Design and Evaluation of a Continuous Data Level Auditing System

Alexander Kogan

Michael G. Alles

Miklos A. Vasarhelyi

Department of Accounting & Information Systems

Rutgers Business School

Rutgers University

180 University Ave

Newark, NJ 07102

Jia Wu

Department of Accounting and Finance

University of Massachusetts – Dartmouth

285 Old Westport Road

North Dartmouth, MA 02747

October 17, 2013[*]

Design and Evaluation of a Continuous Data Level Auditing System

[*] We thank seminar participants at the 2011 Rutgers Continuous Auditing Symposium and the European Accounting Association for helpful comments on an earlier version of this paper. We thank the Editor and anonymous reviewers whose comments help us improve this paper.

SUMMARY: This study develops a framework for continuous data level auditing system and uses a large sample of procurement data from a major health care provider to simulate an implementation of this framework. In this framework the first layer monitors compliance with deterministic business process rules and the second layer consists of analytical monitoring of business processes. A distinction is made between exceptions identified by the first layer and anomalies identified by the second one. The unique capability of continuous auditing to investigate (and possibly remediate) the identified anomalies in "pseudo-real time" (e.g., on a daily basis) is simulated and evaluated. Overall, evidence is provided that continuous auditing of complete population data can lead to superior results, but only when audit practices change to reflect the new reality of data availability.

Keywords: continuous auditing, analytical procedures, population data, auditing practice.

Data availability: The data is proprietary. Please contact the authors for details.

INTRODUCTION

This research note develops a continuous data level auditing system and uses a large sample of procurement data from a major health care provider to simulate an implementation of continuous auditing and investigate how audit could change when using real time complete population data. Overall, evidence is provided that continuous auditing of complete population data can lead to superior results, but only when audit practices change to reflect the new reality of data availability. This research note makes the following contributions:

It develops a framework for continuous monitoring of business processes by internal auditors. In this framework the first layer monitors compliance with deterministic business process rules and the second layer implements analytical monitoring of business processes. In the extant auditing literature (mostly focused on external, not internal auditing), analytical procedures are utilized first in order to identify areas of potential concern for focusing detailed testing on the most relevant samples. This sequence of audit procedures is driven by the objective to reduce the amount of data that the auditor needs to obtain given the high cost of doing so. The change in the order becomes essential in continuous auditing since data today is much more easily accessible to auditors and the automation of tests of detail both enables and necessitates their application to the complete population of business transactions thus eliminating any rationale for sampling.

It introduces a clear distinction between the two types of business process irregularities: those that are violations of deterministic business process rules and those that are significant statistical deviations from the steady state business process behavior. In the note's framework the former are labeled exceptions and the latter anomalies. For example, if a company requires that all purchases be made only from previously authorized vendors, then a purchase order for a vendor not present in the list of authorized vendors constitutes an exception. If a company usually receives between 20 and 30 deliveries during a weekday, then a weekday with only 10 deliveries constitutes an anomaly. No clear distinction between exceptions and anomalies is usually made in the extant literature, and the terms are often used interchangeably. Distinguishing between these two types of irregularities allows for developing appropriate investigation and remediation procedures and properly focusing the efforts of internal auditors.

It bases audit analytics on highly disaggregated business process data down to the level of daily business process metrics. The most disaggregate data utilized in the extant literature tends to be limited to the level of monthly metrics. The problem of choosing the proper level of aggregation for analytical monitoring is identified and analytical models based on daily and weekly process metrics are empirically compared. The analysis shows that there is no universally preferable level of aggregation, but that the choice of aggregation is important, has to be made by the internal auditor, and will depend on the likely distribution of possible errors (with the more concentrated errors better identified by more disaggregated models).

This research note compares the performance of several different analytical models, including the vector autoregressive (VAR) model that is capable of inferring the lags of relationships between business process metrics jointly with the model coefficients. It proposes a way of controlling this model's complexity to prevent possible over-fitting, which the empirical evaluation shows to be successful, with the resulting subset VAR model often providing the best results.

Finally, It simulates and evaluates the unique capability of continuous auditing to investigate (and possibly remediate) the identified anomalies in "pseudo-real time" (which, in the case of daily business process metrics, means on a daily basis). The assumption that an identified business process anomaly can be investigated by the auditors within 24 hours to identify and correct any problems is supported by the current practices in more advanced internal audit departments. The simulation study shows that pseudo-real time error corrections facilitates identification (and therefore correction) of more of the seeded errors, and thus the implementation of continuous auditing can lead to the increased effectiveness of audit procedures. The study shows that the implementation of pseudo-real time error correction provides effectiveness benefits complementary to the use of disaggregated business process data. This result can be used by internal audit departments to reorganize their workflow to enable near-real-time investigation of business process anomalies to improve efficiency and effectiveness of audits.

This paper is organized as follows. The next section provides a review of the relevant literature. Then the paper describes the design and implementation of continuous data level auditing systems, followed by the discussion of how to aggregate transactional data, and how to construct analytical benchmark models using three different statistical methods. The last section compares the

ability of the continuous data level auditing systems to detect anomalies under various settings, introduces the pseudo-real time error correction protocol and discusses the results. The paper concludes by summarizing the results.

LITERATURE REVIEW

This paper draws from and contributes to multiple streams of literature in system design, continuous auditing and analytical procedures.

Continuous Auditing

Groomer and Murthy (1989) and Vasarhelyi and Halper (1991) pioneered the two modern approaches toward designing the architecture of a CA system: the embedded audit modules and the control and monitoring layer, respectively. The literature on CA since then has increased considerably, ranging from the technical aspects of CA (Kogan et al. 1999, Woodroof and Searcy 2001, Rezaee et al. 2002, Murthy 2004, Murthy and Groomer 2004, etc.) to the examinations of the economic drivers of CA and their potential impact on audit practice (Alles et al. 2002 and 2004, Elliott 2002; Vasarhelyi 2002, Searcy et al. 2004). Kogan et al. (1999) propose a program of research in CA. In the discussion of the CA system architecture they identify a tradeoff in CA between auditing the enterprise system versus auditing enterprise data. A study by Alles et al. (2006) develops the architecture of a CA system for the environment of highly automated and integrated enterprise system processes, and shows that a CA system for such environments can be successfully implemented on the basis of continuous monitoring of business process control settings. A study published by the Australian Institute of Chartered Accountants (Vasarhelyi et al, 2010) summarizes the extant state of research and practice in CA. However, few empirical studies have conducted on CA in general and on analytical procedures for CA in particular due to the general lack of data. This paper contributes to the CA literature by providing empirical evidence to illustrate the advantages of CA in close to "real time" problem resolution.

Analytical Procedures

PCAOB Auditing Standard No. 15 (paragraph 21) defines Analytical Procedures (AP) as the "*evaluations of financial information made by a study of plausible relationships among both financial and nonfinancial data*". SAS 56 requires that analytical procedures be performed during the planning and review stages of an audit, and recommends their use in substantive testing in order to limit the subsequent testing of details to areas of detected concern. That sequence is dictated because manually undertaken tests

of detail are costly, and therefore they are resorted to if the account balance based AP tests indicate that there might be a problem.

There is an extensive research literature on analytical procedures in auditing. Many papers discuss various analytical procedures ranging from financial ratio analysis to regression modeling that focus on highly aggregated data such as account balances (Hylas and Ashton 1982, Kinney 1987, Loebbecke and Steinbart 1987, Biggs et al. 1988, Wright and Ashton 1989, Hirst and Koonce 1996). The percentages of errors found using such analytical procedures are usually not high, varying between 15% and 50%. Only a few papers examine analytical procedures for more disaggregated data. Knechel (1985(b)) provides a stochastic model of error aggregation starting with the level of individual transactions up to the level of account balances. Dzeng (1994) compares 8 univariate and multivariate AP models using quarterly and monthly financial and non-financial data of a university, and concludes that disaggregated data yield better precisions in a multivariate time-series based expectation model. Other studies also find that applying AP models to higher frequency monthly data can improve analytical procedure effectiveness (Knechel 1985(a), 1986 and 1988, Chen and Leitch 1998 and 1999, Leitch and Chen 2003, Hoitash et al. 2006). By contrast, Allen et al. (1999) use both financial and non-financial monthly data of a multi-location firm and do not find any supporting evidence that geographically disaggregate data can improve analytical procedures.

This study focuses on operational internal audit and builds a data level CA system which utilizes analytical procedures applied to even more highly disaggregate daily metrics of business processes (BP). It investigates several different probabilistic models of business processes to serve as the audit benchmark: the Linear Regression Model (LRM), the Simultaneous Equation Model (SEM), and Vector Autoregressive Model (VAR). The use of SEM in analytical procedures has been examined by Leitch and Chen (2003), but only using monthly financial statement data. Their finding indicates that SEM can generally outperform other AP models including Martingale and ARIMA.

Vector Autoregressive Model has not been fully explored in the auditing literature. There are a number of studies utilizing univariate time series models (Knechel 1986 and 1988, Lorek et al. 1992, Chen and Leitch 1998, Leitch and Chen 2003), but only one, by Dzeng (1994), which uses VAR. Dzeng concludes that VAR is better than other modeling techniques in generating expectation models, and he specifically recommends using Bayesian VAR (BVAR) models. The VAR model can not only represent the interrelationships between BPs but also capture their time series properties. Although (to the best of our knowledge) VAR has been discussed only once in the auditing

literature, studies in other disciplines have either employed or discussed VAR as a forecasting method (see e.g., Swanson 1998, Pandher 2002). Detailed statistical development of the VAR methodology and related issues can be found in Enders (2004).

DESIGN AND IMPLEMENTATION OF A CONTINUOUS DATA LEVEL AUDITING SYSTEM

Continuous Data Level Auditing System Design

The objective of a CA system designed in this study is to provide close to real-time assurance on the integrity of certain enterprise BPs. As in conventional auditing, such a system can utilize two different types of procedures: those monitoring BP controls and those substantively analyzing BP transactions. As Alles et al. (2006) indicate, BP control monitoring requires that the client possesses a modern integrated IT infrastructure, and faces challenges even then. They also show that today few firms have the type of tight, end to end system integration that continuous control monitoring depends upon. This paper focuses on designing a CA system for the much more common enterprise environment in which data is derived from multiple legacy systems that lack centralized and automated controls. This lack of a control based monitoring system is why the proposed CA system is data-oriented instead, and the provision of assurance is based on verifying transactions and on BP based analytical procedures. Where the IT environment allows, CA would ideally encompass both continuous (data level) assurance (CDA) and continuous control monitoring (CCM).

The architecture of the data level CA system is driven by the procedures it has to implement. While the subject matter it deals with is quite different from that in conventional auditing, one can view its procedures as analogous to automated substantive audit procedures such as detailed transaction testing and analytical procedures. Therefore, the two main components of the CA system are those executing automatic transaction verification and automatic analytical BP monitoring, as shown in Figure 1.

[Insert Figure 1: Architecture of Continuous Data Level Auditing System here]

An important innovation in the proposed architecture of the CA system presented in Figure 1 is the utilization of analytical monitoring as the second (rather than the first) stage of data analysis. In conventional (manual) auditing, testing transactional details is very laborious and has to be based on statistical sampling to control the cost of auditing. Therefore, analytical procedures are utilized

first for the identification of areas of concern to focus the sampling on. In CA, which utilizes automated transaction verification tests, there is no need for sampling, since automated tests can be easily applied to the complete population of business transactions.

The implementation of the transaction verification component of the CA system is based on identifying BP rules and formalizing them as transaction integrity and validity constraints. Every recorded transaction is then checked against all the formal rules in the component, and if it violates any of the rules, then the transaction is flagged as an exception. Every exception generates a CA alarm in the CA system, which it sends to the appropriate parties for resolution. Since the alarm specifies which formal BP rules are violated by the exception, resolving exceptions should be a fairly straightforward task. Once the transaction data is verified, it is in an acceptable form to be screened for anomalies by the analytical monitoring component.

The analytical monitoring component uses as benchmarks stable probabilistic relationships between aggregated BP metrics. The configuration of this component requires the following choices to be made based on the analysis of historical transaction records:

- The choice of BP metrics to monitor such as the cost of the ordered goods, the quantity of the delivered goods, and the dollar amount of payments to vendors;
- The level of aggregation of BP metrics such as daily or weekly;
- The particular AP statistical model to estimate stable relationships between BP metrics such as linear regression, SEM or VAR, and
- The acceptable range of variation for every monitored BP metric.

The chosen AP is used to calculate the expected value for every BP metric. If the difference between the observed and the predicted value of the BP metric is outside the acceptable range of variation, then this value of the BP metric is flagged as an anomaly. The choices described above cannot be made a priory and will differ depending on the circumstances and the past behavior of BPs of a particular enterprise.

Business Data Warehouse

A salient feature of the architecture proposed in Figure 1 is the Business Data Warehouse which serves as the data integration point for the disparate (mostly legacy) systems in the enterprise system landscape. Enterprise systems that support key BPs routinely collect business process data in the unfiltered highly disaggregated form. If the enterprise has implemented an integrated ERP system, then BP data is readily available in the ERP central database. However, the most common current situation is that the enterprise system landscape consists of a patchwork of different systems,

many of which are legacy ones and are often file-based. In such enterprise systems direct real-time access to BP data is highly problematic, if at all possible at any reasonable expense of time and effort. Therefore, a data-oriented CA system usually cannot be cost-effectively deployed in such environment unless the enterprise deploys an overlay data repository commonly known as "Business Data Warehouse". This is a relational database management system specially designed to host BP data provided by the other enterprise systems, including the legacy cycle-focused ones (such as sales processing or accounts receivable). The upload of BP data to the data warehouse usually takes place on the daily basis, and even more often in some enterprise environments.

While the main functionality of a data warehouse is online analytical processing, the CA system developed here relies only on its function as the global repository of BP data. The availability of disaggregated BP data makes it possible for the auditor to access any raw, unfiltered and disaggregated data that is required for the construction and operation of the proposed CA system.

Data Description

Our simulated implementation of the data-oriented CA system focuses on the procurement-related BPs and utilizes the data sets extracted from the data warehouse of a healthcare management business with many billions of dollars in assets and close to two hundred thousand employees. It is a major national provider of healthcare services, with a network composed of locally managed facilities that include numerous hospitals and outpatient surgery centers in the US and overseas.

The organization provided extracts from their transactional data warehouse which, while only a sample limited in time and geography, still encompassed megabytes of data, several orders of magnitude more detailed than anything typically used in a standard audit. The data sets include all procurement cycle daily transactions from October 1st, 2003 through June 30th, 2004.[1] The number of transaction records for each activity ranges from approximately 330,000 to 550,000. These transactions are performed by ten facilities of the firm including one regional warehouse and nine hospitals and surgical centers. The data was first collected by the ten facilities and then transferred to the central data warehouse in the firm's headquarters.

[1] Only older data was available from the client given their privacy concerns. The age of the data has little effect on our research since its purpose is to determine how continuous auditing of population data will change audit practice and not to analyze this particular data set.

Transaction Verification

When auditors have access to population data, the first step is to undertake tests of details to detect violations of key controls. Once that is done the auditor can turn to determining whether there are anomalies that do not violate any established controls but which may be nonetheless indicative of potential problems, such as a proliferation of cash deposits just under the $10,000 federal Cash Transaction Reporting limit.

When implementing the transaction verification component we first identify the following three key BPs in the supply chain procurement cycle: ordering, receiving, and voucher payment. Since this data is uploaded to the business's data warehouse from the underlying legacy system, there are numerous data integrity issues, which have to be identified by the transaction verification component of the CA system before the data is suitable for AP testing. To simulate the functionality of the transaction verification component, we formally specify various data validity, consistency, and referential integrity constraints, and then filter through them all the available transactions.

Two categories of erroneous records are removed from the data sets: those that violate data integrity and those that violate referential integrity.

i. Data integrity violations include but are not limited to invalid purchase quantities, receiving quantities, and check numbers.[2]

ii. Referential integrity violations are largely caused by many unmatched records among different business processes. For example, a receiving transaction cannot be matched with any related ordering transaction. A payment was made for a non-existent purchase order.

An additional step in the transaction filtering phase is to delete non-business-day records. Though some sporadic transactions have occurred on some weekends and holidays, the number of these transactions accounts for only a small fraction of that on a working day. The existence of these transactions violates enterprise rules, and therefore such transactions are in fact exceptions, and should trigger alarms. Additionally, if these non-business-day records were left in the sample, they would inevitably trigger false alarms simply because of low transaction volume, thus introducing noise in our detection models

The client firm considered the list of exceptions identified at this stage as a major source of value added from the project. It is to be anticipated that as legacy systems are gradually superseded

[2] We found negative or zero numbers in these values which cannot always be justified.

by the firm's ERP system with automated controls, the transaction verification component of the CA system will be detecting fewer and fewer problems.

Business Process Based Analytical Procedures

The implementation of the analytical procedure (AP) component of the CA system requires creation of the models of expected behavior to enable anomaly detection. We call such expectation models continuity equations (CE), a term defined in Alles et al (2008): *"… as* ***stable probabilistic models of highly disaggregated business processes****."*

To develop a CE model, a business process metric is calculated over a subset of transactions corresponding to intervals along some important BP dimensions (such as time, region, product, customer, etc.). Since the relationship between the metrics holds only probabilistically, the model also has to specify the acceptable range of variation of the residuals. An anomaly arises when the observed values of the metrics result in residuals which fall outside this acceptable range. Every anomaly generates a CA alarm in the CA system, which it sends to the appropriate parties for resolution.[3] A rating of the seriousness of the anomaly could also be passed along to these parties.

In contrast with an exception, which is associated to an individual transaction, an anomaly is associated with a subset of transactions used to calculate the value of the metric. Therefore, the resolution of an anomaly is not straightforward. Moreover, an anomaly is not necessarily indicative of a problem, since it can be due to statistical fluctuation or have legitimate business reasons. For example, in a business where the weekly amount of orders ranges between $100,000 and $200,000, a week with the total amount of orders worth only $10,000, as well as the following week with the total amount of orders worth $300,000 will both be statistical anomalies. While the investigation of the former may show that the decrease in the orders is due to computer problems in the procurement department, and thus is not nefarious, the investigation of the latter may reveal some unnecessary orders placed by a local manager eager to use up the budget before the deadline. As this example indicates, creating a metric that will prove effective in detecting BP problems is not a trivial task since it must be based on what is "usual" for a BP. The appropriate model will be derived from the population data which has been first filtered to remove exceptions.

[3] For example, while a deposit over $100,000 would automatically trigger money laundering audit, three deposits of $35,000 within in a week could be a suspicious activity or anomaly. It can be legitimate or just a move to avoid the audit.

The next section investigates three probabilistic models that can serve as candidates for the CE benchmarks of the firm's supply chain processes: a Simultaneous Equation Model (SEM), Vector Autoregressive (VAR) models, and a Linear Regression Model (LRM).

ESTIMATION OF CONTINUITY EQUATIONS

Before developing a probabilistic model as a candidate for the CE benchmark of the procurement BP, there are several parameters to be determined that shape the overall approach towards model estimation. These include the choice of BP metric, the extent to which data is aggregated, and how the model estimation changes dynamically as new data is obtained. Once those parameters are chosen, attention can turn to the CE estimation.

Overall Parameters of Continuity Equation Estimation

Choice of Business Process Metric

A critical issue in modeling BPs analytically is the choice of BP metrics. The traditional accounting choice has been the use of financial measures (e.g., dollar amounts), driven in the first place by the reliance on ledger entries as the primary source of data. In a data environment where disaggregate data is available, modeling BPs can also be undertaken using other types of non-financial metrics such as physical measurements or document counts, as well as the traditional dollar amounts of each transaction, or the number of transactions processed. In this study, the primary experiment utilizes the traditional financial metrics, while the transaction item quantity is selected as the BP metric in the additional experiment that shows the robustness of our approach with respect to the choice of BP metrics and demonstrates how APs can utilize multiple metrics to examine transaction flows. [4] Auditing on different metrics would enable auditors to detect a more diverse set of patterns of firm behavior. Once the BP metrics are chosen, the next step is to determine the appropriate degree of aggregation at which to construct the CE based benchmark.

Data Aggregation

The main argument against using aggregated data is that it inevitably leads to a loss of information about individual transactions. Thus, investigating an anomalous aggregated BP metric requires an examination of a large number of transactions that were aggregated to calculate the metric. But aggregation can also make it possible to see more general patterns. There has been

[4] The advanced audit decision support systems used at AT&T (Vasarhelyi and Halper 1991) provided views of different variables if chosen by the auditor including: number of calls, minutes, dollars and modified dollars).

extensive debate in the profession over how and to what extent to aggregate transactional data, and whether to use ledger accounts as a means of summarizing data. In a disaggregated data environment and with the technical ability to process such large data sets, the degree and nature of aggregation is now a choice that is open to auditors to make, rather than one forced on them by technological constraints.

The main statistical argument for aggregation is that it can reduce the variability observed among individual transactions. For example, the transaction amount can differ greatly among individual transactions, as well as the lag time between order and delivery, and delivery and payment. By aggregating the individual transactions, this variance can be significantly reduced, thus allowing more effective detection of material anomalies. Thus aggregation facilitates the construction of more stable models than otherwise would be feasible to derive based on data sets with large variances. Unstable models would either trigger too many alarms or lack the detection power. On the other hand, if individual transactions are aggregated over a longer time period such as a week or a month, then the model would fail to detect many abnormal transactions because the abnormality would be mostly smoothed out by the longer time interval. Thus, the tradeoff is that the more aggregated the metrics are, the more stable the analytical relationships are likely to be at a price of more missed detection. In the meantime, any anomaly involving a metric with higher level of aggregation, requires a more extensive (and expensive) investigation of the larger subpopulation of transactions if an alarm is triggered. Daily and weekly aggregations used in this analysis are natural units of time that should result in a reasonable trade-off between these two forces. Aggregation can be performed on other dimensions besides the time interval, and the choice of the aggregation levels has to be made on a case by case basis considering the inherent characteristics of the underlying transactional data.

This study uses intermediate aggregates of transactions, such as aggregates of transactions of different units in the enterprise, aggregates of transactions with certain groups of customers or vendors. Traditional substantive testing is done either at the most disaggregated level, or at the most aggregated level. Substantive tests of details of transactions are done at the most disaggregated level of individual transactional data, but this is done in order to verify the correctness of that individual transaction rather than to gain a perspective of the overall business process. Tests of details of account balances are applied at the most aggregated level. All standard APs are used for analyzing the account balances or the largest classes of transactions. As the results show, analysis of intermediate aggregates can provide more confidence when making audit judgments about

anomalies and give the auditor a means of thinking about the underlying business process as a whole.

As discussed above, dollar amounts of purchase orders and vouchers are selected as the primary metric for testing. We also use the shipment quantity aggregates in the sample since the dollar value isn't available for shipment. After excluding weekends and holidays and several observations at the beginning of the sample period to reduce noise in the sample, there are 180 days of observations in the data sets for each business process. Summary statistics of the data used in the analysis are presented in Table 1.

[Insert Table 1 here]

Online Model Learning Protocol

One distinctive feature of analytical modeling in CA is the automatic model selection and updating capability that has the potential to assimilate the new information contained in every segment of the data flows and adapt itself constantly.

The online model learning protocol utilized in this study is shown in Figure 2. Each newly updated analytical model is used to generate a prediction only for one new segment of data. In the first step shown in Figure 2, data segments from 1 through 100 are used to estimate the model and to predict the new segment 101. After that, the new segment (101) is used together with the previous dataset to infer the next model. If the size of the previous dataset is small, the new segment is simply added without removing the oldest one from the dataset, as shown in Figure 2. After the dataset used for model inference becomes sufficiently large, it may be preferable to use the "sliding window" approach, in which the size of the training dataset is kept constant, and the addition of a new data segment is combined with removing the oldest one. This model updating procedure is expected to improve prediction accuracy and anomaly detection capability if the business process is changing. The sliding window (also known as moving window or rolling window) approach is commonly used in forecasting when underlying processes are expected to experience structural changes. The choice of optimal size of this window is usually problem dependent, and various approaches have been studied in the literature (see e.g., Pesaran and Timmermann 2007). Since our dataset is limited to only nine months of observations, we keep increasing the size of the training part one segment at a time and never discard the oldest segments. Thus, the approach implemented here should be properly termed "growing window".

