

AD-A270 507



2

DLA-93-P30054

DEFENSE CONTRACT MANAGEMENT COMMAND DATA VALIDATION FILTER

DISTRIBUTION STATEMENT A
Approved for public release
Distribution Unlimited

June 1993



DTIC
SELECTED
OCT 12 1993
S B D

FOR

DEPARTMENT OF DEFENSE
DEFENSE LOGISTICS AGENCY
DEFENSE CONTRACT MANAGEMENT COMMAND
Cameron Station
Alexandria, VA 22304-6100

93-23914

Page 5

INSIGHT THROUGH ANALYSIS

DOR0

CORPORATE RESEARCH

93-23914-8-035

DLA-93-P30054

**DEFENSE CONTRACT MANAGEMENT
COMMAND DATA VALIDATION FILTER**

June 1993

James Russell

CORPORATE RESEARCH TEAM

John S. McKinney

Edward J. Modic

OPERATIONS RESEARCH OFFICE

**DEPARTMENT OF DEFENSE
DEFENSE LOGISTICS AGENCY**

Executive Director (Plans & Policy Integration)

**Cameron Station
Alexandria, VA 22304-6100**



**DEFENSE LOGISTICS AGENCY
HEADQUARTERS
CAMERON STATION
ALEXANDRIA, VIRGINIA 22304-6100**



CAI

FOREWORD

This Defense Logistics Agency Operations Research Office (DORO) report provides information regarding the development and application of a computer based data screening tool. Defense Contract Management Command (DCMC) will use this model initially to validate the monthly unit cost counts and later will begin testing a wide variety of management information data.

Christine L. Gallo

**CHRISTINE L. GALLO
Executive Director
Plans and Policy Integration**

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By _____	
Distribution/ _____	
Availability Codes	
Dist	Avail and/or Special
A-1	

TABLE OF CONTENTS

<u>Section</u>	<u>Title</u>	<u>Page</u>
	FOREWORD	iii
1	INTRODUCTION	1-1
2	METHODOLOGY	2-1
2.1	Statistical Process Control	2-1
2.2	Single Exponential Smoothing	2-1
2.3	Data Validation Model	2-2
2.4	Usage of Model	2-3
3	CONCLUSIONS	3-1
4	RECOMMENDATIONS	4-1
	APPENDIX: USER'S GUIDE	A-1

SECTION 1 INTRODUCTION

The Defense Contract Management Command (DCMC) is becoming increasingly dependent on accurate workload reporting due to the adoption of unit cost based resourcing. DCMC management has identified approximately 150 data elements that will be used to monitor DCMC activity. Eighteen of these data elements feed the unit cost system. Others will be used to track performance. Inaccurate values for these data elements will result in improper unit costs, resource levels, and performance measurements.

DCMC's data problems in the past have included missing data (usually not input on time), partial data reporting, and erroneous data input. Frequently erroneous input will involve one or more extra zeroes in a number. With the advent of the unit cost system, some new data elements are being reported. This introduces other possible errors, for example, using different units of count in different Secondary Level Field Activities (SLFAs), or even within the same SLFA.

The recent sweeping DCMC organizational changes have impacted data accuracy. The realignment of DCMC into five districts and the consolidation of Military Service activities into DCMC are some examples of these changes. Mechanization of Contract Administration Services (MOCAS) data bases have been fragmented for each district during these transitions. This fragmentation has made complete and accurate data collection difficult. Additionally, some interface problems still exist between former Military Service activities and MOCAS.

While some of these problems have been resolved, attention needs to be focused on identifying and correcting data errors. One way to help increase data accuracy is to develop tools that can be used by DCMC personnel during data input and reporting.

SECTION 2

METHODOLOGY

2.1 STATISTICAL PROCESS CONTROL

Statistical Process Control (SPC) is one technique that is a logical choice for the problem of data validation. SPC is widely used in industry to monitor and improve manufacturing processes. Control charts are used to plot a series of measurements for an important characteristic of the output (e.g., hardness of a composite metal or width of a piece of cloth). This chart has a center line (which is usually the average value) along with control limits at 1, 2, and 3 standard deviations above and below the center line. The probability of getting output with only random variation whose measurement falls outside of three standard deviations is only about 1 percent, based on a normal distribution of output values. Certain patterns that appear in this plotted data can also provide useful information. Principles of SPC are used in this model to help determine a reasonable range of values to test data input. The same attributes of SPC that make it such a powerful manufacturing tool can be adapted to also make it a useful tool for data validation.

While SPC is an effective tool, there are only certain types of errors it will catch when validating data input. Errors in magnitude will be effectively highlighted. For example, if the incorrect value 60 is input when 6,000 is the correct value, the incorrect value would be flagged. However, if the incorrect value 60 is input and the correct value is 53, this error would not necessarily be detected. Also, in cases where trends exist, traditional SPC by itself will not effectively accommodate these trends. Any useful data validation model must be able to react and adjust to trends as the Department of Defense continues its downsizing and DCMC workload declines.

2.2 SINGLE EXPONENTIAL SMOOTHING

We increased the effectiveness of our model in handling trends by combining SPC with a forecasting technique called single exponential smoothing (SES). SES is a widely used forecasting technique that predicts a future value by focusing on the most recent actual values. Older values receive (exponentially) smaller weights. For many data sets, we would expect that the latest values in the series would be better predictors for the next period than older values. SES will enhance our model by blending the effects of trend into our range of acceptable values.

The SES technique requires that a smoothing factor between zero and one be selected, depending on the amount of smoothing that is desired in the forecasts. This smoothing factor is called alpha. Smoothing refers to the amount of variation in the forecasts from period to period. Alpha is simply a weight that we assign to place more or less emphasis on the latest data value. Using a small alpha (0.1) will cause forecasts to be very smooth and not very sensitive to changes in the data. A large alpha (0.9) will give forecasts that are not very smooth. The optimum alpha level will lead to forecasts with the least amount of error; this optimum level can vary over time as data characteristics change. The model selects the optimal alpha by minimizing the Mean Absolute Percent Error (MAPE) each time a forecast is made.

2.3

DATA VALIDATION MODEL

By combining SPC and SES techniques, we can create a range of reasonable values for our target data. The SES forecast for a period (in our case, one month) is compared with the actual value for that period. The difference between the forecast and the actual value is the error in the forecast. The average of past forecast errors serves as a mid-line for our pseudo control chart. The range around the mid-line is calculated by multiplying the standard deviation of the forecast errors by 1.96 and -1.96. This range captures approximately 95 percent of expected values, based on a normal distribution. When a new actual value is added to the model, the amount of error in the forecast is compared to the computed range of reasonable errors. If the amount of error falls outside of this range, the model will flag the value. This warns the user that this value is statistically unusual and requires review.

If the user finds that the value is incorrect and changes it in the model, the model will recalculate the amount of error and create a new range for reasonable error values, the next time a forecast is made. If the actual data is zero, the model estimates these as "missing" values, if values for adjacent months are greater than 0. Data that is considered "missing" remains flagged, with the value of 0 (in the data base), but for calculation purposes only a value is substituted. If only one data point is "missing," the estimate used for forecasting will be the average of the values before and after the "missing" value. If two consecutive data points are "missing", interpolation between the two next closest values is used. For example, if the values for January and February 1993 are "missing," and the value for December 1992 is 50 and the value for March 1993 is 62, the model would use 54 for January and 58 for February.

USAGE OF MODEL

The model could be used at either the point of data input and/or after District level data is compiled. The primary consideration for where to use the model was accountability for whether or not changes to the data would be made when required. If the model was only used at the point of input, District level personnel might not be aware if values were verified or corrected. If, however, the model was run at the District level, District personnel would be responsible for seeing that the required changes are made.

The accountability issue doesn't preclude using the model both at the point of input and at the District level but other problems surfaced concerning using the model at the input level. First, there is an effort underway to automate the Unit Cost work counts where no manual input will be required. This will not happen immediately for all work counts, but some may be automated in the very near future. This would mean these automated counts could not be run through the model if it was used at the point of input. The model could validate all work counts, automated or not, if implemented after the District level data was compiled.

There are also differences among the Districts as to who inputs certain work counts. There was also some confusion as to whether all data for a work count for an SLFA was input at the same location. This would require even more customization than would be required even if data was validated at the point of input. However, only five programs and data bases are required if data is validated at the District level.

SECTION 3

CONCLUSIONS

Accuracy in the reporting of unit cost and other management information data is essential for DCMC. Inaccurate reporting will lead to inappropriate workload and resource assessments and reduced efficiency.

Use of this model will increase reporting accuracy and data base integrity. The model should be applied to the unit cost data at the District level after it has been input at the SLFA level. Validating the data at the point of input at the SLFA proved impractical. Many of the monthly counts will soon be extracted automatically from MOCAS and other data systems, which precludes validating the data at the SLFA. Additionally, there will be more control over this validation effort by having a small group responsible for investigating flagged values and making corrections.

Personnel with all levels of data validation expertise will benefit from using this statistical model to highlight certain types of possible errors for further review.

SECTION 4

RECOMMENDATIONS

The Data Validation Filter should be used for validating unit cost work counts by both DCMC headquarters and district level personnel.

After testing and review, the model should be applied to key data elements used to monitor DCMC activity.

Individuals should be designated (by position and name) at each District as well as Headquarters Defense Logistics Agency to aggressively investigate and change (if necessary) all flagged data values.

APPENDIX A
DCMC DATA VALIDATION FILTER (DDVF) USER'S GUIDE

SECTION 1 OVERVIEW

The DCMC Data Validation Filter (DDVF) highlights possible errors in the Unit Cost work counts (there are currently 17 data elements being collected). The DDVF allows for collecting the data being checked (current month) with minimal effort on the part of the user. This required data will be extracted monthly from the Management Analysis Statistical System (MASS) and sent to the Districts along with the data used for the Unit Cost reports. The user will only have to download the data from the Distributed Mini System (DMINS) in the usual manner. **NOTE: the file must be downloaded as a word processing (not binary) file.** Each District will then run a stand alone, menu-driven program on their Personal Computer (PC) which will validate the data and highlight data that should be corrected. Any changes to be made will be made in another part of the model. A third menu selection will allow the user to display the 12 month history (if available) for any work count (one at a time) for any Secondary Level Field Activity (SLFA) (also one at a time).

Because it is expected that the Districts will have to ask their SLFAs for any data corrections, the correction module was designed separately from the validation portion of the model. This allows the user to complete validation, exit the program, and then re-enter the model to make corrections.

SECTION 2 INSTALLING THE MODEL

The model runs on a PC with a minimum of 640 kilobytes of memory, a hard disk drive (not more than 2 megabytes of free memory is necessary), either a 3.5 or 5.25 floppy drive, and connectivity to DMINS to allow file downloading.

The word <enter> will be used in this guide any time the user is required to press the enter key. Double quotes " " will be used to highlight required keystrokes. The actual double quote marks should not be typed.

To install the DDVF model:

1. Turn the computer on.
2. Put the DDVF PROGRAM DISK in the A: drive (Floppy). (Use the 3.5 or 5.25 floppy disks depending on whether the A: drive is a 3.5 or 5.25 drive.)
3. At the C> prompt, create a directory for the model, by typing "md ddvf".
4. Still at the C> prompt, change to the DDVF directory, by typing "cd ddvf".
5. You must be at the C:\DDVF> prompt. Copy the DDVF model and all necessary data bases to the DDVF directory, by typing "copy a:*. *".

6. To run the model, type the file name listed in Table 1 for each District (don't type ".EXE"). For example, for the Southern District, type "DCMDSDFV".
7. To validate new data, remember it has to be downloaded first. Download the designated file will be used for each District. Make sure you download it to the c:\DDVF Directory, and that the downloaded file is named exactly as shown the last column of Table 1. Once the properly named file is in the DDVF Directory, start the model the same as shown in step 6.

Table 1. District Data Filenames

<u>District</u>	<u>.EXE Filename</u>	<u>Downloaded Filename</u>
DCMDS	DCMDSDFV.EXE	ATLPSNS.ASC
DCMDN	DCMDNDFV.EXE	BOSPSNS.ASC
DCMDC	DCMDCDFV.EXE	CHIPSNS.ASC
DCMDW	DCMDWDFV.EXE	LAPNSNS.ASC
DCMDM	DCMDMDFV.EXE	PHIPSNS.ASC

SECTION 3 USING THE MODEL

Performing steps 6 and 7 from Section 2 above will get the user to the main menu. The main menu will have three functions to choose from (four if you count Exit).

Selecting 1 allows the user to validate current work counts. Remember, you have to download the file to the DDVF Directory and name it properly (See steps 6 and 7 in Section 2) before using this menu item. The output of this function is a display of all the work counts for the District. Those values outside the limits calculated by the model are highlighted (the screen color is different) as possible errors.

Menu item 2 is the data correction function. Answering some screen prompts will allow the user to change data for any previous month (which also includes the current month, verified in menu function 1). When prior data is changed, the flag for that data is removed. The next time new data is validated, the changed value is used to calculate the forecast and flags are reset. Therefore, it is possible that a value may be changed (when changed the flag will be removed), but the data may be flagged again when the next month's data is validated. This just means that the changed value is still outside the model's limits. Remember, the model highlights *possible* errors.

Menu function 3 displays and prints historical monthly data by SLFA. The data displayed is the current (most recently validated month) and the previous 6 months, for all the work counts. As on the validation screen, possible errors are highlighted

To print any of the display screens, simply press the "Print Screen" key. Since the highlighted values cannot be printed in a different color, possible errors are highlighted during screen printing with an "*" after the highlighted value (this "*" does not show up during screen display).

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Project (0704-0188), Washington, DC 20503.</small>				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE June 1993	3. REPORT TYPE AND DATES COVERED Final		
4. TITLE AND SUBTITLE Defense Contract Management Command Data Validation Filter			5. FUNDING NUMBERS	
6. AUTHOR(S) John S. McKinney Edward J. Modic				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) HQ Defense Logistics Agency Operations Research and Economic Analysis Office (DORO) c/o Defense General Supply Center Richmond, VA 23297-5000			8. PERFORMING ORGANIZATION REPORT NUMBER DLA-93-P30054	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) HQ Defense Logistics Agency Corporate Research Team (CAILR) Cameron Station Alexandria, VA 22304-6100			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Unlimited distribution; public release			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) This report documents the development and application of a computer based data validation tool for use by the Defense Contract Management Command (DCMC). The model is based on Statistical Process Control (SPC) principles combined with Single Exponential Smoothing (SES) forecasting. Monthly data values that are statistically unusual compared to historical values are flagged for review and possible correction. This kind of model is necessary for the effective implementation of the unit cost based resourcing system. Inaccurate unit cost system data will result in inappropriate workload and resource assessments. This model should eventually be used to validate a much wider array of key management data at the DCMC, Secondary Level Field Activity (SLFA) and District levels.				
14. SUBJECT TERMS unit cost, data validation model, DCMC			15. NUMBER OF PAGES 23	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT	