Department of Pesticide Regulation

Brian R. Leahy
Director

M E M O R A N D U M

Edmund G. Brown Jr.
Governor

TO: Randy Segawa
Environmental Program Manager I
Environmental Monitoring Branch

FROM: Bruce Johnson, Ph.D. *Original signed by*
Research Scientist III
Environmental Monitoring
916-324-4106

DATE: December 4, 2012

SUBJECT: REANALYSIS OF LOST HILLS DATA FOR CONSISTENCY AND ERROR DETECTION

Summary

Procedures used to calculate the period-by-period flux densities in Ajwa and Sullivan (2012) were incorporated into a visual basic program. This program was tested and compared to the published analysis in Ajwa and Sullivan (2012). Amongst the 4 fields, there were 112 flux periods for each of two fumigants: 1,3-dichloropropene (1,3-D) and chloropicrin (PIC), for a total of 224 flux period calculations. The program was successful in duplicating 91 percent (%) of the calculations in the report according to the calculation policy presented in the report. The remaining 9% consisted of 4% mistakes in the report; 2% where according to policy there was insufficient information for flux calculation; 2% where interpolation/substitution was used even though there was sufficient information for calculation; and 1% where a peculiarity in Excel's evaluation of significance for regression with no intercept differed from conventional approaches. Mistakes and inconsistent substitutions/interpolations were recalculated (6%) and the regression significance properly calculated (1%). After including these changes, agreement with the visual basic program was 98% with 2% representing cases where professional judgment was required. The impact of these corrections to Ajwa and Sullivan (2012) was to modestly reduce their published 8 cumulative fluxes (two fumigants over 4 fields), with a maximum reduction of about 10%.

Background

Ajwa and Sullivan (2012) provided results from a flux study for 1,3-D and PIC. This study utilized four fields and two fumigants and provides a rich data set for not only the question of timing of tarp removal, but also questions relating to the best way to analyze period concentration data in order to estimate flux. As a step towards investigating the latter, it is necessary to develop a computer program which embodies the computation policy espoused by Ajwa and Sullivan (2012), which represents one approach to estimating flux using back-

calculation techniques. This effort also provided an opportunity to recalculate the data and fix any errors found in the report.

Methods

The Excel file: "linestroutine3.xlsm" and a visual basic program were used for the analysis. Linestroutine3.xlsm consisted of four blocks of columns. The first block (columns 1 through 15) contained the measured and ISCST3-modeled concentrations for all four fields (Table 1). An indexing system in the first four columns kept track of fumigant (1=1,3-D;2=PIC), field (1-4), period (1 up to 42), monitor (1 up to 16). These were followed by the measured and ISCST3-modeled concentrations, as well as the start and end row for each period. These row numbers were converted into addressing references (Table 1). These addresses made it easy to work on blocks of data which represented monitoring periods using the worksheet function "INDIRECT."

The second block (columns 16 through 35) consisted of ordinary least squares regression analysis with an intercept. This analysis was all performed using worksheet formula. Each row in this section corresponded to period of data (Table 2). The formulas used in this section are outlined in Table 3. Table 3 is vertical in order to conveniently show the formulas for the second row of the linestroutine3.xlsm worksheet, which is the first horizontal row of data analysis. The INDIRECT worksheet function is handy for referencing the data blocks contained in columns 13 and 14 for each period. These statistics correspond to the analysis which is a part of the regression component of the Data Analysis package which is distributed with Excel. The significance level of 0.054 was used because that is the significance level used in Ajwa and Sullivan (2012).

The third block (columns 36 through 54, Table 4) consisted of calculation results from the visual basic program which followed the calculation policy outlined in Ajwa and Sullivan (2012), depicted in Figure 1. The entire program is listed in Appendix 1. It makes use of worksheet regression results in the second block, primarily to make some of the decisions indicated in Figure 1. The principle function used to perform the nonintercept regression is LINEST. This is an array function which gives results in 2 columns by 5 rows. Certain results have to be picked out of this array and added to the current row horizontally in order to avoid overwriting results. Another routine, PER25SULL calculates the 25th percentile of the measured values, which is a component in determining which regression to use when the intercept of the OLS is significant. The worksheet column 7 is used as a workspace for sorting the measured values in order to perform this calculation. The detailed methodology used by Ajwa and Sullivan (2012) for determining the 25th percentile starts by assigning the highest value 100% and the lowest value 0%. The remaining values are assigned percentiles equally space between these two extremes. The 25th percentile is interpolated between the bracketing percentiles.

The analysis policy (Figure 1) includes evaluation of the "significance" of the regression with no intercept. Steel and Torrie (1960) describe regression with no intercept. They state on page 179: "In some situations, theory calls for a straight line that passes through the origin....When there are reservations about the assumption that regression is through the origin, it may be desirable to test this as a hypothesis." The policy in Figure 1 calls for testing for the significance of the intercept and *if it is significant and greater than the 25th percentile of measured values,* regression through the origin is performed. This directly contradicts the recommendation in Steel and Torrie and is, in part, the reason that the Department of Pesticide Regulation (DPR) cannot endorse this calculation policy.

The F value and significance calculation of the regression with no intercept cannot be interpreted in a customary sense. Since the goal of this memorandum and the visual basic computer program was to mimic the calculations in this policy, the logic of the policy in Figure 1 was followed. The statistical tests for the no-intercept case, especially after determining that the intercept is significant, are not meaningful in a conventional sense.

The fourth block (columns 55 through 69) consisted of a copy of the analysis from the Lost Hills report (Ajwa and Sullivan 2012), and a comparison column, which was conditionally formatted to turn pink when the final result from the visual basic program and the final result from Ajwa and Sullivan (2012) differed by more than 0.01. I made notes in these columns as to the nature of the differences.

The procedure was iterative in that I ran the visual basic program, examined differences, then modified the program in order to better reflect the analysis policy in Figure 1, I reran the program and continued focusing on the differences. For example, using the absolute value of the intercept to compare to the 25th percentile of the measured values is not overtly described in Ajwa and Sullivan (2012), but was subsequently confirmed by them (Sullivan personal communication). I did not try to create an algorithm for either interpolating or substituting values when according to the policy in Figure 1 there were insufficient data for analysis (2 or less measured values at 0.1ug/m^3 or greater) because this operation involved some subjectivity. However, the program checked for and flagged these cases in order to facilitate my reconciliation.

Results

Overall agreement between Ajwa and Sullivan (2012) and the visual basic program was 91% of the periods (Table 5). In 9 cases (Table 5, category 1), there were mistakes made in Ajwa and Sullivan (2012). These consisted of two types: in about half the cases the final flux estimate was inexplicably multiplied by a factor of 10, and in the other cases the results of several different flux estimation methods were added together instead of a single one being selected. The visual basic program did not attempt to interpolate or substitute in cases where the policy described

insufficient measured values above 0.1 ug/m^3. It just identified these 5 cases where there were insufficient measured values, according to policy (Table 5, category 3). In 5 other cases (Table 5, category 4), Ajwa and Sullivan (2012) interpolated or substituted for a period, when there was, in fact, a sufficient number of measured values for analysis (3 or more above 0.1 ug/m^3). The fumigant, field and period numbers corresponding to these categories are listed in Appendix 2.

Finally, there were two cases where the LINEST gave a different probability than the VB program (Table 5, category 5). This was due to LINEST using n-2 degrees of freedom for the denominator term, whereas the visual basic program used n-1 degrees of freedom. The data analysis pack used in Excel relies on the LINEST routine and thus reflects the LINEST degrees of freedom used. In an example case (Table 6), the significance level determined by LINEST on the F statistic is 0.0568 (using $F_{1,6}$) which compared to 0.054 (the significance level used in Ajwa and Sullivan) is not significant. However, when 1 over 7 degrees of freedom are used to evaluate the significance level for the same F value, then the significance level becomes 0.0508 and is significant when compared to 0.054, the policy significance reference level (Figure 1). Interestingly, the Excel analysis also provides a t statistic for the estimated coefficient and obtains the same significance level as the $F_{1,7}$ test (Table 6, last row). Steel and Torrie (1960) suggest that the denominator term would use n-1 degrees of freedom for an F test of simple linear regression with no intercept. A degree of freedom is retained because 1 less parameter is being estimated.

I modified analysis in Ajwa and Sullivan (2012) to more accurately reflect the policy of Figure 1 and to correct the mistakes. I changed all of the period estimates listed in categories 2, 4 and 5 (Table 5) to the values estimated by the visual basic program. I did not change category 3 because the visual basic program identified these periods as having 2 or less measured values less than 0.1 ug/m3 and, as such, were either interpolated or substituted according to the policy and professional judgment. Thus the nine mistakes represented by category 2, the five periods where flux was interpolated even though sufficient measured values were available (category 4) and the two cases where Excel evaluated the p value of the regression with no intercept using n-2, instead of n-1, were modified by substituting the visual basic program estimate.

The impact of these substitutions on the cumulative flux was generally small (Table 7). The largest impact was in field 1 for 1,3-D where the relative reduction in cumulative flux was about 10%, from 0.1085 to 0.0980. The other changes were all less than 10% and mostly 3% or less. In all cases, the substituted values resulted in reduction in the cumulative flux.

The purpose in performing this substitution was to obtain a set of numbers which more closely reflected the analysis policy indicated in Figure 1 and to eliminate erroneous calculations. It is not the intent of this memorandum to endorse the calculation policy embodied by Figure 1. In fact, DPR disagrees with utilizing regression through the origin, particularly after determination of a significant intercept. DPR has generally followed a different policy in period flux analyses

(Johnson et al. 2010), which involves using sorting when initial OLS regression is insignificant. Development of the visual basic program in this memorandum which reflected the analysis policy outlined in Figure 1 was a goal of this memorandum. It is a stepping stone towards further study, analysis and comparison of the methods used to calculate period fluxes from fumigant studies.

References

Ajwa, Husein and Davis Sullivan. 2012. Soil fumigant emissions reduction using EVAL barrier resin film (VaporSafe) and Evaluation of Tarping Duration Needed to Minimize Fumigant total Mass Loss. Sponsors USDA-ARS Pacific Area-Wide Project, California Department of Pesticide Regulation. Performing Laboratory Department of Plant Sciences, University of California, Davis 1636 East Alisal Street, Salinas, California 93905, Field Testing Laboratory Sullivan Environmental Consulting, Inc. 1900 Elkin Street, Suite 200, Alexandria, Virginia 22308 Laboratory Study ID HA2011A Study Location Lost Hills, California. DPR 50046-0198 [251539].

Eisenhauer, Joseph G. 2003. Regression through the origin. Teaching Statistics 25(3):76-80.

Johnson, Bruce, Terrell Barry and Pamela Wofford. 2010. Workbook for Gaussian modeling analysis of air concentration measurements. September 1999, Revised May 2010. State of California Environmental Protection Agency, Department of Pesticide Regulation, Environmental Monitoring Branch.

Steel, Robert G.D. and James H. Torrie. 1960. Principles and Procedures of Statistics. McGraw-Hill Book Company, Inc. New York.

Fumigant (1=13d,2=pic)	2Field	3Period	4Monitor	5Model	6Measure	7	8Fumigant	9Field	10Period	11Start Line	12End Line	13Model (X)	14MSR (Y)	15Sample Size
1	1	1	1	3.92	8.17	0.06	1	1	1	2	17	$E$2:$E$17	$F$2:$F$17	16
1	1	1	2	3.23	20.42	0.06	1	1	2	18	33	$E$18:$E$33	$F$18:$F$33	16
1	1	1	3	0.00	0.06	0.06	1	1	3	34	49	$E$34:$E$49	$F$34:$F$49	16
1	1	1	4	3.63	0.08	0.06	1	1	4	50	65	$E$50:$E$65	$F$50:$F$65	16
1	1	1	5	7.13	0.06	0.08	1	1	5	66	81	$E$66:$E$81	$F$66:$F$81	16
1	1	1	6	10.99	0.06	1.97	1	1	6	82	97	$E$82:$E$97	$F$82:$F$97	16
1	1	1	7	13.43	0.06	2.35	1	1	7	98	113	$E$98:$E$113	$F$98:$F$113	16
1	1	1	8	16.21	2.52	2.52	1	1	8	114	129	$E$114:$E$129	$F$114:$F$129	16
1	1	1	9	13.44	2.35	2.64	1	1	9	130	144	$E$130:$E$144	$F$130:$F$144	15
1	1	1	10	10.35	5.02	2.86	1	1	10	146	161	$E$146:$E$161	$F$146:$F$161	16
1	1	1	11	5.06	2.86	5.02	1	1	11	162	177	$E$162:$E$177	$F$162:$F$177	16
1	1	1	12	7.36	1.97	8.17	1	1	12	178	193	$E$178:$E$193	$F$178:$F$193	16
1	1	1	13	4.85	2.64	9.88	1	1	13	194	209	$E$194:$E$209	$F$194:$F$209	16
1	1	1	14	7.10	12.08	12.08	1	1	14	210	225	$E$210:$E$225	$F$210:$F$225	16
1	1	1	15	4.41	9.88	20.42	1	1	15	226	241	$E$226:$E$241	$F$226:$F$241	16
1	1	1	16	5.08	69.82	69.82	1	1	16	242	257	$E$242:$E$257	$F$242:$F$257	16

Table 1. First block of linestroutine3.xlsm showing data for period 1 for 1,3-D in field 1. Contains data and addresses for data. Column 7 is a workspace. The start and end line row numbers are used to build an address in columns M and N for the ISCST3-modeled (column 13 is address, column 5 is the data) and measured (column 14 is the address and column 6 is the data) values using ="E"&K2&":E"&L2, for example, to give the first address in column 13 for the location of data in column 5 (column E). The addressing in columns 13 and 14 makes it convenient to use worksheet formulas for various statistical calculations in columns to the right.

16slope	intcept	stderr	rsq	F	21 fsig	r	sig of r	vary	totssq	avgx	regrssq	residssq	ssqx	seint	tval-intcept	32pvalue int	Using 0.054 for significance level fsig?1,0	Using 0.054 for significance level intsig?1,0	CV measured
-0.85	14.79	17.43	0.05	0.70	0.42	-0.22	0	297.81	4467.17	7.26	213.89	4253.28	296.68	8.54	1.73	0.11	0	0	2.00
0.11	0.08	0.52	0.29	5.69	0.03	0.54	1	0.36	5.39	2.30	1.56	3.84	118.69	0.17	0.46	0.65	1	0	1.75
0.54	2.04	3.05	0.32	6.50	0.02	0.56	1	12.69	190.37	7.93	60.33	130.04	205.65	1.85	1.10	0.29	1	0	0.56

Table 2. Extract of first four rows of linestroutine3.xlsm showing the second block consisting of ordinary least squares analysis of the measured and ISCST3 modeled data. Key worksheet functions utilized the indirect worksheet function (Table 1) to reference blocks of data for analysis.

Table 3. OLS regression calculations. This table is shown vertically, though in the actual worksheet it extends across the second row. Cell N2 (in the first block, see table 1) has the address for the first first period vector of measured values. O2 contains the sample size. The INDIRECT worksheet function is used to let these worksheet functions operate on an address, which changes row by row.

Column #	Column Letter	Column	Formula	Name
16	P	16slope	=SLOPE(INDIRECT($N2),INDIRECT($M2))	slope
17	Q	intcept	=INTERCEPT(INDIRECT($N2),INDIRECT($M2))	intercept
18	R	stderr	=STEYX(INDIRECT($N2),INDIRECT($M2))	std error of regresion
19	S	rsq	=RSQ(INDIRECT($N2),INDIRECT($M2))	r2 value
20	T	F	=S2*(O2-2)/(1-S2)	F value
21	U	21 fsig	=F.DIST.RT(T2,1,O2-2)	significance of F value
22	V	r	=CORREL(INDIRECT(M2),INDIRECT(N2))	correlation coefficient
23	W	sig of r	=IF(O2=16,IF(ABS(V2)>0.497,1,0),IF(O2=15,IF(ABS(V2)>0.514,1,0),IF(O2=14,IF(ABS(V2)>0.532,1,0),IF(O2=8,IF(ABS(V2)>0.707,1,0)))))	significant of corr coeff
24	X	vary	=VAR(INDIRECT(N2))	variance of measured values
25	Y	totssq	=X2*(O2-1)	total sum of squares
26	Z	avgx	=AVERAGE(INDIRECT(M2))	average of modeled values
27	AA	regrssq	=S2*Y2	regression sum of squares
28	AB	residssq	=Y2-AA2	residual sum of squares
29	AC	ssqx	=(O2-1)*VAR(INDIRECT(M2))	sum of squares of x
30	AD	seint	=SQRT(AB2/(O2-2)*((1/O2)+(Z2^2/AC2)))	standard error of the intercept
31	AE	tval-intcept	=Q2/AD2	t value for the intercept
32	AF	32pvalue int	=T.DIST.2T(ABS(AE2),(O2-2))	significance of the intercept t value
33	AG	Using 0.054 for significance level fsig?1,0	=IF(U2<=0.054,1,0)	flag if F significant (here set to 0.054, see text)
34	AH	Using 0.054 for significance level intsig?1,0	=IF(AF2<=0.054,1,0)	flag if intercept significant (here set to 0.054, see text)
35	AI	CV measured	=STDEV(INDIRECT(N2))/AVERAGE(INDIRECT(N2))	coefficient of variation of the ISCST3 modeled values

Table 4. Block 3 computations which are the results from Visual Basic Program analysing the period by period flux based on policies in Figure 1. Final choice represents the the result of the logic in Figure 1.

Column Number	Letter	Title	Description
36	AJ	forced slope (36)	slope -regression with no intercept
37	AK	37	standard error
38	AL	forced slope r1 (38)	r2 for regression
39	AM	forced slope F	F Value'
40	AN	forced slope regss (40)	regression sum of squares
41	AO	forced slope df (41)	degrees of freedom
42	AP	forced slope residss (42)	residual sum of squares
43	AQ	forced slope 'sig' (43)	significance' of forced regression
44	AR	flag if forced regression significant (44)	1 if significan 0 otherwise
45	AS	45 25th %tile	25th percentile of measured values
46	AT	flag if intcept > 25th percentile measured (46)	1 if intercept of OLS regression is > 25th percentile msrd values
47	AU	47avg mod	average of ISCST3-modeled values
48	AV	48avg msrd	average of measured values
49	AW	mean msr/mean mod	ratio
50	AX	50method	final method choice (RATIO AVG, OLS, FORCED)
51	AY	51 OLS slope	copy of OLS slope
52	AZ	52 msrd/mod	average of msrd/average of ISCST3 modeled
53	BA	53 forced slope	forced regression slope
54	BB	54 final choice	numerical value of final choice

Table 5. Assessment of agreement between Visual Basic program and Ajwa and Sullivan (2012) results. Numbers in brackets are line numbers in worksheet . See Appendix 2 for correlation to fumigant, field and period.

Category	Count	Percentage
1. Agreement between Visual Basic program and Ajwa and Sullivan (2012)	203	91
2. Mistake in Ajwa & Sullivan (2012) [8, 9, 29, 63, 64, 97, 136, 150, 171]	9	4
3. Less than 3 samples above 0.1 ug/m3 [67, 94, 179, 195, 215]	5	2
4. Interpolation or substitution used even though 3 or more samples above 0.1 ug/m3 [163, 189, 206, 207, 209]	5	2
5. Excel LINEST function uses 1 less degree of freedom for non-intercept case [62, 78]	2	1
Total	**224**	**100**

Table 6. Results for period 19, 1,3D, Field 2. The ISCST3-modeled and measured results under Model and Measure, respectively in ug/m3. Under "Summary Output" is the result from Excel regression routine when no intercept option is checked. I verified the F value of 5.54. However, the F test used in the ANOVA section uses 1 over 6 degrees of freedom, giving a p value of 0.0568. If the F test used 1 over 7 degrees of freedom, the resulting p value would be 0.0508. In fact, this is the p value which results in this Excel analysis for the t test evaluation in the last line for the X Variable.

Model	**Measure**	SUMMARY OUTPUT					
1.62	0.04						
0.00	0.04	*Regression Statistics*					
0.64	0.97	Multiple R	0.66				
0.99	1.94	R Square	0.44				
3.02	4.76	Adjusted R Square	0.30				
1.02	1.29	Standard Error	1.56				
3.04	1.21	Observations	8				
3.59	0.04						
		ANOVA					
			df	*SS*	*MS*	*F*	*Significance F*
		Regression	1	13.48	13.48	5.54	0.0568
		Residual	7	17.04	2.43		
		Total	8	30.52			
			Coefficients	*Standard Error*	*t Stat*	*P-value*	
		Intercept	0	#N/A	#N/A	#N/A	
		X Variable 1	0.61	0.26	2.35	0.0508	

Table 7. Comparison of cumulative flux as percentage of applied active ingredient before and after making corrections to the flux estimations.

	1,3-D		PIC	
Field	Original	After Corrections	Original	After Corrections
1	0.1016	0.0991	0.0450	0.0428
2	0.1085	0.0980	0.0361	0.0333
3	0.1908	0.1905	0.0999	0.0995
4	0.1663	0.1660	0.1179	0.1147

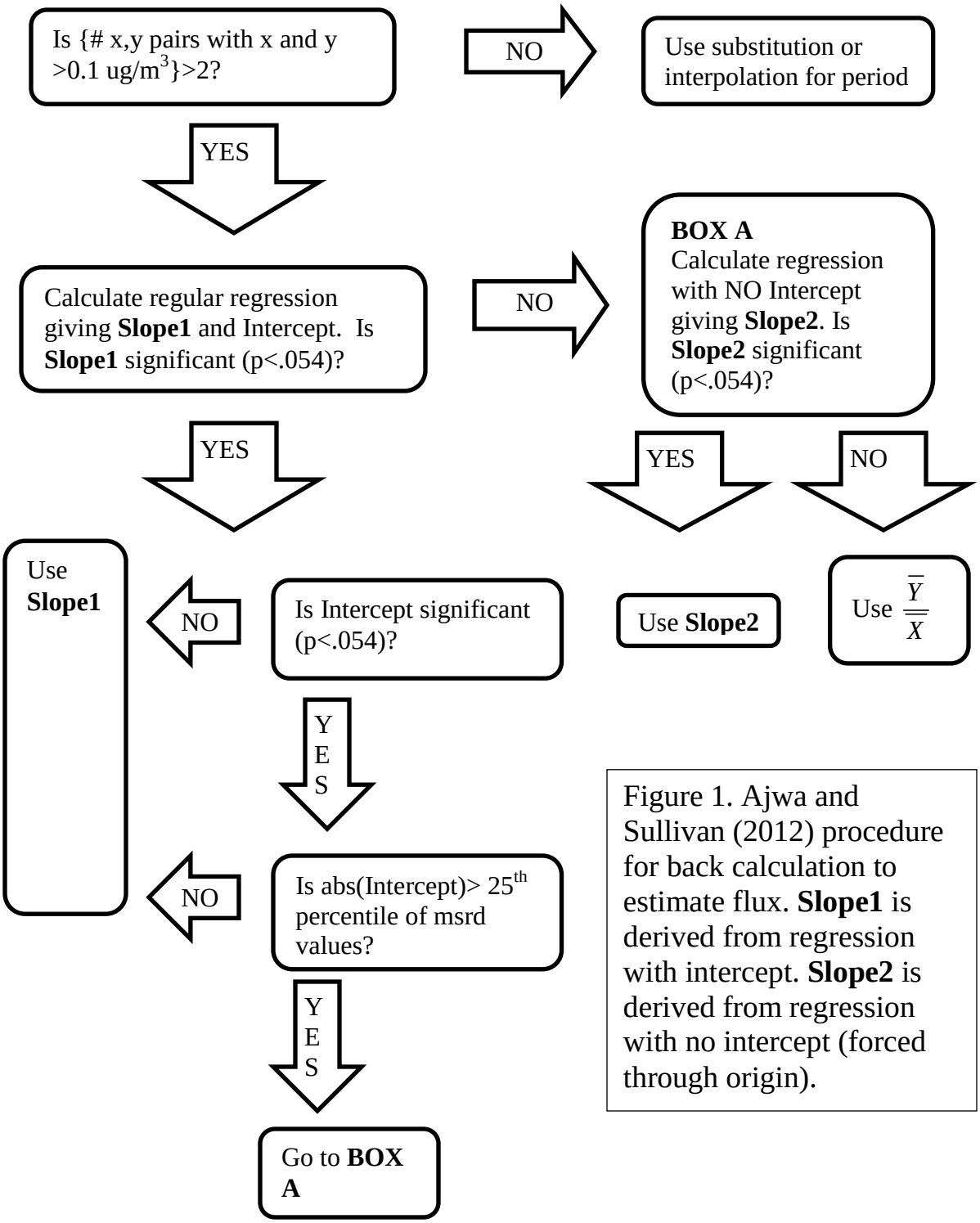

Figure 1. Ajwa and Sullivan (2012) procedure for back calculation to estimate flux. **Slope1** is derived from regression with intercept. **Slope2** is derived from regression with no intercept (forced through origin).

Appendix 1. Visual Basic Program Following Calculation Policy from Ajwa and Sullivan (2012).

Some of the routines included are for testing purposes and are not part of the main function. These testing routine names typically begin with “t” or “test”. The main routine is Sub Sullcompute().

```
Sub sullcompute()
Dim currow As Integer  'currow is the pointer towards which row of the worksheet is being worked on and from
                'because the forced regression is an array formula
                'some rows below the currow are utilized, but then overwritten subsequently
For currow = 2 To 225
Call LRnoconstant(currow)  'calculate and store results of forced regression
Call Procmsrd(currow)  'calculate the 25% percentile of measured values
Call findavg(currow) 'calculate averages and mean msr/mean mod
Cells(currow, 50) = sullogic(currow)  'figure out which estimate to use according to sullivans logic tree
Cells(currow, 68) = insuff(currow) 'figure out if 3 or more msrd values >0.1 ug/m3
Next currow
End Sub '----------------------------------------------------

Function insuff(currow)
'the column containing msrd values is column 6
'the row of the start of the msrd values is in cells(currow,11)
'the row of the end of the msrd values is in cells(currow,12)
'the number of pairs is in (currow,15)
'this function counts each value greater than 0.1ug/m3 and if the value is less than 3, returns "INSUFF"
Dim counter As Integer
Dim i As Integer
counter = 0

For i = 1 To Cells(currow, 15)
If (Cells(Cells(currow, 11) + i - 1, 6) > 0.1) Then
   counter = counter + 1
End If
Next i
If (counter < 3) Then
  insuff = "INSUFF"
Else
  insuff = "OK"
End If
End Function '-----------------------------------------------------

Sub LRnoconstant(currow)
'currow is the current row of work
'
```

```
' This is the beginning of a macro that calculates the regression
'with no constant.  LINEST is an array formula and puts the results into a 2 col x 5 row array
'where the upper left corner is the slope, the array is unlocked, and those results are copied into
' a 1 dim array going to the right from the slope in order to make the values available for testing
' a la sullivan
'

'
  Dim dd, ee As Integer
  Dim rtest As String
  Dim s1, s2, s3, s4 As String
  s1 = "=LINEST(INDIRECT(R"
  s2 = "C14),INDIRECT(R"
  s3 = "C13),FALSE,TRUE)"
  s4 = "AJ"


 dd = currow
'this section is just to get it going, eventually this will be a loop that moves down row by row and
'grabs the x,y data and calculates the forced regression slope and necessary statistics
  'dd = 2
  ee = 5 + dd
  rest = "AJ" + CStr(dd) + ":AK" + CStr(ee)  'the Cstr() function is handy to convert integers to
strings
'    Range("AJ2:AK6").Select
Range(rest).Select  'select the area where the array calculations will go

'    Selection.FormulaArray =
"=LINEST(INDIRECT(R2C14),INDIRECT(R2C13),FALSE,TRUE)"

Selection.FormulaArray = s1 + CStr(dd) + s2 + CStr(dd) + s3  'i think this is what actually does
the forced regression calculation
    actcell = "AJ" + CStr(dd)  'define an active cell in the array formula range
    Range(actcell).Select  ' pick the active cell, this activates the array
    'Range("AJ2").Select  'i think this selects the whole array, because it is a cell IN the array, so
this will have to move along with the array position
    'which will go down row by row as the calculations are performed
    'i need to copy the array values from the array and add them to the current row cells, extending
out the current row
    'in order to be able to process these values later on and not have to preserve the array formual
```

```
    'in order to copy them, i have to destroy the array property, the way you do this is by copying and repasting the
    'array as "values", this destroys the array, keeps the values, and lets me copy them one by one into the appropriate
    'spot in the current row, if you don't destroy the array property then excel gives you an error message that you can't
    'change individual array values when you try to copy just one value out of the array
    Selection.CurrentArray.Select
    Selection.Copy
    Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
        :=False, Transpose:=False  'ok this destroys thearray, but keeps the values
'now copy the values into the current row (currow=dd, remember?)
'these will have to move down as well, row by row, to the next row on each calculation
Cells(dd, 37) = Cells(dd + 1, 36) 'standard error of slope
Cells(dd, 38) = Cells(dd + 2, 36) 'r2 value
Cells(dd, 39) = Cells(dd + 3, 36) 'F value (regss*(n-2)/residss  this is used for sullivan logic
Cells(dd, 40) = Cells(dd + 4, 36) 'regss
Cells(dd, 41) = Cells(dd + 3, 37) 'n-1 this is used to determine signficance of regression
Cells(dd, 42) = Cells(dd + 4, 37) 'residss
'
' fdist Macro
'
    actcell = "AQ" + CStr(dd)
    Range(actcell).Select
    ActiveCell.FormulaR1C1 = "=F.DIST.RT(RC[-4],1,RC[-2])" 'uses relative addressing and calculates 'p' value for F statistic 1,n-1 df
'set flag to 1 if the forced regression is 'significant'
    If (Cells(dd, 43) < 0.054) Then  'sullivan uses 0.54 as his significance level, don't ask me why
      Cells(dd, 44) = 1
    Else
      Cells(dd, 44) = 0
    End If

End Sub '----------------------------------------------------

Sub testpro()
For currow = 2 To 4

Call Procmsrd(currow)
Next currow
End Sub '----------------------------------------------------
```

```
Sub Procmsrd(currow)
'currow is the current row that we are working on
'this subroutine copies the measured values into adjacent column,
' period by period and sorts each block of measured values
' then determines the 25th percentile and puts it into column R
'

'
    Dim i, j, k As Integer
     Dim src, targ As String
     Dim s1, s2, s3 As String
     s1 = "$g$"  'using column g as workspace for sorted values
     s2 = ":$g$"
 'For k = 1 To 224
    k = currow - 1  'this was a convenience so i didn't have to recode all these indices after
subroutine was developed
                'so now k+1=currow, remember that
    src = Cells(1 + k, 14) 'this holds address for msrd values block, we get the address directly
from the worksheet for source
    'but we have to construct address for target range in column G where themeasured values are
going and will be sorted
    'now construct address for where the sorted values will go
    'targ = Cells(2 + k - 1, 16)
    s3 = s1 + CStr(Cells(k + 1, 11)) + s2 + CStr(Cells(k + 1, 12))
    targ = s3
    Range(src).Select
    Selection.Copy   'copy the measured values from F column
    Range(targ).Select
    ActiveSheet.Paste  'paste the measured values into corresponding elements in G column
    Application.CutCopyMode = False
    'now sort the just-copied measured values
    ActiveWorkbook.Worksheets("Sheet2").Sort.SortFields.Clear
    ActiveWorkbook.Worksheets("Sheet2").Sort.SortFields.Add Key:=Range(targ), _
       SortOn:=xlSortOnValues, Order:=xlAscending, DataOption:=xlSortNormal
    With ActiveWorkbook.Worksheets("Sheet2").Sort
       .SetRange Range(targ)
       .Header = xlNo
       .MatchCase = False
       .Orientation = xlTopToBottom
       .SortMethod = xlPinYin
       .Apply
```

```
    End With
    Cells(1 + k, 45) = per25sull(k, Cells(1 + k, 15)) 'this function figures out 25th percentile of
sorted,msrd values
    If (Cells(k + 1, 45) < Cells(k + 1, 17)) Then 'determine if intercept greaterthan 25% tile of
msrd
      Cells(k + 1, 46) = 1
    Else
      Cells(k + 1, 46) = 0
    End If

End Sub '----------------------------------------------------

Sub findavg(currow)
'
' currow is the current row working on
' this routine finds average of modeled and average of msrd values
' puts each into cell and computes avg msr/avg mod
'

Dim s1, s2, s3
    s1 = "AU" + CStr(currow)  'set up the locations where the results will be stored
    s2 = "AV" + CStr(currow)
    s3 = "AW" + CStr(currow)
    Range(s1).Select
    ActiveCell.FormulaR1C1 = "=AVERAGE(INDIRECT(RC[-34]))"  'this is one way to do a
worksheet formula in visual basic
    Range(s2).Select
    ActiveCell.FormulaR1C1 = "=AVERAGE(INDIRECT(RC[-34]))"
    Range(s3).Select
    ActiveCell.FormulaR1C1 = "=RC[-1]/RC[-2]"
End Sub '----------------------------------------------------

Sub tf()
For currow = 2 To 4
Cells(currow, 50) = sullogic(currow)
Next currow
End Sub '----------------------------------------------------

Sub tper25()
'to test the per25sull routine
m = 1
```

```
n = 16
colnum = 4
Cells(13, colnum + 1) = per25sull(m, n, colnum)
n = 15
colnum = 8
Cells(12, colnum + 1) = per25sull(m, n, colnum)
n = 8
colnum = 12
Cells(7, colnum + 1) = per25sull(m, n, colnum)
n = 7
colnum = 16
Cells(6, colnum + 1) = per25sull(m, n, colnum)
n = 6
colnum = 20
Cells(6, colnum + 1) = per25sull(m, n, colnum)
End Sub '--------------------------------------------------

Sub testper25()
Dim p25 As Single
Dim m As Integer
Dim n As Integer
m = 51
n = Cells(m + 1, 15)
p25 = per25sull(m, n)
p = 1
End Sub '--------------------------------------------------

Function per25sull(m, n)
'compared first two 16 values sets 0.07234, 0.064726 OK
'compared first two 15 values sets 0.057888, 2.167125 OK
'compared first two 8 values sets 0.061445, 0.050303 OK
'compared first two 7 values sets 0.076017, 5.752224 OK
'Function per25sull(m, n, colnum)
'colnum is temporary addition to argumetn list for debuggin purposes to
'specify an arbitrary column from test calling routine in order to
'make sure this thing is getting the right estimate, colnum is location of msrd values column
'this routine is based on detailed comparison of my initial 'sullvian' like analysis
' to what sullivan actually got, and i found that the way he assigns probabilities
' is different than how i originally wrote (per25 above), so this per25sull routine
'mimics what sullivan does in order to fidn the 25th percentile and determine whether
'the intercept is greater thanteh 25th percentile
```

```
Dim lower As Single
Dim upper As Single
Dim delta As Single
Dim plow As Single 'this is the lower percentile represented by the lower boundary fo the 25th bracket
Dim lowr As Integer 'this is the index of the lower bound rank (ie 4 when n=16, 4 when n=15, etc)
Dim colnum As Integer
colnum = 7 'this is the column containing the sorted msrd data
'Dim m, n, start As Integer
'this function finds the 25th percentile and interpolates
'm is index of fume,field,period,start,end line (goes from 1 to 224)
'   and corresponds to row m+1 in the worksheet (ie m=currow-1)
'n is the number of pairs in this period and is 7,8,15,16 are the allowable values as of 10/10/2012
'the workbook called U:\Ajwa-Sullivan data disk for Lost Hills emissions study 2011\myreviewstruff\regr-error-study\getalldata\automate-sullivan\[sullivans-percentile-calculations.xlsx]Sheet1
'shows for 16,15,8,7 pairs where the 25th percentile is located between the ranked (sorted) values
'note the values are sorted from smallest to largest so that the 25th percentile is near the top
'16 between 4 and 5 (rank 16 is largest value)
'15 between 4 and 5
'8 between 2 and 3
'7 getween 2 and 3
'delta is the space between ranks in the percentile world 100/(n-1) where n=number of pairs

currow = m + 1
startmsrd = Cells(currow, 11) 'this integer tells what row the sorted measured values start on
delta = 100 / (n - 1)
If (n = 16) Then
lowr = 4
ElseIf (n = 15) Then
lowr = 4
ElseIf (n = 8) Then
lowr = 2
ElseIf (n = 7) Then
lowr = 2
Else
response = MsgBox("Number of pairs out of range: per25sull. Continue? ", 3)
 If response = vbNo Then
   Stop
 End If
```

```
End If

plow = (lowr - 1) * delta 'this should be lower percentile bound of bracket that includes 25th
percentile
per25sull = 0
'per25sull = ((25 - (n - lowr) * delta) / delta) * (Cells(currow + lowr - 1, colnum) - Cells(currow
+ lowr, colnum))
per25sull = ((25 - plow) / delta) * (Cells(startmsrd + lowr, colnum) - Cells(startmsrd + lowr - 1,
colnum))
per25sull = per25sull + Cells(startmsrd + lowr - 1, colnum)
p = 1
End Function ‘--------------------------------------------------

Sub tf2()
For currow = 2 To 4
Cells(currow, 50) = sullogic(currow)
Next currow
End Sub ‘--------------------------------------------------

Function sullogic(currow)
'121101 modified this to use the absolute value of the intercept, i have not yet received confirm
'       from Sullivan on this, but it sure looks like that is what he is doing
‘121105  have received confirmation – they use absolute value
'this takes the relevant information and uses sullivans logic tree
'to come up with 1 of 3 words "OLS", "FORCED", "RATIO AVG"
'DEPENDING ON the various relevant values
'first thing, though, is copy the three estimates into the cells just after where the written
'method is posted, so here is the operation
'col 16 OLS slope into col 51
'col 49 msrd/mod into col 52
'col 36 forced slope into col 53
Cells(currow, 51) = Cells(currow, 16) 'ols slope
Cells(currow, 52) = Cells(currow, 49) 'msrd/mod
Cells(currow, 53) = Cells(currow, 36) 'forced slope
If (Cells(currow, 21) <= 0.054) Then
'original regression, OLS=ordinary least squares, is significant
  If (Cells(currow, 32) > 0.054) Then 'check signifiance level of intercept
    'intercept is NOT significan, go ahead and use OLS slope
    sullogic = "OLS" 'Cells(currow, 16) 'ols REGRESSION MULTIPLIER
    Cells(currow, 54) = Cells(currow, 51) 'OLS is final choice
  Else
```

```
    'intercept IS significant, do further checking by comparing to the 25th%tile of msrd values
'    If (Cells(currow, 17) > Cells(currow, 45)) Then 'is intcp > 25%tile
    If (Abs(Cells(currow, 17)) > Cells(currow, 45)) Then 'is intcp > 25%tile  'use absolute value
brj 121101
      'intercept IS greater than 25th percentile
      If (Cells(currow, 43) <= 0.054) Then 'is forced regression significant
        'forced regression is 'significant'
        sullogic = "FORCED"
        Cells(currow, 54) = Cells(currow, 53)
      Else
        sullogic = "RATIO AVG" 'forced not sig, and OLS intcp > 25th %
        Cells(currow, 54) = Cells(currow, 52)
      End If
    Else
      sullogic = "OLS"
      Cells(currow, 54) = Cells(currow, 51)
    End If
  End If
Else
'when original regression is NOT significant
'is slope of forced regression 'significant'?
  If (Cells(currow, 43) <= 0.054) Then
    sullogic = "FORCED"
    Cells(currow, 54) = Cells(currow, 53)
  Else
    sullogic = "RATIO AVG"
    Cells(currow, 54) = Cells(currow, 52)
  End If
End If
End Function
```

Appendix 2. Summary of differences between final Visual Basic results compared to Ajwa and Sullivan (2012). Category refers to category in Table 5

Category	Row#	Fumigant	Field	Period
2	8	1,3-D	1	7
2	9	1,3-D	1	8
2	29	1,3-D	1	28
2	63	1,3-D	2	20
2	64	1,3-D	2	21
2	97	1,3-D	4	4
2	136	PIC	1	23
2	150	PIC	1	37
2	171	PIC	2	16
3	67	1,3-D	2	24
3	94	1,3-D	4	1
3	179	PIC	2	24
3	195	PIC	3	4
3	215	PIC	4	10
4	163	PIC	2	8
4	189	PIC	3	4
4	206	PIC	4	1
4	207	PIC	4	2
4	209	PIC	4	4
5	62	1,3-D	2	19
5	78	1,3-D	3	5