[Insert Figure 2: Model Updating Protocol here]

Candidate Models for Continuity Equations

Simultaneous Equation Model

In SEM we specify the daily aggregate of order amount as the exogenous variable while the daily aggregates of receiving quantity and payment amount are endogenous variables. Time stamps are added to the transaction flow among the three business processes. The transaction flow originates from the ordering process at time t. After a lag period Δ_1, the transaction flow appears in the receiving process at time t+ Δ_1. After another lag period Δ_2, the transaction flow re-appears in the voucher payment processes at time t+Δ_2. One can utilize various lag sample statistics such as the mode, the median, and the mean, as lag estimates. Our results indicate that the mode estimate works best for the simultaneous equation model. In our sample the mode of the lags equals one day.

We divide our data set into two groups. The first group consisting of the first 100 days is categorized as the training set and used to estimate the model. The second group consisting of the remaining days is categorized as the hold-out set and used to test our model. The simultaneous equation model estimated on the training set is as follows:

$$\begin{cases} receive_t = 0.005 * order_{t-1} + e_1 \\ voucher_t = 87.147 * receive_{t-1} + e_2 \end{cases}$$

where
order = daily aggregate of dollar amounts for the purchase order process
receive = daily aggregate of transaction quantity for the receiving process
voucher = daily aggregate of dollar amounts for the voucher payment process
t = transaction time

The R^2 for the equations are 0.74 and 0.71 respectively, which indicate a good fit of data for the simultaneous equation model. It is important to point out some limitations associated with SEM. First, the lags have to be separately estimated and such estimations are not only time-consuming but also prone to errors. Second, the SEM is a simplistic model. Each variable can only depend on a single lagged value of the other variable. For example, *voucher* can only depend on $receive_{t-1}$ even though there is a strong likelihood that it can also depend on other lagged values of the *receive* variable. Due to these limitations, there is a need to develop a more flexible CE model.

Vector Autoregressive Model

Unlike in the case of SEM, in the case of VAR, no lag estimation is necessary, and every variable can depend on multiple lagged values of the variables in the model. Only the maximum lag period needs to be specified. All possible lags within the period can be tested by the model. We select 13 days as the maximum lag because 90% of the lags of all the individual transactions fall within this time frame.

Again we split our data set into two subsets: the training set and the hold-out set. SAS VARMAX procedure is used to estimate the large VAR model. Despite the fact that this model is a good fit, the predictions it generates for the hold-out sample have large variances.[5] In addition, a large number of the parameter estimates are not statistically significant. We believe the model suffers from the over-fitting problem. Therefore, we apply a step-wise procedure shown in Figure 3 to restrict the insignificant parameter values to zero and retain only the significant parameters in the model. First, we determine a p-value threshold for all the parameter estimates.[6] Then, in each step, we only retain the parameter estimates under the pre-determined threshold and restrict those over the threshold to zero, and re-estimate the model. If new insignificant parameters appear, we restrict them to zero and re-estimate the model. We repeat the step-wise procedure several times until all the parameter estimates are below the threshold, resulting in a Subset VAR model.

[Insert Figure 3: Multivariate Time Series Model Selection here]

The step-wise procedure ensures that all the parameters are statistically significant and the over-fitting problem is largely eliminated. One of the estimated Subset VAR models is expressed as:

$$order_t = 0.24*order_{t-4} + 0.25*order_{t-14} + 1.364*receive_{t-6} + e_o$$

$$receive_t = 0.27*order_{t-3} + 0.49*order_{t-5} + 0.49*voucher_{t-8} + e_r$$

$$voucher_t = 0.539*receive_{t-1} + 0.174*order_{t-9} + e_v$$

Thus, the over-parameterization problem can be resolved by step-wise procedures to transform the general form VAR into Subset VAR. However, it requires auditors' time and judgment to reduce the general form VAR model into Subset VAR model, which is antithetical to the automated nature of the CA system.

[5] The MAPEs for predictions of Order, Receive, and Voucher variables are all over 54%, much greater than the MAPEs of SEM, LRM and the subset VAR model. Refer to Section V for MAPE definition.

[6] If the $p=15\%$ threshold is used , the resulting VAR models have the overall best prediction accuracy.

Recent developments in Bayesian statistics, however, allow the model itself to control parameter restrictions. The BVAR model includes prior probability distribution functions to impose restrictions on the parameter estimates, with the covariance of the prior distributions controlled by "hyper-parameters". In other words, the values of hyper-parameters in the BVAR model control how far the model coefficients can deviate from their prior means and how much the model can approach an unrestricted VAR model (Doan et al. 1984, Felix and Nunes 2003). The BVAR model can relieve auditors from the burden of parameters restriction to derive the Subset VAR model. Thus, in this study we utilize both the BVAR and Subset VAR variants of the VAR model.

Linear Regression Model

In LRM we specify the lagged values of daily aggregates of transaction amounts in the order process and the transaction quantity in receive process as two independent variables respectively, and the voucher payment amount aggregate as the dependent variable. Again, the mode values of lags in individual transactions are used as estimates for the lags in the model (i.e. 2 day lag between the ordering and voucher payment processes, and 1 day lag between the receiving and voucher payment processes). No intercept is used in the model because all the voucher payments are processed for delivered orders.

Again the first 100 days of our data are set as the training subset to estimate the model. The estimated linear regression model is:

$$voucher_t = 0.179 * order_{t-2} + 59.69 * receive_{t-1} + e_t$$

The estimate of the coefficient of the order variable is statistically insignificant ($p>0.0441$) while the coefficient of the receive variable is significant at 99% level ($p<0.0001$).

ANOMALY DETECTION COMPARISON ACROSS CE MODELS

Having developed candidate expectation models for the procurement BP we then determine how well each fits the data and detects seeded errors in order to choose the optimal CE specification. A detected anomaly can only indicate the presence of a problem, and cannot pinpoint the problem itself, while a failed test of detail (for example, a negative confirmation or a reconciliation failure) does, but only if the auditor knows which data items to test. BPs can break down for a variety of reasons, some "real", meaning at the business process level itself, and some "nominal", meaning that even if the integrity of the underlying business process is not compromised, the CE may fail to represent that.

An example of a "nominal" violation would be a slowdown in the delivery of shipments due to changes in macroeconomic factors, which results in a broken CE model due to a shift in the value of the time lag. This is not indicative of a faulty BP, but an inevitable outcome of trying to fit the changing reality into a benchmark constructed using obsolete data. Thus, the auditor's investigation is bound to identify this situation as a false positive, unless the CE model is able to adapt accordingly.

The CE model is expected to signal the presence of anomalies in cases where the underlying BP is compromised, as for example when a strike affects a supplier or when a raw material becomes scarce. The purpose of using CE-based AP tests is to detect these process problems and then to generate a signal for the auditor to investigate the reasons for these anomalies through a targeted investigation of details in as real time as possible. This shows the advantage of using the most disaggregated metrics possible to narrow down the scope of the auditor's investigation as much as possible. On the other hand, the more disaggregated the metrics are the less stable will be the CE relationship. This is the tradeoff between the level of disaggregation of the metrics and the stability of the CEs, as the results of this study demonstrate. It is likely that the stability of relationships will vary widely between companies, processes, products, and times of the year.

Prediction Accuracy of CE Candidate Models

It is desirable for expectation models to make forecasts as close to actual values as possible. Many prior AP studies evaluate expectation models in terms of prediction accuracy (Kinney 1978, Wild 1987, Dzeng 1994, Allen et al. 1999, Chen and Leitch 1998, Leitch and Chen 2003). Following this line of research, we compared the prediction accuracies for the CE models in this study using a measure of prediction accuracy called Mean Absolute Percentage Error (MAPE). Additionally, we compared the CE models on their error detection ability.

MAPE is a commonly used metric of prediction accuracy. It is expected that a good model should have a small MAPE. The training set is first used to estimate each of the models. Then, each estimated model is used to make one-step-ahead forecasts and the forecast variance is calculated. After that, the model is updated based on the new data feeds in the hold-out set and the previous steps are repeated. Finally, all the absolute variances are summed up and divided by the total number of observations in the hold-out sample to compute the MAPE. The results for MAPE of Voucher predictions are presented in Table 2.

[Insert Table 2 here]

The results indicate that as measured by the MAPE metric the prediction accuracies of these models are close. The LRM has the best prediction accuracy (MAPE=0.3877), followed by the Subset VAR model (MAPE=0.3919), though the standard deviation for the LRM is slightly higher than that of the Subset VAR. BVAR model's prediction accuracy is 0.4208. The SEM has the lowest prediction accuracy of 0.4352. These prediction accuracies indicate that the forecasts generated by the expectation models usually differ from the reported amounts by approximately 40%.

There are no universal criteria to determine whether these prediction accuracies are good or not because MAPE values are data dependent. Prior studies (Kinney 1978, Wild 1987, Chen and Leitch 1998) on expectation models indicate that large variances exist in prediction accuracies when different data sets are used. The MAPE values reported in Wild's (1987) study range from 0.012 for Cost of Goods Sold prediction to 7.6 for Cash and Security prediction using the same expectation model. Our conclusion is that by the MAPE metric, all candidate CE models show promise as benchmarks for AP tests.

It is conceivable that if a longer time series were available then the use of the larger window for estimating CE models would result in lower MAPE values if the underlying BPs had low variability. In the case of higher variability BPs, one can attempt to achieve lower MAPE by increasing the level of aggregation of BP metrics. However, while it is desirable to achieve lower MAPE values, the ultimate measure of goodness of CE models for analytical monitoring is their BP problem detection ability.

BP Problem Detection Ability of CE Candidate Models

The rationale for constructing a CE-based AP test is to enable the effective detection of BP problems. In the context of this study, the detection capability of the CE models is measured using two metrics: the number of "false positive errors" and the number of "false negative errors".[7] A false positive error, also called a false alarm or a Type I error, is an anomaly detected by the model that is just a statistical fluctuation and has no underlying BP problem. A false negative error, also called a type II error, is a BP problem not detected by the model as an anomaly. While a false positive error can waste auditor's time and thereby increase audit cost, a false negative error is usually more detrimental because of the material uncertainty associated with the undetected

[7] For the presentation purposes, we also include the tables and charts showing the detection rate, which equals 1 minus the false negative error rate.

anomaly. An effective AP model should keep both the number of false positive errors and the number of false negative errors at a low level.

To compare the anomaly detection capabilities of the CE models under different settings a simulation approach known as "seeding errors" is used that provides the benefits of a controlled experiment. Since anomaly detection is designed to capture problems that are not detectable using transaction verification, in the seeded error simulation the effect of such unknown problems is modeled by modifying some randomly chosen actual values of utilized BP metrics by certain quantities. An AP model that has perfect BP problem detection abilities should identify all the modified values (i.e. those into which the errors were seeded) as anomalous, and should not identify any of the non-modified values as anomalous. This seeded error simulation makes it possible to measure the BP problem detection capabilities of an imperfect AP model by calculating the number of false positive and false negative errors that the model generates.

We test how the error magnitude can affect each AP model's anomaly detection capability with five different magnitudes used in every round of error seeding: 0.1%, 0.5%, 1%, 2% and 4% of the total voucher balance. The entire error seeding procedure is repeated ten times to reduce selection bias and ensure randomness. More specifically, for each round of error seeding in the daily aggregates, we randomly select eight days in the hold-out voucher amount sample to seed errors of a particular size. The original amounts in each of these eight days are then replaced with the seeded error modified amounts. For example, if a 2% error is seeded, the date 168 is selected, and $100,000 payments are made on that day, then $100,000 is replaced with $102,000. We expect a good detection model to identify this seeded error and red-flag the date 168. The same approach is used for the weekly aggregates.

Prior AP studies discuss several investigation rules to identify an anomaly (Stringer 1975, Kinney and Salaman 1982, Kinney 1987, Knechel 1986). A modified version of the statistical rule (Kinney 1987) is used in this study. Prediction intervals (PI), equivalent to a confidence interval for an individual dependent variable, are used as the acceptable thresholds of variance. If the value of the prediction exceeds either the upper or lower limits of the PI, then the observation is flagged as an anomaly.

The selection of the prediction interval is a critical issue impacting the effectiveness of the AP test. The size of the prediction interval is determined by the value of the significance level α. Choosing a low α value (e.g. 0.01), leads to wide tolerance boundaries (i.e. large prediction interval)

and a resulting low detection rate. On the other hand, if a high α value is selected, then the prediction interval will be overly narrow and many normal observations will be flagged as anomalies. To solve this problem, two approaches to select the prediction interval percentages were followed. In the first approach, α values are selected to control the number of false positive errors in various models. More specifically, an α value is selected which is just large enough to yield two false positive errors in the training data set. In the second approach, which is the traditional one, predetermined α values, 0.05 and 0.1, are used for all the expectation models. The α value of 0.05 is used in the tabulated anomaly detection results to compare all of the CE models on a level ground.

