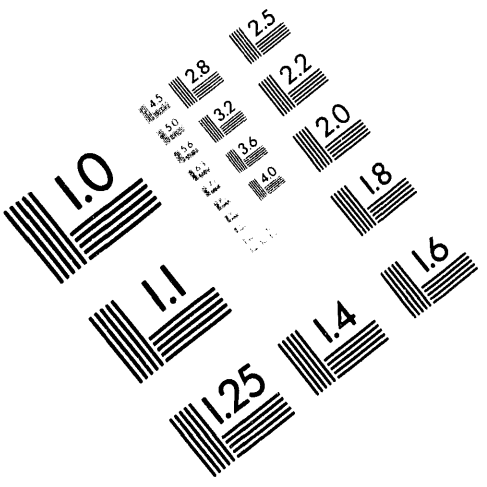
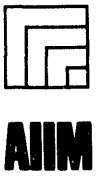
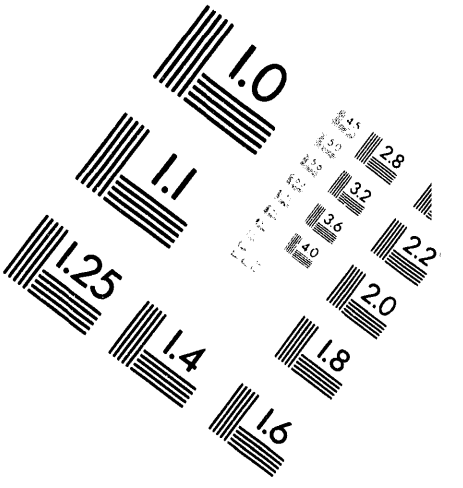
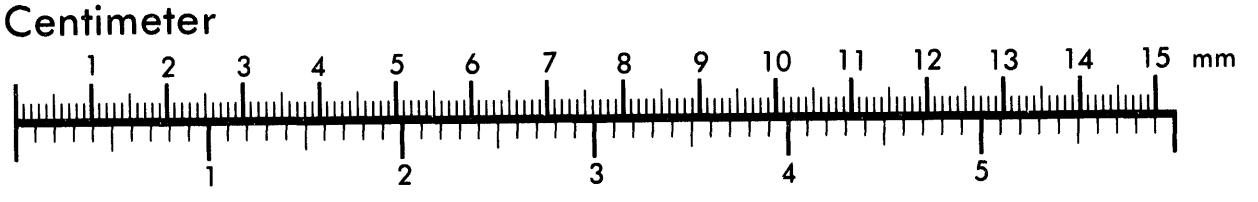
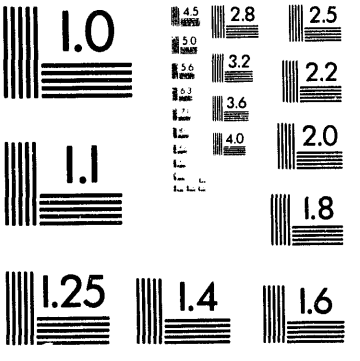


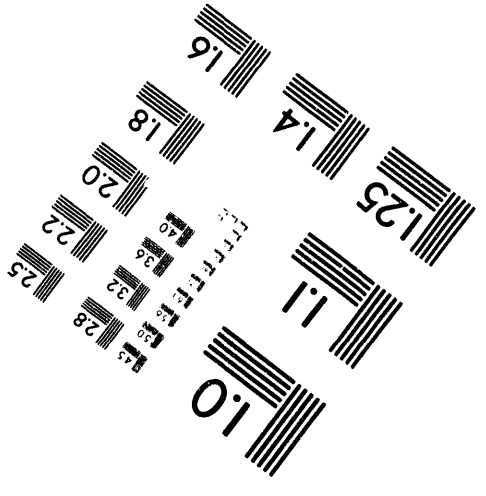


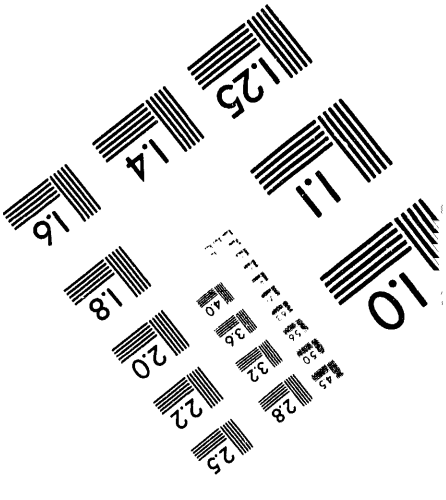
Association for Information and Image Management
1100 Wayne Avenue, Suite 1100
Silver Spring, Maryland 20910
301/587-8202











1 of 1

Conf-9406176--2

Statistical properties of an algorithm used for illicit substance detection by fast-neutron transmission

D.L. Smith, L. Sagalovsky, B.J. Micklich, M.K. Harper and A.H. Novick

Technology Development Division, Argonne National Laboratory
9700 South Cass Avenue, Argonne, Illinois 60439, U.S.A.

ABSTRACT

A least-squares algorithm developed for analysis of fast-neutron transmission data resulting from non-destructive interrogation of sealed luggage and containers is subjected to a probabilistic interpretation. The approach is to convert knowledge of uncertainties in the derived areal elemental densities, as provided by this algorithm, into probability information that can be used to judge whether an interrogated object is either benign or potentially contains an illicit substance that should be investigated further. Two approaches are considered in this paper. One involves integration of a normalized probability density function associated with the least-squares solution. The other tests this solution against a hypothesis that the interrogated object indeed contains illicit material. This is accomplished by an application of the F-distribution from statistics. These two methods of data interpretation are applied to specific sets of neutron transmission results produced by Monte Carlo simulation.

1. INTRODUCTION

Fast-neutron interrogation techniques are of interest for detecting illicit substances such as explosives and drugs because of their ability to identify light elements such as hydrogen, carbon, nitrogen and oxygen.[1] These elements are not easily identified by other non-invasive techniques such as X-ray spectroscopy. Fast-neutron transmission spectroscopy employs a collimated neutron beam from a continuum source as the nuclear probe. Data are acquired by neutron time-of-flight spectroscopy. Comparison is made between the primary spectrum and the spectrum after transmission through the object being interrogated. A favorable neutron source for this purpose can be generated at an accelerator facility via the $^{9}Be(d,n)^{10}B$ reaction, with incident deuterons in the 3-7 MeV energy range.[2] A typical setup is shown schematically in Fig.1 below.

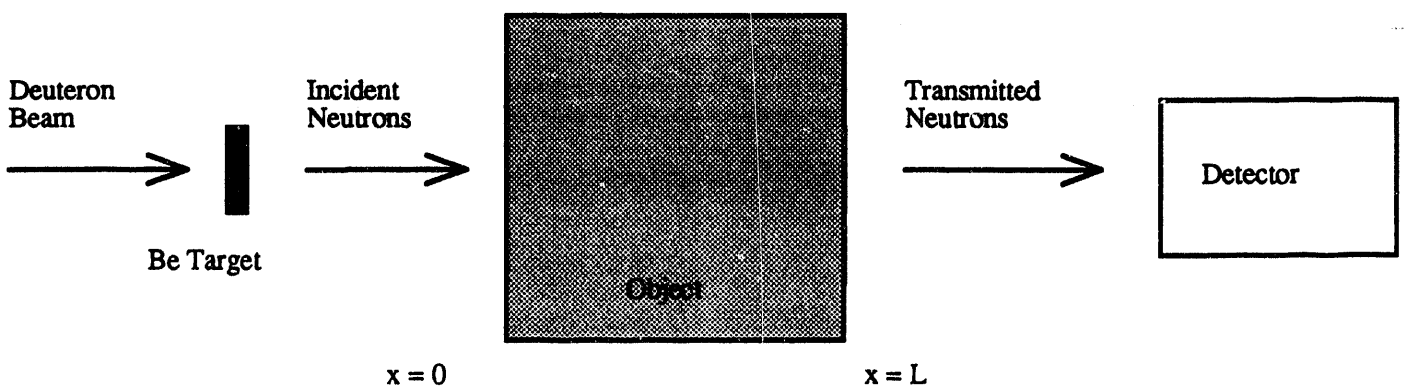


Fig. 1. Schematic diagram of a neutron transmission experiment in which the continuum source of MeV neutrons is produced by deuterons from an accelerator incident on a thick Be metal target. The neutron collimation system is not shown.

Time (or equivalently energy) spectra are collected in a number of bins (n) by the recording device. Typically, n > 100. The relationship between the numbers of incident (N_{oi}) and transmitted (N_i) neutrons for a particular time or energy bin (i) is expressed by means of the formula

$$N_i = N_{0i} \exp\left[-\sum_{k=1}^{m} \sigma_k(E_i) \int_0^L n_k(x)dx\right], \quad (1)$$

where $\sigma_k(E_i)$ is the total cross-section for element k at energy E_i, $n_k(x)$ is the volumetric density of element k along the neutron path through the interrogated object, and $p_k = \int_o^L n_k(x)dx$ is the areal density of element k integrated along that path (x = 0,L) through the object interrogated by the neutron probe. The list of elements to be considered will always include hydrogen (H), carbon (C), nitrogen (N) and oxygen (O) as a minimum set. However, since luggage and cargo containers usually incorporate other elements as well, in various proportions, we have considered simulations involving the addition of such elements as flourine (F), aluminum (Al), silicon (Si), chlorine (Cl), iron (Fe), and copper (Cu).[1] The number of elements (m) included in the analysis is usually ≤ 10. If we define $y_i = ln\ (N_{oi}/N_i)$ and let $(A)_{ik} = \sigma_k(E_i)$, then Eq.1 can be expressed in matrix form by

$$\mathbf{A\,p} \approx \mathbf{y}. \quad (2)$$

The fitting model is governed by the matrix **A** and therefore entails the choice of elements and the individual total cross-section values. The approximate equality in Eq.2 requires an explanation. If the collection of m areal densities symbolized by **p** and the cross-section set forming **A** were known with certainty *a priori* then the transmission parameters **y** could be calculated exactly. However, what is actually known is just the reverse. The vector **y** is determined from experiment and the cross-sections in **A** are derived from evaluated data files. Both have associated uncertainties. A set of solution parameters **p** must be found which best satisfies the over-determined (since $n >> m$) system of linear equations represented by Eq.2. The least-squares solution is very unlikely to yield exact equality for this matrix expression.

In this paper we review the algorithm used to solve Eq.2 and then turn to an examination of probabilistic interpretations for the solutions which it provides. These statistical methodologies are subsequently applied to some realistic examples of simulated neutron-interrogation data.

2. UNFOLDING ALGORITHM

The method of least squares[3] is employed to obtain a solution to Eq.2. This involves finding a vector **p** which minimizes the quantity

$$\chi^2 = (\mathbf{y} - \mathbf{A\,p})^{\dagger}\, \mathbf{V_y}^{-1}\, (\mathbf{y} - \mathbf{A\,p}), \quad (3)$$

where "†" indicates matrix transposition and "$^{-1}$" denotes matrix inversion.

The solution is given by the formulas

$$\mathbf{V_p} = (\mathbf{A}^{\dagger}\, \mathbf{V_y}^{-1}\, \mathbf{A})^{-1}, \qquad \mathbf{p} = \mathbf{V_p}\, \mathbf{A}^{\dagger}\, \mathbf{V_y}^{-1}\, \mathbf{y}, \quad (4)$$

where $\mathbf{V_p}$ is the covariance matrix for areal densities **p** and $\mathbf{V_y}$ is the covariance matrix for the transmission data **y**. The value of χ^2 obtained by substituting this solution into Eq.3 is governed by a chi-square distribution with f= n-m degrees of freedom.[3] The transmission-data covariance matrix $\mathbf{V_y}$ is derived from the expression

$$(\mathbf{V_y})_{ij} \approx \delta_{ij}\,[(1/N_{0i}) + (1/N_i)] + \Sigma_{k=1,m}\; p_k^2\; (C_k)_{ij}\, v_{ki}\, \sigma_{ki}\, v_{kj}\, \sigma_{kj}\ , \quad (5)$$

where $\sigma_{ki} = \sigma_k(E_i)$ is a total cross-section, v_{ki} is the corresponding fractional error, and C_k is the cross-section error correlation matrix. Care must be taken in generating $\mathbf{V_y}$ to insure that it is positive definite.[3] Then it can be inverted and $\mathbf{V_p}$ will also be positive definite. The covariance matrix $\mathbf{V_y}$ is clearly formed by superimposing two components. The first is diagonal and includes random detector-count errors. Larger integrated neutron fluences yield better statistical accuracy. The second term reflects systematic errors in neutron total cross-sections used for the analysis. The required cross-sections σ_{ki}

and their errors are generated from evaluated files, e.g., ENDF/B-VI.[4] It has been found that more consistent results are obtained when transmission-generated cross-sections are employed rather than direct energy-averaged cross-sections, as discussed in ref. 5. Eq.5 is an approximation which evolves from an application of the effective variance method.[6] The impact of errors in the total cross-sections for an individual element obviously depends on the areal density of that element. Since this is not known *a priori*, an iterative approach is required to implement the least-squares method. The first pass is carried out without including cross-section errors. Thus, we set $\mathbf{p} = 0$ in Eq.5. This leads to a solution $\mathbf{p}_a$ which is subsequently used to derive a revised matrix $\mathbf{V}_y$ via Eq.5 for the second pass of this analysis. If no cross-section errors are considered, then a second pass is not required and $\mathbf{p} = \mathbf{p}_a$ is the desired solution.

The expected value of χ^2/f is unity.[3] If it is significantly larger than unity, then one is alerted to the fact that the scatter in the data is inconsistent with the errors represented by $\mathbf{V}_y$, or that the model used for the unfolding is wrong, or that a combination of both factors contributes to the discrepancy. Every attempt should be made to try and trace down these sources of discrepancy. If the problem is not entirely resolved by this procedure, and χ^2/f is still larger than unity, one should "fix" the problem somewhat arbitrarily by simply adjusting the solution covariance matrix $\mathbf{V}_p$ via scaling of all its elements. That is, we make the substitution $\mathbf{V}_p \rightarrow [(\chi^2/f)^{1/2}]\mathbf{V}_p$, leaving the solution values $\mathbf{p}$ untouched. The rescaled matrix will automatically force χ^2/f to unity. This is not a totally satisfying resolution of the problem because it simply spreads unreconciled discrepancies proportionally across all the parameter errors. However, for the purposes of probabilistic interpretation of the results it is necessary to take this step because one cannot proceed with a probabilistic analysis that incorporates a matrix $\mathbf{V}_p$ with too small errors. What should be done if $\chi^2/f < 1$? This is a more perplexing problem. What it signifies is that the scatter in transmission data is smaller than the assigned errors. This is less likely to be traceable to random errors than to the systematic total cross-section errors. One could consider reducing the cross-section errors, but this seems arbitrary if these errors are derived from files and represent an evaluator's educated beliefs on the subject. We have found that quite small chi-square values often emerge from our simulation studies because cross-section uncertainties are not incorporated into the Monte Carlo simulation code MCNP but have generally been considered in our unfolding algorithm.[1,7] To circumvent this dilemma for the present discussion, we have chosen to treat an example in Section 4 which includes random errors but no cross-section errors. As long as the Monte Carlo simulation and unfolding analysis are carried out on equal footing, it is advisable to alter $\mathbf{V}_p$ by adjusting the errors even if $\chi^2/f < 1$ is obtained from unfolding. The technique is again to multiply the solution $\mathbf{V}_p$ by $(\chi^2/f)^{1/2}$ as shown above. From a probabilistic point of view, this is a reasonable choice, and it is likely that the error adjustments will be relatively modest.

3. PROBABILISTIC INTERPRETATIONS

It can be shown from the principles of information theory[3] that if one knows only the mean values $\mathbf{p}$ for a set of parameters $\mathbf{q}$ and their associated covariance matrix $\mathbf{V}_p$, the best choice is a multivariate normal probability density function P in the variable array $\mathbf{q}$, namely,

$$P(\mathbf{q};\, \mathbf{p},\, \mathbf{V}_p) = (2\pi)^{-m/2}\, [\det(\mathbf{V}_p)]^{-1/2}\, \exp\,[-(1/2)\,(\mathbf{q}-\mathbf{p})^{\dagger}\, \mathbf{V}_p^{-1}\, (\mathbf{q}-\mathbf{p})]. \tag{6}$$

This is a normalized function in that $\int P(\mathbf{q};\, \mathbf{p},\, \mathbf{V}_p)\, d\mathbf{q} = 1$. The first probabilistic approach we will consider here is based on Eq.6. Suppose that $\mathbf{p}_0$ represents values of the parameters for an illicit substance. Having obtained a solution $\mathbf{p}$ from unfolding the transmission data by least squares, we ask the question: "What is the risk R_0 that this solution signals detection of $\mathbf{p}_0$?" One answer to this question is given by the formula

$$R_0 = \int_U P(\mathbf{q};\, \mathbf{p},\, \mathbf{V}_p)\, d\mathbf{q}\,, \tag{7}$$

where U encompasses all points outside the m-dimensional closed surface (hyper-ellipsoid) defined by the probability density $\alpha = P(\mathbf{p}_0;\, \mathbf{p},\, \mathbf{V}_p)$, i.e., all points $\mathbf{q}$ belonging to the set $U = \{\mathbf{q}:\, P(\mathbf{q};\, \mathbf{p},\, \mathbf{V}_p) < \alpha\}$. Fig.2 illustrates the geometry for the case $m = 2$.

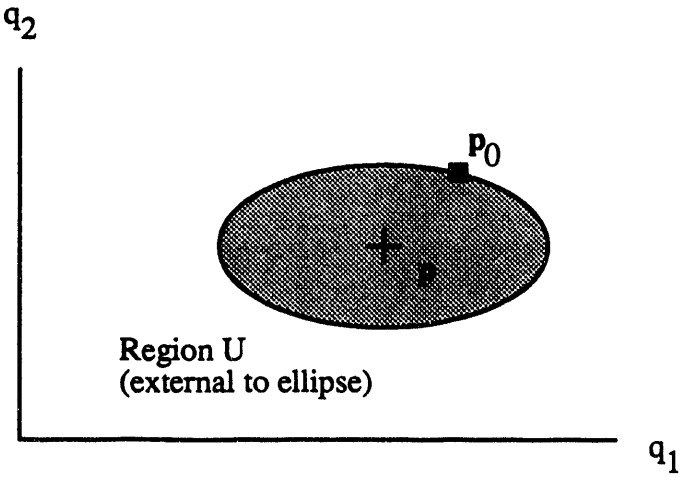


Fig. 2. Schematic diagram of an ellipse in two dimensions. The relationship between the solution array (**p**) and that corresponding to an illicit substance ($\mathbf{p}_0$) is indicated.

R_0 represents a certain probability that what we observe within the interrogated object by obtaining **p** is really an illicit substance defined by $\mathbf{p}_0$. The multi-dimensional integral in Eq.7 can be readily converted into a one-dimensional integral.[8] We obtain

$$R_0 = 1 - \frac{2^{1-m/2}}{\Gamma\left(\frac{m}{2}\right)} \int_0^{r_0} y^{m-1} e^{-y^2/2}\, dy, \tag{8}$$

where

$$r_0^2 = (\mathbf{p}_0 - \mathbf{p})^{\dagger}\, \mathbf{V}_p^{-1}\, (\mathbf{p}_0 - \mathbf{p})\,. \tag{9}$$

We can think of r_0 as a measure of "distance" between the known illicit substance $\mathbf{p}_0$ and the solution **p** in the units of reduced standard deviation. The table below shows the risk factor R_0 for different numbers of elements *m* as a function of r_0.

		Risk Factor			
m	r_0=1	**2**	**3**	**4**	**5**
1	31.731%	4.550%	0.270%	0.006%	0.000%
2	60.653%	13.534%	1.111%	0.034%	0.000%
3	80.125%	26.146%	2.929%	0.113%	0.002%
4	90.980%	40.601%	6.110%	0.302%	0.005%
5	96.257%	54.942%	10.906%	0.684%	0.014%
6	98.561%	67.668%	17.358%	1.375%	0.034%
7	99.483%	77.978%	25.266%	2.512%	0.076%
8	99.825%	85.712%	34.230%	4.238%	0.155%
9	99.944%	91.141%	43.727%	6.688%	0.297%
10	99.983%	94.735%	53.210%	9.963%	0.535%

We must ask whether this approach provides the most useful measure of risk probability. The region U over which integration occurs includes many other points besides $\mathbf{p}_0$. In a sense, what is given here is a very conservative estimate of risk, i. e., an upper bound on the risk that solution **p** could be an illicit substance characterized by $\mathbf{p}_0$.

An approach which appears to give a better statistical measure of risk is to apply the F-distribution test. Let H_0 be the null hypothesis that solution **p** signifies the presence of an illicit substance denoted by $\mathbf{p}_0$. From this point of view we can define the 100 (1-α)% confidence interval for H_0 by the condition

$$r_0^2 = (\mathbf{p}- \mathbf{p}_0)^\dagger \mathbf{V}_p^{-1} (\mathbf{p} - \mathbf{p}_0) \le m F_{m,n-m}(\alpha), \tag{10}$$

where n is the number of measurements and m is the number of elements used in the regression analysis, and $F_{m,n-m}(\alpha)$ is the upper (100 α)th percentile of an F-distribution with m and n-m degrees of freedom. If the condition in Eq. 10 is not satisfied, we must assume that the solution **p** is *not* an illicit substance with the probability for error given by α. If the condition in Eq. 10 holds, we cannot reject H_0 and must examine the situation further. It is important to note that the F-distribution test is designed to test the absence of an illicit substance at a given confidence level. As a practical matter, we suggest using $\alpha = 0.001$ or less. For example, if upon applying the test we find that Eq. 10 is *not* satisfied at a level of α = 0.001, we conclude that the solution **p** is *not* an illicit substance represented by $\mathbf{p}_0$ and the odds of the conclusion being wrong are less than 1 in 1,000.

Simultaneous 100(1-α)% confidence intervals from the regression analysis can be defined for each parameter p_i by

$$p_i \pm [(\mathbf{V}_p)_{ii}]^{1/2} [m F_{m,n-m}(\alpha)]^{1/2}. \tag{11}$$

In applying statistical tests, it is essential to use the optimal number of regression parameters. That is if we use a large number of elements to obtain a solution from Eq. 4 and find out that certain regression coefficients do not differ significantly from zero, we must redo the regression analysis with only the relevant number of elements included. Eq. 11 can serve to determine which elements may be safely dropped from the subsequent analysis. Also, we need to take into account the fact that the derived areal densities **p** are dependent upon the thickness of the object being interrogated. Since what is known about illicit substances are their raw elemental densities, we must design a way to incorporate the size-dependence of the solution into the test. One approach is to use size-independent qualifiers **q** to define the illicit substance. For example, we can choose to use the set $q_k = p_k / [\Sigma_{k=1}^{m} p_k]$ for $k = 1,...,m$, where m is the number of elements used in the final regression analysis. A useful way to identify such qualifiers is via expressions such as O/(H+C+N+O), O/N, H/(N+C), etc. Then, an illicit substance defined by $\mathbf{q}_0$ can be easily converted into a size-dependent vector $\mathbf{p}_0$ against which the solution **p** can be tested. The testing procedure is illustrated by an example in the next section.

4. AN EXAMPLE

Here we apply the statistical tests to the neutron transmission data generated with MCNP for two substances: explosive RDX ($C_3H_6N_6O_6$) and benign melamine ($C_3H_6N_6$). We denote these substances with subscripts 1 and 2, respectively, for testing purposes. We would like to test for the presence of RDX. As mentioned above, we do not consider cross-section errors in this analysis. In our simulations, we used n=193 for the number of time, or energy, bins.

We first apply Eq. 4 to compute regression coefficients p_k using ten elements: H, C, N, O, F, Al, Si, Cl, Fe, Cu. We see that the last 6 elements do not contribute anything to the solution and, according to Eq. 11, the coefficients p_5 through p_{10} are zero within a 99.9% confidence interval. We then rerun the regression calculation using just 4 elements (H, C, N, and O), and scale the resulting covariance matrix $\mathbf{V}_p \rightarrow [(\chi^2/f)^{1/2}] \mathbf{V}_p$ as described above. We get for $m = 4$,

$$\mathbf{p}_1 = (0.0860, 0.04419, 0.0911, 0.0892),$$
$$\mathbf{p}_2 = (0.2223, 0.1120, 0.2316, 0.0000),$$

as the solution values for the integrated elemental densities and

$$V_{p_1} = 10^{-7}\begin{pmatrix} 31.8616 & & & \\ -27.2233 & 114.361 & & \\ -15.8349 & -63.0473 & 101.046 & \\ -2.73119 & -21.2529 & 1.95051 & 22.5448 \end{pmatrix}, \quad V_{p_2} = 10^{-7}\begin{pmatrix} 27.064 & & & \\ -23.1241 & 97.1408 & & \\ -13.4505 & -53.5538 & 85.8307 & \\ -2.31994 & -18.0527 & 1.65681 & 19.1501 \end{pmatrix},$$

for the corresponding symmetric covariance matrices of the two tested substances. The RDX substance we test against has the following qualifiers: H/(H+C+N+O)=2/7, C/(H+C+N+O)=1/7, N/(H+C+N+O)=2/7, O/(H+C+N+O)=2/7, i. e. q_0 = (2/7,1/7,2/7,2/7). We transform q_0 to an equivalent p_0.The Substance 1 will be tested against p_{01}= (0.0887, 0.0444, 0.0887, 0.0887) and the Substance 2 will be tested against p_{02}= (0.1614, 0.0807, 0.1614, 0.1614) in accordance with the formula given in the previous section. We get following values for the "distance" of the tested substances from the "RDX point" from Eq. 9,

$$r_{01} = 2.62 \quad \text{(Substance 1)}, \qquad r_{02} = 126 \quad \text{(Substance 2)},$$

and the corresponding risk levels from Eq. 8,

$$R_{01} = 0.14 \quad \text{(Substance 1)}, \qquad R_{02} < 10^{-14} \quad \text{(Substance 2)}.$$

Note that if we include all 10 elements, the risk level for the Substance 1 becomes about 74% while the risk level for the Substance 2 essentially stays at zero.

We can also test the hypotheses that the Substance 1 is RDX and that the Substance 2 is RDX. As shown in Eq. 10, the test involves comparing r_0^2 with $4F_{4,189}(\alpha)$ for a 100 (1-α)% confidence interval. The values of $4F_{4,189}(\alpha)$ and r_0^2 are shown in the table below for different levels of α.

α	$m F_{m,n-m}(\alpha)$	r_0^2	
		Substance 1	Substance 2
0.1	7.899	6.88	1.58×10^4
0.01	13.68	6.88	1.58×10^4
0.001	19.29	6.88	1.58×10^4
0.0001	24.90	6.88	1.58×10^4
0.00001	30.56	6.88	1.58×10^4

We clearly see that we can reject the Substance 2 as RDX with the odds of being wrong less than 1 in 10^5 while the Substance 1 does not fail the test for RDX even at the 10% level.

5. CONCLUSIONS

The statistical methods of analysis and hypothesis testing presented in this paper are powerful tools for illicit substance detection. Using a multivariate regression model, we can determine from the fast-neutron transmission spectra the most likely integrated elemental densities for the interrogated substance. Coupled with their computed covariance matrix, they provide a basis for testing the substance against a known explosive or drug. Using results from MCNP simulations of transmission experiments, we can see the strengths and limitations of two different statistical methods. In the end, we are left with a probabilistic measure of the risk that an illicit substance is present. Studies of simulated data can show which risk levels are significant and with what confidence the results of testing may be trusted. By placing numerical benchmarks at the disposal of the interrogator, our statistical unfolding algorithm provides an objective framework for making decisions on risk based on probabilities.

6. ACKNOWLEDGEMENTS

The authors acknowledge support from the Federal Aviation Administration Technical Center under contract DTFA03-03-X-00021 and the Counterdrug Technology Assessment Center, Office of National Drug Control Policy, under contract # 6-CO-160-00-195. M.K. Harper was supported by the Student Research Participation Program administered by Argonne's Division of Educational Programs.

7. REFERENCES

1. B.J. Micklich, M.K. Harper, A.H. Novick, D.L. Smith, "Illicit Substance Detection Using Fast-Neutron Transmission Spectroscopy," *Proc. The 8th Symposium on Radiation Measurements and Applications*, Ann Arbor, MI, 1994.

2. J.W. Meadows, "The Thick-Target ^{9}Be(d,n) Neutron Spectra for Deuteron Energies Between 2.6 and 7.0 MeV," Report ANL/NDM-124, Argonne National Laboratory, 1991.

3. D.L. Smith, *Probability, Statistics, and Data Uncertainties in Nuclear Science and Technology*, American Nuclear Society, La Grange Park, Illinois, 1991.

4. P.F. Rose, "ENDF-201: ENDF/B-VI Summary Documentation", *BNL-NCS-17541 (ENDF-201, 4th Edition, ENDF/B-VI)*, Brookhaven National Laboratory, 1991.

5. B.J. Micklich, M.K. Harper, L. Sagalovsky, D.L. Smith, "Nuclear Data Needs and Sensitivities for Illicit Substance Detection Using Fast-Neutron Transmission Spectroscopy," *Proc. International Conference on Nuclear Data for Science and Technology*, Gatlinburg, TN, 1994.

6. Jay Orear, "Least Squares When Both Variables Have Uncertainties", *Am. J. Phys. Vol. 50, No. 10*, pp. 912-916, 1982.

7. J. F. Briesmeister, Ed., "MCNP - A General Monte Carlo N-Particle Transport Code, Version 4A," LA-12625, Los Alamos National Laboratory, 1993.

8. Richard A. Johnson, Dean W. Wichern, *Applied Multivariate Statistical Analysis, 3rd Edition,* Prentice Hall, Englewood Cliffs, NJ, 1992.

DATE

FILMED

7/19/94

END