13 Chapter

Case Studies -- III

Chair: Charles Day, National Agricultural Statistics Service

Anusha Fernando Dharmasena

Clancy Barrett ◆ Francois Laflamme

Gregory D. Weyland

Time-Series Editing of Quarterly Deposits Data

Anusha Fernando Dharmasena, Federal Reserve Board

13
Chapter

Abstract

The Statistical Services Branch in the Division of Information Resources Management at the Federal Reserve Board is responsible for ensuring the accuracy and reliability of deposits data reported by the Federal Reserve Banks. This research attempts to provide a statistical methodology for editing these data using forecasting techniques, to identify "acceptable" and "unacceptable" data. The study will show that changes in the quarterly deposits data are a result of changes in seasonality, the number of respondents, and "micro level" data fluctuations.

These consistent fluctuations in the aggregates have been modeled using regression techniques. The data for this study consists of twelve quarterly deposits items that were summarized by five entity types.

Time-Series Editing of Quarterly Deposits Data

Anusha Fernando Dharmasena, Federal Reserve Board

Introduction

The Deposits Unit of the Statistical Services Branch at the Federal Reserve Board is responsible for editing and refining deposits data reported by the Federal Reserve Banks. One such project is the editing of quarterly deposits data to ensure accuracy and reliability.

Examining the Quarterly Edited Deposits System (QEDS) has became more important as we realize that the present system of data analysis at the micro level is less useful than an analysis of QEDS aggregated data. This research attempts to provide a statistical methodology for editing these data using forecasting techniques, to identify "acceptable" and "unacceptable" data. We assume that the changes in the QEDS data are caused by "micro level" data fluctuations, seasonality, and other macro influences. These consistent fluctuations indicate that the aggregates could be modeled using regression techniques. The close fit of the final model shows that this assumption is true.

The QEDS data consist of the quarterly reported deposit items which are listed below:

- Vault cash
- Total demand deposits
- ATS & NOW accounts
- Total savings
- Small time deposits
- All time deposits
- U.S. Government demand deposits
- Demand deposits due to
- Cash items in process of collection
- Demand balances due from
- Other demand deposits
- Total net transactions
- Total nonpersonal savings and time deposits.

The data have been aggregated by entity type as described below:

- Commercial banks (member banks)
- Commercial banks (nonmember banks)
- Mutual savings banks
- Savings and loans
- Credit unions.

Exploration

The research began by looking at various forms of the dependent variable: the QEDS quarterly value for a given item, together with independent variables that would result in good estimates of QEDS data for a given quarter. To this end many linear regression models were investigated to obtain the best model that fit the data and that helped in the formulation of the final model. This final model serves as the basis for our aggregate data editing procedure.

The data for the initial explorations were taken from a SAS dataset that was created for performing exploratory data analysis with QEDS using SAS/Insight®.

The first attempt was to look at the following multiple linear regression expecting meaningful estimates and useable models:

$$\bar{Y}_t = \beta_0 + \beta_1 \bar{Y}_{t-1} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \beta_6 X_{6t} + \beta_7 X_{7t} + \epsilon_t \quad ,$$

where

- $\bar{Y}_t$ = The percentage change of the QEDS item from quarter to quarter after ensuring that the number of respondents stayed the same for each quarter. This was accomplished by dividing the QEDS value by the number of respondents for each quarter.
- $\bar{Y}_{t-1}$ = The lagged dependent variable (% change from previous quarter)
- X_{2t} = Monetary Aggregates (MA) growth rates
- X_{3t} = Number of banks per quarter
- X_{4t} = Seasonal Dummy 1 representing quarter 1
- X_{5t} = Seasonal Dummy 2 representing quarter 2
- X_{6t} = Seasonal Dummy 3 representing quarter 3
- X_{7t} = Seasonal Dummy 4 representing quarter 4
- ϵ_t = A random error term uncorrelated over time, typically called ***white noise***.

❐ **Results.**--The coefficients for the seasonal factors were significant, indicating strong influences of seasonality. Unfortunately, because the data had to be manipulated to ensure that the same respondents were reporting for two consecutive quarters, some respondents were eliminated from the calculation. This elimination resulted in a sample not completely reflective of the QEDS universe, which, in turn, led to poor models that could not be used to predict aggregate QEDS data.

The next step in the analysis was to examine the explanatory power of a different set of independent variables -- combinations of economic factors. We hoped that these indicators would be linearly related to the deposits data:

$$\bar{Y}_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \epsilon_t \ ,$$

where

- $\bar{Y}_t$ = The percentage change in QEDS -- the dependent variable and the following independent variables for end of quarter reporting dates
- X_{1t} = GNP

Y_{2t} = Interest rates
Y_{3t} = Consumer price index
Y_{4t} = Unemployment rates.

❒ **Results**.--The outcome of the statistical model was disappointing. These leading indicators did not produce strong estimates as hoped, because a correlation analysis of the variables showed the lack of a functional relationship between the dependent and independent variables. Since the postulated model did not describe the data satisfactorily and no fundamental conclusions were recovered from the fitted equation, the model could not be used.

The research continued to use the data described in the first attempt with the dependent variable (Y). One of the new predictor variables (X) reflected the percentage change in the number of respondents from quarter to quarter. In addition, to incorporate an important guide to the properties of time series analysis, we introduced a series of lags of the dependent variable:

$$\bar{Y}_t = B_0 + B_1X_{1t} + B_2X_{2t} + B_3X_{3t} + B_4X_{4t} + B_5Y_{t-1} + B_6Y_{t-2} + B_7Y_{t-3} + B_8Y_{t-4} + \varepsilon_t \; ,$$

where

Y_t = The percentage change in QEDS -- the dependent variable
X_{1t} = MA growth rates
X_{2t} = Seasonal factors
X_{3t} = Percentage change in the number of respondents
X_4 = Number of respondents
$\beta_5\bar{Y}_{t-1}$- $\beta_8\bar{Y}_{t-4}$ = Lag dependent variables.

❒ **Results**.--Although some of the lags were very significant, indicating effects from previous time periods, the overall estimates produced by the model did not produce a good fit because of serial correlation among the residuals.

After much research and model testing, we decided to get data directly from the QEDS archival and to construct quarterly data by item and entity to fit the needs of the project. Thus, the final model that helped in predicting aggregate QEDS data is as follows:

$$¥_t = B_0 + B_1X_{1t} + B_2X_{2t} + B_3X_{3t} + B_4X_{4t} + B_5X_{5t} + B_6X_{6t} + \varepsilon_t \; ,$$

where

$¥_t$ = Dollar difference in levels of QEDS data (Y_t- Y_{t-1}) - the dependent variable
B_0 = Intercept + seasonal dummy quarter 1 + seasonal dummy quarter 2
X_{1t} = MA QEDS estimate
X_{2t} = Number of respondents
X_{3t} = Seasonal dummy quarter 3
X_{4t} = Seasonal dummy quarter 4.

Here the dependent variable reflects the change in levels. The set of regressors for this model was also reduced to reflect seasonal dummies, the MA QEDS estimates, and the number of respondents for the quarter. The QEDS estimates were constructed from the growth rates and the panel shift data that were obtained from MA. These estimates were constructed in a manner similar to that used by MA to obtain the money supply estimates.

- **Results.**--The results from this model fully used the QEDS estimates from MA which included the panel shift data to help explain the variability in our dependent variable. The next section on methodology will cover the model and its results in detail.

The previous discussion was concerned primarily with finding the best model from a group of candidate models using the least squares method for estimation of model coefficients. Implicit in the least squares method are the assumptions that $E(\varepsilon_i) = 0$ and that the ε_i are uncorrelated with homogeneous variance σ^2. In addition, ***normality*** on the ε_i is required for the estimators to attain the property of minimum variance of the class of unbiased estimators. Thus, to address the issue of variances in the dependent variable from observation to observation, the proper estimator of β should take the normality of ε_i into account by weighing the observations in some way that allows for the differences in the results. Sample autocorrelations coefficients were created to measure the correlation among observations at different distances apart. These autocorrelations were used to account for adjustments that take place over time. Therefore, all models were checked for autocorrelation and heteroscedasticity.

Prediction Model

The primary function of the preceding model-building exercise was to determine which regressor variables truly explained the response variable y; the QEDS reported value for a given quarter. The final model had the following regressor variables that were responsible for a significant amount of variation in the dependent variable. These variables are the seasonal dummies, the constructed MA QEDS estimates, and the number of respondents for the quarter.

The main focus of the analysis is to use econometric modeling techniques to make a good prediction of quarterly data. Although an attempt is being made to use a set of mathematical formulas and assumptions to describe this deviation, the uncertainty inherent in statistical prediction methodology will introduce errors. In an attempt to be parsimonious, this scientific methodology will try to capture the systematic behavior of the data and represent the factors that are nonsystematic and cannot be predicted as error terms.

Using the traditional linear regression equation -- the least squares method, the research will attempt to explain the relationship between the dependent variable and the regressors for forty-four quarters as follows:

$$\psi_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \varepsilon_t ,$$

where

ψ_t = The dependent variable -- the initial QEDS reported item measured at time t.

X_{1t} - X_{3t} = Three dummy variables that represent seasonality -- the four quarters of the year. We suspect that seasonality as a qualitative regressor variable will help improve prediction together with other quantitative variables.

X_{4t} = MA QEDS estimates constructed from MA growth rates and panel shift data.

X_{5t} = Number of respondents for each quarter by entity type.

The first step of the model estimation phase was to look at the results of a multiple regression equation to assess the functional relationship between the dependent and independent variables. The search for outliers in the output was the primary goal.

If the statistics indicated that the observation was both an outlier and an influential point, the observation was marked for re-estimation by the regression procedure instead of eliminating it altogether. This process ensured that the number of observations did not diminish while providing better input to the regression process. Estimating the effect of the outlier and removing the effect from the data point will eliminate its adverse influences on the final coefficient estimates.

As the estimation procedure is discussed, output for total demand deposits adjusted (2212) aggregated over member commercial banks (entity 1) will be presented for illustration purposes. Table 1 shows the initial estimation of outlier effects for this item and entity aggregates.

The listing is the first regression procedure that estimates the effects of the outliers Q1, Q7, Q19, and Q43. Each of these outliers is extremely significant, as can be noted in the T-statistic and the appropriate probability. The other independent variables that are significant in this model are the QEDS estimate from monetary affairs, the number of banks, and the seasonal effect of quarter three.

Table 1.--Model Selection and Estimation

Demand Deposits Adjusted (2212) - Commercial Member Banks (Entity 1)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	1.1765984E14	1.6808548E13	888.588	0.0001
Error	36	680976490487	18916013625		
C Total	43	1.1834081E14			

Root MSE	137535.49951	R-square	0.9942
Dep Mean	5024979.59091	Adj R-sq	0.9931
C.V.	2.73704		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-248999	99782.214546	-2.495	0.0173
D3	1	-340885	54426.852218	-6.263	0.0001
F	1	461.528841	106.19117853	4.346	0.0001
QEST	1	0.909282	0.02327381	39.069	0.0001
Q1	1	2136875	157101.47043	13.602	0.0001
Q7	1	2179965	168519.89826	12.936	0.0001
Q19	1	2154042	161564.34373	13.332	0.0001
Q43	1	1519245	146780.11889	10.350	0.0001

As shown in Table 2, outliers (in the dependent variable) have potential to "pull" the regression equation in the wrong direction causing inadequate explanation of the "true" data. In addition, it will not predict future values well. This output shows that observations 1, 7, 19, and 43 have been perfectly estimated from the data and that the next regression is ready for processing once the outlier effects have been removed.

Table 2.--Estimation of Outlier Effects

	Dep Var Obs	Predict QED	Std Err Value	Predict	Residual	Std Err Residual	Student Residual	-2-1-0 1 2	Cook's D
*	1	2281560	2281560	137535.5	-568E-12	.	.		.
	2	2372881	2265323	49724.71	107558	128232.1	0.839	| |* |	0.013
	3	1370009	1868559	72212.91	-498550	117052.6	-4.259	|******| |	0.863
	4	1483171	1385466	58304.64	97705.1	124565.6	0.784	| |* |	0.017
	5	1425684	1338455	58630.46	87228.5	124412.6	0.701	| |* |	0.014
	6	1521271	1437119	58192.14	84152.3	124618.2	0.675	| |* |	0.012
*	7	3674690	3674690	137535.5	-276E-12	.	.		.
	8	3859955	3938499	36784.03	-78543.5	132525.3	-0.593	| *| |	0.003
	9	3636632	3732860	39588.95	-96228.0	131714.6	-0.731	| *| |	0.006
	17	4225384	4275400	32474.31	-50015.9	133646.7	-0.374	| | |	0.001
	18	4260961	4262018	30884.77	-1056.9	134022.9	-0.008	| | |	0.000
*	19	6376338	6376338	137535.5	-483E-12	.	.		.
	20	6739094	6782243	37362.20	-43149.2	132363.4	-0.326	| | |	0.001
	21	6349651	6351019	37091.75	-1368.4	132439.5	-0.010	| | |	0.000
	28	6357825	6314304	30277.27	43521.1	134161.5	0.324	| | |	0.001
	29	6002545	6042162	29553.99	-39617.5	134322.7	-0.295	| | |	0.001
	42	5832947	5720894	42352.31	112053	130852.2	0.856	| |* |	0.010
*	43	6967935	6967935	137535.5	-874E-12	.	.		.

The next step in the model estimation process was to rerun the multiple linear regression model with the estimated observations. Table 3 shows how the regression output was obtained.

Table 3.--Regression Model to Determine Final Estimates

Regression Output Total Demand Deposits Adjusted for Commercial Member Banks

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	1.3836181E14	4.6120603E13	2709.086	0.0000
Error	40	680976490487	17024412262		
C Total	43	1.3904278E14			

Root MSE	130477.63127	R-square	0.9951
Dep Mean	4843385.79272	Adj R-sq	0.9947
C.V.	2.69393		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-248999	88373.959762	-2.818	0.0075
D3	1	-340885	46066.671130	-7.400	0.0001
F	1	461.528841	85.31971721	5.409	0.0001
QEST	1	0.909282	0.01699985	53.488	0.0001

The output lists the final results of the model estimates that were significant for total demand deposits adjusted - for Entity = 1 (commercial member banks). It is clear that this three variable regression model explains ninety nine percent of the variation in the QED's aggregate for commercial member banks apart from outliers.

This regression estimation process was carried out for each combination of entity and item. The model estimates shown in Table 4 on the following page, depicts the observations that needed to be re-estimated and the independent variables that were significant in the final model for each combination.

The next listing examines the error variance over the quarters to ensure consistency.

Heteroscedasticity -- Check for Constant Variance

OBS	D4	D3	D2	D1	ITEM	ENTITY	_TYPE_	_FREQ_	USS	N	Size of Variance
1	0	0	0	1	2212	1	0	11	40983878421	11	61039.39
2	0	0	1	0	2212	1	0	11	76861151631	11	83590.53
3	0	1	0	0	2212	1	0	11	495430154028	11	212224.19
4	1	0	0	0	2212	1	0	11	67701306407	11	78451.67

This listing depicts the size of the variance for the residuals in the column label "size of variance." It is apparent that the variances are synchronized within each quarter with slightly higher variances for the seasonal factor three. The results are consistent with the seasonal effect reflected in the regression model above where D3 or the third quarter dummy variable was significant.

The ARIMA procedure further examines the residuals from the regression model to confirm that time series elements in the dependent variable were considered in construction of the final equation.

The results of the Q statistic (Chi Square = 5.28) clearly indicate that the autocorrelation check for residuals are all highly insignificant. This is evidence that the residuals from the regression model are ***white noise*** and that the model does not suffer from violations of assumptions.

Chi-Squared Check of Residuals

```
                                Autocorrelations

Lag Covariance Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1           Std
  0 1.54767E10     1.00000  |                    |********************|             0
  1 2012229415     0.13002  |                  . |***  .              |      0.150756
  2  -3.0976E9    -0.20015  |                  . ****|     .              |      0.153283
  3  -1.4777E9    -0.09548  |                  .   **|     .              |      0.159112
  4 -1.22818E9    -0.07936  |                  .   **|     .              |      0.160408
  5 -2.36701E9    -0.15294  |                  .  ***|     .              |      0.161298
  6 1606879130     0.10383  |                 .      |**      .            |      0.164561
                                        "." marks two standard errors


                     Autocorrelation Check for White Noise

        To    Chi                     Autocorrelations
       Lag  Square DF   Prob
         6    5.28  6  0.508   0.130 -0.200 -0.095 -0.079 -0.153  0.104
```

Table 4.--Model Estimates

Item	Entity 1 Outliers	Entity 1 Final Model	Entity 2 Outliers	Entity 2 Final Model	Entity 3 Outliers	Entity 3 Final Model	Entity 4 Outliers	Entity 4 Final Model	Entity 5 Outliers	Entity 5 Final Model
0080	1,7,19,43	D3, F, Qest	1,7,19,43	D3, Qest	31,37,43, 44	F, Qest	3,19,27, 43	F, Qest	1,7,19,43	D3, Qest
2212	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	27,33,35 43	D3, F, Qest	11,15,27, 43	D4, Qest	23,35,43	Qest
6917	1,7,19,43	Qest	1,7,19,43	Qest	31,35,42, 43	Qest	3,19,43	D3, Qest	1,7,19,43	Qest
2389	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	27,31,35, 43	F, Qest	1,7,19, 43	F, Qest	1,7,19,43	F, Qest
2697	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	19,31,35, 43	Qest	1,7,19, 43	F, Qest	1,7,19,43	Qest
2604	1,7,19,43	F, Qest	24	D3, F, Qest	27,31,35, 43	Qest	1,7,19, 43	D3, F, Qest	7,19,31, 43	Qest
2280	17,36,40	Qest	21	F, Qest	14	F	15,16,28	F, Qest	17	F, Qest
2698	35	D3, F, Qest	1,7,19,43	Qest	None	Qest	3	Qest	11,12,13, 15	Qest
0020	1,7,19,43,3 9	D3, F, Qest	1,7,19,43	Qest	43	F, Qest	43	F, Qest	19,27,43	Qest
0063	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	25,27,43	Qest	1,3,7,9	Qest	31,43	Qest
2340	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	33,35,43	Qest	11,15,27	Qest	23,43	Qest
2214	1,7,19,43	D3, F, Qest	1,7,19,43	D3, F, Qest	33,35,43	Qest	3,19,27, 43	Qest	1,7,19,42	Qest
6918	1,7,19,43	F, Qest	1,7,19,43	D3, F, Qest	31,39,43	Qest	1,3,11,19	Qest		Qest

D3 = Quarter 3 Seasonal Factor D4 = Quarter 4 Seasonal Factor F = Number of Banks Qest = QEDS MA Estimate

The Current and Proposed Editing Process

The current editing of quarterly deposits data do not use regression analysis -- a collection of statistical techniques that serve as a scientific basis for drawing inferences about relationships among quantities.

Data are currently analyzed at Statistical Services by using SAS/INSIGHT®, an interactive software system that provides extensive statistical capabilities. This system, employs SAS graphical features to display observations that need further investigation. The current editing technique focuses on using box plots and scatter plots, allows the analyst to visualize the data while making decisions about deviant observations. The data for analysis are constructed as follows:

- Aggregated data (especially residuals) for all reported items are analyzed using histograms and box plots to determine historical trends. The current quarter's data are also compared with historical data to determine trends in the current data. Items with abnormal aggregated values are investigated thoroughly at the micro data level.

- Micro data focuses on similar entity types, total deposit levels, and historical fluctuations for certain items. Depending on the variable, the micro data are analyzed using box plots and scatter plots to find trends and unusual data. Institutions with unusual values are referred to the Reserve Banks for verification, explanation, or revision.

This research recommends a different approach to analyze QEDS data. Firstly, it employs panel shift adjustments from quarter three to correctly reflect aggregate deposits data. This allows for a more accurate database from which inferences may be drawn. Secondly, the sample consists of the whole aggregate panel of respondents for a given quarter including the additions to the panel. Finally, in addition to the MA estimate, predictor variables such as the quarter three factor, the quarter four factor and the number of respondents being as significant as they are adds to the models ability to make better predictions. Therefore, this study has been able to develop a statistical methodology for analyzing QEDS data by fine tuning the MA estimates together with other pertinent varibles.

The proposed editing process will be implemented every quarter by compiling a dataset that has been adjusted for panel shift data, to which growth rates have been applied to get MAQEDS estimates. Then using SAS these data will be analyzed using the model estimates to flag deviant data points for the current quarter. Finally, the analyst will use the micro data to rank positive and negative percentage contributors to investigate the individual bank/banks causing the quarters prediction to be off.

Findings and Implementation

This QEDS research project has led to some interesting findings that may prove to be useful:

- The most interesting of all regressor variables was the computed QEDS estimate variable. This predictor was calculated from data reported by another group of institutions who report similar data on a weekly basis. Panel shift information and the growth rate will be applied to obtain observations that represent Monetary aggregates computed at the board. The regression was mostly explained by this variable which was highly significant. Ninety five percent of the models generated for this project included this QEDS estimate in the regression.

- Using quarterly indicator variables in the model allowed for the estimation and significance testing of seasonal effects in predicting QEDS aggregates. It was interesting to find that although QEDS aggregate estimates incorporate Panel shift effects that occur in quarter three (Q3), the significance of the Q3 indicator coefficient indicates that additional information has been obtained from predicting QEDS aggregates. In addition, panel cutoff changes become effective during quarter three -- every three years when cutoffs are reviewed.

- This methodology was also useful in that it allowed the model to take into account the outlier estimation process, which eliminated any distortion in the input data set that finally evaluated the usefulness of the predictors.

- The results from this study varied by "entity type" and item, with similar entity types sharing similar model forms. Entities 1 and 2, which are commercial member and nonmember banks, shared similar models that predicted their aggregates. Entities 3, 4, and 5 which are mutual savings, savings and loans associations, and credit unions, had comparable models explaining their aggregates.

To implement the findings of this methodology, analysts will use the predictive equations to evaluate incoming data. The evaluation process will create regions of acceptability, and any data falling outside these regions will be marked for further examination by analysts, at the micro level. The acceptable regions were computed by adding and subtracting from the predicted value, three times an estimate of the standard error of the regression model. The goal of this analysis is to further investigate the entity and item combinations that are flagged by a 1, since those observations reflect incoming data outside of prediction levels.

The table below is an example of items flagged for investigation for quarter four of 1995 and assists the analyst in focusing on the quarterly data that requires further probing.

DATE	ENTITY	ITEM	Qeds Quarterly Value	Model Estimate	MA Estimate	+/- Tolerance	Flag indicator
960325	3	0080	11822	9883	11200	1642	1
960325	5	0063	939696	821030	842245	116862	1
960325	5	6918	138172	115824	148492	22317	1

Looking at the table above, if the analyst investigates item 0080, 1939 is the difference that needs to be accounted for in this quarters data.

DATE	ENTITY	ITEM	Qeds Quarterly Value	Model Estimate	Difference
960325	3	0080	11822	9883	1939

When reported QEDs aggregates are different from the predicted value, here are some suggestions for probing the data:

- Retrieve data for the past and current quarter for the item in question. Then compute dollar and percentage differences which will be ranked to look at the highest and lowest twenty five rankings. These fifty banks will then be graphed by rank/dollar difference and rank/percent difference to observe irregular patterns. These graphs below are for item 0080 -- Vault Cash for entity 3.

Figure 1 displays the dollar difference for Vault Cash (item 0080) by rank for mutual savings banks. The graph clearly indicates that bank 11 deviates from the general pattern of behaviour of the other banks for this quarter. Therefore, this outlier causes the aggregate data to be inflated and the predicted value to be flagged. The tables below show the data.

Figure 1.--Dollar Differences Ranked for 0080

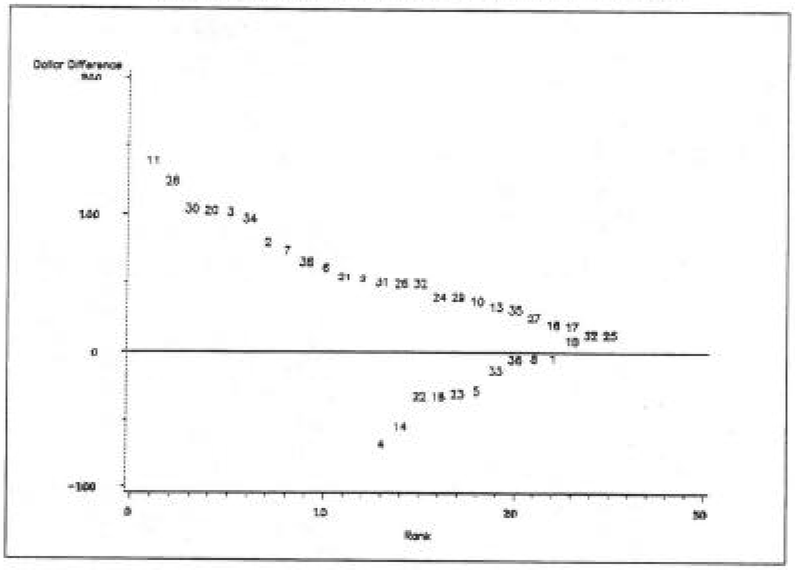
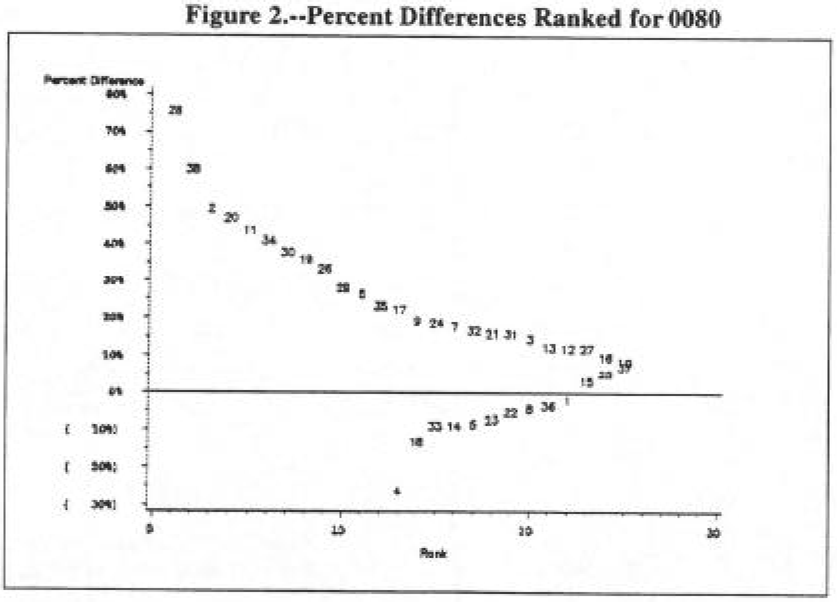


Figure 2 indicates that bank 28 and 38 are high up on the scale of positive percent changes. This indicates that the volume of activity for this bank has increased by a large amount due to structural or other changes. Bank 4 at the lower end of the negative scale should also be investigated. This is another reason for flagging the predicted value for this quarter.

Figure 2.--Percent Differences Ranked for 0080



Dollar Difference Report

Rank	QED00080	Last Quarters Value	Dollar Difference
1	454	316	138
2	288	164	124
3	378	274	104
4	322	219	103
5	803	701	102
:			
:			
:			

Percent Difference Report

17	395	426	-31
16	219	252	-33
15	601	634	-33
14	555	610	-55
13	191	259	-68

Rank	QED00080	Last Quarters Value	Percent Difference
1	288	164	76%
2	176	110	60%
3	242	162	49%

- After using the step above to locate extreme data values, the following statistics will help to identify any pattern in the data for a particular entity and item that may be reflected in the incoming data. In order to obtain an explanation of the general and uniform trend in the data that is not accounted for by an individual or a group of individuals, the following statistics will be helpful:

  - In this example, quantile statistic indicates that most of the differences calculated are positive, which means that there were more increases than decreases from the past quarter.

  - Information on the mean and standard deviation.

Statistics on Dollar Differences for 0080

N	38	100% Max	138	99%	138
Mean	32.65789	75% Q3	62	95%	124
Std Dev	49.14528	50% Med	32	90%	103
Variance	2415.258	25% Q1	-4	10%	-33
		0% Min	-68	5%	-55
		1%	-68		
Range	206				
Q3-Q1	66				
Mode	13				

Conclusion

This research has provided a statistical methodology for editing quarterly deposits data using forecasting techniques, to identify "acceptable" and "unacceptable" data. The assumptions made at the outset have proven to be very useful in building this model. There are many ways to edit "micro level" data and this research has alluded to some of these possibilities. As the model is explored there is no doubt that the prototype can be refined even further.

Bibliography

Pindyck, R. S. and Rubinfeld, D. L. (1991). *Econometric Models and Economic Forecasts*, New York: McGraw-Hill, Inc.

Myers, R. H. (1990). *Classical and Modern Regression with Applications,* California: Duxbury Press.

SAS Institute, Inc. (1991). Time Series Modeling and Forecasting, *Financial Reporting and Loan Analysis*, North Carolina: SAS Institute, Inc. ■

13

Chapter

Experiences in Re-Engineering the Approach to Editing and Imputing Canadian Imports Data

Clancy Barrett and Francois Laflamme, Statistics Canada

Abstract

The large volume of administrative data to be processed on a monthly basis and the need to use limited operational resources more efficiently led International Trade Division of Statistics Canada to the decision to base the new edit system on a combination of micro and macro approaches. The new edit system uses a series of modules that successively handled: high impact records, records in error that belong to aggregates with a high potential error, comparison of current aggregates to historical aggregates, and some special requirements. The presentation will mainly describe in a practical way the detailed mechanism for selecting and manually reviewing records. ■

Data Editing in an Automated Environment: A Practical Retrospective -- The CPS Experience

Gregory D. Weyland, U. S. Bureau of the Census

13
Chapter

Abstract

The advent of computerized data collection has opened the door to a myriad of possible scenarios for editing survey data. This includes the use of edits at the time of data capture, dependent interviewing using longitudinal data, among others, independently or in combination with detailed consistency and allocation edits after data collection is complete.

In January 1994, the Current Population Survey (CPS) became the first Bureau of the Census demographic survey to switch to a completely computerized data collection environment. This was preceded by an almost two year, large-scale test of the new data capture system and the revised processing system it required.

This discussion will review our initial plans for how the data would be edited in the new environment. Then it will review the adjustments and revisions made over the past few years to allow the editing procedures to meet the practical requirements of CPS while still improving the quality of CPS data. ■