H13-15A
COMPARATIVE INVESTIGATION OF SOURCE TERM ESTIMATION ALGORITHMS USING FUSION FIELD TRIAL 2007 DATA – LINEAR REGRESSION ANALYSIS

Nathan Platt and Dennis Deriggi

Institute for Defense Analysis, Alexandria, Virginia, USA

Abstract: The release of hazardous materials into the atmosphere on the battlefield or in populated areas must be considered for future scenarios. Given a warning based on detections at a few sensors, it should be useful to rapidly provide an estimate of the location, time of release, and amount of material released. Such information could lead to refined predictions of the hazardous area and support follow-on actions to investigate the cause and nature of the hazardous release. In September 2007, a short-range test – Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007 (FFT 07) – designed to collect data to support development of prototype source term estimation (STE) algorithms was conducted. A comparative investigation of STE algorithms began in 2008. First, a subset of sensor data from selected FFT 07 trials was provided to participating algorithm developers. Next, developers provided "blind" STE predictions which were then independently compared to the parameters of the actual release. A total of eight different STE algorithm developers participated in this exercise. A total of fourteen full and partial sets of predictions were received with some exercise participants providing multiple sets of predictions based on different algorithms they have been developing. Linear regression analysis considered several variables that might influence results including the number of sensors (four versus sixteen), the release type (instantaneous versus continuous), the time of the release (day versus night), meteorological inputs ("research-grade" inputs versus "simulated" operational inputs), and the number of sources (single versus double versus triple versus quad releases). Both backward and stepwise regressions were conducted to explore the variables that might most affect predictive performance. The results of these analyses are used to ascertain trends among different sets of STE predictions and are presented here.

Key words: *Source term estimation, sensor data fusion, inverse transport and dispersion, backtracking.*

INTRODUCTION

Given a warning based on detection of hazardous materials at only a few sensors, it could be useful to rapidly provide an estimate of the source location, time of release, and amount of material released. Such an estimate can lead to refined predictions of the area affected by the release and can support near-term follow-on actions to investigate the cause and nature of the release. In some cases, refined predictions that could result from such source term estimation (STE) can support tactical decisions (e.g., which roads to travel on and which to avoid). For longer-range situations (tens of kilometres), accurate estimates of the source can allow for improved hazard-area predictions that could better support warnings and possible evacuation, efficient mission-oriented protective posture gear usage, or perhaps medical response.

In September 2007, a short-range (~500 meters), highly instrumented test was conducted at the U.S. Army's Dugway Proving Ground (DPG) (Storwold, 2007). This test, referred to as Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007 (FFT 07), was designed to collect data to support the further development of prototype algorithms. FFT 07 was sponsored by DTRA-Joint Science and Technology Office for Chemical and Biological Defense (JSTO-CBD) and was conceived of and planned within the Technical Panel 9 for Hazard Assessment (TP9) of The Technical Cooperation Program (TTCP) Chemical, Biological, and Radiological Defense (CBD) group, thus making this effort an international (in this case, U.S., U.K., Canada, and Australia) collaboration.

The reasons for conducting FFT 07 were numerous. First, the experiment was meant to provide a set of data that STE model developers can use to improve their algorithms. Next, the collected information could be used to assist in identifying the strengths and weaknesses of the different modelling approaches chosen by the developers. Finally, assessment of STE algorithms using data collected during FFT 07, was meant to help the Department of Defense identify the current state of the STE algorithms in general (the "state of the art").

A comparative investigation of STE algorithms began in 2008. The general method of this investigation was to first provide participating developers with a subset of sensor data that was collected on selected FFT 07 trials. Next, developers provided "blind" predictions that were compared to parameters of the actual release. Phase I of this investigation consisted of 104 individual cases of simulated sensor data that were distributed in September 2008. Table 1 lists the composition of Phase I cases. These cases simulated continuous streams of concentration data for ingestion by STE algorithms. Each case selected for Phase I evaluation was created using available digiPID data.

Table 1. Composition of Phase I cases that were distributed to STE algorithm developers to provide predictions.

Phase I Release Case Composition					
Condition	**All Trials**	**Single**	**Double**	**Triple**	**Quad**
none	104	40	40	16	8
Puff	52	20	20	8	4
Cont	52	20	20	8	4
Daytime	52	20	20	8	4
Nighttime	52	20	20	8	4
Daytime/Puff	26	10	10	4	2
Daytime/Cont	26	10	10	4	2
Nighttime/Puff	26	10	10	4	2
Nighttime/Cont	26	10	10	4	2

Phase I consisted of cases that equally sampled the parameters that were potentially expected to most significantly affect the quality of STE predictions. These parameters included time of day of the tracer release (day or night), type of tracer release (continuous or instantaneous), and number of sensors reporting data (4 or 16). To provide some realism in the meteorological inputs to the STE algorithms, for some cases the developers were provided with surface wind velocity observations and a vertical wind velocity profile from sites up to 1-2 km removed from the tracer releases and sampler grid, instead of the more detailed meteorological observations made at the centre location of the sampler grid itself. An additional sampled parameter that could affect the quality of STE predictions was the number of sources (single, double, triple, or quad). FFT 07 individual puff trials involved multiple (up to 10) realizations. These puffs were released by firing air cannons every few minutes resulting in "trains of puffs" periodically traversing the digiPID grid. Hence, for puff releases, some distributed cases included a single realization of the puff(s), while some of the distributed cases included multiple (up to 10) realizations. The full structure of the Phase I cases and further details on the composition of the phases of the evaluation and construction of the cases are available in Platt, *et al.*, 2008a, 2008b.

A total of eight different STE algorithm developers participated in this exercise. A total of 14 full and partial sets of predictions were received with some exercise participants providing multiple sets of predictions based on different algorithms that they have been developing. Table 2 depicts the organizations that participated in Phase I together with the composition of predicted cases that they provided. Red denotes that a full set of predictions was provided; blue denotes that predictions were provided for at least 50 percent of the distributed cases. Table 3 briefly lists some basic capabilities of each of the STE algorithms that include ability to predict number (e.g., single, double, triple, or quad source) and types of sources (e.g., continuous or instantaneous puff release).

Table 2. Organizations participating in Phase I together with the composition of predicted cases they provided broken down into several categories including release type, time of day, and number of sources. The red font values denote that a full set of predictions was provided; blue font values denote that the predictions were provided for at least 50 percent of the distributed cases.

Composition of the Prediction Sets Recieved									
Organization	Total	Cont	Puff	Daytime	Nighttime	Single	Double	Triple	Quad
Aerodyne	104	52	52	52	52	40	40	16	8
Boise-State	33	14	19	21	12	13	13	4	3
Buffalo / GA	104	52	52	52	52	40	40	16	8
Buffalo / SA	70	34	36	34	36	26	26	12	6
DSTL	35	5	30	20	15	12	14	7	2
ENSCO / Set 1	102	51	51	50	52	39	39	16	8
ENSCO / Set 2	104	52	52	52	52	40	40	16	8
ENSCO / Set 3	42	24	18	19	23	13	15	10	4
NCAR / Variational	38	3	35	20	18	16	14	4	4
NCAR / Phase I	38	3	35	20	18	16	14	4	4
Sage-Mgt	104	52	52	52	52	40	40	16	8
PSU / Gaussian	50	26	24	25	25	18	20	8	4
PSU / SCIPUFF	50	26	24	25	25	18	20	8	4
PSU / MEFA	35	19	16	17	18	13	16	5	1

Table 3. Basic capabilities of each of the STE algorithms that submitted predictions to the Phase I of the exercise.

Algorithm Capabilities		
Organization	Number of Sources	Type
Aerodyne	Multi	Cont/Puff
Boise-State	Single	Cont/Puff
Buffalo / GA	Multi	Cont/Puff
Buffalo / SA	Mostly Single	Cont/Puff
DSTL	Single	Puff
ENSCO / Set 1	Multi	Cont/Puff
ENSCO / Set 2	Single	Cont
ENSCO / Set 3	Single	Cont
NCAR / Variational	Single	Puff
NCAR / Phase I	Single	Puff
Sage-Mgt	Single	Cont/Puff
PSU / Gaussian	Single	Cont/Puff
PSU / SCIPUFF	Single	Cont/Puff
PSU / MEFA	Multi	Cont/Puff

STE ALGORITHM PERFORMANCE TRENDS

The goal of these evaluations is not to declare a "winning" algorithm, but rather to learn by examining the strengths and weaknesses of each of the proposed methodologies, because different approaches may best apply to different sets of tracer release scenarios (i.e., daytime versus night-time, single- versus double- source release, richer information available from simulated sensors versus poorer) or for different specific applications (e.g., near real-time versus forensic). In this way, algorithm developers can learn from each other. The main motivation behind the evaluation matrix is an attempt to trade off the ability to cover the evaluation of a substantial number of potential variables that might influence algorithm performance with the desire to keep the sample sizes large enough to be able to arrive at reasonably robust conclusions. Therefore, we decided to start our analysis with the evaluation of algorithm performance trends instead of analyzing each individual algorithm.

As depicted in Table 3, the individual STE algorithms participating in Phase I evaluations have different capabilities with respect to predicting numbers and types of sources. In order to make a fair comparison among these algorithms, for each individual case for which a prediction was submitted, common metrics that are applicable to all of the algorithms needed to be defined. We selected two metrics: the distance between the averaged predicted and averaged observed source term locations, and the ratio of the observed and predicted release mass from all sources. Figure 1 illustrates the distance metric calculation. From the 14 sets of STE algorithm predictions, 12 algorithms provided enough information to calculate the mass ratio and all 14 algorithms provided enough data to calculate our distance metric.

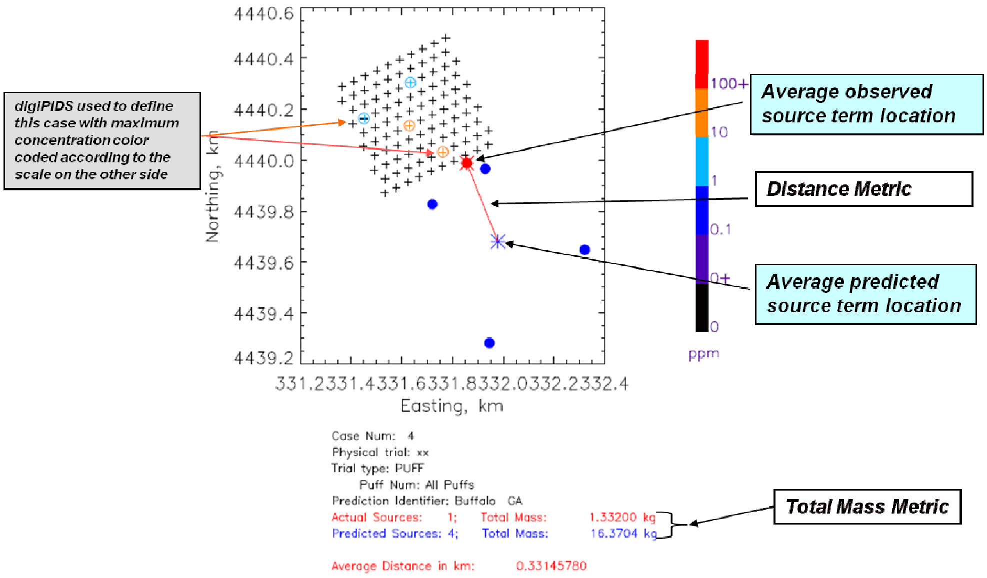

Figure 1. Example of the distance metric computation and total mass calculation used to compare algorithms for individual cases.

LINEAR REGRESSION ANALYSIS DESCRIPTION AND RESULTS

This section describes the use of linear regression to the examination of source term estimation algorithms. In very general terms, this effort was an attempt to determine which of the underlying factors such as diurnal condition, number of release sources, type of release, and several other independent variables, had the greatest effect on the estimation of the mass ratio (the ratio of reported to actual mass) or mean offset distance. Two regression techniques were used in this study: backward regression and stepwise regression.

Backward regression begins with all the independent variables in the regression equation, and then proceeds to eliminate those for which the associated sum of squares is not significant. In contrast, stepwise regression only allows independent variables into the regression equation if their associated sum of squares is significant, and eliminates previously admitted variables if their effect is substantially diminished by other variables in the equation.

Thus, roughly speaking, stepwise regression tests each independent variable to determine whether or not it should enter the regression equation, and again, if it should remain in the equation after others are admitted. Backward regression initially treats all variables as belonging to the equation, then eliminates those whose contribution is substandard. For reference, please see Draper and Smith (1966) and Seber (1977).

We chose the following independent regression variables: 1) "Diurnal," defined as either Night or Day release time; 2) "Met Num," defined as either "Close-in" met corresponding to meteorology obtained at the centre of the digiPID grid, or "Operational" met, which corresponded to using meteorological stations approximately 1-2 km away; 3) "Sources" denoting the number of sources used in the definition of a case (single, double, triple, quad); 4) "Sensors" denoting the number of simulated sensors used in the definition of a case (4 or 16); 4) "Puff/Real," defined as: "-1" if case is constructed from a continuous trial, "0" if case is constructed using single realization of a puff trials, and "1" if case is constructed using multiple realizations of a puff trial. The "Puff/Real" independent variable is expected to succinctly represent two distinct parameters that could affect quality of STE predictions: continuous vs. instantaneous/puff releases and single vs. multiple releases from the same location.

The list of dependent regression variables included: "Mean," defined as the distance between averaged predicted and averaged observed source term locations for the individual case as shown in Figure 1, and "Mass Ratio," defined as a ratio of predicted to observed total mass of the material used to define a particular case.

The results for stepwise and backward regressions are summarized in Tables 4 and 5, respectively. To simplify viewing these tables, the coloured background in the table entries is coded according to which independent variable is called by the particular significant factor.

Table 4. Table of significant factors for backward regression. This table is divided into two sections, one for each dependent variable. Each section contains the proportion of variance explained by regression (adjusted R2), independent variables selected by backward regression, standard coefficient for that variable, unstandardized coefficient, and significance level. All computations were performed using SPSS 15.0 with a removal criterion of 10 percent significance as determined by the appropriate partial F-test. Background cell colours are used to designate and group individual variables that were considered significant by the regression analysis.

model	dependent	R2	significant factor	significant factor	significant factor
ENSCO 3	Mass Ratio	0.379	Puff Real (0.51, 2.49, 0)	Sources (-0.447, -1.9, 0.001)	
Buffalo SA	Mass Ratio	0.273	Sources (-0.348, -0.723, 0.002)	Met Num (0.235, 0.632, 0.031)	Diurnal (0.231, 0.508, 0.029)
DSTL	Mass Ratio	0.254	Puff Real (-0.567, -287.1, 0.001)	Sources (-0.376, -75.9, 0.026)	
ENSCO 2	Mass Ratio	0.221	Puff Real (0.37, 1.3, 0.0)	Sources (-0.32, -0.93,0)	Sensors (0.17, 0.074, 0.06)
PSA Gaussian	Mass Ratio	0.209	Puff Real (0.46, 0.059, 0.01)	SourceS (-0.407, -0.037, 0.02)	
PSU SCIPUFF	Mass Ratio	0.203	Sources (-0.5, -0.011, 0.035)		
Buffalo GA	Mass Ratio	0.172	Sources (-0.365, -2.376, 0)	Puff Real (0.183, 1.417, 0.044)	Diurnal (0.177, 1.224, 0.051)
ENSCO 1	Mass Ratio	0.15	Puff Real (0.398, 14.64, 0)		
Aerodyne	Mass Ratio	0.096	Puff Real (0.262, 0.852, 0.006)	Sensors (-0.212, -0.089, 0.026)	
NCAR Phase I	Mass Ratio	0	constant		
NCAR Variation	Mass Ratio	0			
SAGE Mgt August	Mass Ratio	0			
Boise State	Mass Ratio		NO DATA		
PSU MEFA	Mass Ratio		NO DATA		
model	dependent	R2	significant factor	significant factor	significant factor
DSTL	Mean	0.67	Puff Real (-0.725, -1.105, 0)	Sources (0.212,0.129, 0.056)	
NCAR Phase I	Mean	0.266	Sources (0.534, 0.09, 0.001)		
NCAR Variation	Mean	0.204	Sources (0.475, 0.09, 0.003)		
ENSCO 3	Mean	0.148	Sources (-0.366, -0.031, 0.015)	Sensors (0.258, 0.003, 0.08)	
PSA Gaussian	Mean	0.102	Sources(0.306, 0.055, 0.029)	Puff Real (-0.254, -0.057, 0.069)	
SAGE Mgt August	Mean	0.083	Sources (0.303, 0.204, 0.002)		
ENSCO 1	Mean	0.043	Met Num (0.228, 0.009, 0.021)		
ENSCO 2	Mean	0.04	Sensors (-0.173, -0.002, 0.076)	Met Num (0.169, 0.017, 0.083)	
Aerodyne	Mean	0.033	Sensors (-0.206, -0.003, 0.036)		
Boise State	Mean	0	constant		
Buffalo GA	Mean	0	constant		
Buffalo SA	Mean	0			
PSU MEFA	Mean	0	constant		
PSU SCIPUFF	Mean	0	constant		

Table 5. Table of significant factors for stepwise regression. As in the previous Table 4, Table 5 is divided into two sections, one for each dependent variable. Each section contains the proportion of variance explained by regression (adjusted R2), independent variables selected by stepwise regression, standard coefficient for that variable, unstandardized coefficient, and significance level. All computations were performed using SPSS 15.0 with an entry criterion of 5% and an elimination criterion of 10% significance as determined by the appropriate partial F-tests. Background cell colours are used to designate and group individual variables that were considered significant by the regression analysis.

model	dependent	R2	significant factor	significant factor	significant factor
ENSCO 3	Mass Ratio	0.379	Puff Real (0.51, 2.49, 0)	Sources (-0.447, -1.9, 0.001)	
Buffalo SA	Mass Ratio	0.273	Sources (-0.348, -0.723, 0.002)	Met Num (0.235, 0.632, 0.031)	Diurnal (0.231, 0.508, 0.029)
DSTL	Mass Ratio	0.254	Puff Real (-0.567, -287.1, 0.001)	Sources (-0.376, -75.9, 0.026)	
PSU SCIPUFF	Mass Ratio	0.203	Sources (-0.5, -0.011, 0.035)		
ENSCO 2	Mass Ratio	0.201	Puff Real (0.37, 1.3, 0)	Sources (-0.32, -0.93, 0)	
ENSCO 1	Mass Ratio	0.15	Puff Real (0.398, 14.64, 0)		
Buffalo GA	Mass Ratio	0.125	Sources (-0.365, -2.376, 0)		
Aerodyne	Mass Ratio	0.096	Puff Real (0.262, 0.852, 0.006)	Sensors (-0.212, -0.089, 0.026)	
NCAR Phase I	Mass Ratio	0			
NCAR Variation	Mass Ratio	0			
PSU Gaussian	Mass Ratio	0			
SAGE Mgt August	Mass Ratio	0			
Boise State	Mass Ratio		NO DATA		
PSU MEFA	Mass Ratio		NO DATA		
model	dependent	R2	significant factor	significant factor	significant factor
DSTL	Mean	0.641	Puff Real (-0.807, -1.23, 0)		
NCAR Phase I	Mean	0.266	Sources (0.534, 0.09, 0.001)		
NCAR Variation	Mean	0.204	Sources (0.475, 0.09, 0.003)		
ENSCO 3	Mean	0.101	Sources (-0.35, -0.03, 0.023)		
SAGE Mgt August	Mean	0.083	Sources (0.303, 0.204, 0.002)		
ENSCO 1	Mean	0.043	Met Num (0.228, 0.009, 0.021)		
Aerodyne	Mean	0.033	Sensors (-0.206, -0.003, 0.036)		
Boise State	Mean	0			
Buffalo GA	Mean	0			
Buffalo SA	Mean	0			
ENSCO 2	Mean	0			
PSU Gaussian	Mean	0			
PSU MEFA	Mean	0			
PSU SCIPUFF	Mean	0			

The regression outcomes were ranked in decreasing order of their respective ("adjusted") coefficients of determination. The coefficient of determination is referred to as R2 and is the square of the correlation between observations and values returned by the regression equation. It is equal to the proportion of the variance in observed data that can be "explained" by regression (G. Seber, 1977). The adjusted R2, which determined the ordering, takes into account the number of variables in the model and is equal to $1 – (1 - R2)(n-1)/(n -p - 1)$, where p is the number of independent variables in regression equation and n is the number of observations. The point of using the adjusted R2 is to force models to be economical by penalizing excessive numbers of independent variables. This is in contrast to the (unadjusted) R2, which increases with the number of independent variables.

Two types of regression coefficients are presented in this paper: standardized and unstandardized. The former refers to regression coefficients obtained after transforming all data so that the dependent variable and all independent variables have a mean of zero and a standard deviation of 1.0. In some sense, this treats all data as being on an equal footing. The unstandardized coefficients are the result of performing regression without this transformation. For each model listed in Tables 4 and 5, both types of coefficients appear in parentheses after each independent variable that was selected by the regression process. The level of significance – that is the probability of the same or more extreme observations under the null hypothesis that this coefficient was zero (loosely speaking, the probability that the coefficient is zero), also appears in parentheses after the coefficient. Models with gray backgrounds in Tables 4 and 5 are those for which there were no data or which regression was not significant.

With respect to predicting miss distance between predicted and observed STE location, the regression analysis indicates:

1. "Diurnal" (Day/Night) and "Met Num" variable representing "Close-In" versus "Operational" met options is not a significant variable for both backward and stepwise regressions for almost all algorithms.
2. "Sources" variable representing number of sources used in the definition of a case is a significant predictor of algorithm performance for six algorithms. Six algorithms are called by stepwise regression and four algorithms are called by backward regression.
3. "Sensors" regression variable representing number of sensors (4 vs. 16) used in the definition of the case is a significant predictor of algorithms performance for only three algorithms. This indicates that most STE algorithms do not benefit from having a larger number of sensors.
4. "Puff Real" regression variable is a significant predictor for algorithm performance for two algorithms using backward regression and one algorithm using stepwise regression.

With respect to mass ratio dependent variable, regression analysis indicates:

1. "Diurnal" (Day/Night), "Met Num" (Close-In/Operational met) and "Sensors" (4 vs. 16) are not significant variables for most algorithms for both backward and stepwise regression.
2. "Sources" independent regression variable representing number of sources used in the definition of a case is a significant predictor of algorithm performance for seven algorithms.
3. "Puff Real" regression variable is a significant predictor for algorithm performance for seven algorithms.

We would like to caution that regression analysis results should serve as a guide for further investigation of which algorithm/variable combinations are important. For instance, the regression analysis does not tell if the algorithm performed as expected with respect to a given variable.

Phase II of this exercise is being planned to start in FY 10 and incorporate: 1) lessons learned from Phase I, 2) the addition of "bar-sensor" input data stream, and 3) the use of a simulated environment to supplement the field trial data.

ACKNOWLEDGEMENTS

This effort was supported by the Defense Threat Reduction Agency with Dr. John Hannan as the project monitor and Institute for Defense Analysis professional development program. The authors would like to thank Edward Argenta, Donald P. Storwold, Jr., and John White of the Meteorology Division, West Desert Test Center, U.S. Army Dugway Proving Ground for providing access to initial FFT 07 test results and for providing useful review of the evaluation plan. The authors also would like to thank Paul Bieringer and Jon Hurst from the National Center for Atmospheric Research for helping design and coordinate a common output format for the STE predictions. Finally, we would like to thank all STE participants who provided predictions. The views expressed in this paper are solely those of the authors.

REFERENCES

Draper, N. and H. Smith 1966: Applied Regression Analysis, Wiley.

Platt, N., Warner, S. and S.M. Nunes, 2008a: *Plan for initial comparative investigation of source term estimation algorithms using FUSION field trial 2007 (FFT 07)*, Institute for Defense Analyses Document D-3488.

Platt, N., Warner, S. and S.M. Nunes, 2008b: *Evaluation plan for comparative investigation of source term estimation algorithms using FUSION field trial 2007 data*, Croatian Meteorological Journal, Proceedings of the 12th International Conference on Harmonization within Atmospheric Dispersion Modelling for Regulatory Purposes, Part 1: Oral Presentations, Vol. 43, p. 224-229.

Seber, G. 1977: Linear Regression Analysis, Wiley.

Storwold, D.P, 2007: *Detailed test plan for the Fusing Sensor Information from Observing Networks (FUSION) field trial 2007 (FFT 07.* West Desert Test Center, U.S. Army Dugway Proving Ground, WDTC Document No. WDTC-TP-07-078.