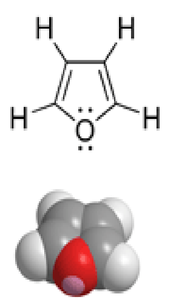
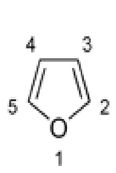
SAND2015-4190C

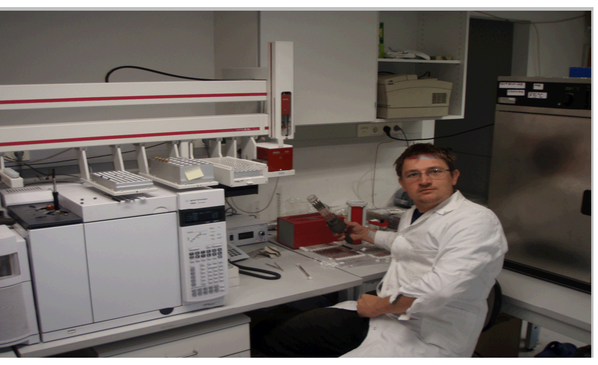
A Method for Automated Data Extraction and Peak Identification from Large GC-MS Data Sets Using Multivariate Analysis

Joshua Coon[1], Mark Van Benthem[1], James Hochrein[1], Curtis Mowry[1]

1. Sandia National Laboratories, Albuquerque New Mexico, 87123.







Sandia National Laboratories

Exceptional service in the national interest

Overview:

Objectives:

- **Use multivariate analysis, automated spectral peak identification to improve analysis of non-ideal data sets**
- **Bundle analysis tools into a user-friendly Graphical User Interface (GUI) for distribution**

Introduction:

Peak identification for GC-MS data sets in materials aging studies is often difficult due to both non-ideal data collection conditions and the challenges of dealing with large data sets.

Non-ideal data collection conditions:

- Retention shifts
- Dirty Column
- Different Instruments
- Collected over several years

Non-ideal data sets mean:

- Increased expertise level needed for peak identification
- Decreased confidence in chemical identifications

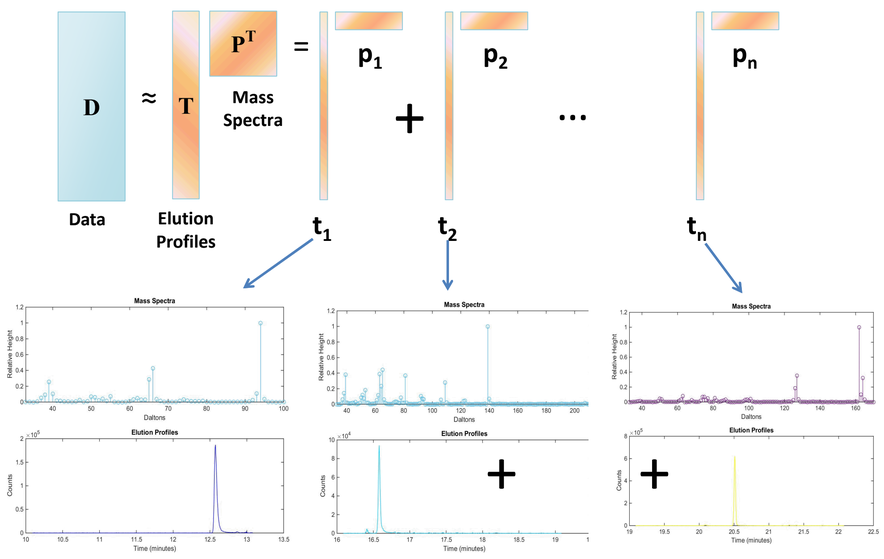
Data overload:

- Typical Data sets =7500 times X 350 masses X 50 samples = **131,250,000** elements
- Assume each sample has 100 species and 1 minute per identification
- ~ 3 hours per data set (including “copy-and-paste” work), 2 sets per day maximum
- Total parsing time is 100 days or **5 months** of work

Methods:

Multivariate Analysis:

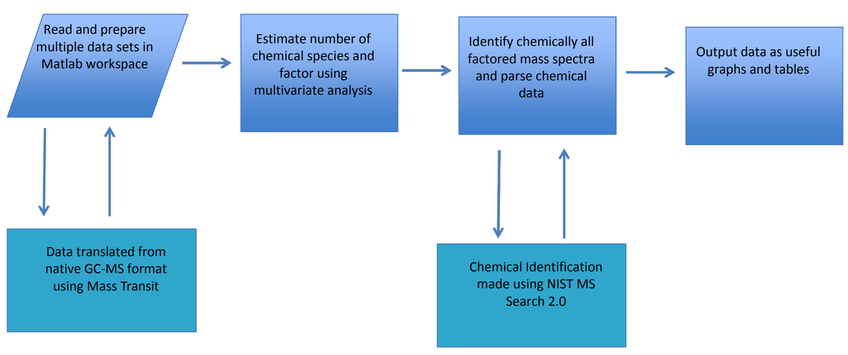
- Factors full rank data matrix into outer product of individual elution profiles and mass spectra using a least squares algorithm



Data Analysis GUI:

- Packages multivariate tools into accessible format for use by non-experts

GUI Workflow:



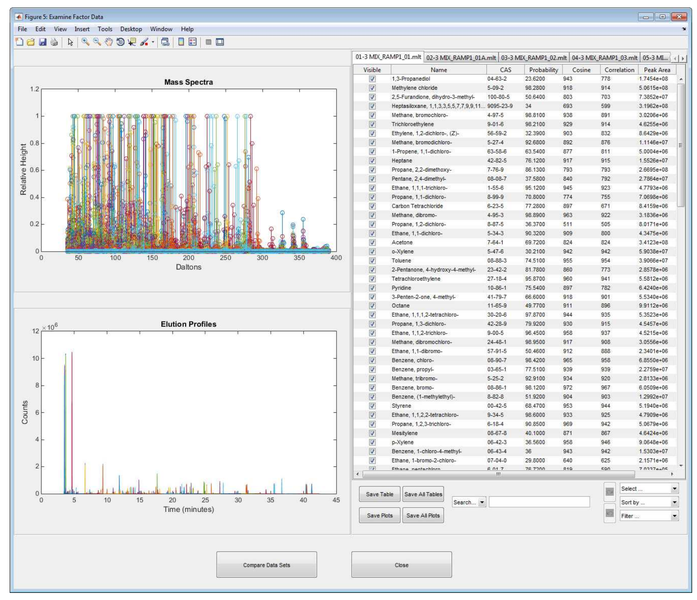
User-friendly GUI Features:

- Search species by name or mass
- Data statistics tooltip
- Save tables and figures
- Data set comparisons through clustering

Results:

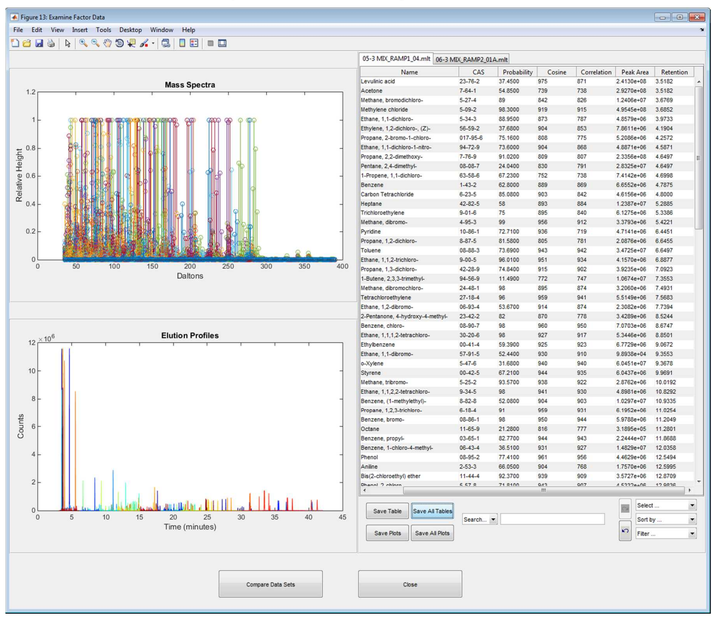
Processing Large Number of Data Sets:

- Five data sets processed in less than 5 minutes
- ~15 hours of work using AMDIS



Processing Non-Ideal Data Sets:

- Two data sets with ~5 minute oven ramp time difference
- Parsed without parameter adjustment



Accuracy of Species Identification:

- GUI species identifications compared with human expert analysis
- Identifications that agree are highlighted in green

Name	Prob	Cosine	Correlation	Counts	RetentionTime	
N-Nitrosodimethylamine		91.98	675	938	2.59E+06	6.17
Pyridine		62.86	722	914	5.10E+06	6.44
3-Penten-2-one, 4-methyl-		67.47	915	926	6.58E+06	7.36
2-Pentanone, 4-hydroxy-4-methyl-		87.07	831	887	3.39E+06	8.53
Phenol		72.98	883	901	5.13E+06	12.57
Bis(2-chloroethyl) ether		91.06	914	943	4.12E+06	12.88
Phenol, 2-chloro-		67.92	913	924	5.02E+06	12.99
Benzene, 1,3-dichloro-		34.97	952	954	1.33E+07	13.80
Benzene, 1,3-dichloro-		35.01	895	931	7.76E+06	14.30
Bis(2-chloroisopropyl) ether		50.93	855	891	4.95E+06	14.78
1-Propanamine, N-nitroso-N-propyl-		96.98	875	947	3.75E+06	15.19
Phenol, 3-methyl-		40.33	952	954	1.97E+07	15.23
Ethane, hexachloro-		98.93	904	913	5.70E+06	15.38
Benzene, nitro-		96.96	906	916	4.27E+06	15.63
Methanamine, N-(1-phenylethylidene)-		55.31	714	748	1.69E+06	16.04
Isophorone		88.37	919	924	4.51E+06	16.41
Phenol, 2-nitro-		88.12	894	921	3.21E+06	16.58
Phenol, 2,4-dimethyl-		29.16	937	942	6.35E+06	16.81
Methane, bis(2-chloroethoxy)-		90.64	928	957	4.22E+06	17.07
Phenol, 2,4-dichloro-		42.63	924	936	5.44E+06	17.27
Benzene, 1,3,5-trichloro-		43.86	951	953	7.59E+06	17.47
Naphthalene		59.5	957	960	8.84E+06	17.65
p-Chloroaniline		43.89	873	914	2.41E+06	17.87
1,3-Butadiene, 1,1,2,3,4,4-hexachloro-		95.04	647	801	3.46E+05	18.00
1,3-Butadiene, 1,1,2,3,4,4-hexachloro-		98.72	926	929	7.49E+06	18.02
Phenol, 4-chloro-3-methyl-		50.1	916	932	5.33E+06	19.15
4-Chloroaniline, N-isopropylidene		83.44	746	790	2.53E+06	19.35
Naphthalene, 2-methyl-		48.1	941	942	1.88E+07	19.40
3-Methylbenzothiophene		42.83	625	827	1.30E+06	19.59
Hexachlorocyclopentadiene		97.64	894	900	3.03E+06	19.78
Phenol, 2,4,5-trichloro-		31.6	914	919	1.14E+07	20.09
Naphthalene, 2-chloro-		62.36	961	970	9.21E+06	20.51
o-Nitroaniline		45.04	864	911	6.38E+06	20.77
Quinoline, 1,2-dihydro-2,2,4-trimethyl-		54.24	690	831	8.22E+05	21.18
Dimethyl phthalate		86.77	962	969	7.14E+06	21.20
Benzene, 1,3-dinitro-		71.04	864	910	6.43E+06	21.24
Benzene, 2-methyl-1,3-dinitro-		91.49	887	915	5.15E+06	21.31
Acenaphthylene		61.1	913	965	1.11E+07	21.41
Acenaphthene		78.39	956	960	1.07E+07	21.78
Phenol, 4-nitro-		60.41	664	858	3.67E+06	22.10
Dibenzofuran		87.87	816	845	1.32E+07	22.14
Phenol, 2,3,4,6-tetrachloro-		57.18	916	929	1.18E+07	22.42
Diethyl Phthalate		84.96	947	953	7.47E+06	22.66
Fluorene		76.16	910	912	1.14E+07	22.85
Benzene, 1-chloro-4-phenoxy-		85.36	929	958	8.48E+06	22.87
p-Nitroaniline		71.88	837	909	3.86E+06	22.93
Diphenylamine		69.05	944	953	9.42E+06	23.11
Azobenzene		92.21	935	948	7.89E+06	23.19
Benzene, 1-bromo-4-phenoxy-		98.1	917	931	7.60E+06	23.86
Benzene, hexachloro-		75.78	932	938	9.58E+06	23.95
Phenol, pentachloro-		78.78	671	837	5.37E+06	24.47
Anthracene		42.44	954	958	2.36E+07	24.78
Carbazole		64.63	942	960	1.06E+07	25.22
Dibutyl phthalate		45.05	921	952	8.36E+06	25.92
Pyrene		51.09	923	947	2.80E+07	27.22
Benzyl butyl phthalate		90.59	816	901	8.66E+06	29.52
Hexanedioic acid, bis(2-ethylhexyl) ester		64.64	763	911	8.60E+06	29.80
[1,1'-Biphenyl]-4,4'-diamine, 3,3'-dichloro-		91.77	773	866	7.88E+06	31.22
Benz[a]anthracene		50.91	923	977	3.44E+07	31.37
Bis(2-ethylhexyl) phthalate		37.98	728	876	1.06E+07	31.53
Di-n-octyl phthalate		26.7	705	855	1.20E+07	34.28
Benzo[k]fluoranthene		26.53	888	965	4.26E+07	35.58
Benzo[k]fluoranthene		26.73	851	964	2.33E+07	36.69
Dibenz[a,j]acridine		68.19	677	863	3.05E+07	40.50

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration