectance spectra.

1.4. Research Questions

1. Is there a difference in the spectral reflectances of the surface soil from the polluted and non-polluted areas?
2. Are there distinct spectral features in the field spectral data that differentiate the surface soil of polluted areas from non-polluted areas?
3. Can these spectral features be used in the prediction of polluted areas of surface soil? If so, how reliable are these predictions?
4. Is there a gradient in the reflectance values of soil taken at different depths which indicate the migration of pollution towards the surface?

1.5. Hypothesis

Null hypothesis: There is a noticeable difference between the spectral signatures of surface soil from polluted areas and non-polluted areas.
Alternative hypothesis: There is no noticeable difference between the spectral signatures of surface soil from polluted areas and non-polluted areas.

1.6. Research Approach

The research makes use of spectral measurements of soil together with contamination data obtained from drilling to investigate if there are distinguishable spectral features associated with hydrocarbon contamination of soil from pipeline leakages. Visual assessment and descriptive statistics are used to investigate the spectra. Logistic regression analysis is used to further investigate the spectra and to develop predictive models. The predictive models are visually assessed taking into consideration the location of the pipeline and proximity to drilling sites with relatively high pollution levels. The thermal spectra and radioactivity measurements are also investigated for confirmation of detectable pollution levels in the surface soil.
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

2. Materials

2.1. Study area

The study area is along a 20km pipeline which transported benzene condensates from Grijpskerk in Groningen Province to Anjum in Friesland Province, in the north of the Netherlands (see Figure 1). The pipeline was ‘sweating’ benzene condensates into the ground at some of the connection points, between the nine meter segments from which the pipeline is constructed. The location of the pipeline is known but the location of the connection points is unknown. The pipeline is approximately two meters below the surface. The soil is predominately clay and the land use in the area is largely agriculture, with planted pasture predominating. The terrain is relatively flat, interrupted only by slight undulations, dikes and canals.

Figure 1: Location of study area in northern Netherlands. The red line represents the pipeline. The specific fields that were visited are indicated with pink labels. Town names are given in cream labels – map courtesy of DHV internal report (2006, unpublished).
The study area was selected as it is part of an on-going project, which includes the investigation of the soil pollution levels, remediation measures, and the effect of the pollution on vegetation reflectance spectra. It has been shown that the vegetation reflectance spectra in the study area show anomalies associated with the pollution (van der Meijde et al., 2005). Based on pollution data available, sites along the pipeline were selected for study during the research process.

2.2. Data available

Since this study was part of an ongoing project there were some data already available. Much of the previous and ongoing research in this project is focused on the vegetation. Since the project is being done in conjunction with a private company, there are often restrictions on data available or on accessibility to the study area. Thus preferential datasets were not available or even possible to collect.

There were pollution level data available from drilling campaigns which were done in 2002/03. The locations of the drilling sites were determined by the private company and thus may not be preferred sites that would assist with the research. An additional drilling campaign was undertaken in September 2007 and the data was made available in October 2007, together with 2003/04 drilling data. Only two of the drilling sites correspond between the drilling campaigns. The depth of the drilling samples vary from 0.5m to 6.5m. The 2007 drilling data supplies both information on the value of benzene condensate contamination, as well as the category of contamination at the specific drilling point. The 2002/03 drilling data supplies only information on the category of contamination at the specific drilling point.

Two previous field campaigns were carried out, one in 2004 and one in 2005 to coincide with the hyperspectral sensor overflight. In 2004, spectral data were collected using the ASD field spectrometer and soil conductivity data was collected using the EM31. This data was collected for vegetation only. In 2005, spectral data was collected using the ASD field spectrometer (including one site that was bare soil) and for the calibration of the hyperspectral imagery. The drilling data for the bare soil site was not released by the private company. The data collection for the various years targeted different areas along the pipeline.

There was HyMap hyperspectral imagery available from 2005, which was flown specifically for this project. The imagery has been corrected to reflectance values and geo-corrected by the data provider, the German Aerospace Center (DLR).
Additionally, vector data was available for the study area, including the location of the pipeline.

2.3. Data collected

Due to the available data not being collected with a soil reflectance study as the objective, not all necessary data was available or available for the same locations. Thus it was necessary to collect additional field data. Two field campaigns were conducted specifically for this study. A field campaign was conducted on 23 and 24 August 2007, with a further one on the 23 October 2007.

Since no recent pollution data was available for the August 2007 field campaign, appropriate sample locations were determined in consultation with the private company, which would also correspond with their September 2007 drilling campaign. Two sites were selected of expected high pollution levels, one of which corresponded with the location of a 2002/03 drilling site. Unfortunately, the September 2007 drilling campaign data revealed that the one site showed practically no pollution in 2007 and the other site was not sampled during the drilling campaign.

It was thus necessary to conduct another field campaign, which took place on 23 October 2007. The drilling pollution data from the September 2007 campaign only showed a few locations of intermediate pollution levels and none of high pollution level, with most of the locations showing below intervention levels of pollution. Of the five locations that measured intermediate levels of pollution, two were excluded due to their location within remediation sites (and thus subjected to substantial soil disturbance) and another due to its proximity to the road. Thus only two sites remained available for measurement.

2.3.1. August 2007 field campaign

In August 2007, spectral data was collected using the ASD field spectrometer. Additionally, soil moisture measurements were taken using a soil moisture probe sensor and radioactivity measurements were taken using a gamma ray spectrometer. Two sites were visited, namely Field 1 and Field 9 (according to the project naming conventions; see Figure 1). Field 1 has the coordinates 206457m E; 593396m N and 206496m E; 593358m N for its centre line and Field 9 has the coordinates 208685m E; 591786m N and 208664m E; 591769m N for row 1. The coordinate system used is the standard Dutch national grid, namely Rijksdriehoekstelsel.
Sampling grids were set up as shown in Figures 2 and 3. Field 1 had a pasture grass vegetation cover. Thus the grass was scraped away to reveal the soil beneath but with the exclusion of as much root material as possible. The measurements were taken at a soil depth of less than 1cm. The ASD spectrometer measurements were taken immediately after scraping to reduce water loss and possible pollution from the sample areas. The measurements were taken with a contact probe to avoid illumination variations due to time of day and cloud cover. The instrument was calibrated at the start of every row/transect and after every battery change using a white reference. Five measurements were taken at each sample point. Due to continual laptop battery failure (four batteries were used) only one transect of soil measurements was taken, namely row 2m. Vegetation measurements were taken for the entire grid for a related study.

Figure 2: Sampling grid for Field 1. The 2007 drilling site pollution level was less than the target value. Soil measurements were only taken for row 2. The sample names were designated based on row number, followed by the measured position along that row. The measurements started from the 15m location and ended at 45m.
Figure 3: Sampling grid for Field 9. There was no drilling site for 2007. Soil measurements were along two rows. The sample names were designated based on row number, followed by the measured position along that row. The measurements started from the 0m location and ended at 25m and 30m respectively for rows 1 and 2.

At each sample point in the grid, radioactivity measurements were taken with the gamma ray spectrometer. The sample time for each measurement was 100 seconds. Additionally, soil moisture measurements were taken at each sample point using a soil moisture probe sensor. Three measurements were taken at each sample point.

Field 9 had a maize vegetation cover. Since maize is planted in distinct rows, there was sufficient bare soil exposed for measurements. However, to ensure that the measurement surface was uniform and smooth, the surface was scraped before spectral measurements were taken. Again, the ASD spectrometer measurements were taken immediately after scraping to reduce water loss from the sample areas. The measurements were taken with a contact probe to avoid illumination variations due to time of day and cloud cover. The instrument was calibrated at the start of every row/transect and after every battery change using a white reference. Five measurements were taken at each sample point. Full measurements for both the rows in the sample scheme were recorded. At each sample point, radioactivity
measurements were taken with the gamma ray spectrometer. The sample time for each measurement was 100 seconds. Due to one of the probes breaking on the soil moisture probe sensor, no soil moisture measurements were recorded for Field 9.

### 2.3.2. October 2007 field campaign

On the 23 October 2007 another field campaign was undertaken, using the information obtained from the September 2007 drilling campaign. Due to the ASD spectrometer no longer being operational, it was necessary switch to the GER 3700 spectrometer. The GER 3700 spectrometer available has not been setup for field use and thus it was necessary to collect samples for analysis in the laboratory. The GER 3700 spectrometer was selected as a substitute since it has a similar spectral range to the ASD spectrometer. It has some additional bands in the short wavelengths.

---

**Figure 4:** Sampling grid for Field 4b West. The 2007 drilling site pollution level was between intervention and 10x intervention level. The location of the 0.5m and 0.9m depth samples (samples 21 and 22 respectively) were at sample 2. The numbers of the samples sites are the same as those used to name the samples for the analysis process.
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 5: Sampling grid for Field 4b East. The 2007 drilling site pollution level was less than the target value (indicating no pollution). The location of the 0.5m and 0.9m depth samples (samples 38 and 39 respectively) were at sample 24. The numbers of the samples sites are the same as those used to name the samples for the analysis process.

The numbers of the samples sites are the same as those used to name the ASD spectrometer has a bandwidth of 10nm with a sampling interval of 2 nm. The GER 3700 spectrometer has bandwidths and intervals ranging from 1.5nm to 9.5nm.

The two sites that were sampled were Field 4b and Field 13 (according to the project naming conventions; see Figure 1). Field 4b had two areas that were sampled within the field, which were referred to as Field 4b East and Field 4b West. The coordinates for Field 4b West were 206563m E; 593305m N and 206550m E; 593291m N for its centre line and for Field 4b East they were 206617m E; 593258m N and 206608m E; 593248m N for its centre line. The coordinates for Field 13 were 205027m E; 597051m N and 205064m E; 597035m N for its centre line. The coordinate system used is the standard Dutch national grid, namely Rijksdriehoekstelsel.

Sampling grids were set up as shown in Figures 4, 5 and 6. The grids were designed to sample around the drilling site, the pipeline and away from the pipeline to try and
Figure 6: Sampling grid for Field 13. The 2007 drilling site pollution level was greater than 10x intervention level. The location of the 0.5m and 0.9m depth samples (samples 62 and 63 respectively) were between sample 44 and 45. The numbers of the samples sites are the same as those used to name the samples for the analysis process.

ensure adequate coverage of the polluted areas and then also include clean areas for the same field, with the assumption that there should not be any pipeline related pollution at large distances from the pipeline. A Garmin GPS (Global Positioning System) 12XL was used to locate the drilling site. The grids were orientated making use of the location of the drilling site and the permanent markers for the location of pipeline trench. The GPS reading for Field 13 did not seem to be accurate. It is also possible that the coordinate of the drilling point was incorrect. Thus the trench markers were primarily used for this site. Both Field 4b and Field 13 had pasture grass as their vegetation cover. Cows were out to pasture in Field 13 at the time of data collection. Field 4b showed no signs of being used as pasture but rather for the harvesting of the grass.

For both fields, soil samples were collected from a depth of about 15cm, which is below the predominant root level. Additionally, some soil samples were collected at
depth using an auger to extract the soil. Auger samples were taken in both Field 4b and Field 13, at sample sites where drilling indicated the pollution was high and at points where pollution was insignificant. Samples were taken at depths of 0.5m and 0.9m. The samples were collected in aluminium containers to avoid any chemical reactions should benzene condensates be present in the sample. The containers were sealed with masking tape to reduce the loss of any benzene condensates and water. The outside temperature was below 10 degrees Celsius and the samples were stored in a refrigerator once back in the laboratory. Care was taken when handling the samples to ensure that the containers were maintained upright and that they were unsealed for the shortest practical time. While in the field, in situ radioactivity measurements were taken with the gamma ray spectrometer at each sample point. The sample time for each measurement was 100 seconds. No soil moisture measurements were taken, as the measurements from the previous field campaign showed more within site variability than between sample sites.

Due to laboratory availability, the samples were only measured on 30 October 2007 to 1 November 2007 using the GER 3700 spectrometer. The measurements were done in the spectrometry dark room at ITC, where the room has darkened walls and surfaces and no windows. The samples were placed on Petri dishes, where their surfaces were smoothed using a spatula to maintain uniformity and reduce the effects of shadowing. The sample covered most of the area of the Petri dish with only the area close to the edges not being covered. The Petri dishes were initially measured with no sample present and had a negligible reflectance. The samples were illuminated with an external light source, specifically designed for this purpose. The background surface was a black mat with a negligible signature. The instrument was calibrated with a white reference panel at regular intervals, particularly after the lamp had been switched off due to an interruption in measurements. The lamp was allowed to warm up again before calibration took place. The standard instrument setup settings were used, with the averaging set to the power of 4 and the detector matching set to reflectance. Each sample was measured five times and then moved and rotated to compensate for any possible surface or soil composition irregularities, with a further five measurements being taken.

The Fourier Transform Infrared (FTIR) spectrometer was a newly acquired instrument at ITC, which measures reflectance in the thermal range of the electromagnetic spectrum. At the time the samples were collected, only a small amount of soil could be measured due to the limitations of the sample compartment of the FTIR spectrometer. An integrating sphere, which would allow a larger amount of soil to be measured was on order, but delivery continually delayed. Initially it was
decided to await the arrival of the integrating sphere before measurements be done on the soil samples, since the possible amount of benzene condensates would be relatively small. The greater the amount of soil available for measurement, the more likely it was to obtain a signature indicating pollution. However, due to the continual delay, it was decided to make some sample measurements with the built-in sample compartment. It was uncertain if these small samples of soil would render any information when measured using the FTIR. Thus only a few samples were selected.

Diffuse reflectance infra-red Fourier transform (DRIFT) measurements were taken using the FTIR on 13 and 14 December 2007 on selected samples, based on their expected levels on contamination, in other words whether the sample was potentially clean or polluted. Again, care was taken when handling the samples to ensure that the containers were maintained upright and that they were unsealed for the shortest practical time. Samples were taken from the bottom half of the container to ensure the least possible change from the soils' original composition at time of sample collection. The sample of soil filled the measurement container and was levelled using a spatula to reduce shadowing. One measurement was taken per built-in sample compartment of soil and this was repeated for the sample site, resulting in two measurements per sample site. A total of six sample sites were measured. These were samples 2 and 44 as they were most likely to be polluted; samples 31 and 53 which were most likely to be clean and samples 22 and 63 which were taken from a depth of 0.9m and were likely to be polluted. The instrument was calibrated with a mirror before each set of measurements for a sample site. The settings for the instrument were a resolution of 4, 500 scans, an aperture of 4mm and with a phase resolution of 16.
3. Methods

The principal datasets used for the study were the GER and FTIR spectral measurements and the gamma ray spectrometer measurements together with the pollution data. A subset of the GER spectral measurements was generated which included measurements for only the most polluted, at depth and clean sites. This allowed for a more manageable dataset, as well as the ability to examine spectra for extremes of pollution levels.

3.1. Data preparation

The spectra generated by the GER 3700 spectrometer using SVC software (from Spectra Vista Corporation, Poughkeepsie, New York) were in ASCII format. It was thus necessary to import these spectra into software that would enable visualisation and manipulation. The software selected was ENVI version 4.0 by Research Systems, Inc. The spectra were imported and compiled into spectral libraries. Band 614 of wavelength 2214.52nm was found to have an error and was thus excluded from the spectra as a bad band.

The FTIR spectrometer data were generated using OPUS spectroscopy software (from Bruker Optik GmbH, Ettlingen, Germany) and were in a comma delimited text format. The OPUS carbon dioxide and water vapour atmospheric compensation was applied to the spectra. The spectra were imported into ENVI and compiled into a spectral library for visualisation and data processing.

The gamma ray spectrometer data was generated in tab delimited text format. This was imported into Microsoft Office Excel 2003 (from Microsoft Corporation, Redmond, Washington).

The GER and FTIR spectral data and the gamma ray data were also imported into SPSS version 15.0 (from SPSS Inc, Chicago, Illinois) for statistical analysis, which included box plot production and regression analysis.
3.2. Description of datasets used

For ease of interpretation, the data were separated based on whether the soil samples were expected to be polluted or clean. For many of the samples this was uncertain, and thus a reduced dataset was created of samples where it was most likely to be expected that the samples were either clean or polluted. The samples were also selected from both Field 4b and Field 13. The sample sites included in these datasets are sensor specific and are shown in Tables 2, 3 and 4 as the beginning of each subsection respectively. These datasets (one each for GER, FTIR and gamma ray) became the primary datasets for analysis, from which the other datasets were created. A summary of the datasets used in this study is given in Table 1.

Table 1: Summary of datasets used during the analysis process.

<table>
<thead>
<tr>
<th>Dataset #</th>
<th>Dataset description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GER reflectance</td>
</tr>
<tr>
<td>2</td>
<td>GER continuum removed</td>
</tr>
<tr>
<td>3</td>
<td>GER first derivatives</td>
</tr>
<tr>
<td>4</td>
<td>GER continuum removed no depth</td>
</tr>
<tr>
<td>5</td>
<td>GER continuum removed 499-2500nm</td>
</tr>
<tr>
<td>6</td>
<td>GER continuum removed 499-2346nm</td>
</tr>
<tr>
<td>7</td>
<td>GER continuum removed 499-2346nm no depth</td>
</tr>
<tr>
<td>8</td>
<td>GER continuum removed for Field 4b 499-2346nm no depth</td>
</tr>
<tr>
<td>9</td>
<td>GER smoothed continuum removed 499-2346nm</td>
</tr>
<tr>
<td>10</td>
<td>GER noise</td>
</tr>
<tr>
<td>11</td>
<td>GER continuum removed 499-2346nm depth only</td>
</tr>
<tr>
<td>12</td>
<td>GER reflectance no depth (for CIE 1964)</td>
</tr>
<tr>
<td>13</td>
<td>GER reflectance depth only (for CIE 1964)</td>
</tr>
<tr>
<td>14</td>
<td>FTIR reflectance</td>
</tr>
<tr>
<td>15</td>
<td>FTIR reflectance no depth</td>
</tr>
<tr>
<td>16</td>
<td>FTIR continuum removed</td>
</tr>
<tr>
<td>17</td>
<td>FTIR continuum removed no depth</td>
</tr>
<tr>
<td>18</td>
<td>GER continuum removed 499-2346nm no depth with FTIR continuum removed no depth</td>
</tr>
<tr>
<td>19</td>
<td>FTIR first derivatives</td>
</tr>
<tr>
<td>20</td>
<td>Gamma ray radioactivity</td>
</tr>
</tbody>
</table>
3.2.1. GER subsets

The primary GER dataset (Dataset 1) was used for the assessment and analysis of the reflectance spectra (see Table 2). Three measurements for each sample were included in the dataset. However, absorption features may be very subtle and not discernible with visual inspection. In addition, differences in brightness that may not be related to pollution, were clearly observable in the spectra and make it more difficult to compare the spectra. Thus, in order to overcome these concerns, the spectra were normalised using the continuum removal method (Dataset 2). The continuum removal function in ENVI was used to normalise the dataset. Since no consistent differences were observed in either the normal or continuum removed spectra, it was decided to explore the first derivatives of the reflectance spectra as they provide a more sensitive analysis. The derivatives highlight regions where the rate of change of reflectance with wavelength is high (Smith et al., 2004). This may however not only enhance signal but possibly noise as well. The first derivatives of the reflectance spectra were calculated for Dataset 1 (Dataset 3) using the IDL ‘deriv’ function in ENVI Spectral Math.

Table 2: Samples selected for the GER primary dataset (Dataset 1) for assessment and analysis based on their expected pollution and location.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pollution</th>
<th>Location</th>
<th>Depth (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>12</td>
<td>clean</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>18</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>21</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.5</td>
</tr>
<tr>
<td>22</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.9</td>
</tr>
<tr>
<td>24</td>
<td>clean</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>30</td>
<td>clean</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>31</td>
<td>clean</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>44</td>
<td>polluted</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
<tr>
<td>45</td>
<td>polluted</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
<tr>
<td>52</td>
<td>clean</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
<tr>
<td>53</td>
<td>clean</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Since continuum removal normalises the spectra and is not expected to enhance noise, preference was given to the analysis of various continuum removed spectra datasets. Due to there being considerably more sand visible in the samples taken at depth, it was thought that this could be influencing the regression outcomes. Thus the measurements from auger soil samples, namely samples 21 and 22, were then
excluded from the continuum removed Dataset 2 to form Dataset 4. There was a large amount of noise in the shorter wavelengths of the GER measurements, which is likely as a result of low instrument sensitivity in this wavelength region. It was therefore decided to exclude these from the analysis. Continuum removed spectra were calculated from 499.47nm through to 2500.04nm. Initially, the samples from depth were still included in this dataset (Dataset 5). This allowed for a comparison with the results from Dataset 2. There was also some noise in the longer wavelengths of the GER spectra, which is also likely a result of low instrument sensitivity in this wavelength region. It was thus decided to exclude these from the continuum removed dataset and observe what influence they have on the regression results. This new dataset only included bands from 499.47nm through to 2346.75nm (Dataset 6).

As already discussed, the depth samples visibly had more sand present which possibly had an effect on the regression analysis. The samples from depth (samples 21 and 22) were then excluded from Dataset 6 to produce Dataset 7. Field 13 was quite severely trampled by cattle at the time the samples were collected. It was therefore decided to exclude these samples from the previous dataset in case they were obscuring true patterns in the data. This new dataset consisted of only samples for Field 4b, namely samples 2, 12, 18, 24, 30 and 31, and still excluded the depth samples (Dataset 8).

In an effort to differentiate noise from signal, the following procedure was applied. The IDL smooth function was run on the continuum removed spectra, from 499.47nm through to 2346.75nm (i.e. Dataset 6) to create Dataset 9. The parameter used was a width of three bands. The smoothed spectra (Dataset 9) were then subtracted from the continuum removed spectra (Dataset 6). The resulting dataset (Dataset 10) should essentially contain only noise. This procedure to enhance the noise was used so as to enable analysis on the noise, not to gain a predictive result but to test whether the noise results resemble any of the previous results.

In order to explore the data from the depth samples, a new dataset was created which only included the samples 2, 21, 22, 44, 45, 62 and 63. Samples 2, 21 and 22 were from Field 4b and were taken at depths of 0.15m, 0.5m and 0.9m respectively. Samples 44, 45, 62 and 63 were from Field 13 and were taken at depths of 0.15m, 0.15m, 0.5m and 0.9m respectively. Again three measurements for each sample were included. The spectra were subset to include only the bands 499.47nm through to 2346.75nm and the continuum removal function was applied to the spectra (Dataset 11).
The CIE 1964 colour function required reflectance spectra as the input data. Since the surface samples differed in composition from the depth samples, these were excluded from the Dataset 1 to create Dataset 12. In addition, a dataset to explore the colour differences in the depth data was created. Like Dataset 11, it included the samples 2, 21, 22, 44, 45, 62 and 63, but used the reflectance spectra instead of continuum removed spectra and included the whole range of bands (Dataset 13).

3.2.2. FTIR subsets

The primary FTIR dataset (Dataset 14) was used for the assessment and analysis of the reflectance spectra (see Table 3). Two measurements for each sample were included in the dataset. The auger samples from depths 0.9m were then excluded from the dataset (Dataset 15). This was due to there being notable differences in the composition of the soil at these depths. Again, since brightness differences were apparent in the FTIR spectra, the spectra from Dataset 14 (thus including the depth samples for comparative purposes) were then normalised using ENVI’s continuum removal function to create a new dataset (Dataset 16). The depth samples were similarly removed from Dataset 16 to create Dataset 17. In order to weigh the FTIR data against the GER data, the sensor specific continuum removed datasets, which excluded depth samples, were combined. However, in order for the datasets to be merged, they needed to have the same samples. Thus some of the samples were removed from the GER dataset. Thus Dataset 4, with only two measurements for samples 2, 31, 44 and 53, was combined with Dataset 17 to form Dataset 18. The first derivatives of the FTIR reflectance spectra (Dataset 14) were computed using the IDL ‘deriv’ function to form Dataset 19.

Table 3: Samples selected for the FTIR primary dataset (Dataset 14) for assessment and analysis based on their expected pollution and location.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pollution</th>
<th>Location</th>
<th>Depth (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>22</td>
<td>polluted</td>
<td>Field 4b</td>
<td>0.9</td>
</tr>
<tr>
<td>31</td>
<td>clean</td>
<td>Field 4b</td>
<td>0.15</td>
</tr>
<tr>
<td>44</td>
<td>polluted</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
<tr>
<td>53</td>
<td>clean</td>
<td>Field 13</td>
<td>0.15</td>
</tr>
<tr>
<td>63</td>
<td>polluted</td>
<td>Field 13</td>
<td>0.9</td>
</tr>
</tbody>
</table>
3.2.3. **Gamma ray subsets**

The primary gamma ray dataset (Dataset 20) was used for the assessment and analysis of the radioactivity measurements. This dataset included the same samples as the GER primary dataset except that there were no depth samples as radioactivity measurements are volumetric (see Table 4). In addition there was only one measurement per sample, except for sample 12 where there were two. The dataset included percent potassium, uranium (in ppm), thorium (in ppm), the total count of radiation, and the ratios of potassium over thorium and uranium over thorium as variables.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Pollution</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>polluted</td>
<td>Field 4b</td>
</tr>
<tr>
<td>12</td>
<td>clean</td>
<td>Field 4b</td>
</tr>
<tr>
<td>18</td>
<td>polluted</td>
<td>Field 4b</td>
</tr>
<tr>
<td>24</td>
<td>clean</td>
<td>Field 4b</td>
</tr>
<tr>
<td>30</td>
<td>clean</td>
<td>Field 4b</td>
</tr>
<tr>
<td>31</td>
<td>clean</td>
<td>Field 4b</td>
</tr>
<tr>
<td>44</td>
<td>polluted</td>
<td>Field 13</td>
</tr>
<tr>
<td>45</td>
<td>polluted</td>
<td>Field 13</td>
</tr>
<tr>
<td>52</td>
<td>clean</td>
<td>Field 13</td>
</tr>
<tr>
<td>53</td>
<td>clean</td>
<td>Field 13</td>
</tr>
</tbody>
</table>

3.3. **Visual assessment**

The first stage in data analysis was to assess the spectra visually. For the GER and FTIR data, the normal spectra as well as the continuum removed spectra and first derivatives were examined. The spectra were visualised using line graphs which plotted the reflectance values along the y-axis and the band wavelengths along the x-axis. The spectra were examined across the entire spectrum as well as at specific wavelengths where known features were expected.

Box plots were produced for each variable, comparing the polluted with the clean spectra. For the GER and FTIR data, the variables were the spectral bands. For the gamma ray spectra, the variables were percent potassium, uranium (in ppm), thorium (in ppm), the total count of radiation, and the ratios of potassium over thorium and uranium over thorium. The box plots show the median, first and third quartiles, the
minimum and maximum observations as well as any outliers or extreme values. The box plots were examined for all variables as well as at specific wavelengths where known features were expected.

3.4. Logistic regression

Due to the lack of representation of pollution levels from the drilling pollution data, particularly as a continuous variable, standard multivariate regression analysis would be an inappropriate method of analysis. Logistic regression is ideally suited for binary data, where in this case the interest is in whether or not there is pollution. The various datasets were analysed in SPSS software. The spectra were imported into a SPSS data file and then analysed using the binary logistic regression function. The dependant variable is the presence or absence of pollution, where 0 is the absence and 1 represents the presence of pollution. The covariates are the spectral bands, or in the case of the gamma ray data, the variables listed in section 3.2.3. SPSS estimates the regression coefficients through an iterative maximum likelihood method.

The forward conditional stepwise method of logistic regression was chosen, as it adds one variable per step and is not initially restricted by the number of cases versus variables. This is particularly important in this instance, as there are a large number of variables and relatively few cases. The maximum number of iterations was increased to 100 so as not to be a limiting factor. Logistic regression has the option of whether or not to include a constant in the equation. The constant is the slope of the log odds ratio regression line. For this study the models included a constant. The probability for stepwise entry was 0.05 and for removal 0.10.

The output of the regression analysis includes the constant and coefficients for the regression equation. The regression equation is written as

$$P = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \ldots + b_i x_i)}}$$

where

- $P$ is the probability of a 1 occurring
- $e$ is the base of the natural logarithm (approx. 2.718)
- $b_0$ is the constant
- $b_i$ is the coefficient of variable $x_i$
- $b_i$ is the coefficient of variable $x_i$
Another of the regression analysis output is the Hosmer and Lemeshow (Chi square) Goodness-of-Fit test statistic, which is used to help evaluate whether the model adequately describes the data. If this statistic is greater than 0.05, then the null hypothesis which states that there is no difference between the observed and predicted values of the dependent is not rejected. This only implies that the model explains the variance of the dependent (i.e., pollution) to a significant degree and that the estimates fit the data at an acceptable level (Garson, 2008). Thus it is preferable if the Hosmer and Lemeshow Goodness-of-Fit test statistic is above 0.05.

In addition, the regression analysis output includes a table with the Rao's efficient score together with the significance of each variable. In this case, if the significance value is above 0.05, the accepted null hypothesis states that the coefficient of the variable is zero and the variable is excluded from the model (Garson, 2008). Thus for a variable to be included in the model, it should have a variable significance value that is below 0.05.

The classification table, another output, indicates to what extent the model correctly predicts the values of the samples that were used to develop the model. Two or more classification tables are included in the regression analysis output. The first is for a model which has no variables but only a constant. The second (or later) table is for the final model which includes both the constant and variables. The percentage of correctly predicted samples for the second table can be compared with the first. Since the first table does not include variables, differentiation is not based on any information. The second table should have a higher percentage than the first otherwise the model should be rejected (even if the percentage in the second table seems relatively high).

It is possible that the analysis may find more than one solution (model) which predicts equally well as the solution that is finally presented in the output. This is reported as the solution is not unique.

3.5. Prediction

The regression models were applied to the full set of samples, namely samples 1 to 63, before the reduction of the dataset to only what was considered polluted and clean. This was done in Microsoft Excel 2003. Again three measurements for each sample were included. The prediction results for the three measurements were averaged for the sample site. These averaged datasets were imported into ArcMap™ 9.2 (ESRI, Redlands, California) and the predicted pollution, or lack thereof,
visualised. These outputs were assessed, taking into account proximity to the pipeline and the drilling sites, and the pollution levels of the drilling sites.

A regression model was applied to the 2005 HyMap image of the study area. This was done in ENVI, making use of the software’s continuum removed and spectral math functions. The results were also visualised and assessed in ArcMap™.

3.6. **CIE 1964 colour matching**

The CIE 1964 colour matching function was applied to the data. The CIE 1964 colour matching function transforms the reflectance spectra into x and y coordinates which when plotted on the CIE chromaticity diagram, allocates a colour which simulates the colour that is seen by the human eye. The function also outputs an albedo value for the sample (van der Werff, 2006). The input data is the reflectance spectra. A script written by van der Werff (2006) was applied to the data using IDL (ITT Corporation, Boulder, Colorado) and ENVI. The output is a comma delimited text file which was imported into Microsoft Excel 2003 for analysis and visualisation.
4. Results

4.1. GER data

4.1.1. Reflectance spectra

The GER 3700 reflectance spectra for the soil samples (Dataset 1) were visually inspected for any obvious differences between the polluted and clean samples. The naming convention used for the spectra is the sample number followed by the field name from whence the sample was collected. It was expected that distinct differences between the polluted spectra and the clean spectra would be visible at various wavelengths. In particular, it was expected that absorption features would be apparent in the polluted spectra at 1730nm and 2310nm and not in the clean spectra. As mentioned in section 1.1, hydrocarbons are known to absorb light at these wavelengths (see Table 5). In addition, absorption features associated with mineral alterations were also possibly expected to be observed in the polluted samples. Upon careful inspection of the reflectance spectra, it was observed that there were no consistent differences visible between the polluted and clean soil spectra over the full wavelength range (see Figure 7) as well as around the hydrocarbon specific wavelengths at 1730nm and 2310nm (see Table 5 and Figures 8 and 9).

Table 5: Characteristic absorption wavelengths in the visible/near infrared part of the spectrum for hydrocarbons (Cloutis, 1989).

<table>
<thead>
<tr>
<th>Wavelength (nm)</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1730</td>
<td>First overtone of the C-H stretch</td>
</tr>
<tr>
<td>2310</td>
<td>Combination band</td>
</tr>
</tbody>
</table>

Variations in brightness were observed in the spectra (see Figures 7, 8 and 9). Brightness variations can be ascribed to many factors, including variations in moisture, mineral composition, grain size, organic matter content and possibly hydrocarbon content (de Jong and Epema, 2001; van der Meer et al., 2001b; Winkelmann, 2005). Winkelmann (2005) found in her laboratory experiments that pollution of soils with different hydrocarbons causes a decrease in the overall reflectance of those soils. This expected decrease in reflectance in the polluted soil samples was not discernible for the current dataset. The brightness differences appeared to be random between the polluted and clean spectra. When visually
Figure 7: GER reflectance spectra for polluted and clean soil samples (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.

Figure 8: GER reflectance spectra for polluted and clean soil samples at the hydrocarbon absorption feature wavelength 1730nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 9: GER reflectance spectra for polluted and clean soil samples at the hydrocarbon absorption feature wavelength 2310nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.

 inspecting the actual soil samples, differences in colour and soil composition were observed. The variations in brightness of the samples can probably be attributed to moisture and compositional differences.

Some absorption features may be subtle, which may partially account for no consistent differences being observed between the polluted and clean spectra during visual assessment. Consequently, box plots were generated for each band comparing the spectra from polluted samples (identified as 1) with clean samples (identified as 0). Each box plot was assessed for separability between the polluted spectra and clean spectra. It was expected that by evaluating the distribution of reflectance values for the polluted and clean samples at each band, that bands would be identified where there are no overlaps between the polluted and clean spectra. However, none of the bands displayed completely discrete distributions. Some bands showed partial overlaps in the reflectance distributions between polluted and clean spectra. Band 1057.23nm is an example of the distribution found in the bands that showed partial overlaps (see Figure 10) and perhaps has the most separable distribution amongst the bands. The majority of the bands however exhibited large or complete overlaps between the distributions. The box plot for band 406.76nm is just one example of the distributions for most of the bands (see Figure 10). The
spectra also overlap considerably for the bands where hydrocarbon absorption features are expected for the polluted samples, namely 1733.82nm and 2310.13nm.

Figure 10: Box plots of GER reflectance spectra for polluted and clean soil samples (where 0 is clean and 1 is polluted). Band 406.76nm is representative of most of the bands, displaying almost complete overlap in the distributions. Band 1057.23nm is representative a few bands which show only partial overlaps in the distributions, with band 1057.23nm perhaps being the most discrete. The box plots for bands 1733.82nm and 2310.13nm illustrate the overlaps in the distributions in the hydrocarbon absorption bands.

The next stage in the analysis process was to use logistic regression to further explore the data. It was expected that logistic regression may be able to discern slight differences in the spectra that were not visually apparent. It was expected that
the bands that were identified with partial overlaps in the visual assessment process would be included in the regression equation, or at least have high score statistics with low significance values. In addition, it was expected that the bands that were included in the regression equation or those low significance values would correspond with known hydrocarbon absorption features.

Initially, the GER reflectance spectra (Dataset 1) were analysed using forward conditional stepwise logistic regression. Of all the bands available, only two were included in the final regression equation, namely 1057.24nm and 406.76nm. The solution was not unique. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 1. Since this is greater than 0.05, it only implies that the model explains the variance of the dependent (i.e. pollution) to a significant degree and that the estimates fit the data at an acceptable level (Garson, 2008).

The model predicts 100 percent of the samples used to create the model correctly. Comparing this to the 50 percent of the initialisation model, where there is only a constant and no variables, it would indicate that the model should not be rejected. Of the 645 bands available, 280 bands have relatively high score statistics with individual significance values below 0.05. Since the values are below 0.05 the null hypothesis, which states that the coefficient of the variable is zero, is rejected and thus the variable is not excluded from the model (Garson, 2008).

The band 1057.24nm the first variable selected for the regression equation, showed only partial overlaps in the distributions visualised in the box plot (see Figure 10). Of the two hydrocarbon absorption bands, 1733.82nm has a significance value below 0.05, whereas band 2310.13nm does not. Comparing this to what was observed in the box plots, the distributions of band 1733.82nm do not overlap as completely as for band 2310.13nm (see Figure 10).

4.1.2. Continuum removed spectra

The absorption features may be very subtle. In addition, differences in brightness, which were clearly observable, make it more difficult to compare the spectra. Thus, in order to overcome these concerns, the normalised continuum removed dataset was investigated (Dataset 2). Upon careful inspection of the spectra, it was observed that none of the differences were consistently present in the polluted and clean spectra (see Figure 11). To enhance the visualisation of the spectra, a smoothing function was applied to each spectrum. The smoothing function helps to compensate for
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 11: GER continuum removed spectra for polluted and clean soil samples (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.

Figure 12: GER continuum removed smoothed spectra for polluted and clean soil samples at the hydrocarbon absorption feature wavelength 1730nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples. The large absorption feature around 1780nm in all the samples is associated with organic matter (Kumar et al., 2001).
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 13: GER continuum removed smoothed spectra for polluted and clean soil samples at the hydrocarbon absorption feature wavelength 2310nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples. The absorption features around 2280nm and 2350nm in all the samples are associated with organic matter (Kumar et al., 2001)

random noise. However, even after smoothing, the differences in the spectra still appear to be random with no consistent differences emerging (see Figures 12 and 13).

Box plots were generated for the continuum removed spectra (Dataset 2). It was expected that bands would be identified where there is a clear separation between the distribution of clean and polluted samples. No bands showed completely discrete distributions of polluted spectra compared to the clean spectra. Bands 1280.33nm, 1345.95nm and 1313.48nm displayed the least overlaps (see Figure 14). The remainder of the bands showed large or complete overlaps in the distributions of polluted and clean spectra. This is illustrated in the box plots for bands 852.48nm and 1870.56nm in Figure 14.

Since continuum removal normalises the spectra, preference was given to analysis of continuum removed datasets. A description of the datasets can be found in section 3.2. Logistic regression analysis was conducted on the continuum removed datasets. For all continuum removed datasets, the model solution was not unique, the Hosmer
and Lemeshow Goodness-of-Fit test statistic was 1 and the model predicts 100 percent of the samples used for the model correctly. A summary of the bands included in the different regression equations based on GER datasets is given in Table 6 (at the end of section 4.1.3).

![Box plots of GER continuum removed spectra for polluted and clean soil samples (where 0 is clean and 1 is polluted). Bands 1280.33nm and 1345.95nm are examples of where there are the least overlaps in the distributions. Bands 852.48nm and 1870.56nm are examples of the distributions observable in the rest of the bands.](image)

The continuum removed spectra for the entire range of GER spectra (Dataset 2) were analysed using logistic regression. The results of the logistic regression included four bands, namely 1280.32nm, 1345.94nm, 1635.08nm and 1237.93nm.
Bands 1280.33nm and 1345.95nm were also identified as having the least overlaps when examining the box plots for these bands. Only 38 of the 645 bands have significance values below 0.05. Neither of the expected hydrocarbon absorption bands (see Table 5) was included in this list of variables with individual significance values below 0.05.

The measurements from auger soil samples, i.e. those at depth, were then excluded from the samples (Dataset 4). This was due to there being considerably more sand visible in the depth samples, which it was thought could be influencing the regression outcomes. The results of the logistic regression on this new dataset only included two bands in the final regression equation, but they were two that were also included in the previous results, namely 1345.94nm and 1635.08nm. The bands with a significance value below 0.05 were 144 in number. Again, neither of the hydrocarbon absorption bands, namely 1733.82nm and 2310.13nm, was included in this list.

There is a large amount of noise in the shorter wavelengths of the GER measurements. It was thus decided to exclude these from the analysis. Continuum removed spectra were calculated from 499.47nm through to 2500.04nm, including the depth samples for comparative purposes (Dataset 5). Logistic regression on these spectra included the 862.96nm, 586.46nm, 1194.43nm and 776.76nm bands in the equation. Sixty-six of the included 525 bands had significance values of less than 0.05. The hydrocarbon absorption bands had significance values greater than 0.05.

There is also some noise in the longer wavelengths of the GER spectra. It was thus decided to exclude these from the continuum removed calculation and observe what influence they have on the regression. Continuum removed spectra were calculated from 499.47nm through to 2346.75nm (Dataset 6). The logistic regression rendered the same four bands, namely 862.96nm, 586.46nm, 1194.43nm and 776.76nm, for the final equation. The bands with significance values below 0.05 are also the same as the previous regression, except that 2425.75nm is obviously now not included.

The samples from depth were then excluded. As already discussed, the depth samples visibly had more sand present which possibly has an effect on the regression analysis. The regression was recalculated on this new dataset, namely continuum removed spectra from 499.47nm through to 2346.75nm and excluding depth samples (Dataset 7). Only one band was included in the final regression equation, namely 678.21nm. This does not correspond with any of the bands selected previously in the regression analyses. Of the 508 bands available, 147 bands
have significance values below 0.05. This does not include either 1733.82nm or 2310.13nm (the two hydrocarbon absorption bands).

Field 13 was quite severely trampled by cattle at the time the samples were collected. It was thus decided to exclude these samples from the analysis in case they were obscuring true patterns in the data. The logistic regression analysis was recalculated, using only the data from Field 4b but still excluding the depth samples and only for bands 499.47nm through to 2346.75nm (Dataset 8). The final regression included only one band in the equation, namely 676.74nm. Again, this is not a band that has been included in one of the previous results from the logistic regression analyses. Of the 508 bands available, 126 bands have significance values less than 0.05.

4.1.3. **First derivatives**

Since no consistent differences were observed in either the normal or continuum removed spectra, it was decided to explore the first derivatives of the reflectance spectra (Dataset 3) as they provide a more sensitive analysis. The derivatives highlight regions where the rate of change of reflectance with wavelength is high (Smith *et al.*, 2004). This may however not only enhance signal but possibly noise as well.

![Figure 15: First derivatives of GER reflectance spectra for polluted and clean soil samples (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.](image-url)
Figure 16: First derivatives of GER reflectance spectra for polluted and clean soil samples at 1730nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.

Figure 17: First derivatives of GER reflectance spectra for polluted and clean soil samples at the hydrocarbon absorption feature wavelength 2310nm (spectra names consist of sample number and field number). No consistent differences were observed between the clean and polluted samples.
It was expected that there would be visible differences between the first derivatives of the polluted and clean spectra, especially at the hydrocarbon absorption wavelengths of 1730nm and 2310nm. However, when the first derivatives were visually assessed, no consistent differences between the clean and polluted spectra were observed (see Figures 15, 16 and 17).

Figure 18: Box plots of the first derivative of GER reflectance spectra for polluted and clean soil samples (where 0 is clean and 1 is polluted). Bands 600.92nm and 1385.58nm are examples where there are only partial overlaps in the distribution of the reflectance values. Bands 1057.23nm and 1280.32nm are examples of where there are large or complete overlaps in the distributions.
Box plots were created for the first derivatives of the reflectance spectra. With the increased sensitivity of the first derivatives, it was expected that slight differences in the spectra may become evident. The distributions for band 600.92nm appear discrete (see Figure 18). However, there are a number of extreme values which fall into the overlap areas in the distributions. Band 1385.58nm (see Figure 18) is an example of the partial overlaps that were observed in some of the bands. Bands 1057.23nm and 1280.32nm are representative of the large to complete overlaps that were observed in the distributions (see Figure 18). In addition, band 1057.23nm showed less overlaps in distribution in the reflectance spectra dataset (Dataset 1) and band 1280.32nm had less overlaps in the continuum removed dataset (Dataset 2) than is observed in the derivatives dataset (Dataset 3).

### Table 6: Summary of the bands selected for the regression equations by the regression analyses for the different GER datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Band 1</th>
<th>Band 2</th>
<th>Band 3</th>
<th>Band 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflectance (Dataset 1)</td>
<td>406.76nm</td>
<td>1057.24nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuum removed (Dataset 2)</td>
<td>1237.93nm</td>
<td>1280.32nm</td>
<td>1345.94nm</td>
<td>1635.08nm</td>
</tr>
<tr>
<td>Continuum removed no depth (Dataset 4)</td>
<td>1345.94nm</td>
<td>1635.08nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuum removed 499-2500nm (Dataset 5)</td>
<td>586.46nm</td>
<td>776.76nm</td>
<td>862.96nm</td>
<td>1194.43nm</td>
</tr>
<tr>
<td>Continuum removed 499-2346nm (Dataset 6)</td>
<td>586.46nm</td>
<td>776.76nm</td>
<td>862.96nm</td>
<td>1194.43nm</td>
</tr>
<tr>
<td>Continuum removed 499-2346nm no depth (Dataset 7)</td>
<td>678.21nm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuum removed for Field 4b 499-2346nm no depth (Dataset 8)</td>
<td>676.74nm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First derivative (Dataset 3)</td>
<td>600.92nm</td>
<td>1850.06nm</td>
<td>2417.19nm</td>
<td></td>
</tr>
<tr>
<td>Continuum removed 499-2346nm no depth with FTIR continuum removed no depth (Dataset 18)</td>
<td>657.71nm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The logistic regression was conducted on the first derivative dataset (Dataset 3). The regression equation included three bands, namely 600.92nm, 1850.06nm and 2417.19nm. All three of these bands showed only partial overlaps in the distributions of polluted and clean samples. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 1 and the model predicts 100 percent of the samples correctly. The model was not unique. The bands with significance values below 0.05 number 137 but do not include either of the two known hydrocarbon absorption bands.

A summary of the bands included in the different regression equations based on GER datasets is given in Table 6. Each new regression analysis renders different resulting bands that are included in the regression equation. This is unexpected as the regression analyses are based on the same base dataset (Dataset 1). It indicates that perhaps the differentiability of the samples using these bands is not notable. It was thus decided to evaluate the bands based on their significance as calculated for each regression. It was hoped that some sort of pattern would emerge. When comparing the bands that have a significance value below 0.05 from each regression analysis, only three bands consistently emerge. These are 1280.32nm, 1337.89nm and 1361.93nm.

4.1.4. Noise

Since there were no bands that distinctly differentiated the polluted and clean spectra and since the regression analyses rendered different bands, there was some concern that what the regression may be identifying as information may in fact be purely noise in the data. In an effort to try and differentiate noise from signal, the following procedure was applied to the data. A smooth function was run on the continuum removed spectra, from 499.47nm through to 2346.75nm (Dataset 6). The parameter used was a width of three bands. The smoothed spectra (Dataset 9) should retain sufficient of the original signal but with the satisfactory removal of noise (see Figure 19; O’Haver, 2006). The continuum removed dataset was used as it is normalised and as most of the regression analyses were run on continuum removed data. The smoothed spectra (Dataset 9) were then subtracted from the continuum removed spectra (Dataset 6). The resulting dataset (Dataset 10) should essentially contain only noise.
This new noise-only dataset (Dataset 10) was used in a logistic regression analysis. The bands that were selected for the regression equation were 987.32nm, 949.15nm, 923.28nm and 713.42nm. All these bands appear at some point in the lists of bands with significance values of less than 0.05 for the different datasets. Upon inspection of the bands that have a significance value of less than 0.05 for the noise dataset, it was found that a few of the bands included in the regression equations for the different datasets (see Table 6) were also considered noise. These were 676.74nm, 1280.32nm and 1345.94nm. Of the three bands that emerged when comparing the bands that consistently have significance values below 0.05 for all the regression analyses, 1280.32nm is considered as noise. This would imply that bands 1337.89nm and 1361.93nm are valid information bands. However it would appear that the information is not sufficient for differentiation purposes. When evaluating the bands with significance values below 0.05 per regression dataset with the bands with significance values below 0.05 for the noise dataset, between nine and fifteen percent of the bands are considered noise.

4.1.5. Prediction

There is insufficient data available to validate the different regression equations. However, in an attempt to evaluate if any of the equations could possibly be useful, the full datasets were used with the various prediction models and these outputs
The prediction model was also applied to the spectra collected for Field 9 to appraise whether the patterns of pollution presence predicted would be what was expected taking into account proximity to the pipeline and 2002/03 drilling pollution levels. The results were not what were expected. Some samples from close to the pipeline were predicted as clean while some samples further away were predicted as polluted.
Figure 21: Prediction of polluted and clean sites for Field 4b East using the regression equation from the continuum removed 499-2346nm with no depth samples for only Field 4b regression analysis. Sites 23, 25, 33 and 35 were expected to be predicted clean since there is below target value of pollution detected at the drilling site.

Furthermore, this equation was used on the 2005 HyMap image of the study area. The results were less than satisfactory, with the majority of the image being predicted as polluted, including sites which have drilling information showing they are clean.

4.1.6. CIE 1964 colour matching

Many of the bands which are included in the regression equations or have individual significance values below 0.05 are within the visible part of the electromagnetic spectrum, which is between approximately 400nm to 700nm. In addition, van der Werff (2006) found differences in the albedo of the contaminated soils when studying macroseepages and Winkelmann (2005) found brightness differences in her laboratory samples that were contaminated with different concentrations of hydrocarbons. Taking this into consideration, the CIE 1964 colour matching function was applied to the data. The CIE 1964 colour matching function transforms
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 22: Prediction of polluted and clean sites for the Field 13 using the regression equation from the continuum removed 499-2346nm with no depth samples for only Field 4b regression analysis. Site 49 was expected to be predicted clean due to its distance from the pipeline. Site 43 was expected to be polluted due to its proximity to the drilling site which has a pollution level greater than ten times the intervention value.

The spectra into x and y coordinates which when plotted on the CIE chromaticity diagram, allocates a colour which simulates the colour that is seen by the human eye. The function also outputs an albedo value for the sample.

The function was applied to the reflectance spectra (Dataset 12). The x and y coordinate output for the clean and polluted soil samples were compared, as show in Figure 23, there is no consistent pattern or shift in the colour between the two sets of samples. When comparing the albedo values for the clean and polluted samples, the values overlap in range, with the clean samples having albedo values between 537.97 and 627.40, and the polluted samples having values between 549.42 and 556.46. Thus no distinguishing difference between the colour or between the albedo of the clean and polluted samples were found.
4.1.7. Depth

The pipeline is 2m below the surface. It is expected that the pollution concentrations would be higher the closer to the pipeline and decrease in concentration with increasing proximity to the surface. The depth spectra (Dataset 11) were investigated to see if there was any evidence of a gradient of pollution. Initially the spectra were assessed visually. No obvious gradients in the spectra were observed when examining the normal spectra. However, the normalised continuum removed spectra showed gradients in the 600-700nm range and the 0.9m samples differentiable from the 0.5m and 0.15m (15cm) samples in the 900-1300nm range (see Figure 24).

Although there are only seven samples with three readings per site included in Dataset 11, box plots were generated to see if any distribution patterns emerge. Upon inspection of the box plots for the continuum removed spectra, it was found that bands 618.34nm through to 648.94nm clearly show different distributions per sample depth along a gradient. There is a relationship between the spectral distribution gradient and the depth gradient and similarly, it is expected that there is a relationship between the depth gradient and the pollution level gradient. In addition,
bands 868.95nm to 1246.50nm differentiated between the 0.9m samples and the 0.5m and 0.15m samples but not between the 0.5m and the 0.15m samples. This confirms, and refines, what was observed in the visual assessment of the spectra. Examples of the spectra distribution are given in Figure 25. However, there were noticeable visual differences in the soils that were sampled at the different depths, including colour changes in the clay and observable presence of light coloured sand. Thus the observed differences in spectra may be purely due to compositional differences.

Since the depth dataset is not a binary dataset, logistic regression was not an appropriate form of analysis. Multiple linear regression was used for the regression analysis. The assumption is made that with increasing depth, that is the closer in depth to the pipeline, the higher the concentration of pollution that should be expected. Thus depth is used as an indicator of pollution for the purposes of the regression analysis. As already noted above, the number of samples in this dataset is small but is being used as an indicator of possible information. The output regression equation included 20 variables, which are given in Table 7. None of these band correspond with the bands that came out of the previous regression analyses (see

---

**Figure 24: Comparison of continuum removed spectra of depth samples, showing corresponding gradient in spectra with gradient in depth at wavelength range 600-700nm and the 0.9m samples differentiable from the 0.5m and 0.15m samples in the 900-1300nm range (spectra names consist of sample number and field number).**
Figure 25: Box plots of the continuum removed spectra for the depth samples showing the distribution of the spectra based on depth of the sample. Band 625.60nm is an example of the distributions found for the range of bands 618.34nm through to 648.94nm, where there is a gradient for the distribution which corresponds with the depth and thus expected pollution gradient. Band 897.53nm is an example of the distributions found for the range of bands 868.95nm to 1246.50nm, where the shallower samples are different from the 0.9m sample but not different from each other.

Table 7: Bands included in the multivariate linear regression equation for the depth dataset.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Band (nm)</th>
<th>Variable</th>
<th>Band (nm)</th>
<th>Variable</th>
<th>Band (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>646.02</td>
<td>( x_9 )</td>
<td>1839.60</td>
<td>( x_{15} )</td>
<td>599.47</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>580.69</td>
<td>( x_{10} )</td>
<td>2272.59</td>
<td>( x_{16} )</td>
<td>519.23</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>899.04</td>
<td>( x_{11} )</td>
<td>985.78</td>
<td>( x_{17} )</td>
<td>631.44</td>
</tr>
<tr>
<td>( x_4 )</td>
<td>976.63</td>
<td>( x_{12} )</td>
<td>618.34</td>
<td>( x_{18} )</td>
<td>2253.46</td>
</tr>
<tr>
<td>( x_5 )</td>
<td>1870.56</td>
<td>( x_{13} )</td>
<td>593.69</td>
<td>( x_{19} )</td>
<td>754.63</td>
</tr>
<tr>
<td>( x_6 )</td>
<td>2234.11</td>
<td>( x_{14} )</td>
<td>914.17</td>
<td>( x_{20} )</td>
<td>611.08</td>
</tr>
</tbody>
</table>

Table 6 at the end of section 4.1.3). In addition, bands 599.47nm and 985.78nm have significance values of below 0.05 from the noise dataset, indicating that these bands may be predominantly noise.
The CIE 1964 colour matching function was also applied to the soil samples taken at depth. When comparing the x and y coordinate output for samples taken at 0.15m, 0.5m and 0.9m, a progressive shift to the right in the x coordinate was observed with increasing depth (see Figure 26). This would indicate the soil became increasingly redder with depth. Although it was assumed that pollution increased with increasing depth, the colour changes were most likely due to observed compositional changes in the soil. When examining the distribution of the samples per field in the y coordinate direction, an increase in the y was observed with increasing depth for Field 13. However, a slight decrease in the y was seen for Field 4b. Thus there was an inconsistent shift in the y coordinate with increasing depth, which reflects the blue-green part of the chromatography diagram. The albedo values for the different depths appeared to show a slight pattern of increasing with increasing depth. However, the distribution of the albedo values did overlap considerably, with the albedo values ranging from 80.59 to 81.06, 81.03 to 81.16 and 81.05 to 81.37 for 0.15m, 0.5m and 0.9m depths respectively.

Figure 26: CIE 1964 colour matching function x and y coordinates plot comparing the 0.15m, 0.5m and the 0.9m samples. There is a shift in the x coordinate direction which would indicate a change in the red colour with a change in depth. This is mostly likely attributed to compositional changes and not due to pollution.
4.2. FTIR data

4.2.1. Reflectance spectra

The FTIR spectra were expected to display distinct differences between the polluted and clean samples. This is due to hydrocarbons, together with other aromatic compounds, showing relatively strong absorption features in multiple regions in the mid-infrared portion of the spectrum. The feature at 3100-3000cm\(^{-1}\) (3225-3333nm) due to C-H stretching is particularly strong (Clark, 1999; Stuart, 2004). Table 8 shows the main characteristic regions for aromatic compounds. However, upon inspection of the FTIR reflectance spectra (Dataset 14), no consistent differences between the polluted and clean samples were observed (see Figure 27).

<table>
<thead>
<tr>
<th>Wavelength (nm)</th>
<th>Wavenumber (cm(^{-1}))</th>
<th>Assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>3225-3333</td>
<td>3100-3000</td>
<td>C-H stretching</td>
</tr>
<tr>
<td>5000-5882</td>
<td>2000-1700</td>
<td>Overtone and combination bands</td>
</tr>
<tr>
<td>6350-6993</td>
<td>1600-1430</td>
<td>C=C stretching</td>
</tr>
<tr>
<td>7843-10000</td>
<td>1275-1000</td>
<td>In-plane C-H bending</td>
</tr>
<tr>
<td>11111-14492</td>
<td>900-690</td>
<td>Out-of-plane C-H bending</td>
</tr>
</tbody>
</table>

Figure 27: FTIR reflectance spectra for polluted and clean soil samples (spectra names consist of sample number and field number). No consistent differences in the polluted and clean spectra were observed.
Although there were relatively few samples, box plots were generated for each FTIR band comparing the distribution of the spectra from polluted samples (identified as 1) with clean samples (identified as 0). Each box plot was then studied for separability between the polluted spectra and clean spectra. It was expected that for the normal spectra, there would be some dissimilarity between the polluted and clean samples. Special attention was given to the bands within the known absorption regions given in Table 8. However, these regions did not display discrete distributions between the polluted and clean spectra. Other bands did show only slight partial overlaps between the spectra. Band 530.40 nm is an example, displaying almost no overlap (see Figure 28). However, since there were so few samples in the dataset, no conclusions can be drawn. Band 2460.09 nm is another example of the distributions that were observed in the spectra, with only partial overlap in the spectra distributions (Figure 28).

![Figure 28: Box plots of FTIR reflectance spectra for polluted and clean soil samples at 530.40 cm\(^{-1}\) (18853.70 nm) and 2460.09 cm\(^{-1}\) (4064.89 nm) (where 0 is clean and 1 is polluted). There are only partial overlaps in the distributions. However, the dataset consists of only a small number of samples.](image)

Further analysis of the FTIR data was done by using logistic regression. A summary of the outcomes are given in Table 9 at the end of section 4.2.3. It was expected that the regression analysis would detect differences in the bands that were not necessarily observable during the visual assessment. It should however be noted that there were insufficient samples for the regression analysis to be considered statistically sound. The standard accepted sample size is 30 or more (Crawley, 2002). This dataset only has six sample sites with two measurements per sample. The
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

4.2.1. Analysis

The analysis should rather be considered as preliminary for future analysis. Initially, the reflectance spectra (Dataset 14) were used as an input dataset for the logistic regression analysis. Only one band was included in the final regression equation, namely $1039.58 \text{ cm}^{-1}$ (9619.27 nm). This band falls within the range of 1275-1000 cm$^{-1}$ which corresponds to the in-plane C-H bending related to hydrocarbons (see Table 8). The Hosmer and Lemeshow Goodness-of-Fit test statistic was 0.889 and the model predicted 85.7 percent of the samples used in the regression correctly. However, since the model with only the constant and no variables predicted 71.4 percent of the samples correctly, the final model is only a slight improvement. The solution was unique. All the bands had significance values of below 0.05.

The auger samples which were taken from depths of 0.5m and 0.9m were then excluded from the analysis (Dataset 15). This was due to there being notable differences in the mineral composition of the soil at these depths. Band $530.40 \text{ cm}^{-1}$ (18853.70 nm) was the only band included in the regression model equation. The solution was not unique. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 1 and the model predicted 100 percent of the samples used in the regression correctly. One hundred and four bands of the 4148 available had significance values below 0.05. Fifty four of these bands were within the range 1275-1000 cm$^{-1}$ which corresponds to the in-plane C-H bending related to hydrocarbons (see Table 8). This may be an indication that there is pollution in the samples, but would require additional investigation.

4.2.2. Continuum removed spectra

Brightness differences were observed in the reflectance spectra, which may make the identification of subtle absorption features or patterns in the spectra difficult. The spectra were thus normalised using the continuum removal method (Dataset 16). It was expected that differences between the continuum removed polluted spectra and the continuum removed clean spectra would be visible, particularly at the known characteristic absorption bands (see Table 8). Upon inspection of the spectra, it was observed that the differences between the clean and polluted samples appear to be random with no consistency. In addition, no absorption features were observed in the polluted spectra at the expected bands (see Figure 29).

Box plots were produced for the FTIR continuum removed spectra. Since the continuum removal process normalises the data, it was expected that the polluted samples would be more clearly differentiable from the clean samples. Again,
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 29: FTIR continuum removed spectra for polluted and clean soil samples, with the characteristic infrared bands of aromatic compounds marked in grey (after Stuart, 2004; spectra names consist of sample number and field number). No consistently different features were observed between the clean and polluted samples.

Figure 30: Box plots of FTIR continuum removed spectra for polluted and clean soil samples at 1500.55 cm\(^{-1}\) (6664.22 nm) and 4479.47 cm\(^{-1}\) (2232.41 nm) (where 0 is clean and 1 is polluted). The spectra were not differentiable in the regions of known hydrocarbons (e.g. 1500.55 cm\(^{-1}\)) but rather in other parts of the spectrum.
particular attention was given to the regions of known absorption features (see Table 8). As exemplified by two bands in Figure 30, the spectra were not differentiable in the wavelengths expected but rather in other regions of the spectrum.

The normalised dataset (Dataset 16) was then analyzed using logistic regression. The output regression equation only made use of the $4479.47\text{cm}^{-1}$ (2232.41nm) band. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 0.900 and the model predicted 83.3 percent of the samples used in the regression correctly. However, since the model with only the constant, and no variables, predicted 66.7 percent of the samples correctly, the final model is only a slight improvement. The solution was unique. Of the 4146 bands, only 49 had significance values of below 0.05. None of these 49 bands fell within the expected hydrocarbon absorption ranges given in Table 8.

The depth samples were then excluded from Dataset 16 to see if these were influencing the analysis (Dataset 17). The regression equation made use of two bands, which were $4398.47\text{cm}^{-1}$ (2273.52nm) and $4446.69\text{cm}^{-1}$ (2248.86nm). The Hosmer and Lemeshow Goodness-of-Fit test statistic was 1 and the model predicted 100 percent of the samples used in the regression correctly. The solution was not unique. Only eight of the bands had a significance level below 0.05, none of which fell within the ranges of absorption features related to hydrocarbons (see Table 8).

Finally, a GER dataset of continuum removed spectra between 499nm and 2346nm and excluding depth samples was combined with the FTIR dataset of continuum removed spectra which also excluded the depth samples (Dataset 18). A logistic regression was then run on this new dataset. Only one band was selected used in the regression equation, namely 657.71nm. This band is from the GER dataset. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 1 and the model predicted 100 percent of the samples used in the regression correctly. The solution was not unique. Of the 148 bands that had significance values below 0.05, only eight of these were from the FTIR spectra. Surprisingly, all eight of these bands fell in the range 2230nm ($4399.43\text{cm}^{-1}$) to 2275nm ($4409.08\text{cm}^{-1}$), which overlaps with the GER spectral range. However, these wavelengths did not have low significance values for the GER bands. This could indicate that there is no useful information in these wavelengths or that there is noise in one of the instruments at these wavelengths.
4.2.3. First derivatives

Given that no consistent variations between the polluted and clean samples were observed for either the normal or continuum removed spectra, the first derivatives were calculated. As mentioned previously, the first derivative is more sensitive, particularly with regard to the rate of change of reflectance with wavelength (Smith et al., 2004). It was expected that the first derivative spectra for the polluted samples would be differentiated from the clean samples, particularly in the aromatic compound absorption regions (see Table 8). The first derivatives of the spectra (Dataset 19) showed some apparent differences between the polluted and clean samples, especially at the in-plane C-H bending region of 1275-1000 cm\(^{-1}\) (see Figure 31). These results were explored further by looking at the distribution of reflectance values with regard to clean and polluted samples.

The apparent differences that were observed when visually assessing the first derivatives of the spectra, were further explored using box plots. Box plots were in fact computed for all the bands of the first derivatives, but with special attention given to the bands highlighted in the visual assessment. It was expected that there would be distinct differences in the distribution of the polluted spectra compared with the clean. The variations observed were not as distinctive as were expected and still showed some overlaps in their distribution. This is shown in Figure 32.

Figure 31: First derivative of FTIR reflectance spectra for polluted and clean soil samples, with in-plane C-H bending band marked in grey (after Stuart, 2004; spectra names consist of sample number and field number).
INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES

Figure 32: Box plots of the first derivative of FTIR reflectance spectra for polluted and clean soil samples at 1000.04 cm\(^{-1}\) (9999.60 nm) and 1171.70 cm\(^{-1}\) (8534.61 nm) (where 0 is clean and 1 is polluted).

The derivatives dataset (Dataset 19) was then analyzed using logistic regression. The output regression equation only made use of the 3450.50 cm\(^{-1}\) (2898.13 nm) band. The Hosmer and Lemeshow Goodness-of-Fit test statistic was 0.900 was not given but the Cox and Snell \(R^2\) was 0.75 and the Nagelkerke \(R^2\) was 1.0. The model predicted 100 percent of the samples used in the regression correctly. The solution was not unique.

Table 9: Summary of the bands selected for the regression equations by the regression analyses for the different FTIR datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Band 1 - wavenumber</th>
<th>Band 1 - wavelength</th>
<th>Band 2 - wavenumber</th>
<th>Band 2 - wavelength</th>
</tr>
</thead>
<tbody>
<tr>
<td>FTIR reflectance</td>
<td>1039.58 cm(^{-1})</td>
<td>9619.27 nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dataset 14)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTIR reflectance no depth</td>
<td>530.40 cm(^{-1})</td>
<td>18853.70 nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dataset 15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTIR continuum removed</td>
<td>4479.47 cm(^{-1})</td>
<td>2232.41 nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dataset 16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTIR continuum removed no</td>
<td>4398.47 cm(^{-1})</td>
<td>2273.52 nm</td>
<td>4446.69 cm(^{-1})</td>
<td>2248.86 nm</td>
</tr>
<tr>
<td>depth (Dataset 17)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FTIR first derivatives</td>
<td>3450.50 cm(^{-1})</td>
<td>2898.13 nm</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Dataset 19)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Each regression analysis renders different bands that are included in the final regression equations (see Table 9). This would indicate that there is no clear band from the FTIR data from which the polluted and clean spectra may be separated.

4.3. Gamma ray spectrometer data

The radioactivity measurements from the gamma ray spectrometer were assessed (Dataset 20). The relative amounts of potassium (K), uranium (U) and thorium (Th) were examined, as well as the total count of radiation and the ratios of potassium to thorium and uranium to thorium. Hydrocarbons have two possible effects on the radioactivity of the soil. Hydrocarbons can increase the acidity of the soil and thereby destroy clay minerals resulting in the leaching out of most of the potassium and uranium, but causing little change in the amount of thorium. Otherwise, hydrocarbons may form a redox electrochemical cell, with low concentrations of uranium and thorium (International Atomic Energy Agency, 2003).

It was expected that the relative amount of potassium, as compared with the thorium, would be less in the polluted sample sites. Another possible expectation would be that the relative amounts of uranium and potassium would be less in the polluted sites. No conspicuous differences between the distributions of the polluted and clean measurements were observed. Box plots were generated to visualise the distribution of the data (see Figure 33).

A subset of the gamma ray spectrometry measurements was created which corresponded to the same sample sites used in the majority of the analyses, for which it was known whether the sites were polluted or clean. The gamma ray spectrometry dataset was analysed using logistic regression analysis. However, none of the variables had significance values less than 0.05 and so no solution could be computed. This was expected, as no differences between the polluted and clean sample site measurements were observed when examining the data previously with box plots.
Figure 33: Box plots of the gamma ray spectrometer measurements for polluted and clean soil samples sites (where 0 is clean and 1 is polluted). The distributions for the variables overlap partially or completely.
5. Discussion

This section will discuss the results found in the previous section and relate these to the research questions posed in section 1.

5.1. GER spectrometer data

5.1.1. Surface samples

Visual assessment was performed on the reflectance spectra, continuum removed spectra and first derivatives of the reflectance spectra, comparing the polluted and clean samples. No clear consistent differences between the polluted and clean spectra were observed. Although it was expected that some differences in the spectra would be observed, it is not unreasonable to find no distinguishable differences when visually assessing the spectra. Some differences may be very subtle and thus not distinguishable when visually examining the spectra on a computer monitor. In addition, from the 2007 drilling information, it is known that the pollution levels were only as high as 720µg/l for polluted drilling sites. At least two sites, of the sites that were measured in 2003/02, had values higher than 32000µg/l. Consequently the amount of contaminant in the sample is expected to be relatively low. If this is the case, the reflectance signal from the pollutant would be low, in relation to the reflectance signal from the soil itself. Winkelmann (2005), in laboratory studies, found that in clay soils contaminant levels as high as 3.0 wt.% (which is approximately 50000000µg/l) could only be detected spectrally in some cases and with no apparent dependence on soil moisture. It can be deduced that the pollution levels in the clay soils of the study area are most likely too low for spectral detection.

Making use of the output data from the CIE 1964 colour matching function, the albedo and colour of the clean and polluted surface samples were evaluated. The range of albedo values for the polluted samples fell within the range of values for the clean samples. There is no difference in overall brightness between the polluted and clean soil samples. However, Winkelmann (2005) observed that fuel hydrocarbon contaminated soil samples have a decrease in their overall reflectance when compared with uncontaminated soil samples. It is possible that the pollution levels are too low for there to be a detectable change in the albedo.
In addition, there was no distinct pattern in the CIE 1964 colour \(x\) and \(y\) coordinates differentiating the polluted from the clean samples. The soils from both fields are very similar in colour, with no colour difference which distinguishes polluted soil from clean soil. It was expected that there might be a possibility of mineral alteration due to the presence of pollution. This could result in colour changes in the soil (van der Meer et al. 2001b). However, no evidence of colour change was found in this study. The pipeline has only been leaking subsequent to its construction in 1996 and mineral alteration may require longer time periods. The results of the analysis of the gamma ray spectrometry data would seem to confirm this. Changes in radioactivity are associated with mineral alterations (refer to section 4.3) and since no differences were observed, it would indicate that alterations have not yet occurred. From this we can deduce that there are no observable differences in the spectral reflectances of the polluted and non-polluted bare soil (refer to research question 1).

Box plots were produced for the reflectance spectra, continuum removed spectra and first derivatives of the reflectance spectra, comparing the distribution of the reflectance values per band for the polluted and clean samples. There were some bands where the distributions showed a difference between the polluted and clean samples. However these differences were typically insufficient to differentiate between the polluted and clean samples as there were overlaps in the range of reflectance values. In addition, there was little consistency between the reflectance spectra, continuum removed spectra and the derivatives in the bands that showed differences. This would indicate that there are no features that differentiate between the polluted and clean spectra.

Although the entire range of bands was examined, particular attention was paid to the bands around 1730nm and 2310nm, where it is known that hydrocarbon absorption features occur. However, there was no evidence of absorption features at these wavelengths in the polluted spectra. This would imply that the levels of pollution in the soil were too low to be detected in the spectra.

The regression analyses revealed a number of bands that may be used to discriminate between polluted and clean soil samples. It is not unreasonable to expect that since the datasets are based on the same initial dataset that there should be some consistency with the output from the regression analyses. However, there is little consistency in the bands that were selected, with practically every regression analysis using different bands in the regression equation (see Table 6 at the end of section 4.1.3). In order to try and explain this lack of consistency, these bands were checked against the noise dataset. Three of these sixteen bands, namely 676.74nm,
1280.32nm and 1345.94nm, have significance values less than 0.05 for the noise dataset, thus indicating that they are noise and not information.

In addition, when comparing the bands that were considered as significant at 0.05 in the analyses, only three bands consistently emerged. These were 1280.32nm, 1337.89nm and 1361.93nm. However, band 1280nm was considered noise and not information after testing. Interestingly, the range of 1100nm to 1300nm is associated with the C-H stretch second overtones which occur for all biochemical constituents (Kumar et al. 2001). Bands 1337.89nm and 1361.93nm are not associated with known absorption features for aromatic hydrocarbons. These bands were also not found to be associated with any of the expected mineral alterations (van der Meer et al., 2001a). As neither of these bands were selected for regression equations (but were significant), it can be inferred that their contribution to separating the polluted samples from the clean may not be considerable. Nevertheless these two bands should be further investigated to confirm whether they may be useful for the differentiation of polluted and clean soil. This study found no distinct spectral features that differentiate polluted soil from non-polluted soil (refer to research question 2).

Statistical assessment of the prediction models was not possible as not enough data was available. However, the models were visually assessed for pollution distribution patterns that would be expected, taking into consideration the drilling pollution information and proximity to the pipeline. Of the available models, the one based on Field 4b’s continuum removed spectra between 499nm and 2346nm with no depth samples appeared to best represent the expected pollution distribution. This model only used band 676.74nm. This band is in the visible part of the spectrum and in the red range. However, when the colour of the polluted and clean samples was evaluated with the CIE 1964 colour matching function, no pattern was discernible based on the x and y colour coordinates. This would suggest that there may not be sufficient information from this band to be useful for predictive purposes. This study found no substantiation that there is sufficient information in the spectra to facilitate the identification of polluted soils using predictive models (refer to research question 3). The models, however, could not be tested for reliability due to lack of available data.

5.1.2. Depth samples
The spectra for three depths, namely 0.15m, 0.5m and 0.9m, were compared in order to establish if there is any detectable pollution at depth and if there is a migration
gradient of pollution upwards. Although the sample size was small, both visual assessment and the box plots showed a gradient in the continuum removed spectra for the bands 618.34nm to 648.94nm. The CIE 1964 colour matching function also showed a gradient in the x coordinate, with the deeper soils shifting in colour towards red. This corresponds with the bands above which are in the red part of the electromagnetic spectrum. These results may indicate the presence of pollution. Van der Werff (2006) found a shift towards the red part of the CIE chromaticity diagram in his study of soils affected by macroseepages in Paradfürdö, Hungary.

In addition, the albedo values for the different depths appeared to show a slight pattern of increasing with increasing depth, but with overlaps in their distribution. This might indicate that albedo increases with increasing pollution. In contrast to Winkelmann’s findings (2005; refer section 4.1.1), van der Werff (2006) observed an increase in the CIE 1964 albedo values when looking at soils affected by macroseepages compared with those not affected at a site in the Upper Ojai Valley, California. Although the changes in albedo and colour may be due pollution, they could also be the result of compositional changes in the soil, which were visually observed when collecting the samples. There was also a separation of the 0.9m samples from the other two depths in the range 868.95nm to 1246.50nm. This is quite a broad range and may not be diagnostic.

Although inconclusive, there appears to be a gradient in the reflectance values taken at different depths. This may be due to compositional changes. It may, however, be due to the presence of pollution or mineral alterations as a result of pollution. If this is the case, it may be an indication that the pollution is migrating upwards towards the surface (refer to research question 4).

5.2. FTIR spectrometer data

Using the FTIR spectra, it was expected that they would discriminate between polluted and clean soil samples, as there are such strong aromatic compound absorption features in this wavelength range (see Table 8 in section 4.2.1). This was not the case. After visually inspecting the spectra and examining the box plots, the FTIR spectra appear to have equivalent or better information for differentiation between the polluted and clean samples than the GER spectra. However, there are still overlaps in the distributions of polluted and clean spectra. In addition, the sample size is small. It is not unreasonable to expect that with an increased sample size the distributions may increase and thus overlap to a greater extent, rather than improve in separability.
The regression analyses rendered different bands for the equations depending on which processing had occurred and whether the depth samples were included. In addition, when a regression analysis was done which included both the FTIR and GER spectra, only one band was used in the regression equation and this was from the GER spectra. This would confirm that there may be no information in the FTIR spectra. This could be due to the pollution levels being too low for detection with the FTIR spectrometer. However, it may be due to the samples being stored for seven weeks before measurements were undertaken. Although every possible precaution was taken to prevent loss of contaminant, it is possible that it might have vaporised during the storage and handling of the samples.

5.3. Gamma ray spectrometer data

The radioactivity data from the gamma ray spectrometer did not display any differences between polluted and clean sample sites. This may be due to a number of reasons. The pollution levels may be insufficient to cause detectable changes in radioactivity. The pollution may not have caused leaching of the potassium as the time frame (less than twelve years) may be too short. The nature of the pollution may not have been appropriate to result in a redox electrochemical cell and the time frame may not be long enough. The lack of differences observed in the CIE 1964 colour matching supports that these mineral alterations may not have occurred. Moreover the gamma ray spectrometer measures radioactivity for a volume of soil. The distribution of pollution and thus any changes in radioactivity which may result may be very localised. With the volumetric measurements, slight localised changes in radioactivity may be obscured by the bulk measurement of non-altered soil that is measured together with the polluted soil.
6. Conclusions and recommendations

6.1. Conclusions

The spectral responses of soil were investigated with regard to pollution by contaminants from pipeline leakage, specifically benzene condensates. Visual assessment and statistical analysis were conducted on the GER spectra, FTIR spectra and gamma ray radioactivity measurements. No consistent differences were observed between the spectra of the polluted and clean soil samples. In particular, no consistent differences were observed at the wavelengths which are associated with known hydrocarbon absorption features. Additionally, no consistent colour or albedo differences were observed. Moreover, there were no differences in the radioactivity measurements between the polluted sites and clean sites. This could indicate that the levels of pollution may be too low for detection. It may also signify that mineral alterations may not yet have occurred. It could also suggest that these methods may be inappropriate for the detection of pollution from pipeline leakages. However, according to the drilling data, the pollution levels for the study sites were low at the time of measurement. Thus it is more likely that the pollution levels were too low for detection. As the study area has clay soils, this is supported by Winkelmann’s (2005) laboratory findings that high levels of hydrocarbon pollution were necessary for spectral detection in clay soils.

After analysis and comparison of the results of the different datasets used in the study, it was shown that some of the bands which were selected in the regression analysis or which had low individual significance values could be noise. This may be due to noise from the measuring equipment. These bands should not be considered as possible information bands for the differentiation of polluted soils from clean soils. After a regression analysis of a dataset that combined both the GER spectra and the FTIR spectra, it was found that the GER spectra contributed more than the FTIR spectra to the differentiation of polluted samples from clean samples. This was unexpected as there are strong hydrocarbon-associated absorption features in the thermal part of the electromagnetic spectrum. However, the soil samples were stored for seven weeks before the measurements and it is possible that the contaminant may have vapourised during this time. Bands 1337, 89nm and 1361.93nm were found to consistently have low significance values in the various regression analyses. These bands are not associated with hydrocarbon absorption
features or with any expected alterations. However, they should be explored further as they may be associated with a yet unknown alteration that is a result of the benzene condensate pollution. They were also not shown to be noise.

The analysis of the samples which were taken over a range of depths showed a gradient in the distribution of the spectral values for polluted versus clean samples in the red part of the electromagnetic spectrum. The findings from the CIE 1964 colour matching supported this observation. There was a shift in the x coordinate towards the right with increasing depth, which indicates an increase in the redness of the soil. This may be as a result of the pollution in the soil, mineral alterations due to the pollution in the soil or, most likely, natural compositional changes associated with depth. In addition, an increase in the albedo values was observed with increasing depth. However, the number of samples in the depth dataset was particularly small and thus these results are not conclusive.

Perhaps the greatest limitation for this study is the lack of appropriate data. Due to insufficient current data on the location and levels of pollution, only a small number of samples were collected and even fewer could actually be used for analysis. The sample sizes for the analysis datasets were below 30, which is considered the minimum size to be statistically sound (Crawley, 2002). Additionally, there were insufficient data to verify the prediction models that were obtained as a result of the regression analyses. Thus, the results from this study are not conclusive and should be used to direct further studies.

6.2. Recommendations

Recommendations for further studies would include the exploration of bands 1337.89nm and 1361.93nm and whether they are associated with any particular substance, and whether the presence of this substance may be the result of alterations due to hydrocarbon pollution. In addition, the observations from the depth analysis should be explored further to establish if they are due to the natural composition of the soil horizon or due to the presence of pollution. Chemical analysis of the soil should add further information, particularly with regard to the presence or absence of the contaminant, as well as to mineral alterations.

It is recommended that the results from the depth dataset be expanded upon. It would be of interest to conduct a study which investigates the migration of the pollutant in the soil, particularly with regard to its migration towards the surface.
It is recommended that future studies ensure that the datasets consist of more than 30 samples. In addition, these studies should preferably have adequate confirmed sites of known pollution levels, with sufficiently high levels of pollution. Furthermore, it would be of interest to investigate sites of different soil types and with different types of hydrocarbon pollution to evaluate how these factors influence what can be detected in the spectra. It is recommended that for the FTIR measurements, the orbital sphere be used to maximise the amount of soil being measured and thereby maximise the chance of detecting contaminant in the soil sample.
7. REFERENCES


Degermenci, O., 2001, ‘Title: From the Caspian Sea region to the European market - Feasibility of crude oil and natural gas transport via longdistance pipelines or routes through the Caspian Sea and their further continuation links westward’, Oil Gas – European Magazine, Vol. 27, No. 3, pp. 7.


INVESTIGATION OF SOIL REFLECTANCES FOR DETECTING HYDROCARBON PIPELINE LEAKAGES


Winkelmann, K.H., 2005, On the applicability of imaging spectrometry for the detection and investigation of contaminated sites with particular consideration given to the detection of fuel hydrocarbon contaminants in soil, Unpublished PhD Thesis, Brandenburg University of Technology.

8. APPENDIX

8.1. CIE chromaticity colour matching

Figure 34: The CIE chromaticity diagram (from van der Werff, 2006)