Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water

R&D Technical Report E1-109/TR

M O'Connor, R Buxton

Research Contractor:
Centre for Intelligent Environmental Systems,
Staffordshire University

Publishing Organisation
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury,
BRISTOL, BS32 4UD.

Tel: 01454 624400 Fax: 01454 624409
Website: www.environment-agency.gov.uk

ISBN: 1844322653

The views expressed in this document are not necessarily those of the Environment Agency. Its officers, servants or agents accept no liability whatsoever for any loss or damage arising from the interpretation or use of the information, or reliance upon views contained herein.

Dissemination Status

Internal:	Released to Regions
External:	Publicly Available

Statement of Use
The software developed through this project will be used by the National Library Service to support legal cases against suspected polluters. It will improve the reliability and defensibility of evidence given in court.

Research Contractor
This document was produced under R&D Project E1-109 by:
WRc, Frankland Road, Blagrove, Swindon, Wiltshire, SN5 8YF

Project Manager
The Environment Agency's Project Manager was:
Wayne Civil, National Laboratory Service, Starcross laboratory**,** Staplake Mount,
Starcross, Exeter, Devon, EX6 8PE

Further copies of this report are available from:
Environment Agency R&D Dissemination Centre, c/o
WRc, Frankland Road, Swindon, Wilts SN5 8YF

tel: 01793-865000 fax: 01793-514562 e-mail: publications@wrcplc.co.uk

CONTENTS

Page

APPENDICES

LIST OF FIGURES **Page**

LIST OF TABLES

EXECUTIVE SUMMARY

This Technical Report documents the findings of National R&D project E1-109 “Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water”. Inspired by the conclusions of an earlier project (E1-050) the principal aim was to develop techniques to match Gas Chromatogram traces of fuel oils, leading to an oil type classification and a successful prosecution of the polluter.

Several “proof of method” software systems were developed to a standard that allowed the research team to decide on a reliable methodology. The software was split into two separate components one to tackle the problem of oil type matching, the other to address the question of source matching a specific pollution incident. There are detail differences between the separate components but the central approach adopted was similar in each.

A large number of laboratory-produced gas chromatogram (GC) traces were provided by the Environment Agency (the Agency) [National Laboratory Service] personnel. Some of these were the results of controlled weathering experiments that gave an indication of how an oil type’s chemical composition altered with time. The GCs were later used to define oil type templates that mimic the changes and allow the type matching software to obtain a match in spite of the effects of weathering. A number of case study GC files were also provided by the Agency. These were files relating to past pollution incidents and were used to confirm the accuracy of the software during a period of testing.

An early version of the type matching software has been demonstrated at the Agency laboratories in Starcross. At a later date, the Centre for Intelligent Environmental Systems (CIES), based at Staffordshire University, also hosted a hands-on workshop where agency personnel were given the opportunity to operate both sections of the software. The response of the Agency personnel has generally been positive. Some feedback from this session has been incorporated into the latest version of the software.

The research completed during this program has proved the Artificial Intelligence (AI) approach to be valid when applied to the complex pattern matching problems that form the basis of the intellectual challenge here. There are several areas of the work covered by the project that warrant further investigation.

Key words: oil spill, inland waters, pattern recognition, gas chromatogram, prosecution, type-matching, source-matching.

1. INTRODUCTION

1.1 Background to the study

The work described in this Technical Report was completed as part of a collaborative research project between Staffordshire University and the Environment Agency, entitled "Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water" (National R & D Project E1-109).

The Environment Agency's National Laboratory Service (NLS) use Gas Chromatography to test oil samples. The apparatus produces a graph of signal intensity against time, called a Gas Chromatogram (GC), which can be used to identify particular oil types. GCs for oil samples typically contain a number of distinctive "major peaks", each of which represent a particular chemical component common in hydrocarbon fuels (for example *n*-alkanes). The distinctive patterns of these major peaks enable identification of particular oil types, such as *diesel* or *unleaded petrol*, by comparison with known 'template' patterns typical of the oil type. If a sample GC has a pattern of major peaks that is similar to a standard oil type then a type match has been achieved.

When a pollution incident is reported, NLS staff collect a 'field sample' for analysis and attempt to find the pollution source by comparing the field sample GC with those produced by samples of potential sources (here referred to as 'reference' samples). In this case, there is no 'template' pattern with which to compare the field sample; instead, the 'field sample' is compared directly with each of the 'reference' samples. NLS experts attempt to establish the identity of polluters by looking for patterns in the GC that are common to the field sample and a reference sample representing a suspect.

Currently, NLS staff have to match GC traces purely by eye. This means that matches are potentially very subjective, as there is no standard way of quantifying the closeness of a match. From a legal standpoint this means that obtaining convictions can be a difficult process involving numerous expert witnesses with opposing views. It is clear that a reliable and robust pattern matching system would be valuable to the Environment Agency in their attempts to obtain successful convictions against polluters. In order to assess the possibility of producing such a system, a feasibility study was commissioned. The origins of this project can thus be traced back to National R & D Project E1-050, the results of which were published in National R & D Technical Report E72 (Walley, Robotham & O'Connor, 1999). The feasibility study was carried out by the Centre for Intelligent Environmental Systems (CIES), a research group based in the School of Computing at Staffordshire University. The group, headed by Prof. W. J. Walley, has been applying artificial intelligence (AI) techniques to environmental issues for several years and has produced a number of successful systems for the Environment Agency based on these techniques. It was felt that an AI approach to the problem would be appropriate, given the complex nature of the task.

The main objectives of the feasibility study E1-050 were:

- To determine the feasibility of using neural network techniques of pattern recognition to compare the gas chromatograms (GC) or infrared spectra (IR) of spilled oils with those of standard oils from possible sources of pollution.

- To investigate whether the use of neural network techniques of pattern recognition might increase the reliability and defensibility of oil identification relative to visual methods.

Initial inspiration for the investigation of pattern matching techniques as a suitable method for type and source matching came from the reports of successful *neural network* systems developed for similar problems relating to off shore oil spills and olive oil; Technical Report E72 (Walley, Robotham & O'Connor, 1999) provides a comprehensive literature survey. Artificial neural networks are parallel computing devices consisting of many interconnected simple processors (Callan, 1999) that work collectively to solve problems, and can provide a solution to many complex pattern matching problems. However, a network needs to be *trained* using a dataset containing examples, often several thousand (Mallach, 1994), of the patterns similar to those it is expected to recognise later. The nature of the problem, particularly source matching, is such that there is no practical way of capturing the amount of information necessary to train a neural network to recognise a particular source GC. This means that it is unlikely that a neural network solution could ever be realised. For crude oil spills at sea a wealth of information exists on the characteristics of oils from different regions, making the collection of the large amounts of data required for training neural networks a realistic prospect. Crude oils produce very distinctive GC traces, and there are relatively few potential sources. For refined oils typical of an inland spill situation, it is highly unlikely that a large database of suitable GC traces could ever be compiled. The scope of the feasibility study was therefore widened to include other possible methods.

Several essential characteristics of the problem were identified:

- The pollutant may be one of several types of refined oil (eg. diesel, unleaded petrol etc.), each of which has its own characteristic GC 'fingerprint'.

- A sample taken in the field may have 'weathered' or degraded to the extent that the GC fingerprint is very different from that of the original oil. Thus weathering/degradation makes the matching of field and reference samples more difficult.

- Samples of potential sources are only acquired after the pollution incident has taken place.

- The type of refined oil responsible for the spill can be identified from key features in its GC fingerprint.

- Very little use is currently made of information contained within the mass of poorly-resolved GC peaks or the unresolved complex mixture (UCM), although both appear to contain useful source-specific information that could be used for source identification.

GC fingerprinting methods are machine sensitive, so it is important to analyse field and source samples using the same testing machine and the same methodology, or to use some method to take the consequent variability into account.

Based upon the findings of the feasibility study it was decided that a program of further research should be undertaken and the problem should be addressed in two stages. These stages are referred to as 'type-matching' and 'source-matching' respectively. The first stage would identify the broad oil type (e.g. diesel), whilst the second would attempt to identify the specific source from a set of potential sources. It was suggested that the main peaks would not provide sufficient information for a reliable source match. It was suggested that the areas between these peaks could provide enough additional information to allow a reliable match. The main conclusions outlined in the executive summary of Technical Report E72 were:

- The problem of identifying a refined oil on an inland water is different from and considerably more difficult than that posed by a crude oil spill at sea.

- A proven computer package (EUROCRUDE) is available that has been used to determine the source of crude oil spills at sea.

- There were no appropriate computer packages that Agency staff could use for chemical fingerprinting.

- Techniques were available that could potentially be used to develop a source identification package.

- It is probable that information contained between the major peaks could be used in source matching. Evidence of this was provided by a small pattern recognition exercise carried out using 20 volunteers, where results pointed to the validity of using the intermediate points in matching.

The work described herein was carried out by the Centre for Intelligent Environmental Systems (CIES) in the School of Computing at Staffordshire University, initially under the supervision of Dr. Charles R. Day, and later, the Centre Manager Mark O'Connor. Full time assistance was provided by Research Associate Rob Buxton. Initially, the Environment Agency's Project Manager was Dr. David Gazzard; in the later stages of the project Mr. Wayne Civil undertook this responsibility.

1.2 Objectives of E1-109

The objectives of project E1-109 were directly based upon the findings of E1-050. A three stage approach was adopted. The overall objectives of these stages were:

1. to produce an automated oil-type classification ('type-matching') system;

2. to produce an automated source identification ('source-matching') system; and,

3. to produce an integrated system (Oil Pollution Diagnostic System, OPDS) which could be used by the NLS staff in a litigation situation.

Each stage had several sub-objectives, summarised below.

Stage 1

- Agree a list of oil-types that the system should be able to identify. This should coincide with the most common oil types encountered by NLS personnel.

- Obtain further information with regard to International Standards information on oil composition.

- Acquire the necessary information to construct a database of free sample signatures of each required oil type. The unweathered "free" samples should provide a starting point in the construction of a recognition system.

- Acquire information on the differences in traces due to instrument and test protocol variability.

- Acquire the information to allow the construction of a database of weathered. Several GCs of each oil type would be needed so that the effect of weathering could be determined accurately.

- Develop a number of prototype oil-type classifiers, each using different statistical, neural network or other pattern matching techniques. A comparison of relative performance would enable a methodology to be derived. The testing needed to include both free and weathered samples so that an accurate appraisal could be made.

- Based upon the results achieved. Produce a prototype oil type classifier with enough functionality so as to be demonstrable at a workshop session.

Stage 2

- Compile a Case Study database. It was suggested that this should contain a minimum of 20 incidents. Also, the database should reflect the variety of oil types represented in the Stage 1 databases.

- Develop the source matching system based on the techniques learned from the type matching (Stage 1) part of the project. Testing would use the case study information described above.

- Integrate both systems to produce a single piece of software.

- Provide the user with interfaces that were easy to use, whilst still conveying the complex information needed.

Stage 3

- Evaluate any special requirements that may be needed in a court room situation.

- Demonstrate the system to key NLS personnel at a 'hands-on' workshop; obtain feedback on the prototype from the potential end-users.

- On the basis of workshop feedback, determine any enhancements which could provide a basis for further work outside of the scope of this contract.

1.3 Methodology

Several type-matching methods were explored during the early stages of the project. These provided a good indication of the techniques that were likely to be needed. Some methods established simple matches expressed merely as the numbers of peaks matched. Problems shared by most of the early methods included an over reliance on retention times that could make the methods inaccurate when slight apparatus variances or moderate amounts of weathering were encountered. One such system was a type of Bayesian classifier that was considered. CIES has past experience in developing successful systems using Bayesian techniques, in particular the RPBBN river pollution system currently being developed for the Agency. However, whereas the RPBBN system uses a complex decision-making structure called a Bayesian Belief Network, the classifier uses a simpler approach, but still bases its decision on probabilistic reasoning. It should be emphasized that none of the earlier methods were sufficiently reliable to warrant development into a proposed method per se. The main areas of weakness were an inability to match more than two main oil types (although information had only been provided for Diesel and Unleaded Petrol at the time development) and an inability to cope well with the effects of weathering (only limited weather related data was available at this time). Also, an overall match was expressed simply in terms of the number of peaks that the method managed to match. There were also interface design problems that would have made the systems difficult to use by anyone but the designer.

Using the performance of the earlier methods as a guide, a number of issues were identified that needed to be addressed if the final type-matching system was to succeed:

- The method must be able to recognise the type of oil in cases where the sample has been subjected to some degree of weathering that may have occurred to the sample.

- Techniques must be derived that can cope with the effects of apparatus-based retention time variation.

- A novel system of expressing the quality of a match between a sample and a template pattern must be developed. This must be objective, but flexible enough to allow the user to detect anomalies that may arise during the match.

1.4 Summary of outcomes

The overall aims of the project were achieved. A combined source and type matching system was successfully demonstrated to Agency personnel at a half-day workshop held at Staffordshire University on May 29th 2003. The workshop demonstrated the system to key Agency personnel and provided the opportunity for the development team to receive feedback with a view to further enhancement of the software.

Prior to the workshop a program of testing served to confirm the success of the method. The type-matching component successfully matched 99 out of 100 test sample files, and the source-matching component matched 47 out of 52 field sample files. Further details relating to the testing and results are given in Sections 3.5 and 3.8.

2. PROJECT DATA

2.1 Introduction

The project required a large quantity of good quality laboratory-sourced GC data from the Agency. Data was required for both type-matching and source-matching. The type-matching data consisted of free samples of various different oil types, together with GCs for these samples that were artificially weathered under laboratory conditions. The source-matching data consisted of GCs produced for past pollution incidents, that had been archived as case studies.

2.2 Type-matching data

A type-matching data set was produced over several months by NLS staff at the Environment Agency's laboratories, initially at Fobney Mead, Reading, and later at Starcross. The aim was to determine (under controlled conditions) the way that different oil types degrade with time. Although it is straightforward for humans to differentiate between the patterns that represent two particular oil types, this simple task provided a suitable means to deliver and test methods and software. In order to gain an insight into the effects of weathering a set of laboratory controlled experiments were set up. Weathering is a term that relates to the changes of the chemical composition of an oil due to its interaction with the immediate environment. Weathering is a time-dependent procedure. Usually, lighter components are affected quicker than heavier, more stable ones. Initially the experiments covered two oil types (Diesel and Unleaded Petrol), two suspension media (soil and water) and two lighting conditions (light and dark). The idea being that the laboratory staff would test the samples at regular intervals over a number of months and the GCs generated would contain detailed information on how the structure of the oil had evolved during this period. The changes are mainly due to natural degradation caused by exposure to the elements and light (where applicable). Later in the contract some information on two further oil types (Paraffin and White Spirit) was provided and the requirement for separate dark and soil-based templates was abandoned. It was decided that, given the complexity of the modelling process, the main emphasis would be placed upon developing a set of models (templates) for the required oil types with respect to water samples. Even at this early stage it was obvious that the oil type templates had to be dynamic, in that their characteristics should change to mimic the effects of weathering observed. The following list shows the numbers of Oil Type files received.

Diesel/Light/Water	56 files
Diesel/Dark/Water	51 files
Diesel/Light/Soil	49 files
Unleaded Petrol/Light/Water	41 files
Unleaded Petrol/Dark/Water	53 files
Unleaded Petrol/Light/Soil	24 files
Paraffin	15 files
White Spirit	4 files

The type templates were derived from the sample GC files supplied by the NLS. Firstly the files were analysed by eye. It was obvious that several “features” or distinctive peaks occurred with sufficient consistency to allow them to be isolated. **Fig 2.1** shows a set of distinctive graph peaks derived from the raw data contained in the text files.

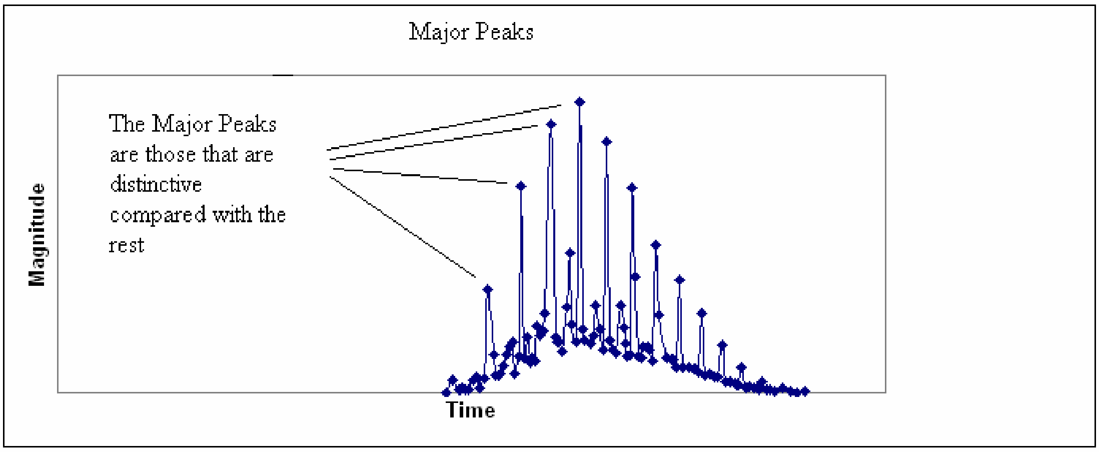


Figure 2.1 Major Peaks

The peaks were considered distinctive when their magnitude (peak height) was clearly in excess of those peaks adjacent to it. Because the samples were processed using the same equipment (and identical test protocols) it was observed that there was minimal ‘drift’ in the retention time when the same ‘feature’ was compared across a number of sample files. Once isolated the selected “features” of each file used in the template were summed for their magnitude allowing the normalisation of the set of distinctive peaks.

The normalised height h_{it} of the i^{th} major peak, after weathering time t, is given by

$$h_{it} = H_{it} \Big/ \sum_{i=1}^{n} H_{it} \qquad \text{Eq 2.1}$$

where

H_{it} = the height of the i^{th} principal peak after weathering time t, and
r_i = the rescaled retention time of the i^{th} principal peak.
T = time at which weathering tests were terminated.

Now, for the purpose of developing a model to predict the effect of weathering on a free sample, it is necessary to derive a set of parameters α_{it} that define the normalised height of each weathered sample “feature” h_{it} as a ratio of the corresponding height of the free sample “feature” h_{i0} (i.e. when $t = 0$), as follows:

$$\alpha_{it} = \frac{h_{it}}{h_{i0}} \qquad \text{Eq 2.2}$$

Fig 2.2 shows how the patterns of "features" change with the weathering time. Each curve is a best fit polynomial for the set of normalised "feature" magnitudes. By fitting curves to the plotted α_{it} values it is possible to derive equations of the form:

$$\alpha_{it} = f_i(t) \qquad \text{Eq 2.3}$$

where f_i is a polynomial function describing the impact of weathering on the i^{th} "feature". Hence, by substituting a value for t in the polynomial function an approximation of the relative states of all "features" can be derived.

Similarly a reference standardised retention time (on the scale 0 to 1) was derived for each "feature" based on the position of the feature in the pure, "free" sample.

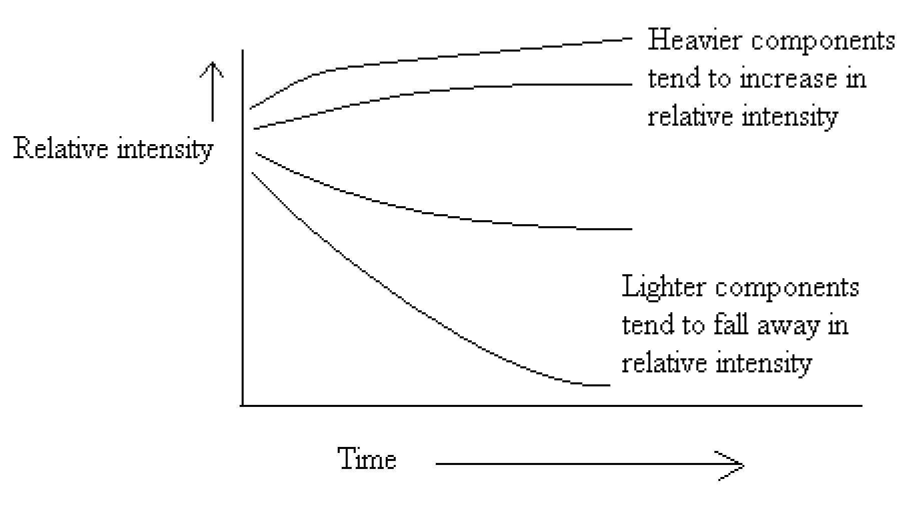

Relative intensity
Heavier components tend to increase in relative intensity
Lighter components tend to fall away in relative intensity
Time

Figure 2.2 Typical weathering curves showing how lighter parts of the mixture gradually drop in relative intensity compared to heavier fractions.

The Gas Chromatograms were supplied as plain text files. The files take the form of a header containing detailed information relating to the timing and duration of the session, followed by lines of information relating to each peak. Each line provides the following information about the peak:

- Peak number(simply a reference to the relative position of the peak in the sequence of readings)
- Retention time (in minutes)
- Peak type, this takes the form of a two letter code relating to the exact method used by the apparatus to derive the peak area
- Peak width
- Peak area
- Peak height (signal strength)

- Percentage of total trace area

Several soil tests were repeated after an evaluation of the data showed that several of the distinctive features showed an unexpected increase in intensity when prior tests had led us to expect a steady drop in relative values. The exact cause of the problem was never fully explained, although several samples were stolen from the NLS laboratories and interference by outsiders is a possible cause.

An example of the GC files is given on page 10, (note it is usual to have several dozen lines of peak information not merely eight as shown here)

```
Data File C:\HPCHEM\2\DATA\030901\003F0301.D Sample Name: 0309013
HP5890 FID/FPD 10/24/01 11:19:02 AM Wayne Civil

Dark-Petrol on water

======================================================================
Injection Date  : 9/3/01 1:37:58 PM              Seq. Line :    3
Sample Name     : 0309013                             Vial :    3
Acq. Operator   : Wayne Civil                          Inj :    1
                                                 Inj Volume : 1 µl
Acq. Method     : C:\HPCHEM\2\METHODS\OIL.M
Last changed    : 8/2/01 9:17:56 AM by Wayne Civil
Analysis Method : C:\HPCHEM\2\METHODS\OIL.M
Last changed    : 10/24/01 11:15:36 AM by Wayne Civil
                  (modified after loading)
WEATHERED SAMPLE DATA ACQUISTION
======================================================================
                      Normalized Percent Report
======================================================================

Sorted By             :      Signal
Multiplier            :      1.0000
Dilution              :      1.0000

Signal 1: FID1 A,

======================================================================
                         Summed Peaks Report
======================================================================

Signal 1: FID1 A,
======================================================================
                      Final Summed Peaks Report
======================================================================

Signal 1: FID1 A,
======================================================================
                         Area Percent Report
======================================================================

Sorted By             :      Signal
Multiplier            :      1.0000
Dilution              :      1.0000


Signal 1: FID1 A,

Peak RetTime Type  Width     Area      Height     Area
  #   [min]        [min]  counts*s   [counts]      %
----|-------|----|-------|----------|----------|--------|
   1   2.821 PV    0.1922 2.61048e5  2.08345e4   0.09482
   2   3.559 VV    0.2696 2.65132e8  1.80253e7  96.30064
   3   3.951 VV    0.1135  854.41058  104.36880  0.00031
   4   5.049 BP    0.1238 1.70997e4  1960.56274  0.00621
   5   6.496 BP    0.1639 2948.48193  291.87094  0.00107
   6   6.961 VB    0.1945 1535.02502  119.01247  0.00056
   7   7.819 BP    0.1496  797.44690   74.08600  0.00029
   8   8.465 BV    0.1737 1.27926e4  1052.90759  0.00465
```

For development purposes, superfluous header information was removed from the files, and they were manually converted to comma-delimited (.csv) format to enable easier file manipulation.

2.3 Source-matching data

The information supplied by the NLS to aid in the development of the source-matching software was based on a series of archive case studies. Originally it was envisaged that the data (plain text files as before) would relate to a sample recovered from the pollution incident, and several alternative reference samples (ie. possible sources of the pollution), enabling a decision to be made based upon the comparison of a set of results. However many of the case studies received only contained one ‘reference’ sample. Where only one potential source existed, samples of the same oil type were added to the data set so that each case had at least two reference samples to choose from. The GC files were the same format as in the type-matching; an additional variable was added to each line to identify which peaks were considered to be ‘major’ peaks. A total of 202 source-matching GC files have been received, a large majority of which are classed as Diesel/Heating Oil.

3. SOFTWARE DEVELOPMENT

3.1 Background

In the early stages of the project the exact mechanism by which a match may be obtained was undecided. Several systems had been produced to allow a match of crude products (see Section 1.1) but no recognised techniques existed for the matching of refined products on inland waters. A number of approaches were discussed and developed, and the final selected techniques - which were then carried forward to the stage of software production - are described in the following Sections. There is a high degree of commonality between the type- and source-matching components. To avoid repetition, features common to both components are described first. Features unique to each component are then described in separate sections. The type-matching system was developed first, since it was considered the least problematic of the two issues to resolve. The techniques were refined by developing this component first, allowing a more efficient design process to be employed on the source matching component. The type-matching system went through several evolutionary stages before production of a fully operational prototype. These stages are discussed here.

3.2 Common features

In both the type- and source-matching components of the system a sample GC is to be compared to some known example file. For type-matching this will be an oil type 'template' as described in Section 2.2; for source matching, it will be a 'reference' sample as described in Section 2.3. The methods that each component will then employ to establish and measure the closeness of the two patterns is almost identical. The main difference between the two is that the type-matching component matches the major peaks and the source-matching component matches the peaks lying between the major peaks. Once a set of peaks has been chosen then the basis of a match is the average rescaled Euclidean distance between each pair of peaks.

The use of Euclidean distance (or a rescaled Euclidean distance) to establish the similarity between two patterns composed of distinct features forms the basis of many neural network pattern recognition systems that are applied in several different problem domains. These include TEXTAL a system for identifying organic proteins, NNIR a system that retrieves images from database dependent upon their content and TNA a system designed to detect explosives in airport baggage. The distance between a pair of peaks *m* and *n*, *D(m,n)*, is easily calculated (Eq.3.1).

$$D(m,n) = \sqrt{((t_m-t_n)^2+((h_m-h_n)\alpha)^2)} \qquad \text{Eq 3.1}$$

where:

t_n and t_m are the retention times for the peaks to be matched.

h_n and h_m are the peaks heights for the peaks being matched.

α is a modifying factor that addresses the difference in scale between retention times and peak/feature heights.

Peak selection was an issue common to both components. The key questions were:

- How are the 'major' peaks defined in the sample GC ?

- What is the process by which a sample peak is matched to a type template feature or reference sample peak ?

The identification of the major peaks is crucial to the operation of both components. In the type-matching component the peaks were directly compared to the values contained within the oil type templates. In the source-matching component the major peaks served as markers, enabling the system to concentrate on the smaller peaks lying between these points. The major peaks of a reference sample GC were identified by comparing the magnitude of each peak to those on either side of it. If the magnitude of the peak in question was more than 2.5 times the average of the four peaks that are adjacent to it, then it was defined as a major peak. An example is given below. The figure of 2.5 was found, in practice, to result in reliable identification of peaks.

Consider the EA GC data file below:

```
 10   7.862 BP    0.1747 6236.68701   592.98181  0.00224
 11   8.461 BV    0.1693  976.49310    85.46042  0.00035
 12   9.151 VV    0.1389  659.64136    71.99341  0.00024
 13   9.515 VV    0.0780 4635.10498   932.21265  0.00167
 14   9.761 VV    0.1076  733.14417   107.39464  0.00026
 15  10.030 VV    0.1033 3766.57983   554.37738  0.00136
 16  10.347 VV    0.1065 6683.00488   944.71783  0.00240
 17  10.933 VV    0.1306 2163.82178   246.16637  0.00078
 18  11.109 VV    0.0768 2089.94604   429.62192  0.00075
 19  11.381 VB    0.0708 1.06178e4  2439.59741  0.00382
 20  12.037 VV    0.1052  589.52356    84.65962  0.00021
 21  12.224 VV    0.0644 1867.66138   449.80182  0.00067
 22  12.411 VV    0.0692 9562.60547 2096.70386  0.00344
 23  12.566 VV    0.0563 3075.05054   809.88025  0.00111
 24  12.660 VV    0.0590 2259.29590   560.94391  0.00081
 25  12.785 VV    0.0885 1342.60828   216.09915  0.00048
 26  13.059 VV    0.0627 6618.85254 1654.00073  0.00238
 27  13.336 VV    0.0586 1.56122e4  4273.26758  0.00562
 28  13.480 VV    0.0572 1737.69067   492.38974  0.00063
 29  13.593 VV    0.0675 7666.79736 1882.49133  0.00276
 30  13.805 VV    0.0604 8468.81348 2039.74707  0.00305
 31  14.020 VV    0.0571 9284.29883 2400.96240  0.00334
 32  14.167 VV    0.0782 1.25114e4  2204.34253  0.00450
 33  14.363 VP    0.0616 1544.83032   363.71082  0.00056
 34  14.615 VV    0.0447 3.28901e4  1.17169e4   0.01184
 35  14.713 VV    0.0595 7315.77881 1798.74622  0.00263
 36  15.034 VV    0.1395 1.54383e4  1559.59155  0.00556
```

```
37  15.344 VV    0.0640 1.69867e4  4130.77148  0.00611
38  15.451 VV    0.0535 5306.83496 1491.25964  0.00191
39  15.537 VV    0.0550 1.47886e4  4014.49634  0.00532
40  15.721 VV    0.0654 1.84101e4  4023.30566  0.00662
```

The file is in fixed format but the order of information is consistent with that explained in Section 2. Hence, the peak heights are represented by the sixth figure on every line. This information is plotted graphically in Fig 3.1.

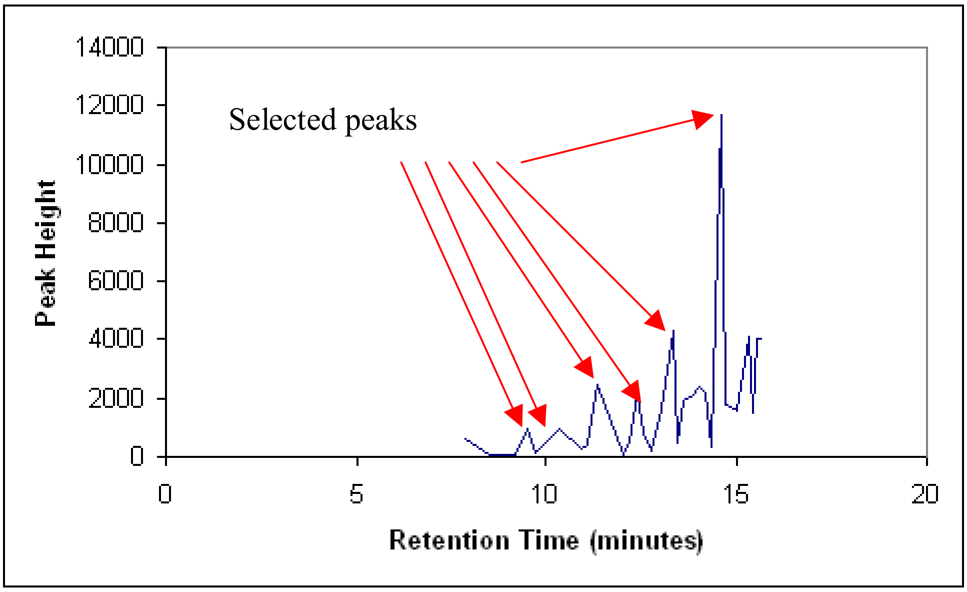


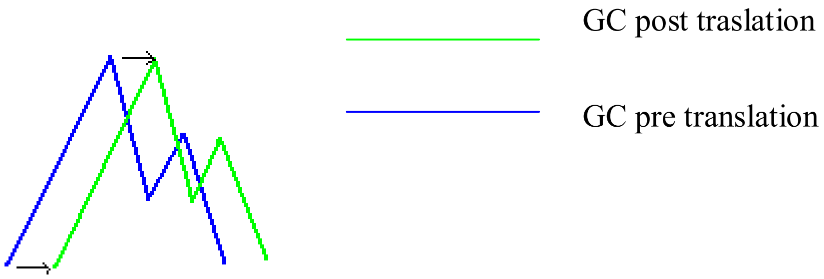
Figure 3.1 GC fragment

If the calculation outlined above is applied to every peak height listed in the file fragment peak numbers 13, 16, 19, 22, 27 and 34 are identified as being ‘Major Peaks’.

It is important that the reader fully understands the process by which the overall ‘match’ is derived. The ‘Difference’ figure displayed by the software is derived by averaging all the peaks deemed ‘matched’. As a result, the overall difference is very much dependent on the closeness of one pattern to the other. A set of design features, common to both components, was established at an early stage of software development whilst these remained broadly unchanged throughout the course of the project, several were modified following consultations with Agency personnel. As has been shown in Section 2.2 the raw GC data files provided by the Agency contain positional information (times) and magnitude information (height/area). The measurements provided are absolute in nature. Before the matching process can begin several modifications are needed to this raw data so that a more reliable match could be attained.

When considering the positional (retention time) information the main apprehension was that the retention time values could be affected by the apparatus calibration, apparatus age, the test protocols used or a combination of all three. If the position of the peaks to be matched vary due to the problems mentioned above then the problem presented to the software system becomes even more testing. The Agency was asked to provide a small number of GCs using alternative apparatus so that a direct comparison could be made and some idea of the possible temporal variation could be judged. One method used to help reduce the effect of possible variations between GC files being matched was to rescale the retention times of all peaks so that they lie between 0 and 1. In this way the system considers the relative position of a peak rather than the absolute position. When the results of the alternative apparatus tests were studied they showed that although, to the human observer, the patterns could be identified as being similar, the peak positions were subject to changes in position *and* relative position. When the weathered information files are studied closely the displacements are still there, however, the effects are many times smaller. In order to counteract the temporal variations observed it was decided that the system should have the ability to modify the GC, in this way the system would be matching based on a more level playing field.

Two transformations are available to help align the sample pattern with a template or reference sample, a further two may be needed to achieve a match in the vertical orientation of the patterns. The two horizontal transformations are a 'translation' or a 'stretch', both of which are applied to the sample peaks.



Fixed difference between GC peaks before and after translation regardless of their position in the trace

Figure 3.2 Translations applied to a GC fragment

If a translation is applied to a sample GC trace then each peak has it's position changed by a chosen amount. There is no change in the distance between points appearing in the trace. The relative positions of peaks in the trace remain the same.

Each peak in the GC is subject to a displacement that is proportional to its distance from the origin (zero point)

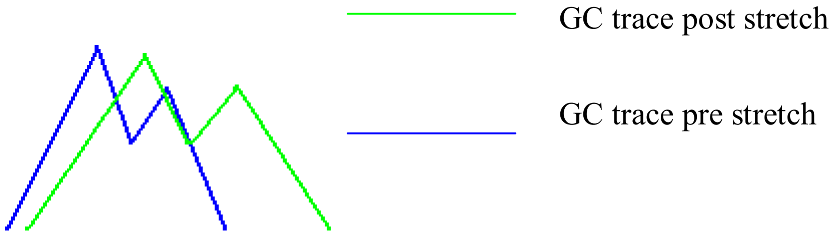


Figure 3.3 Horizontal stretch applied to a GC fragment

A horizontal stretch is applied to the sample trace. This is a more complex transformation than the translation. Experience has shown that in most cases the best horizontal alignment is achieved by combining stretches and translations.

The raw peak height information also needs to be considered as a possible cause of matching inaccuracies. The peak heights in the file are dimensionless but are based on the detector signal intensity. But the absolute values can easily be affected by the amount of pollutant in the sample that is being tested. Obviously, the higher the amount of material the higher the peak height readings. In order to reduce the effect of the sample size on the peak heights all the peaks are normalised based on an initial estimate of the major peaks in the sample file.

If the GC is stretched vertically the magnitude of the peaks is changed in proportion with the peaks height above the baseline

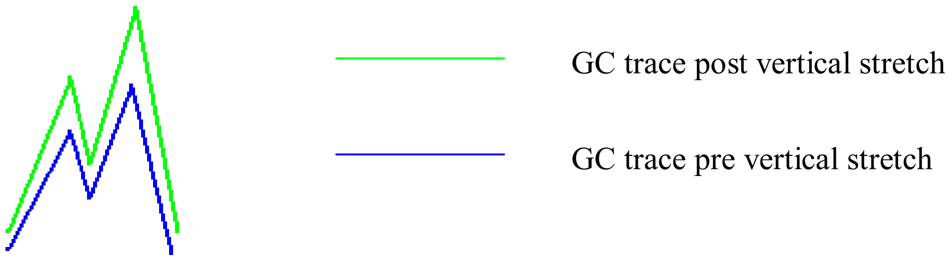


Figure 3.4 The vertical stretch

In addition to the rescaling effect caused by the normalisation of the peak heights it was found necessary to include a facility to stretch the sample GC vertically in order to counter variations in the normalised peak heights.

The final type of transformation relates only to type-matching. The effect of weathering applies to the oil type template. As discussed in section 2.2 the template information includes information relating to the weathering characteristics of the particular oil. Substituting a ‘time’ in a polynomial weathering equation gives the template approximation of the major peak heights.

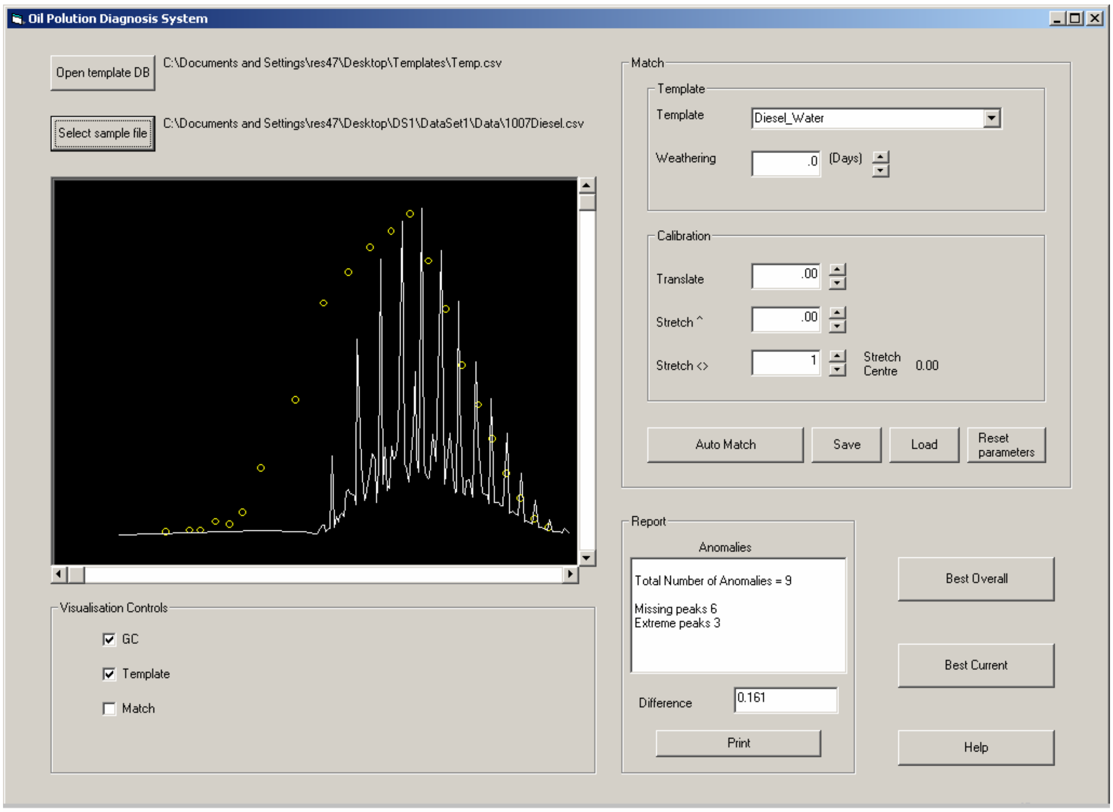
3.3 Type matching user interface development

The type- and source-matching software systems were developed using Microsoft Visual Basic Version 6 (VB6). VB6 is suitable for rapid application development since it allows the programmer to produce functional prototype user interfaces within a short development time period. The development process is described here in approximate stages (although it should be noted that there was considerable overlap between these stages). To make description easier development prior to the initial meeting and software demonstration at Starcross (22nd October 2002) is referred to as Phase 1. Phase 2 refers to development from that time up to the software demonstration and workshop held at Staffordshire University (29th May 2003) . Phase 3 refers to subsequent developments based on Agency feedback from the workshop.

3.3.1 Phase 1

The Phase 1 type-matching system was the first version of the system to have any useful functionality, since prior to this the questions concerning the matching techniques to be employed were largely unresolved. When considering the user interface many functions existed on the Phase 1 version that were carried over to the later versions. As in later versions the user could utilize the system in one of three ways.

Firstly they could undertake a manual matching session. This is where the visual display screen (Fig 3.5) is used in conjunction with the “Difference” textbox and the “Anomalies” display to give an indication to the user of the quality of the match achieved. The sample is represented by the unbroken graph line, the template features are represented by circles. A well-aligned match is achieved when the distance between the features and the major peaks in the sample is minimized. The alignment is achieved by applying transformations to the sample, and weathering effects to the template. The transformations available are translations, where the user applies a fixed shift to the retention time, a horizontal stretch where the shift increases as the retention time increases or a vertical stretch where the magnitudes of the sample peaks are increased in proportion to their heights. In order to simulate a degree of weathering applied to a template the user specified a weathering time in days and this action changes the template characteristics accordingly. At any time during the session the user could view the best match achieved, they could change the template to attempt to achieve a better match or they could view the best match achieved for the particular template selected.



C:\Documents and Settings\res47\Desktop\Templates\Temp.csv

C:\Documents and Settings\res47\Desktop\DS1\DataSet1\Data\1007Diesel.csv

Template: Diesel_Water

Weathering: .0 (Days)

Translate: .00

Stretch ^: .00

Stretch <>: 1

Stretch Centre: 0.00

Total Number of Anomalies = 9

Missing peaks 6
Extreme peaks 3

Difference: 0.161

☑ GC
☑ Template
□ Match

Figure 3.5 Phase 1 software, prior to meeting

Secondly, the user could let the software automatically find an optimum match from the available type-templates. This process took several minutes to achieve because of the computationally expensive process of the software repeatedly applying transformations to the sample, and weathering effects to the all the templates until a set number of transformations had been achieved. At this point the best match was displayed.

Thirdly the user could start by matching the sample manually and finish the match automatically in order to “fine tune” the result.

Fig 3.5 shows the system as presented to the Agency personnel at Starcross laboratories. This system had a substantial proportion of the functionality described for the current system and was used in the initial testing program.

3.3.2 Phase 2

The Phase 2 system benefited greatly from the results of a meeting with Agency personnel at Starcross laboratories. Fig 3.6 shows the main screen. The user interface has several minor improvements over the Phase 1 version (Fig 3.5). Most of the changes implemented in the Phase 2 version were to do with usability requirements and key functionality issues. These are discussed in more detail below.

Modifications were made to the three main (Diesel, Unleaded and Paraffin) templates in order to improve their accuracy. A program of testing prior to the meeting had indicated that improvements could be made, the tests indicated that the templates were more likely to produce a misclassification or an inaccurate weathering prediction in certain areas. Closer examination showed that that in these areas one or more of the curves that made up the oil type template deviated from the results of the weathering experiments. The accuracy in these areas was increased by modifying the polynomial weathering equations so that they more closely fitted the curves dictated by weathering data.

Agency personnel specifically requested that the system should have a facility to match using unmodified retention times. As previously mentioned the retention times of the peaks in the sample are rescaled upon loading to a scale of 0 –1. Whilst Agency staff felt that this was useful they suggested that in some cases laboratory staff might also require to use the retention times in minutes. Hence, the software was modified extensively to allow the user to switch between the original “real” retention times and the rescaled values.The Agency staff also asked if an alternative to “Days” could be provided when expressing the extent of the perceived weathering. When giving an opinion on the extent of weathering in a report Agency staff tend to use expressions like “fairly fresh” or “slightly weathered”. They expressed an interest in using a measure of weathering based on a percentage scale relating to the amount of volatile mixture still present in the sample. In order to add this feature to the software the Agency provided the development team with information on the ‘percent removed’ for a limited number of oils compared with time. Based upon this limited information the development team were able to add a facility that automatically converts the weathering time to a percentage.

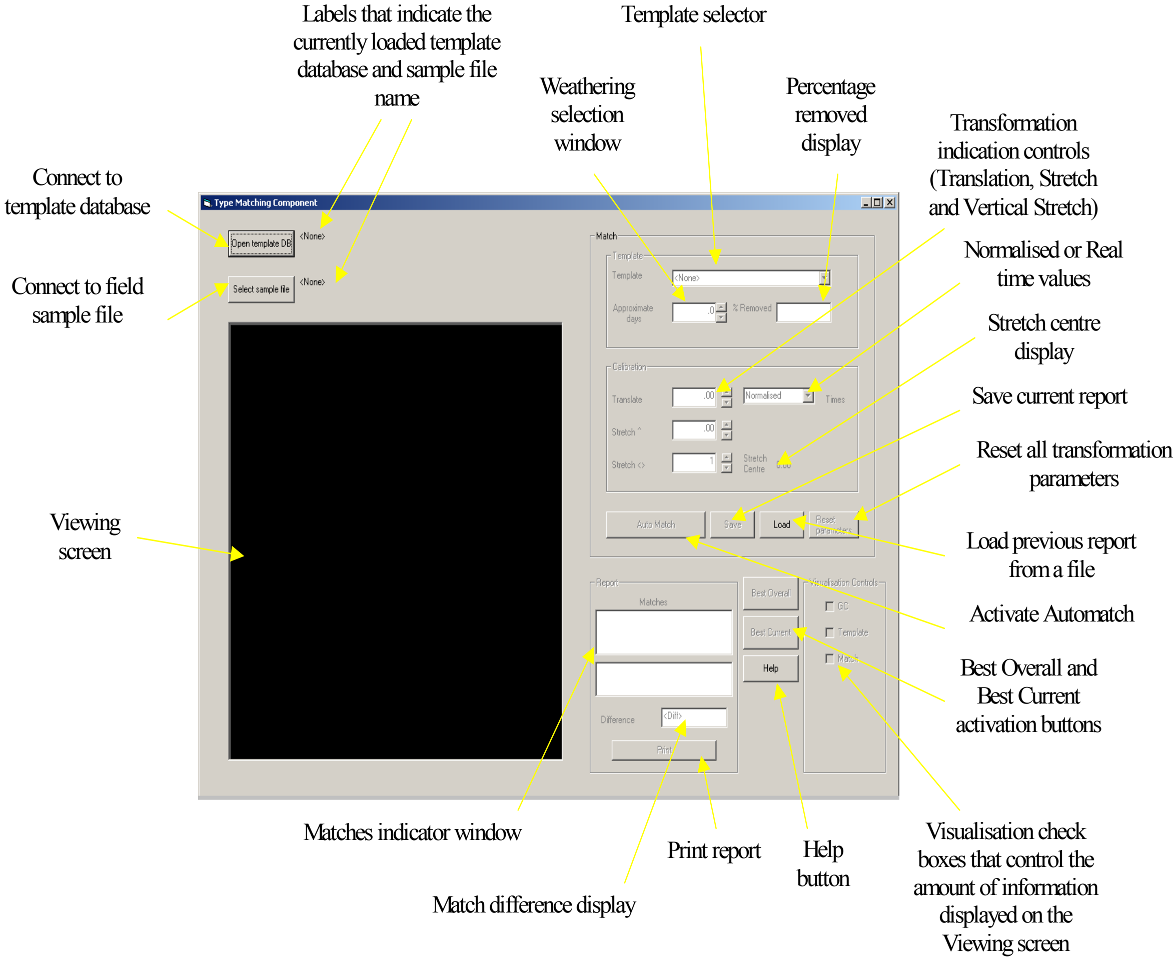

Labels that indicate the currently loaded template database and sample file name
Template selector
Weathering selection window
Percentage removed display
Transformation indication controls (Translation, Stretch and Vertical Stretch)
Connect to template database
Connect to field sample file
Normalised or Real time values
Stretch centre display
Save current report
Reset all transformation parameters
Viewing screen
Load previous report from a file
Activate Automatch
Best Overall and Best Current activation buttons
Matches indicator window
Print report
Help button
Visualisation check boxes that control the amount of information displayed on the Viewing screen
Match difference display
Type Matching Component
Open template DB
<None>
Select sample file
<None>
Match
Template
Template
<None>
Approximate days
.0
% Removed
Calibration
Translate
.00
Normalised
Times
Stretch ^
.00
Stretch <>
1
Stretch Centre
0.00
Auto Match
Save
Load
Reset parameters
Report
Matches
Difference
<Diff>
Print
Best Overall
Best Current
Help
Visualisation Controls
□ GC
□ Template
□ Match

Figure 3.6 OPDS Phase 2 ‘Main’ screen functions

3.4 Using the type-matching system

Fig 3.6 shows the system prior to the loading of any sample or template files. The visual display is blank and several of the controls are disabled. Common sense dictates that the user must load a sample and template file before they can undertake any kind of matching, since a minimum of two patterns is needed to form a constructive match.

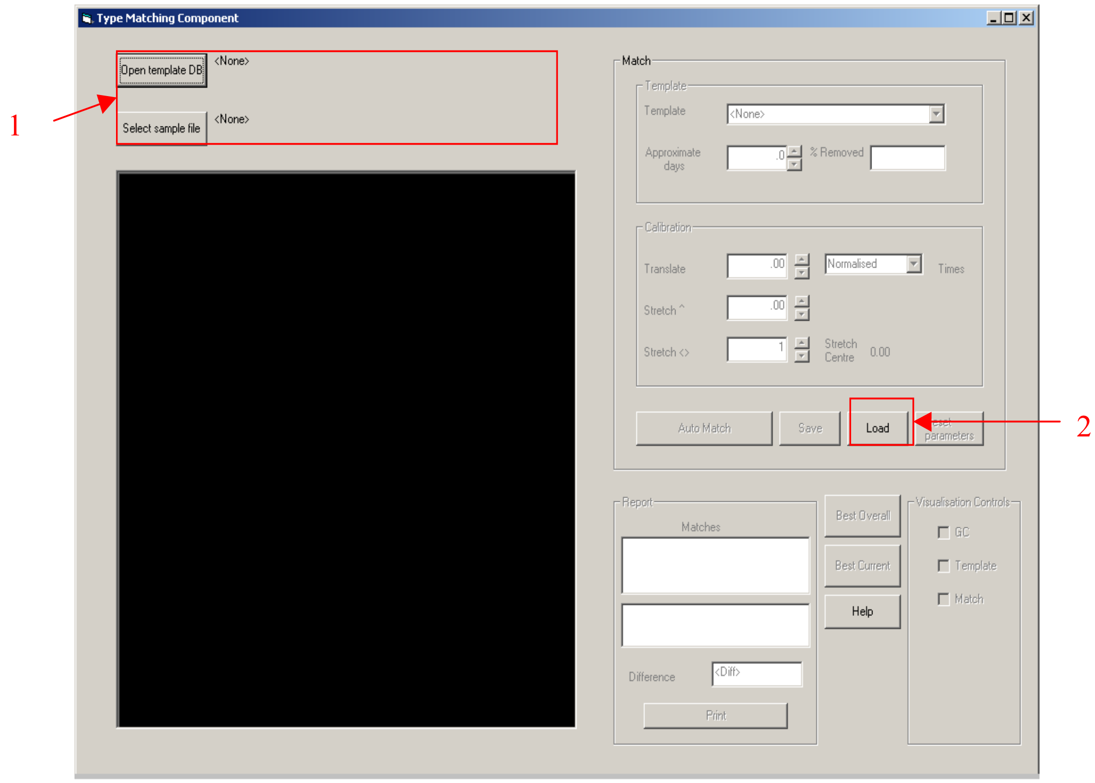
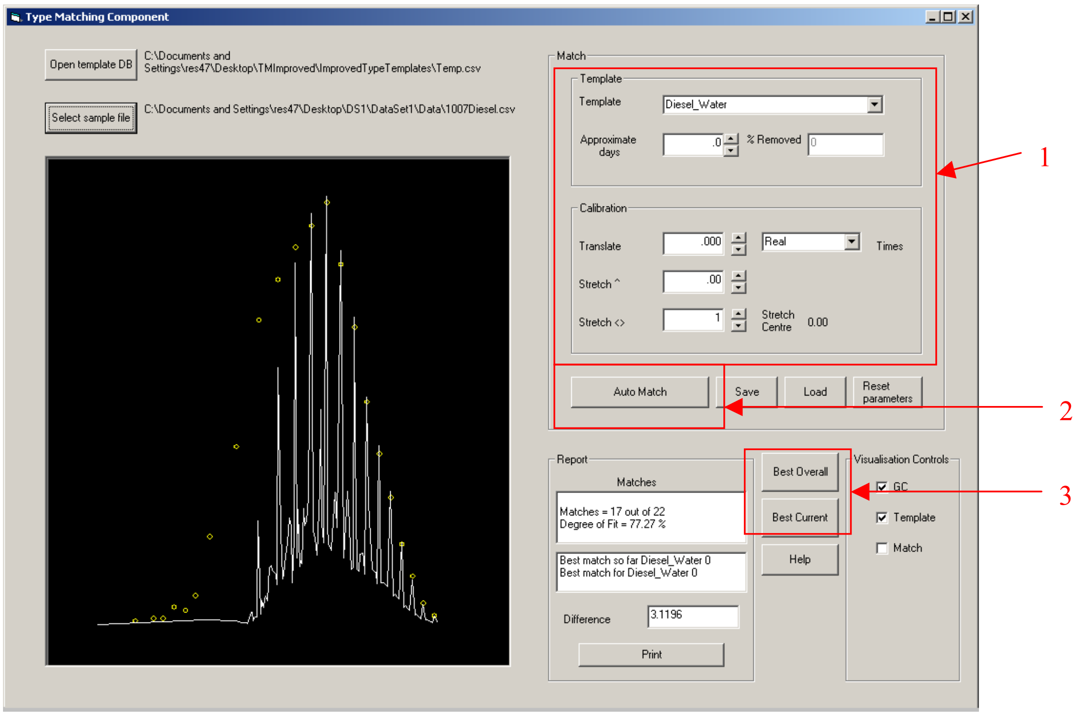


Figure 3.7 Type-match loading options

The user has two main options. The most common procedure would be to open a template database and load a suitable sample file (Fig 3.7 area 1). Here the user navigates to an appropriate template file and sample file using a windows-type dialogue box that most users should find familiar. Once the user has connected to a suitable template file and sample file the name of each file is displayed by the appropriate label next to the appropriate button ("Select sample file" and "Open template DB"). The sample and template files can be loaded in any order. When the second file is loaded the system automatically establishes an 'initial match'. In doing this the system undertakes a complex set of procedures, explained below.

Firstly the system defines the 'eligible' peaks for matching in the sample. This is achieved by considering each peak in relation to those within a predefined time cut-off either side of the peak in question. The considered peak's magnitude is compared to the standard deviation of the heights of the collected peaks. The peak is considered 'eligible' if its height exceeds two standard deviations. Hence we have a collection of 'eligible' peaks that we can use to establish an initial match. The calculation of the match involves the system considering each feature (for the selected template) in relation to each eligible sample peak. The best matches according to the modified Euclidean distance are saved, and then any ‘double selections’ ie. cases where a particular sample peak is matched to more than one template feature are sought out and reallocated. The total ‘difference’, ie. all the match distances totalled is then divided by the total number of matched features. Please note this is not automatically the total number of template features because some maybe none-matched. If none-matched features are present a ‘penalty’ for each such feature is added to the total prior to the division.

The second option available to the user when loading is to load a session that has been previously saved (Fig 3.7 area 2). Activating this control allows the user to navigate to a session file (.csv type) and loading one of these files recreates the saved session, allowing the user to modify the match using the transformations, if need be.



Type Matching Component

Open template DB: C:\Documents and Settings\res47\Desktop\TMImproved\ImprovedTypeTemplates\Temp.csv

Select sample file: C:\Documents and Settings\res47\Desktop\DS1\DataSet1\Data\1007Diesel.csv

Match

Template

Template: Diesel_Water

Approximate days: .0

% Removed: 0

Calibration

Translate: .000

Real

Times

Stretch ^: .00

Stretch <>: 1

Stretch Centre: 0.00

Report

Matches

Matches = 17 out of 22
Degree of Fit = 77.27 %

Best match so far Diesel_Water 0
Best match for Diesel_Water 0

Difference: 3.1196

Visualisation Controls

☑ GC

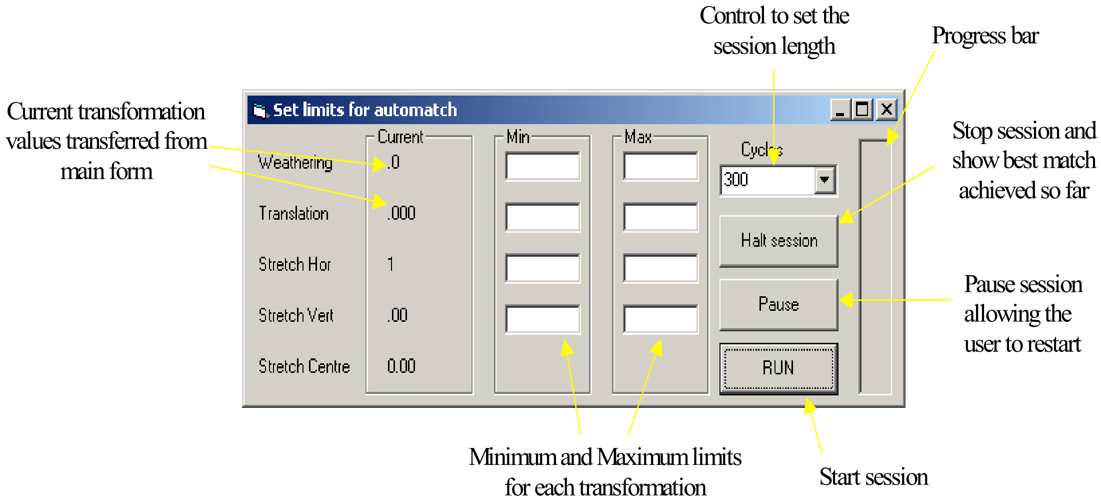
☑ Template

□ Match

Figure 3.8 Type-match session options

The user can use the manual controls (Fig 3.8 area1) to orientate the sample peaks and the template features so that a match is attained. The controls included in this section include the template selection mechanism allowing the user to select any of the templates contained within the database for a manual match, the transformation controls, the weathering control and the control that determines whether the retention times in minutes, or a set of index values, are used for the match. When the user alters a value in one of the transformation (translation, stretch or vertical stretch) windows the system automatically recalculates the sample retention times and/or peaks heights. Then it finds the best match for each template feature (out of all possible peaks), it removes any double selections as before and displays the new information on the Viewing screen. When the template weathering figure is changed the process of altering the feature heights to imitate the weathering characteristics of the particular oil type is done by substituting the ‘days weathered’ amount in the polynomial equation that describes that particular oil type (see Section 2 for information on templates). In this way some features are reduced in relative intensity and some are increased depending on the time selected and the nature of the oil that has been modelled.

The user can also decide to initiate an ‘Automatch’ session (Fig 3.8 area 2). Pressing the ‘Automatch’ button displays a small form (Fig 3.9) that allows the user to set limits on the transformations that can be used during the automatic matching session.



	Current
Weathering	.0
Translation	.000
Stretch Hor	1
Stretch Vert	.00
Stretch Centre	0.00

Cycles: 300

Current transformation values transferred from main form

Control to set the session length

Progress bar

Stop session and show best match achieved so far

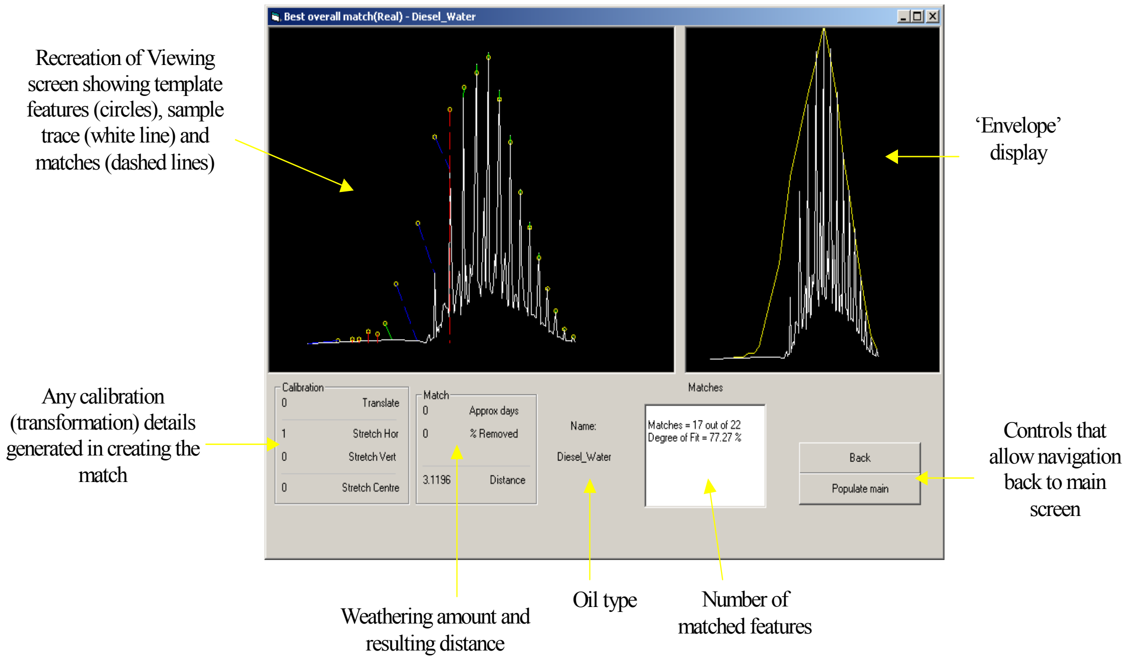
Pause session allowing the user to restart

Minimum and Maximum limits for each transformation

Start session

Figure 3.9 The ‘Limits’ screen

The user can also choose to view the ‘Best Overall’ or the ‘Best Current’ match for that particular template type (Fig 3.8 area 3).



Calibration:

0	Translate
1	Stretch Hor
0	Stretch Vert
0	Stretch Centre

Match:

0	Approx days
0	% Removed
3.1196	Distance

Name: Diesel_Water

Matches = 17 out of 22
Degree of Fit = 77.27 %

Recreation of Viewing screen showing template features (circles), sample trace (white line) and matches (dashed lines)

‘Envelope’ display

Any calibration (transformation) details generated in creating the match

Controls that allow navigation back to main screen

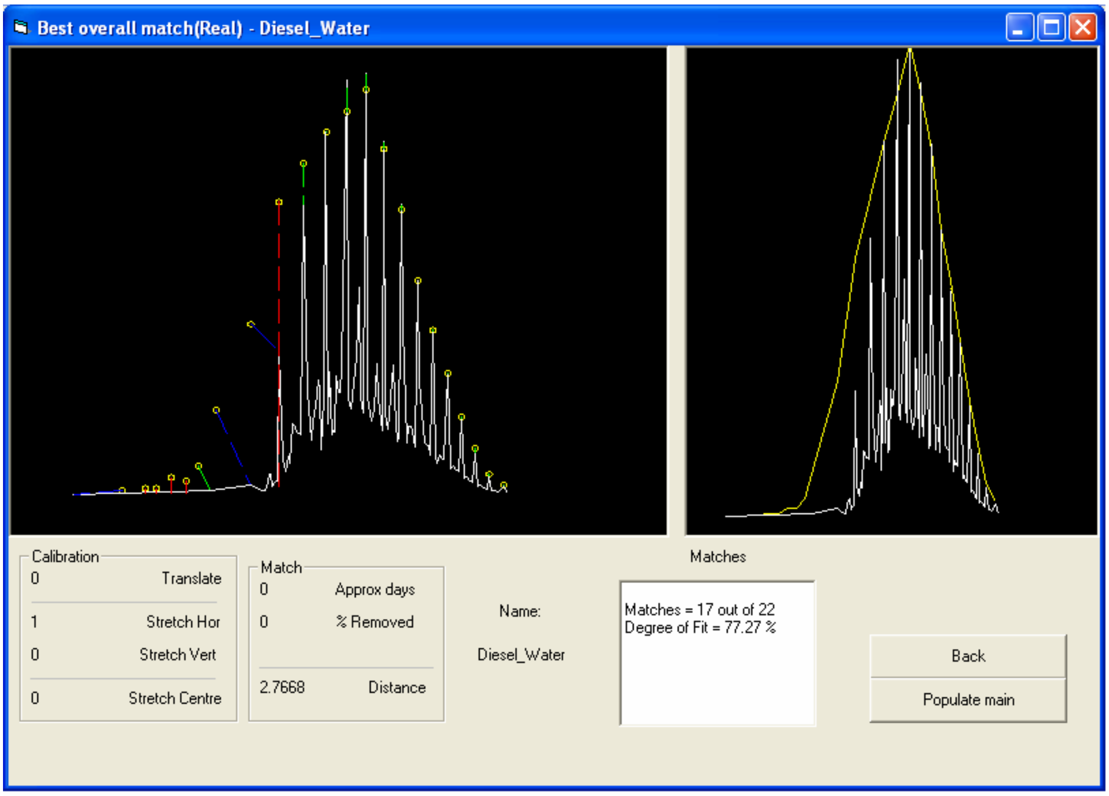
Weathering amount and resulting distance

Oil type

Number of matched features

Figure 3.10 ‘Best Overall’ screen (‘Best Current’ identical layout)

The user can activate the 'Best Overall' screen at any time during a manual session in order to view the best available match. This screen is also automatically displayed at the conclusion of an 'Automatch' session, and when the user loads a previously-saved session file. Unlike the main screen the 'Best Overall' screen has two viewing displays. The first reproducing the display from the main screen but with 'matches' added. The second allows the user to decide on the quality of the match by displaying the 'envelope' created by connecting the template features, and overlaying the sample trace. The idea of this secondary display is to give the user a better chance of identifying a mismatch. This type of display achieves this by giving the overall pattern more of an emphasis over the individual matches that are prevalent in the main display. The user should choose to accept or ignore the advice of the system based upon the balance of the displays.



Best overall match(Real) - Diesel_Water

Calibration
0 Translate
1 Stretch Hor
0 Stretch Vert
0 Stretch Centre

Match
0 Approx days
0 % Removed
2.7668 Distance

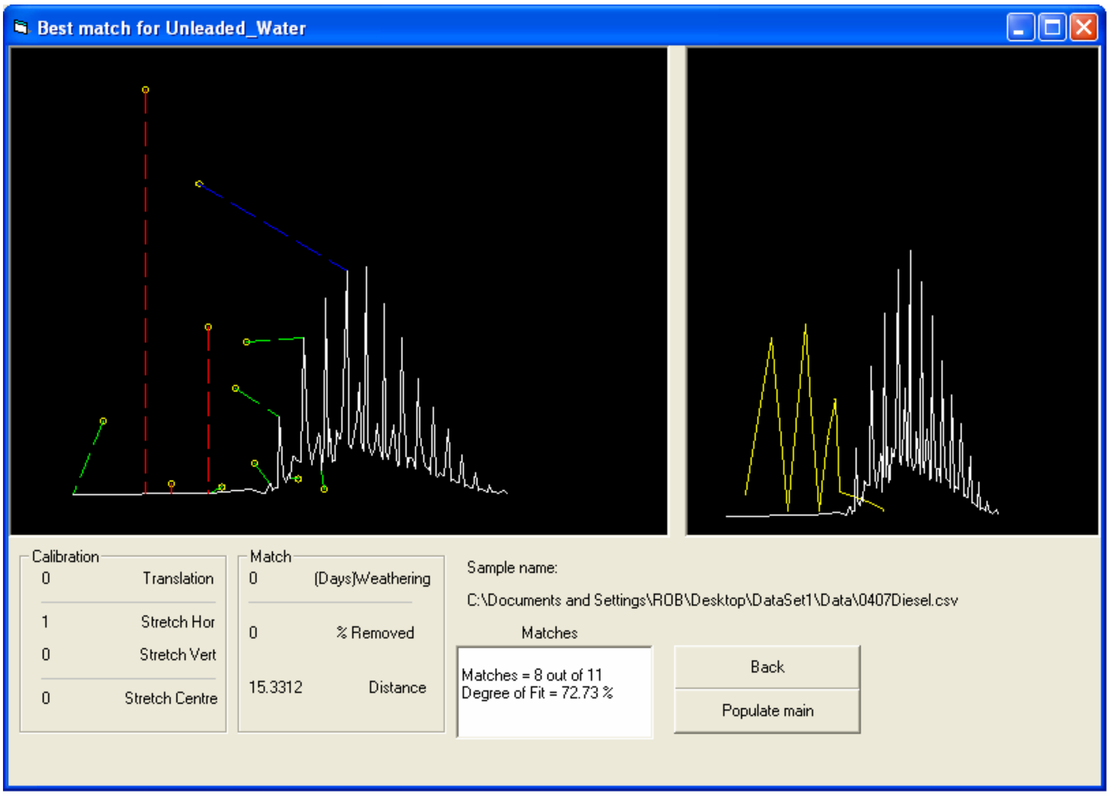
Name:
Diesel_Water

Matches
Matches = 17 out of 22
Degree of Fit = 77.27 %

Figure 3.11 A 'good' match

Figure 3.11 illustrates the type of match that the user should be seeing at the end of a session. The individual matches (shown by the right hand display) are generally acceptable at the higher end of the scale however, the substandard matches of the earlier features indicate that some weathering would generate a better match in this case. The envelope in the left hand display is well matched indicating that the template choice is good.

When this is compared with Fig 3.12 the envelope is displaced compared with the sample trace indicating that it would be unlikely that any reasonable amount of transformation could result in an acceptable match.



Best match for Unleaded_Water

Calibration

0	Translation
1	Stretch Hor
0	Stretch Vert
0	Stretch Centre

Match

0	(Days)Weathering
0	% Removed
15.3312	Distance

Sample name:

C:\Documents and Settings\ROB\Desktop\DataSet1\Data\0407Diesel.csv

Matches

Matches = 8 out of 11
Degree of Fit = 72.73 %

Figure 3.12 A 'poor' match

3.5 Type match testing

A program of informal testing has been ongoing since the early stages of development. This is crucial in locating areas of the software affected by "bugs". Prior to the meeting at the Starcross laboratories the Phase 1 software was subjected to a more complete series of tests that produced extremely encouraging results, further confirming the suitability of the overall pattern-matching concept. The results of these tests are shown in Table 1. The testing results presented in Table 1 represent the performance of the system as at 30/10/02. Testing was as thorough as possible at that stage, although there were only a small number of samples available for some oil types.

		Sample								
		D	D/HO mix	Other D mix	U	P	WS	LFO	LFO/P mix	K
No. of Samples		54	56	7	43	15	4	4	10	1
% match		100	96.4[h]	85.7[h]	83.7	93.3	100	-[j]	80[i]	-[j]
Template[b]	D	0.3068	1.0400	1.7336	2.6131	2.7485	2.2203	1.7636	1.8761	1.9832
	U	5.3617[c]	4.2237	4.5885	1.1068	5.6314	5.1714	2.8079	2.4896	4.2897
	P	3.5474[d]	2.8906	3.5869	3.5462	0.8788	3.7890	1.8995	1.4714	2.6294
	WS	9.8653[e]	9.4989	10.218	9.2616[f]	9.2549[g]	0.3887	4.7295	4.4780	8.1155
w'thering[a]		41.056	-[k]	-[k]	35.48	40.22	21.34	-[k]	-[k]	-[k]

Table 1 Results of Phase 1 testing of type-matching software

Notes on Table 1:

D: diesel, HO: heating oil, U: unleaded petrol, P: Paraffin, WS: white spirit, LFO: light fuel oil, K: kerosene

a) Weathering error calculated on correct matches only; see Eq. 3.2 for weathering error evaluation

b) Actual differences × 100 (for easier formatting and comparison) – a relatively low figure signifies a better match.

c) One diesel file could not be matched with the unleaded template

d) One diesel file could not be matched with the paraffin template

e) Six diesel files could not be matched with the white spirit template

f) Three unleaded petrol samples could not be matched with the white spirit template

g) Two paraffin samples could not be matched with the white spirit template

h) Diesel mixtures matched with diesel regarded as 'correct' match

i) Paraffin mixtures matched with paraffin regarded as 'correct' match

j) No 'correct' match for light fuel oil & kerosene (no templates available)

k) No weathering information available for mixtures, light fuel oil or kerosene

In Table 1 "No of Samples" refers to the number of test cases used relating to this particular oil type. Some of the files were the type-matching laboratory weathered files, some were from the case study data provided for the source matching. The "% match" gives the percentage of samples correctly matched to the appropriate oil template. No templates existed for Light Fuel Oil or Kerosene, so it was impossible for these to be correctly matched however, the system should still pick the closest template pattern to the sample pattern regardless. In cases where no template matches the sample pattern the difference value quoted in the best match is relatively large compared to that usually seen by the user, allowing them to check the results further. "Template" shows the average difference of all the appropriate matches. "Weathering" gives the average weathering error calculated according to Eq 3.2. The case study files had no precise weathering information provided, hence this measure could only be calculated for the laboratory weathered files. Table 1 shows that even at this relatively early stage of development results were promising. Although there were misclassifications in some cases these were usually accompanied by an unusual amount of translation, stretch etc. and as such could be readily separated from more valid results.

A measure of the 'severity of error' for a weathering time evaluation has been proposed. The actual error $w_A - w_P$ for actual weathering time w_A and predicted weathering time w_P is only of limited value; for example, an error of 5 days could be regarded as severe in a case where there was no weathering, but insignificant for a case with 100+ days of weathering. Instead, a rough measure of severity of error can be calculated as:

$$E = 100\left(\frac{\ln\left|w_A - w_P\right| + 1}{\ln\left(w_A + w_P\right) + 1}\right) \qquad \text{Eq 3.2}$$

where $w_A \neq w_P$, with $E = 0$ when $w_A = w_P$. This severity varies between 0 (when the actual and predicted weathering times are the same) and 100 (when there is no weathering but some is predicted, or the sample is weathered but no weathering is predicted), with higher values indicating more severe errors. We have found that values above 70 could be regarded as indicative of ‘significant’ error, whilst 40 or below shows a good evaluation - although these guidelines are based simply on our subjective judgement.

Further testing has been undertaken on the Phase 2 software with the joint aims of confirming the successful integration of the new software elements and evaluating the effectiveness of the modifications. The results are shown in Table 2. The testing of the Phase 2 system used a smaller set of files. Only the laboratory weathered files allowed an accurate measure of weathering error (Eq 3.2) to be calculated. Since the principle reasoning behind this testing was to confirm the increased performance of the modified template weathering equations and time was of the essence, only these files were used at this stage.

	Sample			
	D	U	P	WS
No. of Samples	49	34	14	3
% match	100	97.1	100	100
w’thering[a]	30.87	27.76	20.55	37.23

Table 2 Results of Phase 2 testing of type-matching software

It can be seen that modifications to the weathering equations for the main templates (Diesel, Unleaded and Paraffin) and to the software code have increased the accuracy of the type matching process considerably. The results concerning the three main oil types have all shown an increased accuracy of match both in terms of misclassifications and weathering error.

3.6 Source matching

As described in section 3.2, both components follow the same general methodology. The match is based on a modified Euclidian distance (see Equation 3.1), averaged over the number of matched peaks. Unlike the type-matching software however, the match is applied to the areas between the main alkane peaks (Fig 3.13). The process of source matching is computationally more expensive because the number of matches it has to achieve is far more. An average type-matching template may have 15 features to match, a source-matching reference sample may have 90 or more peaks that have to be matched. The principle reason for this is that there are always far more peaks in the GC file than there are identifiable major peaks.

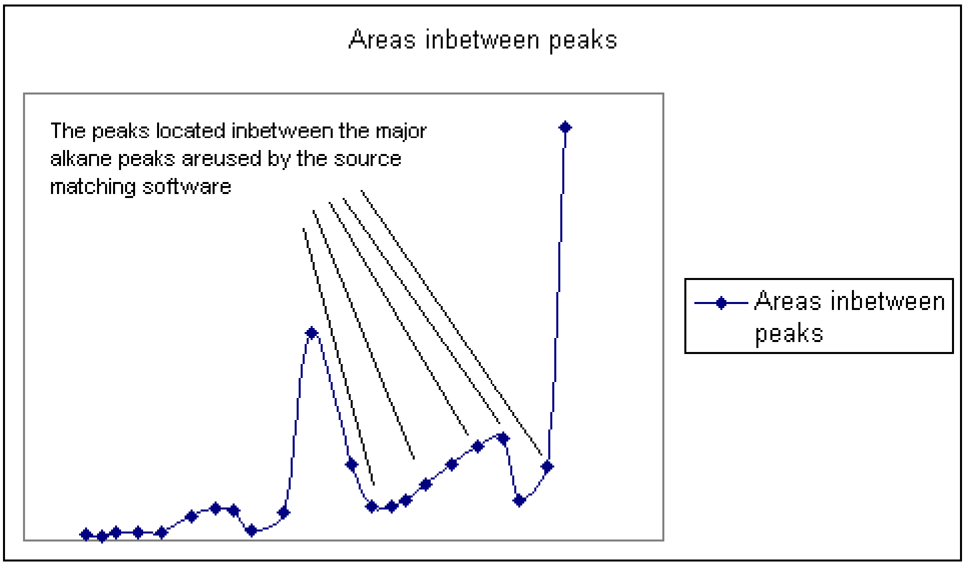


Figure 3.13 GC trace illustrating the areas between the major alkane peaks

3.7 Using the source-match system

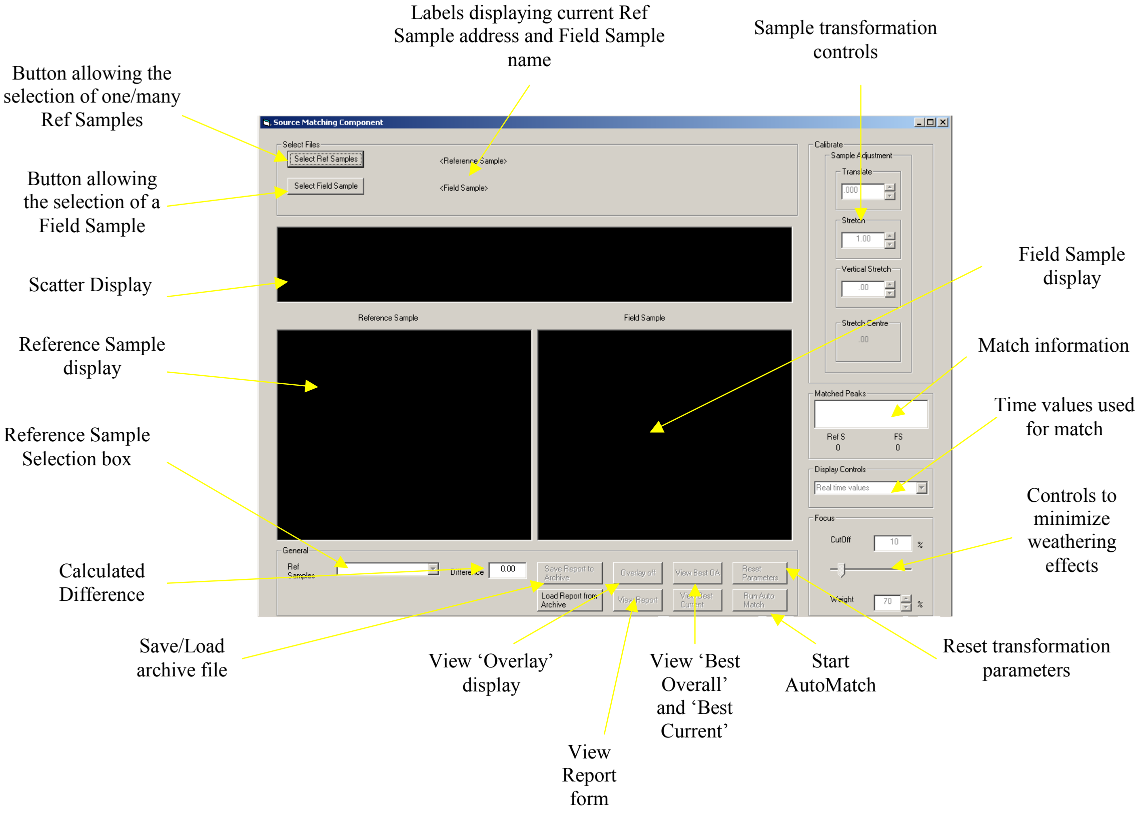


Figure 3.14 The main source-match functions

There are some key differences between the visual design of the type-matching and the source-matching that are worth noting. The most important is that the default form (Fig 3.14) has 3 distinctive viewing screens instead of one. The main reason for this is that the added complexity of the visual match that the system is trying to portray makes it difficult to use a single display without it becoming cluttered. There was a high probability that this degree of clutter could distract the user, impacting on the overall system performance. Hence, the information is split into three components. There is a separate display for the Reference and the Field sample, also there is a 'Scatter' display.

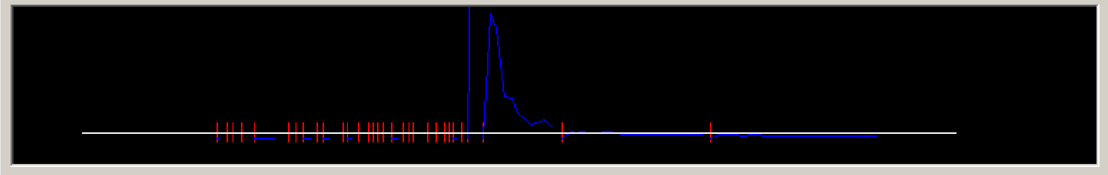


Figure 3.15 'Scatter' display

The purpose of the 'Scatter' display is to allow the user to establish quickly where the major deviations in the two patterns lie. Major differences in the matched peaks are shown by the blue line's distance from the base line. Peaks that are not matched are shown by a red dash breaking the base line.

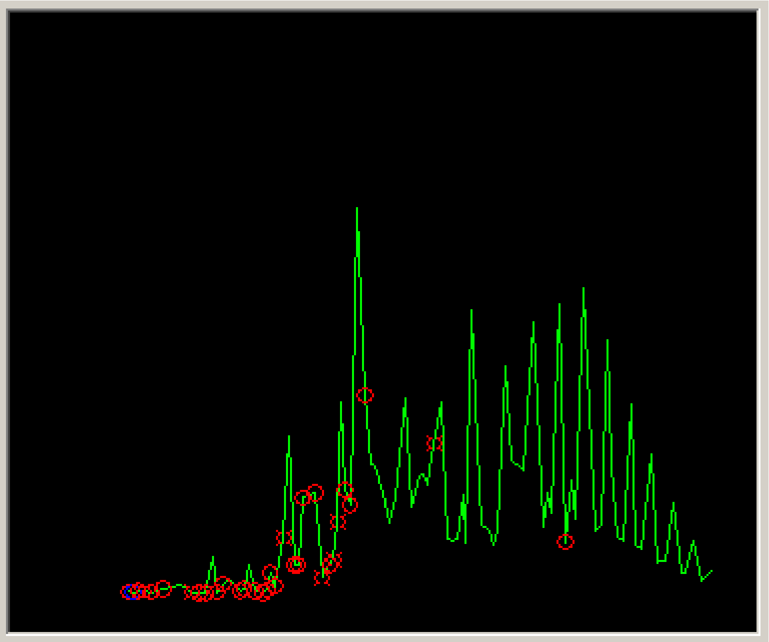


Figure 3.16 The Reference Sample display

The selected Reference sample is displayed as a green GC trace. Red circles indicate the presence of unmatched reference peaks. Blue circles indicate a peak the match is sub-standard compared to the rest.

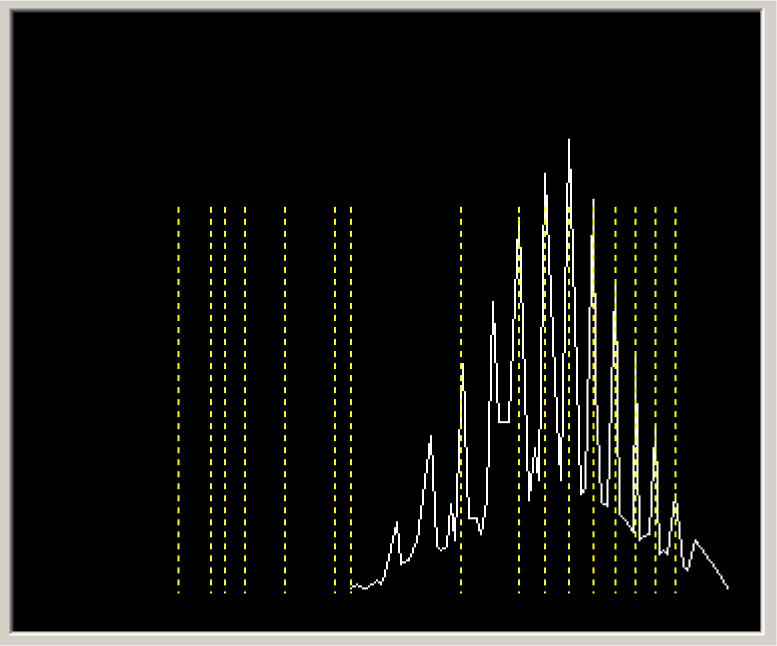
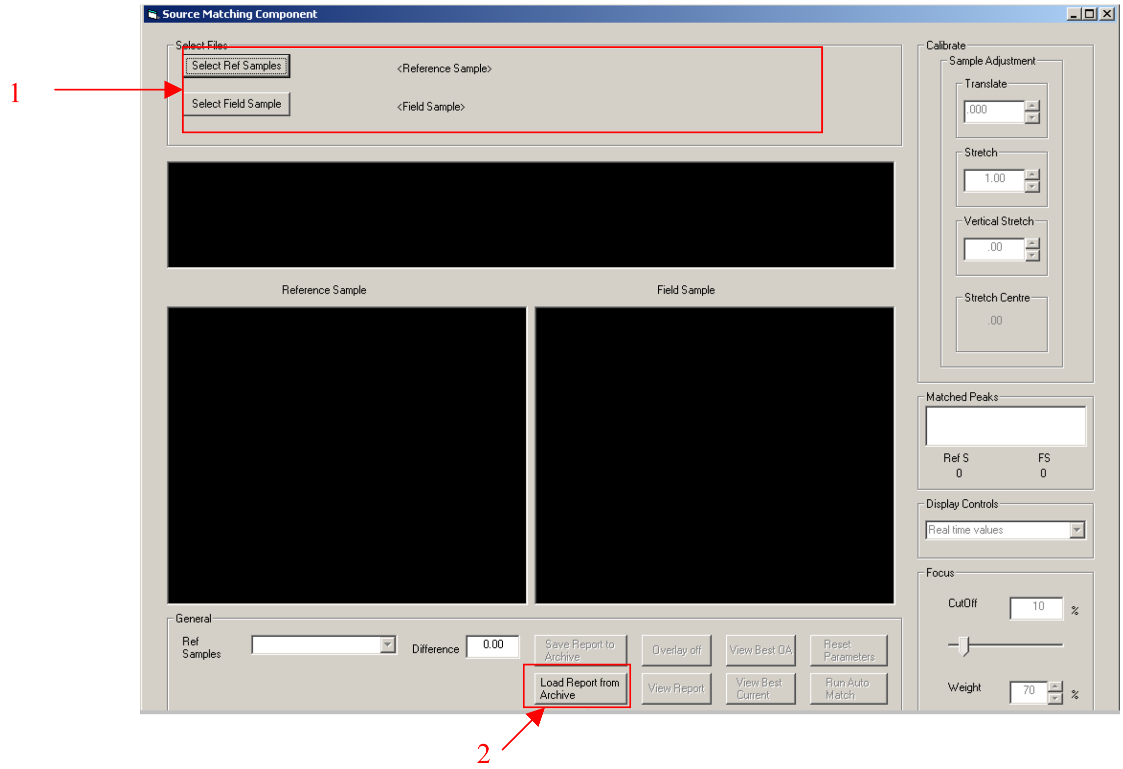


Figure 3.17 The Field Sample display

The Field sample display (Fig 3.17) shows the field sample as a white GC trace. The yellow dashed lines indicate the position of the major peaks as defined by the reference sample file. This allows the used to accurately align the two traces whilst minimizing the confusion incurred by overlaying one GC trace on another permanently. The height of the dashed lines indicate the height of the highest reference 'Major Peak'.

In order to start a source-matching session the user must load one or more reference samples and a field sample, since, as in the type-matching system before, no meaningful match can be established without a sample and a reference sample of some description.

The user can navigate to a reference file or collection of reference files and load these together with a single sample file (Fig 3.18 area 1) or the user can load an archived report file (Fig 3.18 area 2). If the first option is chosen the system creates an initial match in a similar manner to the type-matching component. Once the GCs for consideration are loaded the user can start to match the Field sample to the reference sample(s). There are a number of matching options available to the user. The user can choose to match a sample manually (Fig 3.19 area 1), controlling the process themselves. Or automatically(Fig 3.19 area 2), where the software carries out the match. Or a combination of both. When a manual match is carried out the operator uses the graphical displays on the main screen to align the reference sample GC with the field sample GC. A number of features allow the operator to best decide on the validity of a match. Firstly, on the main screen a textbox marked "Difference" gives the user a "snapshot" measurement of the difference between the patterns at that time. A 'match information' display gives the number of matched peaks. The scatter display (Fig 3.14) gives a visual indication of where major variations in the two patterns are. Some users may also choose to use the "Overlay" facility (Fig 3.19) to get a better idea of how the patterns compare directly**.**



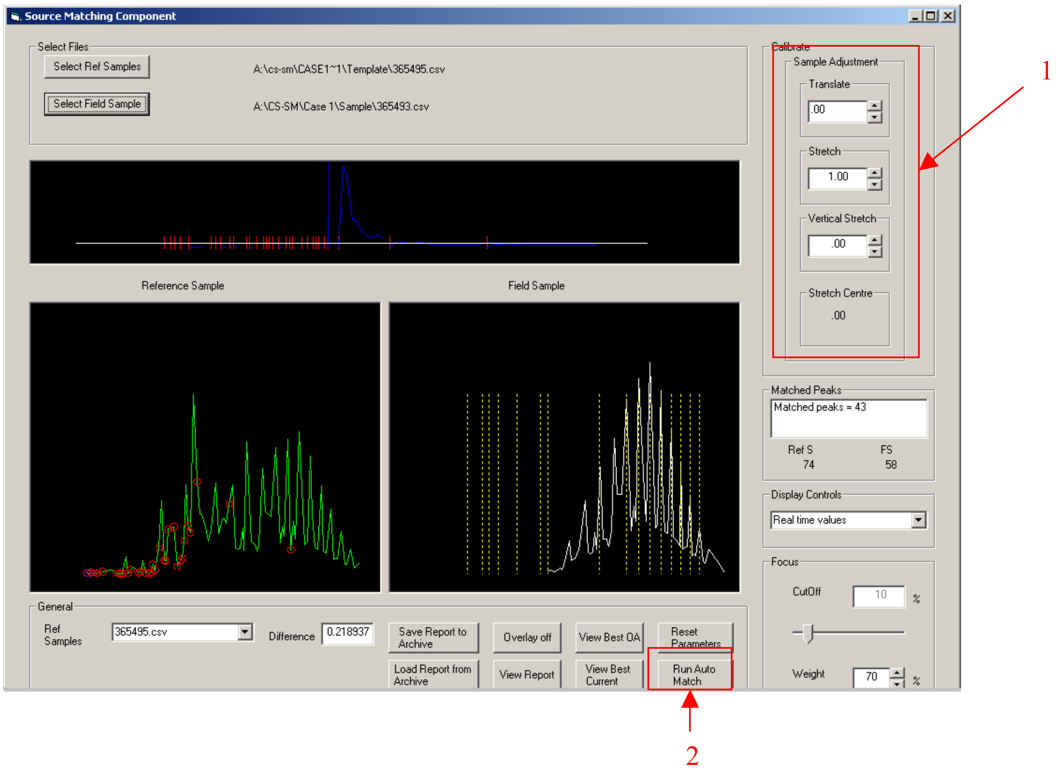
Ref Samples
Difference 0.00
Translate 000
Stretch 1.00
Vertical Stretch .00
Stretch Centre .00
Ref S 0
FS 0
Real time values
CutOff 10 %
Weight 70 %

Figure 3.18 Loading options

In order to orientate the two GCs whereby achieving the best match the user can apply combinations of transformations including translations, vertical stretches or horizontal stretches. The source matching system has no facility to directly allow for the effects of weathering on a sample. However the user can reduce the effects of weathering on a match by reducing the significance of matches achieved in the earlier part of the GC trace.

If the user chooses to let the software carry out an automatic match then it automatically applies transformations to the sample, operating within user-defined limits (Fig 3.20), until a set number of transformations have been undertaken. When the session is complete the system presents the user with the best match.

The user could also decide to start the match manually and fine tune it automatically afterward, the reverse also applies.



A:\cs-sm\CASE1~1\Template\365495.csv

A:\CS-SM\Case 1\Sample\365493.csv

Translate: .00

Stretch: 1.00

Vertical Stretch: .00

Stretch Centre: .00

Matched peaks = 43

Ref S 74 FS 58

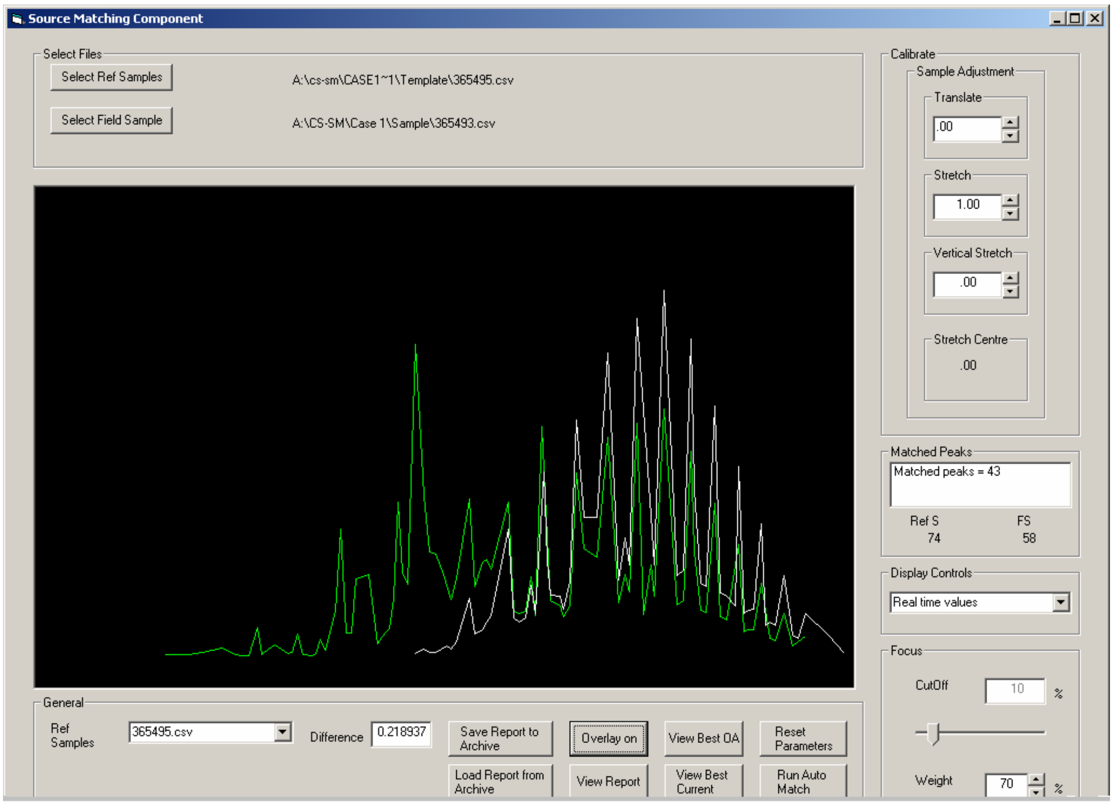
Display Controls: Real time values

CutOff 10 %

Weight 70 %

Ref Samples: 365495.csv Difference 0.218937

Figure 3.19 Matching options



A:\cs-sm\CASE1~1\Template\365495.csv

A:\CS-SM\Case 1\Sample\365493.csv

Translate: .00

Stretch: 1.00

Vertical Stretch: .00

Stretch Centre: .00

Matched peaks = 43

Ref S 74 FS 58

Display Controls: Real time values

CutOff 10 %

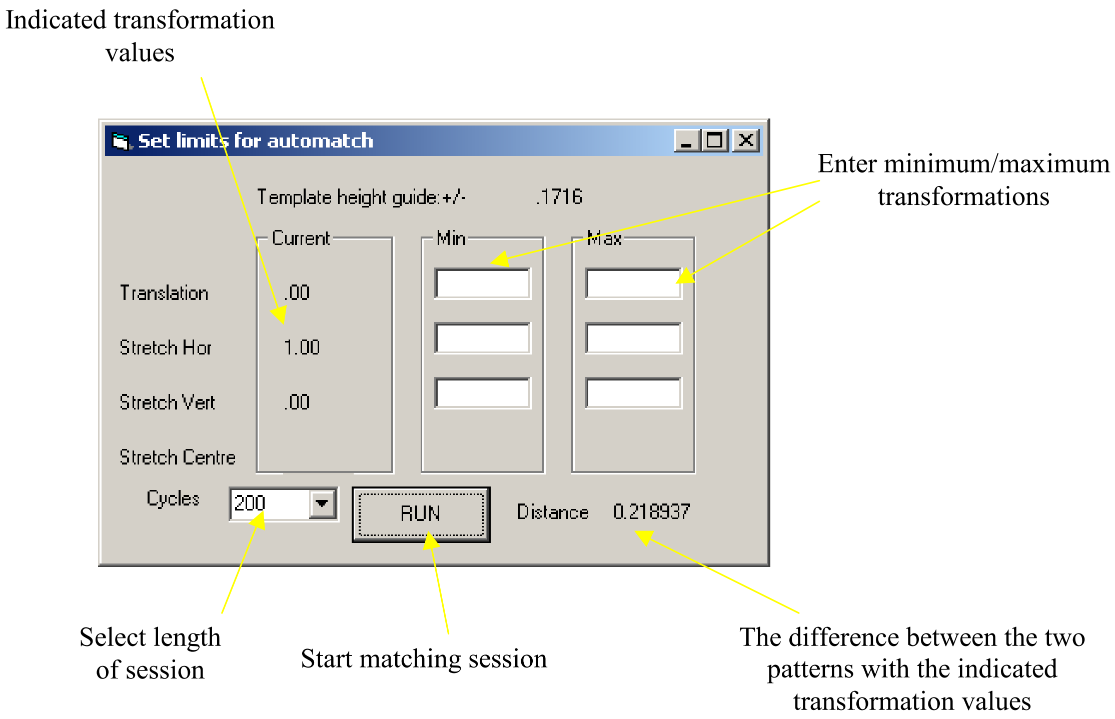
Weight 70 %

Ref Samples: 365495.csv Difference 0.218937

Figure 3.20 'Overlay' view

Indicated transformation values

Enter minimum/maximum transformations



Template height guide:+/- .1716

Translation .00

Stretch Hor 1.00

Stretch Vert .00

Stretch Centre

Cycles 200

Distance 0.218937

Select length of session

Start matching session

The difference between the two patterns with the indicated transformation values

Figure 3.21 Limits for Source Match

3.8 Source match testing

Testing was carried out using Case Study information provided by Agency personnel. The Case Studies represented real pollution incidents that they felt were indicative of the sort of problems they would expect the system to encounter. Initially work was needed on the data in order to increase the number of reference samples per case. It was felt that this would provide the system with a sterner test. The results of testing using the modified data is shown in Table 3.

	Sample				
	D/HO mix	Other D mix	HLB	LFO/P mix	K
No. of Samples	39	2	1	9	1
% match	92.3	100	100	77.8	100

Table 3 Results of source match testing

As can be seen in Table 3 the case study data provided by the Agency at the time of testing is heavily weighted on the side of Diesel/Heating oil mixtures. As contractors we are dependent upon the Agency providing us with a representative data set with which to test the software. Since the data is based upon real incidents this has been assumed to be the case. However, it is particularly surprising that no Unleaded petrol incidents were included in this period since much information was provided about this oil type in order to perfect the Unleaded type template.

3.9 Integration of type-matching and source-matching components

The type and source matching components were developed as totally separate systems, making the preliminary development process an easier task. It was desirable to merge the two systems prior to the workshop forming the OPDS, however, thus making demonstrations and exercises easier for the novice user. The process of combining two complex pieces of software can be extremely risky. However, a simple, and safer option was implemented. The two components were linked via a simple form allowing the user to select one or the other quickly. The hence the two executable versions of the components remain separate allowing for further development in the future, but they are "called" by a third controlling executable file.

4. SUMMARY AND GENERAL DISCUSSION

4.1 Background

The work described herein came as a result of a feasibility study carried out by Prof. W.J. Walley and M.A. O'Connor (E1-050). The aim of the study was to determine the applicability of pattern matching techniques to the twin problems of type and source-matching oil spills when they occur in inland waters. The Centre for Intelligent Environmental Systems, based at Staffordshire University has a history of successfully applying novel AI techniques to problems concerning the environment. However, this was the first time the centre had diversified into the developing field of Environmental Forensics.

As part of the feasibility study Prof. Walley and Mark O'Connor reached the conclusion that pattern matching techniques are wholly applicable to the problems mentioned above. This conclusion was reached as a result of a thorough literature search. In addition, it became apparent that no software appropriate for matching oil GCs in an inland spill situation existed in the public domain at that time. Hence, the CIES were awarded a two-year contract to develop a suitable methodology or methodologies and create prototype systems capable of demonstrating these principles.

Hence, the OPDS system is not only the product of work carried out on this project but also the findings of the feasibility study (E1-050) carried out earlier.

4.2 Key findings of earlier projects

These findings were all thoroughly discussed in technical report E-72;

- Two software packages existed for the matching of oil samples to source oils. One was Eurocrude, which was specifically aimed at the problem of offshore oil spills. A second was Pirouette which was very general in it's approach, a third was MatchFinder which has been withdrawn from the market.

- There are other systems developed and tested by researchers, but none of these are commercially available.

- Neural Networks had not, at that time, been used for pattern comparison in a commercially available software system. Neural Nets were used in MatchFinder but not as a matching mechanism.

- Some researchers have compared the performance of Neural Networks against statistical methods and domain experts when applied to problems of matching chemical trace data and in certain cases have found them to be superior in their matching performance. However, the validity of some of these results is questionable.

- Gas chromatography seemed to be the most appropriate analytical technique.

- In order to pattern match GC traces it was normal to reduce the trace to a set of key “features” and use normalisation to reduce the apparent effect of weathering/degradation. Other techniques had been tried, however.

- No other research had been done specifically on the use of pattern matching applied to the problem of oil spill on inland waters.

4.3 Key findings of this project

The requirements of the contract have been fully met. In fulfilling the requirements set out in the contract documentation the following findings have been made:

- There are distinctive features, some of which correspond to the main alkane peaks, that can be located by an automated system and used to provide a reliable type match. The system used to do this attempts to mimic the way a human expert picks relevant pattern information from a complex scene and uses this to establish a match. Although people find complex pattern matching problems relatively simple to resolve and can intuitively match complex visual images to similar examples, automated systems find the resolution of such problems to be a huge task. In the context of this the achievements of the development team in tackling the many problems associated with this project cannot be overlooked.

- Sets of complex curves can be used to describe the weathering characteristics of the distinctive features detected. These form patterns that are distinctive for different types of oils. The curves were derived from analysing laboratory data provided by theAgency. The data was the result of an extensive set of tests carried out on different oils in different media and under different lighting conditions. Sets of polynomial equations matched to these curves make up type templates that can accurately model how an oil type degrades with time.

- The areas between the main peaks can be isolated and form a reliable basis for a robust match. When source-matching different information is needed. The major peaks are of less importance because these carry information relating to the oil type, not the specific information relating to the particular oil in question. The areas between the main peaks contain more subtle information that is suitable for a more detailed source-matching process.

- An automated system can be used to match a field sample and a reference sample by concentrating on the areas between the peaks.

- A modified Euclidian distance measure can provide the basis of a robust type/source match decision support system. Modified Euclidean distances have been used in neural networks research as a basis of a match between two patterns of information.

4.4 Summary of outputs delivered

- Oil Pollution Diagnostic System (version 1.0), plus an outline user guide.

- R & D Technical Report.

4.5 Operational value of the system

The approach adopted by the CIES has led to a robust and useful prototype system that could be adopted by the Agency as a basis of obtaining more successful court actions in respect of detected oil spill incidents. However, it must be stressed that the true value of this system is that it opens the way for a more coordinated national and possibly international approach to the detection and identification of oil spill criminals. It is suggested that in addition to the technical areas addressed in "Future Possibilities" that some consideration should be given to the possibilities of integrating the OPDS into a coordinated strategy for tackling these problem areas.

4.6 Future Possibilities

- Development of more templates for other common oil types, allowing better coverage of classification.

- Investigate thoroughly the effect of light on the weathering process of oils, how does this affect the accuracy of the templates? The contractor already has some laboratory data relating to oils weathered in the dark.

- Investigate the possibilities of producing templates for oils on soil.

- Development of software to automatically create template files from a set of laboratory sourced GC (.txt) textfiles.

- Development of software to give enhanced analysis capabilities with regard to a type and source match session.

- Development of the software so that laboratory (.txt) files can be used directly for matching without manual modification/validation.

- Enhance the performance of the software so that matching sessions are reduced in time, but with equal or improved accuracy.

- Enhance the visual displays with particular regard to Courtroom requirements.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Recommendations

It is recommended that a further program of work be undertaken. The testing undertaken so far has indicated that the techniques and methods developed by the CIES are valid in addressing the problems of type and source-matching inland oil spills. However, given the complex nature of the problems facing the development team allied with the time constraints of the project there are many aspects of the work that would benefit from further investigation/development work. Several of these are discussed below:

- Development of more oil type templates. This is a critical area. At present we have data on four oil types. A template database containing a more comprehensive list of oil types is obviously preferable if the system is to function reliably over a period of time. This would require a determined program of laboratory work to provide sufficient reliable weathered GC data for all the additional oil types necessary.

- Investigate the possibility of producing templates for oils based on alternative lighting/media. This would give the system more flexibility in dealing with samples collected in areas affected by diminished light conditions.

- Develop a facility within the software that would allow the user to automatically create a type template from a number of GC files.

- Develop a facility in the software to allow the direct use of (.txt) GC files, thus reducing the time spent in manual conversion/validation.

- Enhance the speed of the software by recoding computationally intensive components using "C++" language. This would be a major task but would benefit the users greatly allowing the system to operate at far greater efficiency.

- If proper use is to be made of the full potential of the system an organised staff-training program needs to be planned and implemented. This will require considerable input from both contractor and client in order to achieve desired results.

- As with every prototype unforeseen errors/effects created by the software are bound to occur. The contractor should offer all necessary assistance in alleviating these problems.

- Further input will be needed from the client in more clearly defining the exact courtroom requirements of the system. This is an area that has not been fully discussed as yet.

It should be noted that this list is by no means an exhaustive one and additional areas for consideration maybe added by Contractor/Agency staff.

5.2 Conclusions

After a period of investigation a robust pattern matching methodology has been developed that has been successfully applied to the problems of type and source-matching inland oil spills. Based upon these notional methods a prototype integrated system (OPDS) has been successfully demonstrated to Environment Agency personnel. There is still vast scope for further development, however, both in the presentation and functionality of the system.

It can be concluded that although the questions of type-matching and source-matching applied to inland oil spills provide substantial challenges to the AI "state of the art", the approach adopted by the CIES researchers has successfully resulted in a robust methodology and prototype software system that has been looked upon favourably by Agency personnel at a prototype demonstration and a successful "hands on" workshop.

6. ACKNOWLEDGEMENTS

The authors wish to thank David Gazzard, Wayne Civil and the NLS staff at the Environment Agency’s Laboratory at Starcross for their help in providing data for the development of the techniques discussed here.

7. REFERENCES

Callon R., (1999) *The Esssence of Neural Networks*. Prentice Hall Europe.

Mallach E. G., (1994) *Understanding Decision Support Systems and Expert Systems*. Irwin.

Walley W. J. , Robotham P. W. J., O'Connor M. A., (1999) *Use of Pattern Recognition to Identify the Source of an Oil Spill on an Inland Water*. R&D Technical Report E72.

APPENDIX A

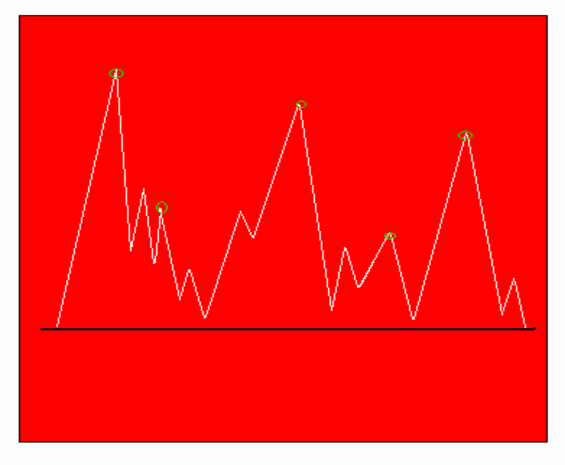
PROTOTYPE USER GUIDE

Oil Pollution Diagnostic System

Type Matching and Source Identification System

Version (Prototype)

User Guide



Contents

1.1 Background

The Centre of Intelligent Environmental Systems at Staffordshire University has developed the Oil Pollution Diagnostic System (OPDS), in fulfilment of an Environment Agency funded research project. A feasibility study (E72) carried out by Prof. Bill Walley and Mark O'Connor in 1999 confirmed that it was possible that a computer-based pattern matching approach could be applied to the twin problems of Type Matching a pollution incident, and, more importantly, Source Matching. As a direct result of the work mentioned above the CIES was awarded a two-year contract (E1-109) to develop an approach, and as a result, create a prototype software system that could demonstrate the validity of the chosen methodologies. It should be emphasized that this document describes the prototype, later versions may have detail differences.

1.2 Pattern Recognition

OPDS attempts to mimic some of the processes used by human experts in interpreting visual data and creating a match. Although the process of Type Matching and Source Matching differ slightly in the details of their operation the general approach is identical. A human "expert" has past experience to allow them to develop an exemplar or standard template from memory, or alternatively from a visual summary of information provided. During the comparison process the expert may break the pattern down into many smaller fragments, or features, that allow the direct comparison of one feature to it's direct counterpart. The data is represented in the form of a Gas Chromatogram that allows a visual representation as a standard graph with time on the x – axis and magnitude on the y – axis. The closeness of a sample pattern to an example or template is often expressed in terms of a distance. Usually the smaller the distance the better the match.

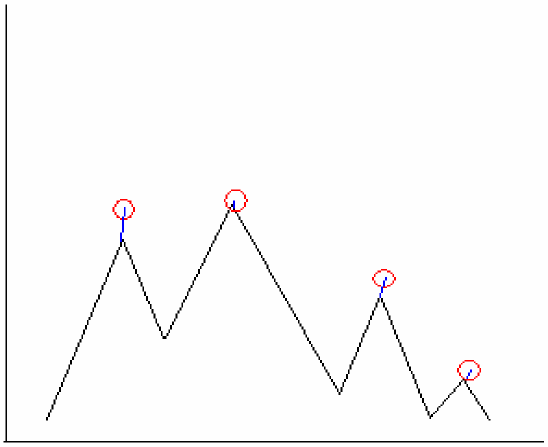


Figure 1.2.1 Template features matched to sample peaks.

The way that this distance maybe derived is illustrated (Fig 1.2.1). The black pattern is a simplified version of a sample GC trace; the circles represent the position of expected peaks according to the template. The connecting lines represent the differences between the two. In a nutshell, the aim of the matching process is to minimize the total connecting line length. The Type Matching software uses exactly this approach. The Source Matching component uses a similar approach but with subtle differences.

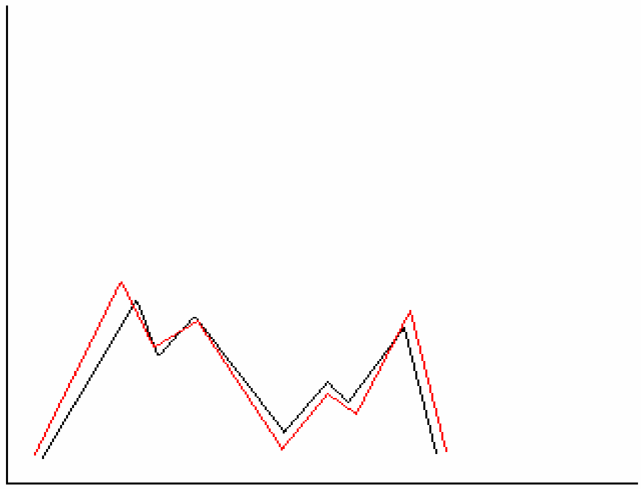


Figure 1.2.2 Similarities of patterns

The Source Matching software concentrates on the areas between the major peaks, the subtle patterns contained within these areas are considered to be a more robust set of features upon which to base a reliable match. It is accepted, however, that inaccuracies in the template definitions of the major peaks will inevitably mean that some of these are included in the patterns. Hence, the matching process in the Source Matching software concentrates on collections of features rather than isolated points.

1.3 Type Matching Theory

In Type Matching the user is attempting to categorize the pollutant as “Diesel”, “Unleaded” etc. As mentioned previously, this is achieved by assessing the closeness of a match between the sample GC and a set of templates. The construction of the templates is a complex process. The templates have to be dynamic, in other words they have to change with time in order to mimic the changes exhibited by an oil when it interacts with the surrounding environment. The Environment Agency laboratories provided the gas chromatogram data needed to assess this by periodically testing a sample of “raw” pollutant as it is naturally “weathered” over a period of several months. The templates each have a number of characteristic features represented by a set of distinctive peaks. The features vary in relative magnitude as exposure to the environment degrades the pollutant. Generally the lighter elements (with a lower retention time) tend to degrade quickest, the heavier components remaining relatively stable throughout the process. A set of equations were derived for each oil type considered. The equations describe how the relative intensity of the features change with time. The controls that affect the weathering times substitute values in these equations in order to generate the representative set of features for a particular oil at a particular time. The best match is generally the template where the modified feature heights are closest to the chosen sample peaks.

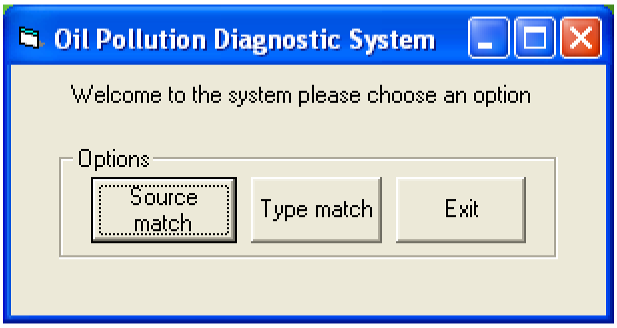
1.4 Source Matching Theory

In source matching the user is trying to ascertain the probable origins of a pollution incident. When source matching the comparison is between an incident sample and one or many candidate reference samples. Unlike type matching no single features are singled out for matching. The match is based on most of the GC trace, concentrating on the areas between the major alkane peaks.

1.5 Starting the System

Insert the CD provided. The CD should spin up automatically allowing the user to follow the simple installation wizard. At the end of the installation process the user can either run the system straight away, or wait until later and run the application from the "Programs" menu in Windows.

The resulting "Start" screen (Fig 1.5.1) allows the user to access either software component easily.



Oil Pollution Diagnostic System

Welcome to the system please choose an option

Figure 1.5.1 Start Screen

2.1 Manual Type Matching

Traditionally all software systems have a "main" screen and OPDS is no exception. The Type Matching main screen is illustrated in Fig 2.1.1.

When the system is started most of the controls are disabled, however, the user has several options when starting a session. It is assumed that the most common choice would be to load a template or set of templates, followed by a suspect sample. This is initiated by clicking the "Open template DB" button to load a set of templates, or the “Select sample file” button to load a sample. In either case both are needed to precede any further with a matching session.

When the user has loaded both a templates (or set of templates) and a sample (both are displayed in the graphics window) several options are available.

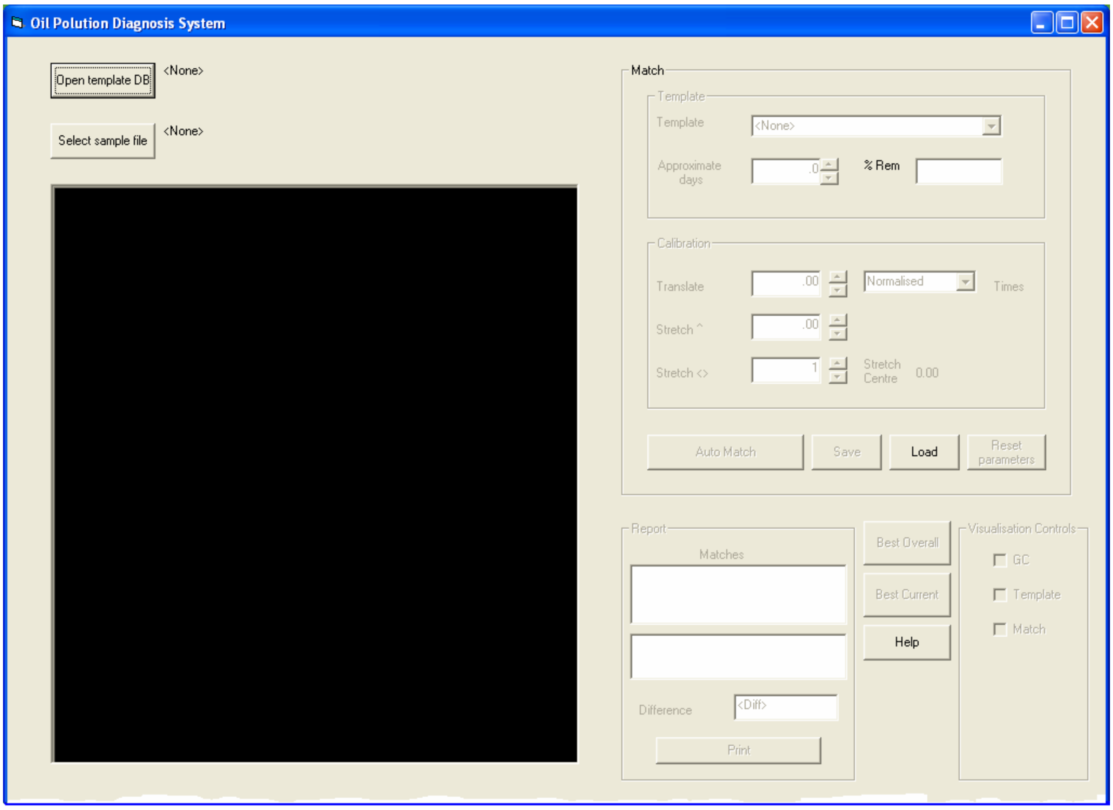
The user may choose to change the template for another contained within the database. This achieved by selecting an alternative template from the combo box in the “Template” frame.

The user may decide to apply some degree of weathering to the displayed template by manually altering the figure in the “Approximate days” text box, see 2.6. If this is done the “% Rem” window is automatically updated to the nearest 10%.

The user may decide to apply some kind of transformation (translation (2.3), stretch (2.4) or vertical stretch (2.5)) to the sample if it is obvious that this will provide a better match.

The user may wish to view the best match achieved in report form. If so, clicking on the “Best Overall” button activates a report screen (2.6) that allows the user to review the best overall match more closely.

Alternatively the user may wish to view the best match achieved for the current template, this is displayed when “Best Current” is activated (2.7).



Open template DB <None>

Select sample file <None>

Template <None>

Approximate days 0

% Rem

Translate .00 Normalised Times

Stretch ^ .00

Stretch <> 1

Stretch Centre 0.00

Difference <Diff>

Figure 2.1.1 The Type Matching screen at the start of a session.

The user can customize the main display at any time by using the check boxes in the “Visualisation Controls” frame.

The user may want to print a hard copy of the current display.

The user can specify the use of either real retention time values or rescaled “normalised” times. Where the times are rescaled the maximum value is 1.00. Real values are the original retention time values expressed in minutes.

Less likely options are that the user will want to load an archive file (2.9) of a previously saved session .The “Load” button opens a windows dialogue box that allows the user to navigate to the desired archive file. Alternatively the user may require some kind of help file although this is not currently operational.

2.2 Translation

The system features most likely to be needed by the user when creating a manual match are contained in the top right hand side of the screen, bounded by frames named “Match” and “Calibration”. There are four possible transformations. The first of these is Translation. A translation is a modification of the original sample retention times that involves a small addition or subtraction. These can be applied to “Real” or “Normalised” cases, although the translations in the “Real” cases should be numerically smaller. The modification is uniform; this means that it affects each point (peak) in the sample data equally.

2.3 Stretch

A stretch is a more complex transformation. This affects points more if they are farther away from the origin of the stretch.

2.4 Vertical Stretch

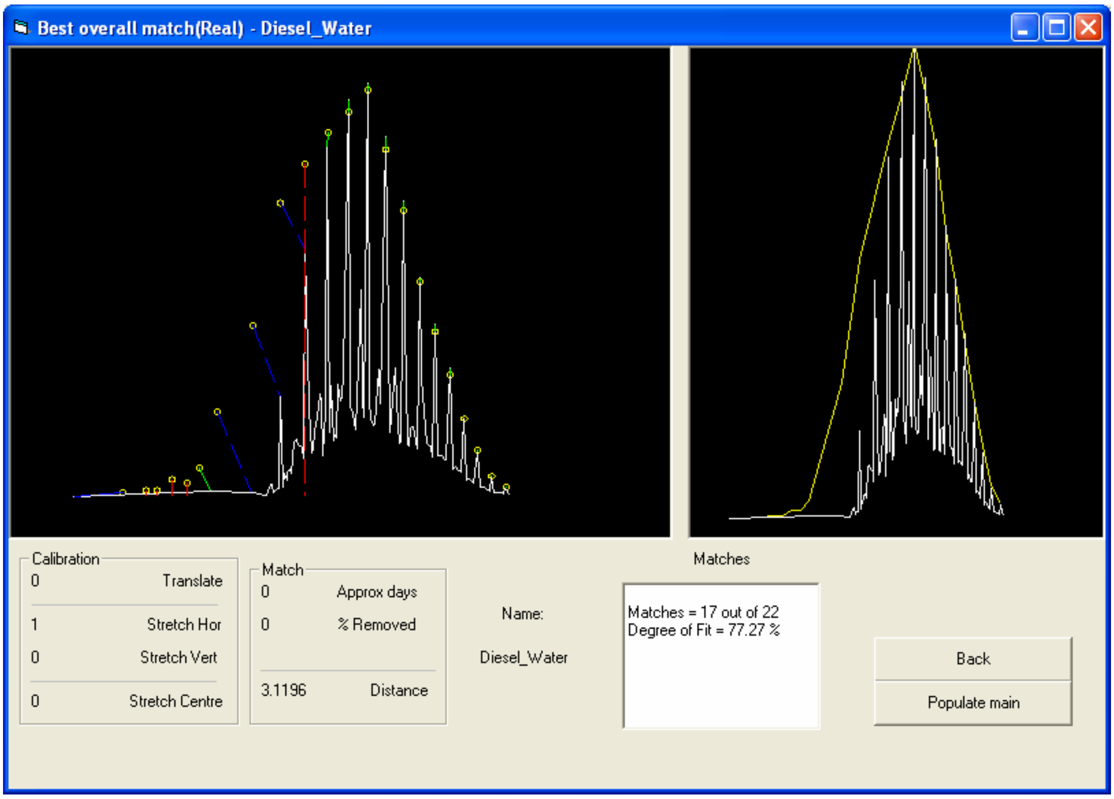
As above except in the vertical plane.

2.5 Weathering

All the transformations discussed previously refer exclusively to the sample. The translations etc. do not alter the template in any way. When type matching it is advantageous, however, to attempt to simulate the natural degradation imposed upon the major features of a particular oil type. Based upon laboratory information collected over a period of time (by the UKEA) the CIES have produced a set of mathematical models that simulate the changes in the GC traces for 3 main oils. For the purposes of the calculations this is expressed in “days” of weathering. Because the time factor can depend on other conditions a “Percent Retained” conversion is also provided giving an approximate percentage rounded to the nearest 10%.

2.6 Best Overall

The “Best Overall” report screen is illustrated in Fig 2.6.1.


Best overall match(Real) - Diesel_Water

Calibration
0 Translate
1 Stretch Hor
0 Stretch Vert
0 Stretch Centre

Match
0 Approx days
0 % Removed
3.1196 Distance

Name:
Diesel_Water

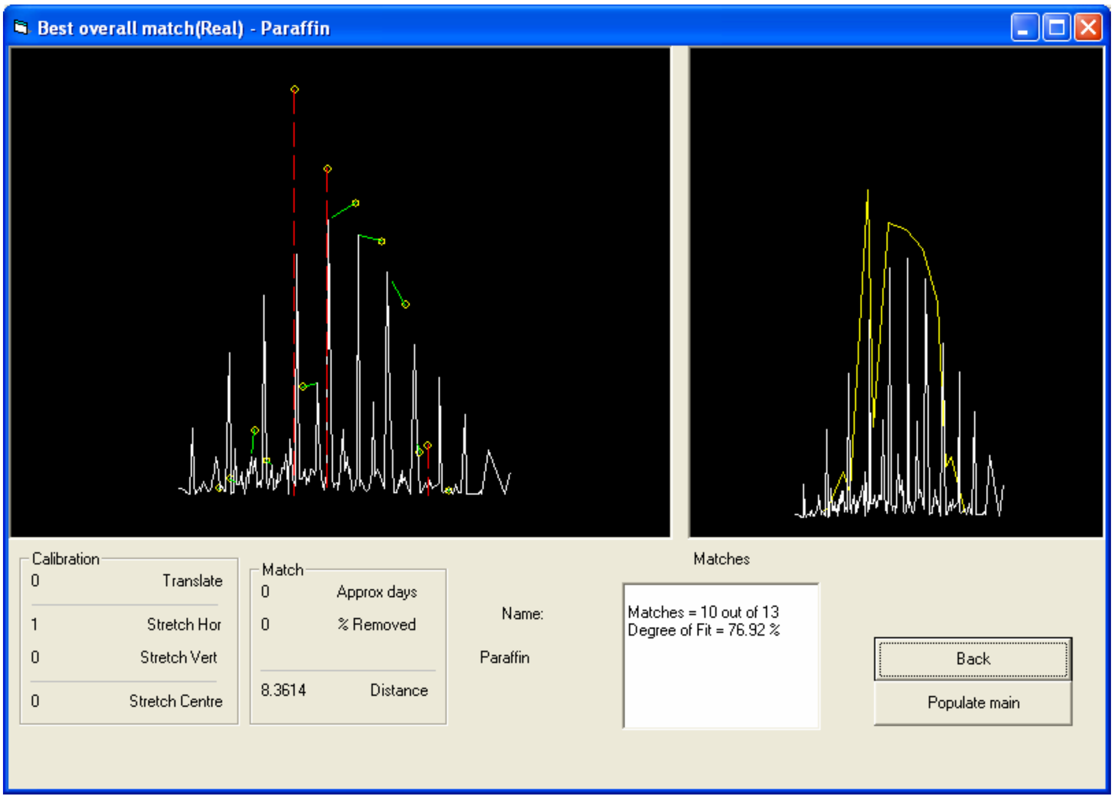
Matches
Matches = 17 out of 22
Degree of Fit = 77.27 %

Figure 2.6.1 The Report Screen showing the Best Overall match so far.

On the left is displayed a graphic of the best match showing acceptable matches (green), “Extreme” matches (blue) and non-matches in red. The yellow circles represent the template features; the sample is shown in white as before.

The right hand graphic displays the match in the form of a template “envelope” compared to the sample. This display can help the user determine whether the match achieved is genuine. By way of illustration consider the example (Fig 2.6.2). Because the program considers the individual template features in isolation the initial match (prior to any manual or auto matching) may include features matched to sample peaks that happen to coincide with the areas expected by the template, but which, on closer inspection are wrong. These cases, the envelope display allows the user to recognise that the match is faulty. The user may then sample the Best Current displays from the other templates available to establish a more acceptable match (Fig 2.6.3).

The user can choose to go directly back (“Back”) to the main screen (Fig 2.1.1) or can transfer the parameters displayed to the (“Populate main”) main screen for further consideration or possible modification.



Best overall match(Real) - Paraffin

Calibration

0 Translate

1 Stretch Hor

0 Stretch Vert

0 Stretch Centre

Match

0 Approx days

0 % Removed

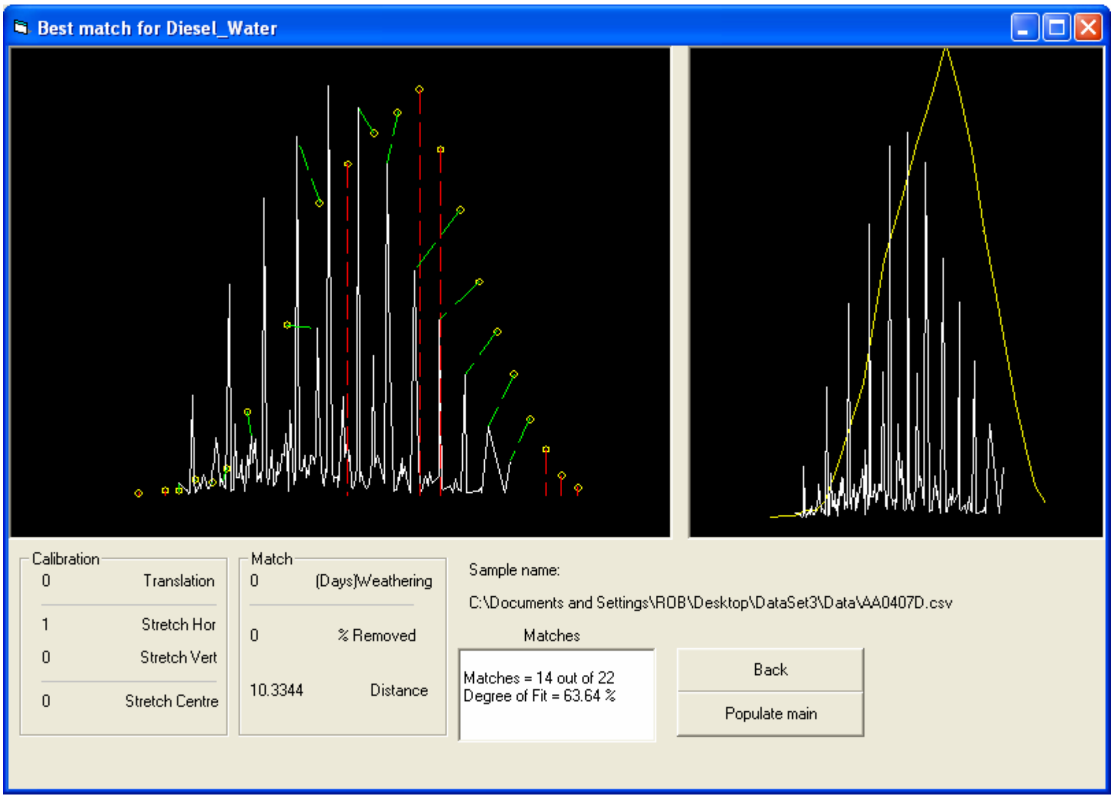
8.3614 Distance

Name: Paraffin

Matches

Matches = 10 out of 13
Degree of Fit = 76.92 %

Figure 2.6.2 Example of an incorrect match



Best match for Diesel_Water

Calibration

0 Translation

1 Stretch Hor

0 Stretch Vert

0 Stretch Centre

Match

0 (Days)Weathering

0 % Removed

10.3344 Distance

Sample name:
C:\Documents and Settings\ROB\Desktop\DataSet3\Data\AA0407D.csv

Matches

Matches = 14 out of 22
Degree of Fit = 63.64 %

Figure 2.6.3 a closer match showing a more similar envelope profile (on the right).

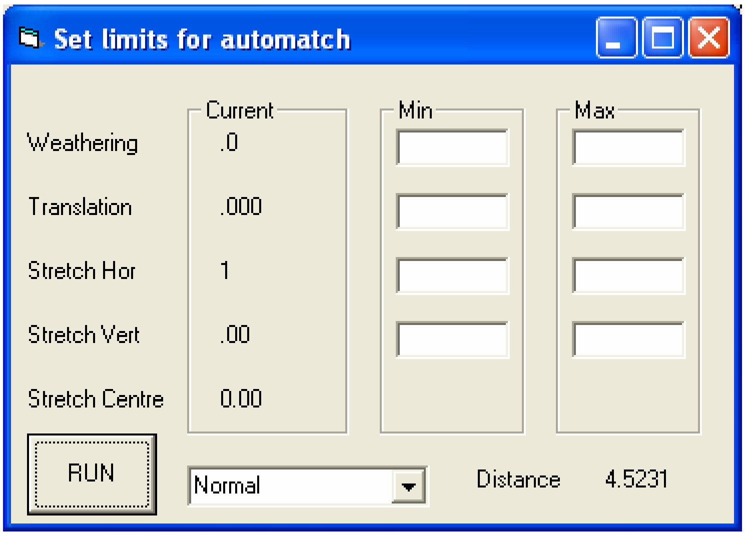
2.7 Best Current

This report screen displays the best match achieved for the particular template being considered by the user at the time. Hence, this maybe the same as the best overall, it is equally likely however, that it will display an inferior match.

2.8 Automatic Matching

The system has the facility to match the sample totally automatically. The user can initiate this by clicking on “Auto Match” on the main (Fig 2.1.1) screen. A small screen requiring the user to set limits on the match is displayed.

The screen displays any values already entered for the currently displayed template. The user sets the maximum and minimum limits for the session. The combo box at the bottom of the screen allows the user to control the duration of the matching session (the duration in terms of running time is also affected by the number of templates the system has to consider). If all the boxes are populated then the user can initiate the session by clicking on “RUN”.



Set limits for automatch

	Current	Min	Max
Weathering	.0		
Translation	.000		
Stretch Hor	1		
Stretch Vert	.00		
Stretch Centre	0.00		

Normal

Distance 4.5231

Figure 2.8.1 Screen allowing the user to set limits for an Automatic Matching Session

At the end of the session the Best Overall (Fig 2.6.1) screen displays the best match found during the session. This may not be the best possible match, but should be close to the best achievable given the constraints imposed by the user when setting up the limits. In some cases the user may wish to attempt to improve the match achieved by using “Populate Main” and attempting some further manual calibration. Hence, it should be emphasized that the user should use discretion when choosing limits. It maybe possible in future versions for a set of default values to be used if the user decides not to bother. However, it must be understood that no one set of values will guarantee optimum performance in all cases.

2.9 Loading a Session File

The user may decide to investigate a match previously saved (2.10) by clicking “Load” on the main screen (Fig 2.1.1). A window allows the user to navigate to a directory or “archive” of saved files. Loading a file allows the user to modify the match if desired and resave or simply to view the previous match as a reference.

2.10 Saving a Session File

The user may wish to save a match to a file for future reference (2.9). Clicking “Save” on the main screen (Fig 2.1.1) allows the user to save the file in a desired location.

3.1 Manual Source Matching

The first screen that faces the user when the Source Matching option is chosen is the main screen shown below.

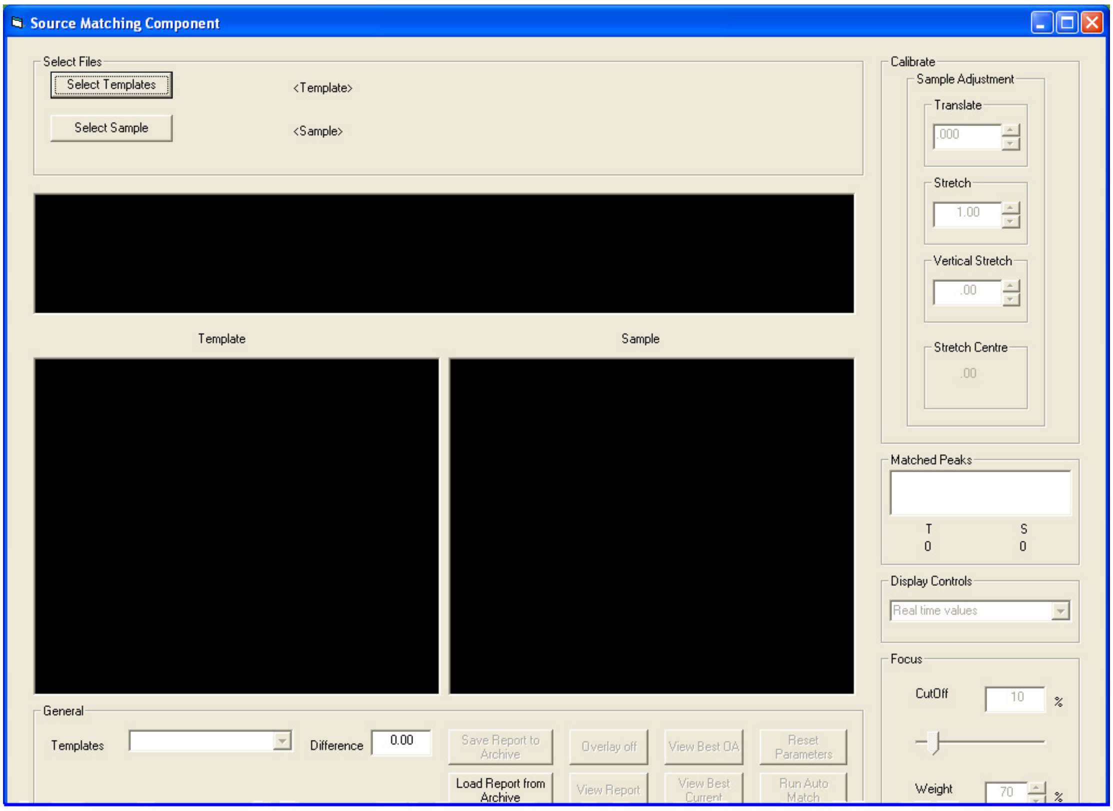


Figure 3.1.1 The main Source Matching screen

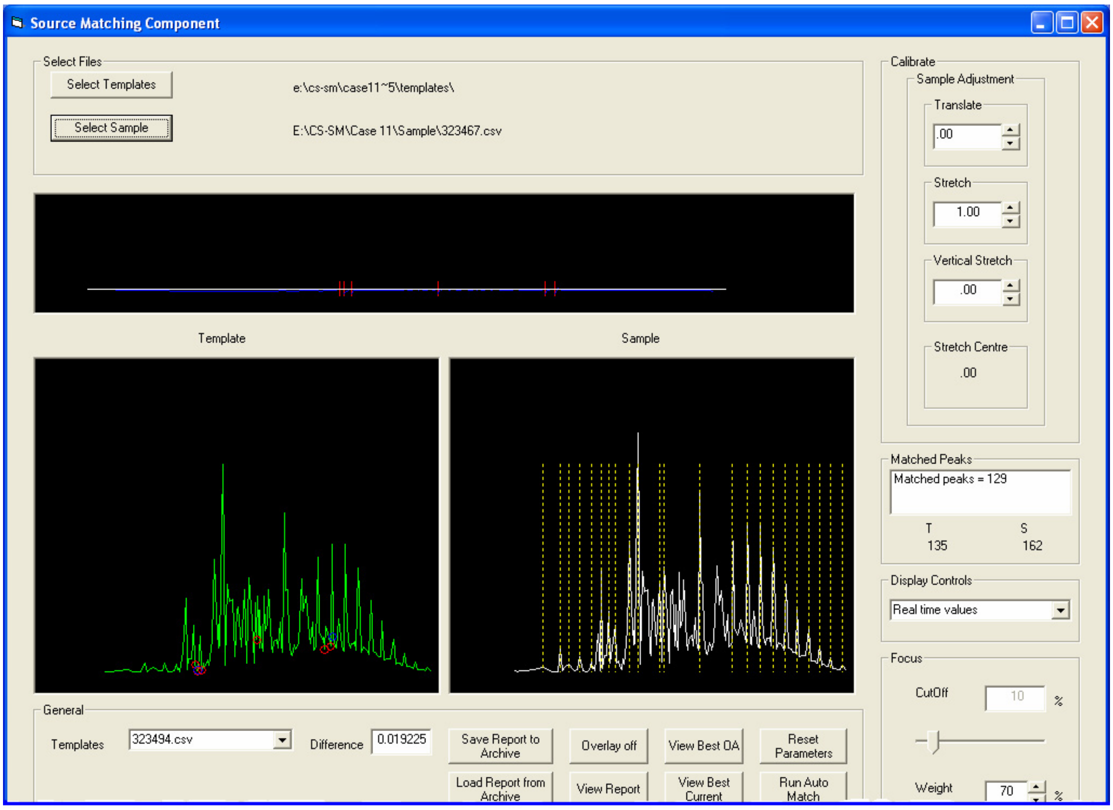
Significantly the appearance and “feel” of the Type Matching software have been retained allowing a novice user to quickly move from one interface to the other with the minimum of additional instruction. As previously discussed in 2.1 the user can load a template (“Select Templates”), or set of templates (a more useful matching session) and compare them to a sample (“Select Sample”). The user should be aware, however, that the time taken for the session to complete depends on the number and size of the files involved.

The form houses three connected graphic windows. Each has a slightly different function. At the top is a “scatter” plot showing the relative variation between matches (blue lines), and the positions of unmatched peaks (red dashes). Below this and to the left is a template plot. This provides a view of the selected template showing the position of matched and unmatched peaks (red circles). To the right is the sample plot. This gives an outline view of the sample. Dotted lines define the relative position and magnitude of the template’s main features (as defined by the template file); this is provided to aid the user in orientation during a manual match.

Many of the features are similar to those in the Type Matching component. The user can view the best overall match, the best match for the current template and load or save archive information. Several differ slightly however, and these are discussed in the next section.

3.2 Overlay

In normal mode the graphics are split between three windows. However, it was considered useful to let the user have the facility to overlay the sample and the template on the same screen. Hence clicking the “Overlay off” (Fig 3.1.2) button produces a screen (see Fig 3.2.1) that replaces the three view graphics with a single large overlaid plot showing a direct comparison between the template and sample.



Source Matching Component

Select Files

Select Templates: e:\cs-sm\case11~5\templates\

Select Sample: E:\CS-SM\Case 11\Sample\323467.csv

General

Templates: 323494.csv

Difference: 0.019225

Calibrate

Sample Adjustment

Translate: .00

Stretch: 1.00

Vertical Stretch: .00

Stretch Centre: .00

Matched Peaks

Matched peaks = 129

T 135

S 162

Display Controls: Real time values

Focus

CutOff: 10 %

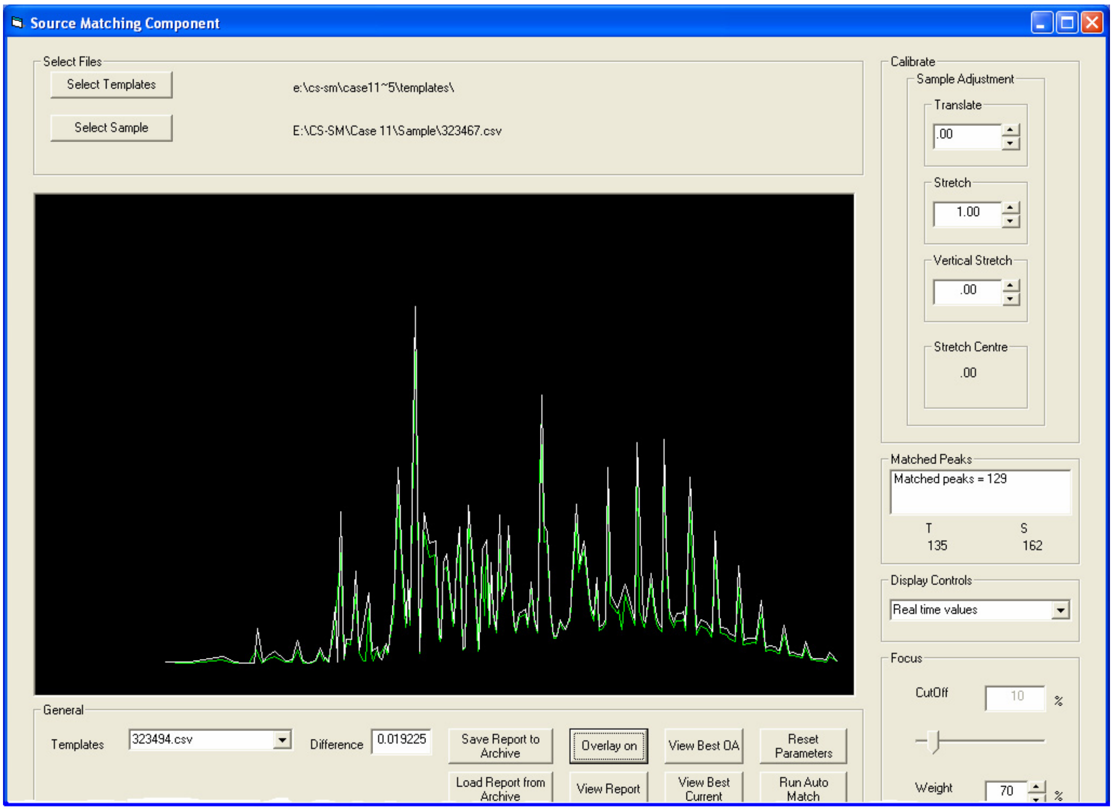
Weight: 70 %

Figure 3.1.2 Main Source Matching screen after loading templates and sample

3.3 Report

In addition to the “Best Overall” and “Best Current” screens the Source Matching component has a dedicated report form (Fig 3.3.1) where the user may enter comments in a text field. In addition, the user may view a larger version of the graphic by clicking on the image. The report can be saved to an archive of the user’s choice by clicking on “Save Report to Archive” (Fig 3.1.2).

In a similar fashion to the Type Matching component the user may manually calibrate the sample (translation etc.), or they may choose to click “Run Auto Match” and let the software initiate a match.



e:\cs-sm\case11~5\templates\

E:\CS-SM\Case 11\Sample\323467.csv

Translate .00

Stretch 1.00

Vertical Stretch .00

Stretch Centre .00

Matched peaks = 129

T 135 S 162

Real time values

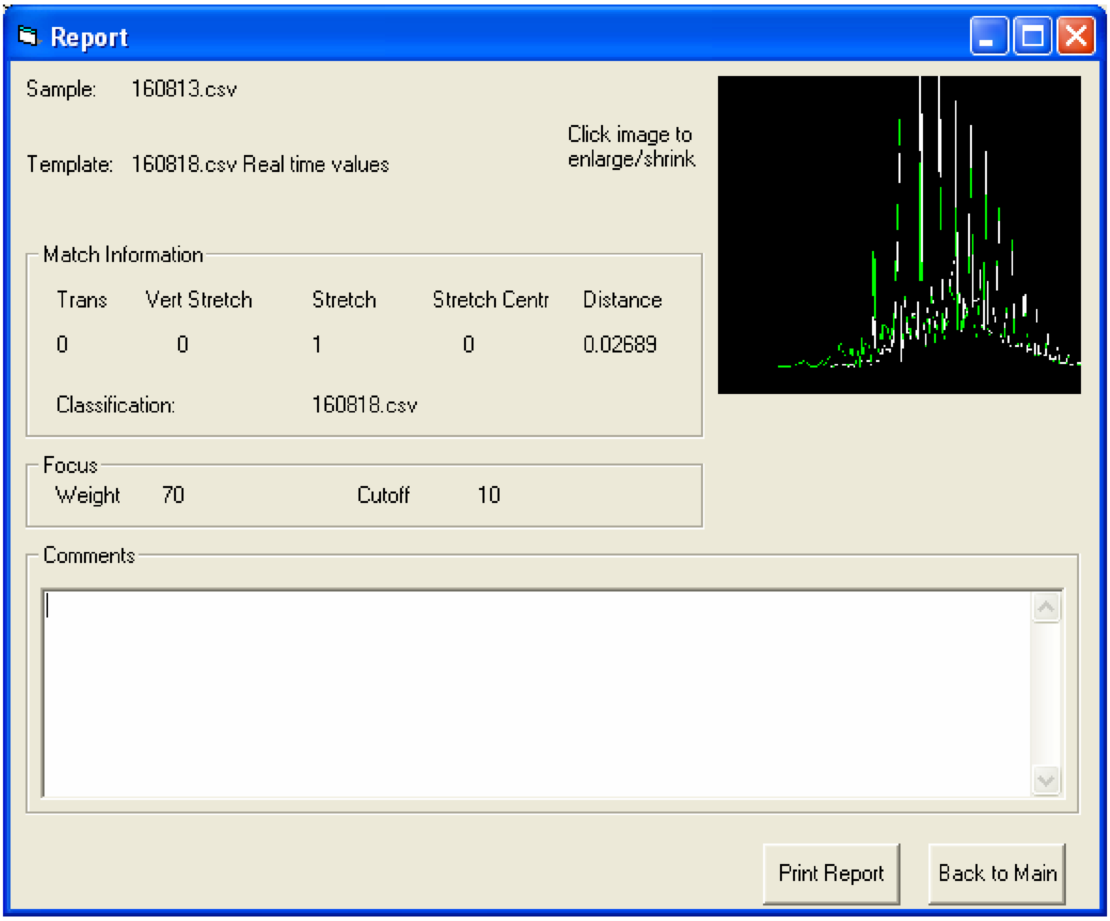
CutOff 10 %

Weight 70 %

Templates 323494.csv

Difference 0.019225

Figure 3.2.1 The screen with “Overlay” activated



Sample: 160813.csv

Template: 160818.csv Real time values

Match Information

Trans	Vert Stretch	Stretch	Stretch Centr	Distance
0	0	1	0	0.02689

Classification: 160818.csv

Focus

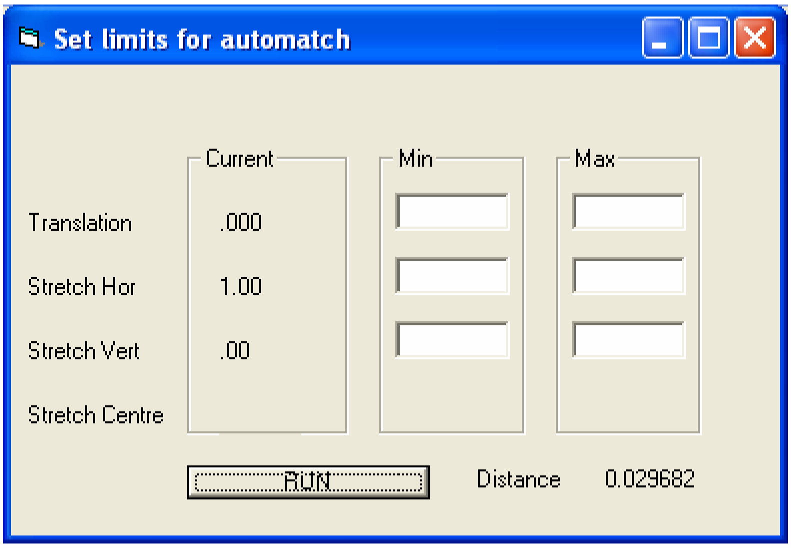
Weight 70 Cutoff 10

Comments

Figure 3.3.1 Report screen

3.4 Source Matching Auto Match

If "Run Auto Match" is selected a screen requiring the user to define limits for the standard transformations in the matching session is displayed (Fig 3.4.1). The purpose of this screen is similar to that discussed previously. The practical limitations imposed on the number of matching transformations that can be undertaken during one session mean that the imposition of "sensible" limits to the individual transformations is essential. At present the user must enter values to define the limits prior to the initiation of the matching session. It is anticipated that in later versions a suitable set of default values may be automatically entered if the fields are left blank. However, the importance of the limits to the success of a match should not be underestimated. Users should be encouraged to define suitable limits for the particular session. Guidance on deciding suitable limitation values is given later (4.2).



Set limits for automatch

	Current	Min	Max
Translation	.000		
Stretch Hor	1.00		
Stretch Vert	.00		
Stretch Centre			

RUN

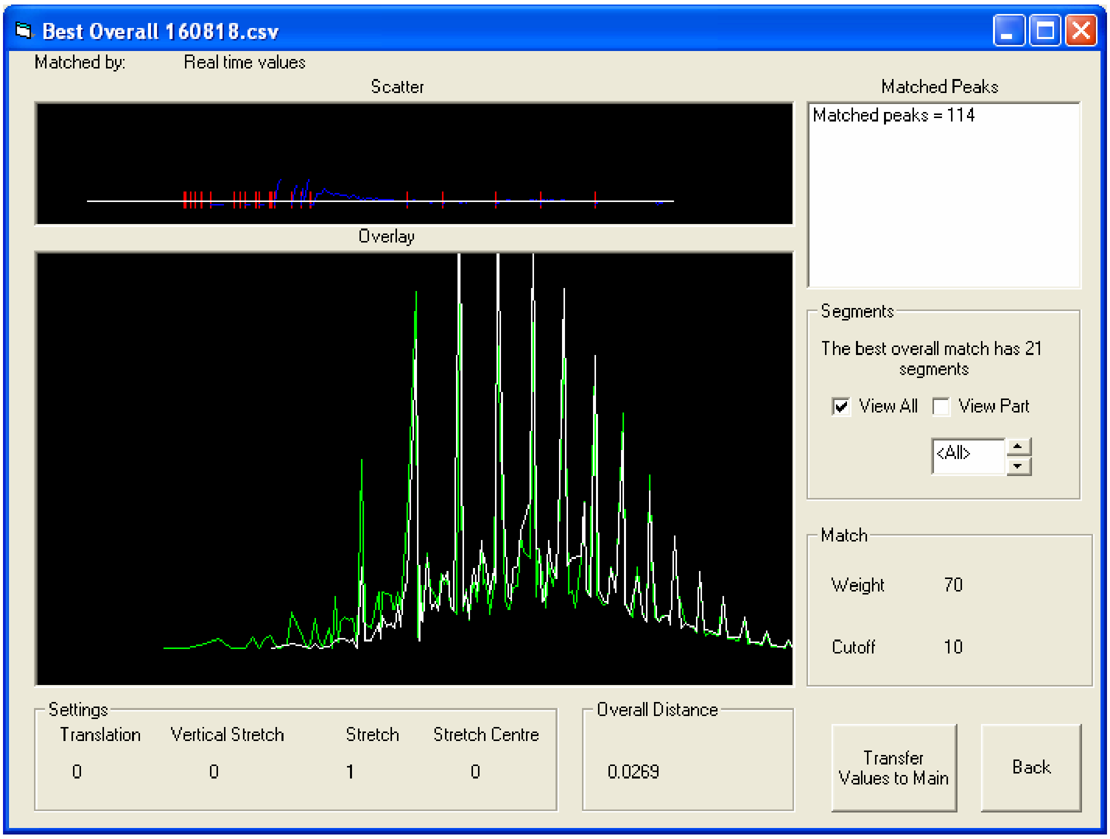
Distance 0.029682

Figure 3.4.1 The screen used to set limits for a matching session

It is anticipated that in future versions of the software some provision may be made that allows the user to select the number of cycles over which the session is to be run. Once the session is complete the system displays the best match achieved by loading the best overall screen (Fig 3.4.1).

If the user ticks the “View Part” box it is possible to view each segment (area between the main peaks) in turn.

The user may choose to transfer the information to the main and attempt to modify the match achieved by the system.



Best Overall 160818.csv

Matched by: Real time values

Matched peaks = 114

Segments

The best overall match has 21 segments

☑ View All □ View Part

<All>

Match

Weight 70

Cutoff 10

Settings

Translation 0

Vertical Stretch 0

Stretch 1

Stretch Centre 0

Overall Distance 0.0269

Figure 3.4.1 The screen displayed at the end of a matching session

3.5 Setting the Focus

Two other features peculiar to the source-matching component are the "Cutoff" and "Weight" features encapsulated in the "Focus" pane (bottom right). These are provided to help the source-matching system overcome any possible interference caused by weathering. There is not enough information present in a single sample to allow a model to be developed as in the type-matching system. Hence, a simpler solution has been developed. The "Cutoff" control (default setting 30%) allows the user to focus on a particular section of the GC trace. In this way the user can reduce the relative importance of the first part of a weathered trace in favour of more robust components later on.

4.1 Type Matching Limits

The smaller the range of values between limits the more likely it is that a close match will be achieved. During testing values of -0.5(min) and 0.5(max) "Real", -0.2, 0.2 "Normalised" were used exclusively for translation and 0.9(min), 1.1(max) for horizontal stretch. The limits the user places on the weathering times should be carefully chosen. During testing twenty day wide "slots" were used and this degree of variation tended to give good results. If the user is not certain in any way concerning the degree of weathering a wider "window" may be needed. It should be noted, however, that this will likely result in sub optimal accuracy. However, a "general" search of this type may allow the user to run a second match using a smaller window based upon the indications provided by the session with highly spaced limits. It was found that limits of –0.3 and 0.3 were suitable in most cases for the vertical stretch.

4.2 Source Matching Limits

It was found during the testing period that limits for translation of –0.02 and 0.02 (normalised) and -0.05 and 0.05(real) were adequate in most cases. For horizontal stretch 0.95 and 1.05 worked well and for vertical stretch –0.25 and 0.25 were generally sufficient.

5.1 Type Matching Files

The type matching sample files taken by the system are comma delimited .csv files. The fields of data in order are :-

1. Peak number
2. Retention time
3. Peak type
4. Peak width
5. Peak Area
6. Peak Height
7. Relative % of total area

The type matching template file is also a .csv file. This file has information about each Oil Type (Diesel, Unleaded etc. on each separate line). The fields of data in order are:-

1. Template ID number
2. Template name
3. Number of features
4. Retention time of final peak
5. Constant used in initial normalisation process

6.-16. Constants used in determining % retained

17.- onwards data relating to weathering characteristics of each oil.

5.2 Source Matching Files

The source matching sample files are comma delimited and have the same fields as the type matching sample files.

The source matching template files have an additional field:-

8. Boolean variable indicating weather or not the peak is classed as a major peak.