Before this methodology can be used to compare the CE models, another critical issue needs to be addressed that only arises in a CA setting: real time error correction.

Pseudo-Real Time Error Correction

An important distinction between CA techniques and standard auditing that was explored in this project is what we call "Pseudo-Real Time Error Correction". In a real world CA environment when an anomaly is detected, the auditor will be notified immediately and a detailed investigation, if relevant, can be initiated. In theory, the auditor can then have the ability to complete the investigation and (if necessary) to correct the error before the next round of audit starts. The duration of the audit round will determine the precise meaning of "pseudo-real time". This duration will depend on the natural rhythm of the business process and the utilized CA methodology. In more advanced internal audit departments, such as the one in a large South American bank (Aquino et al. 2013), the monitoring and resolution of identified anomalies is accomplished on the daily basis. For example, if the amount of returned checks or payment cancelations in a branch exceeds a predetermined value, an investigation is initiated to reveal by the next day if this occurrence is due to any problem, or just an unusual fluctuation. In such cases "pseudo-real time" refers to 24 hours. In other circumstances "pseudo-real time" can be as long as a week, if it turns out that the weekly BP metrics are most suitable for analytical monitoring.

Whether this technical possibility can or will be carried out in practice depends both upon the speed at which error correction can be made and the more serious issue of the potential threat to auditor independence of using data in subsequent tests that the auditor has had a role in correcting. These issues clearly require detailed consideration, but doing so is beyond the scope of the current study. What we focus on here is quantifying the benefits of pseudo-real time error correction in a CA environment, i.e., the technical implication for AP in CA if errors are indeed

detected and corrected in pseudo-real time. Specifically, in the error correction process, if a seeded error is detected, we substitute the seeded error with the true value (the original value in our data set). The new error free data is used to update the CE model and continue the anomaly detection process. For comparison purposes, we test how our candidate CE models work with and without pseudo-real time error correction. Unlike in CA, anomalies are detected but usually problems are not corrected immediately in traditional auditing. To simulate this scenario, we don't correct any errors we seeded in the hold-out sample even if the AP model detects them.

One can argue that it is possible in principle to utilize error correction in traditional auditing. However, since traditional auditors will have to investigate all the anomalies at the end of the audit period, the time pressure of the audit is likely to prevent them from rerunning their analytical procedures and conducting additional investigations over and over again to detect additional anomalies after some previously detected problems are corrected.

[Insert Table 3 here]

The results in Table 3 show that, in terms of BP problem detection, all of the four CE models have excellent detection capability when the error size equals to 2% or 4% of voucher account balance. With the error correction protocol, the CE models' false negative error rates are close or equal to zero. When the error size is at 1% level, SEM continues to exhibit excellent detection capability with the false negative error rate at 0.025 followed by LRM at 0.2125. BVAR and Subset VAR have higher than 50% false negative error rates. If the error size drops to 0.1% and 0.5% level, all of the models have high false negative errors. Overall, the results show lower false negative error rates for all CE models with error correction, especially when the error magnitude is 1% or more. This finding is consistent with the prior expectation. As for the false positive rates, all of the CE models except SEM have close or equal to zero error rates. It is not surprising to see that SEM generates higher false positive error rates because it has relatively low false negative rates at 0.1% and 0.5%. BVAR, Subset VAR and LRM with error corrections have fewer false positive errors than their counterparts without error corrections. The non-correction SEM model has better false positive rate than the correction model. It means that the error-correction SEM models can detect more anomalies but at a cost of triggering more false alarms. However, a further investigation reveals that the false alarms are mostly caused by certain records in the holdout sample, which are called in this study "original anomalies". These original anomalies are very likely to be caused by measurement errors since our dataset consists of un-audited operational data. This measurement

error problem with non-audited data is also reported by previous studies (Kinney and Salamon 1982, Wheeler and Pany 1990). Because the non-correction model would not correct those undetected errors, the impact of original anomalies, whose values remain constant, would be eclipsed by the increase in seeded error magnitude. Therefore, the non-correction model would trigger fewer false alarms when the magnitude of seeded error increases. On the other hand, the impact of original anomalies would not decrease as the error-correction model would correct all detected errors.

In summary, the error-correction models have better BP problem detection performance than the non-correction models. Thus, the real time error correction protocol can improve the detection performance, and the enabling of this protocol is an important benefit of CA.[8]

Disaggregated versus Aggregated Data

This study examines if the use of disaggregated data can make CE models perform better. Data can be aggregated on different dimensions, and we compare the efficacy of CE based AP tests using temporal disaggregation.

In the temporal disaggregation analysis we examine the differential anomaly detection performances using weekly versus daily data through three different approaches. First, errors are seeded into the weekly data in the same fashion as in previous daily simulations. We only use the weekly aggregates assuming that daily data are not available. In our second and third approach, we follow prior studies (Kinney and Salamon 1982, Wheeler and Pany 1990) in the choice of methods to seed weekly errors into the daily sample. In the best case scenario, the entire weekly error is seeded into a randomly selected day of a week. In the worst case scenario, the weekly error is first divided by the number of working days in a week and then seeded into each working day of that week. In addition to the different aggregation levels, the error-correction and non-correction models are again compared to verify if the previous findings still hold. Due to the scope of this study we only use a single α value 0.05 in all models.

[Insert Table 4 here]

[8] The stochastic nature of both the underlying business process and error seeding imply that the improvement in performance is not a deterministic outcome. For example, if the value of the business metric of the day randomly chosen for error seeding happens to be abnormally low, then the increase resulting from adding the error will bring that daily value closer to the expected one, and the subsequent model update may benefit more from utilizing the uncorrected value rather than the corrected one. Overall, as the results show, the error correction does improve the performance, but this is true only in the probabilistic sense.

The results presented in Table 4 are generally consistent with our expectations. Weekly aggregation results in Table 4 indicate that the detection capability of all CE models suffers when using the weekly aggregates. Compared with daily aggregates, the false negative error rates for weekly aggregates are very high even when the error magnitude reaches 2% level. This shows the benefit of disaggregated modeling and supports our expectation that continuous auditing at the daily level can detect more anomalies than that at the weekly or monthly levels.

All the CE models perform the best using the best case scenario data. The false negative error rates are close to or equal to zero for all CE models at 2% and 4% error size. At 1% error magnitude, the SEM and LRM exhibit strong detection capability. However, all the models have the poorest anomaly detection performance using the worst case scenario data. This result is not surprising because the weekly error is spread evenly into each day making the seeded error act as a systematic error which is almost impossible to detect (Kinney 1978). Across the three types of sample tests (weekly, weekly best case and weekly worst case), we find that error correction models are still better in BP problem detection than the non-correction models.

[Insert Table 5 here]

Table 5 presents false positive errors for the tests using weekly, weekly best case and weekly worst case data. We find that weekly aggregates trigger more false alarms than daily aggregates for all CE models except BVAR. The weekly best case data gives the lowest false positive error rates for all CE models except SEM. The test results from the worst case scenario suggest that very few false alarms are triggered but it is achieved at the cost of low detection rates.[9]

CONCLUSION

The purpose of this paper is to demonstrate how audit practice may change when auditors have access to real time population data, how to use real world data to develop APs for CA, and compare different analytical procedures in a CA context.

Our research shows that there can be significant changes in the role and sequence of audit procedures. When data access is not a constraint, tests of detail can be carried out first on the complete population data to find exceptions to controls and for transaction verification. Then APs can be used, again, on the complete population data, to find anomalies. The use of APs in this way

[9] We conducted a large number of robustness tests including the use of non-financial metrics (transaction item quantity), and aggregating data on the geographic dimension. Results are similar to those reported in the paper. Full details are available from the authors.

can result in more effective audits, but that depends on the construction of empirical models of BPs so that normality can be defined from which anomalies are detected. A variety of parameters shape the development of these CE models, including the choice of metrics, the degree of aggregation, and how the models are updated dynamically.

This study shows that while there are differences in the predictive ability and detection performance of various CE models, all models perform reasonably well and no single model performs better on all aspects. From this two important conclusions can be drawn:

First, the choice of a particular model across the candidate CE models is less important than the fact that all models yield fairly effective AP tests. Because of its automated nature, it is quite feasible and even desirable for the continuous data level audit system to use benchmarks based on multiple CE models instead of being forced to select only one, as would be necessary in a more manual system. For example, the SEM can be used first to detect anomalies because it has a low false negative error rate. Subsequently, the BVAR and the LRM can be used to remove the false alarms from the SEM-detected anomalies because these two models have relatively low false positive error rates.

Our second conclusion from the fact that all the CE models yield reasonably effective analytical procedures is that when auditors have access to complete transaction data, the richness of that disaggregate data combined with the reorganization of auditing workflow to implement pseudo-real time error correction makes BP problem detection robust across a variety of expectation models. In other words, it is the nature of the data that serves as audit evidence that is the primary driver of audit effectiveness, with the selection of the specific AP a second order concern—not because the audit benchmark is not important, but because auditing at the process level makes anomalies stand out much more obviously in the data.

REFERENCES

Allen R.D., M.S. Beasley, and B.C. Branson. 1999. Improving Analytical Procedures: A Case of Using Disaggregate Multilocation Data, *Auditing: A Journal of Practice and Theory* 18 (Fall): 128-142.

Alles M.G., G. Brennan A. Kogan, and M.A. Vasarhelyi. 2006. Continuous Monitoring of Business Process Controls: A Pilot Implementation of a Continuous Auditing System at Siemens. *International Journal of Accounting Information Systems*, Volume 7, Issue 2 (June): 137-161.

Alles M.G., A. Kogan, and M.A Vasarhelyi. 2008. Putting Continuous Auditing Theory Into Practice: Lessons From Two Pilot Implementations. *Journal of Information Systems*, Vol. 22, No. 2, pp. 195-214.

_____. 2002. Feasibility and Economics of Continuous Assurance. *Auditing: A Journal of Practice and Theory* 21 (Spring):125-138.

_____.2004. Restoring auditor credibility: tertiary monitoring and logging of continuous assurance systems. *International Journal of Accounting Information Systems* 5: 183-202.

Alles, M.G., A. Kogan, M.A. Vasarhelyi, and J.D. Warren. 2007. *Continuous Auditing, A portfolio assembled for the Bureau of National Affairs*. Washington, DC.

_____ and J. Wu. 2004. Continuity Equations: Business Process Based Audit Benchmarks in Continuous Auditing. *Proceedings of American Accounting Association Annual Conference.* Orlando, FL.

American Institute of Certified Public Accountants. 1988. Statement on Auditing Standards No. 56: Analytical Procedures. New York.

Auditing Concepts Committee of the American Accounting Association. 1972. Report of the committee on basic auditing concepts.

Aquino C. E. de, E. Miyaki, N. Sigolo, and M.A. Vasarhelyi. 2013. A balancing act. Internal Auditor (April): 51-55.

Bell T., F.O. Marrs, I. Solomon, and H. Thomas 1997. *Monograph: Auditing Organizations Through a Strategic-Systems Lens*. Montvale, NJ, KPMG Peat Marwick.

Biggs, S., T. Mock, and P. Watkins. 1988. Auditors' use of analytical review in audit program design. *The Accounting Review* 63 (January): 148-161

Brown, Carol E. Jeffrey A. Wong, and Amelia A. Baldwin. 2006. Research Streams in Continuous Audit: A Review and Analysis of the Existing Literature. *Collected Papers of the Fifteenth Annual Research Workshop on: Artificial Intelligence and Emerging Technologies in Accounting, Auditing and Tax*. pp. 123-135. Washington, DC, USA, August 5, 2006.

Chen Y. and Leitch R.A. 1998. The Error Detection of Structural Analytical Procedures: A Simulation Study. *Auditing: A Journal of Practice and Theory* 17 (Fall): 36-70.

_____. 1999. An Analysis of the Relative Power Characteristics of Analytical Procedures. *Auditing: A Journal of Practice and Theory* 18 (Fall): 35-69.

CICA/AICPA. 1999. *Continuous Auditing*. Research Report, Toronto, Canada: The Canadian Institute of Chartered Accountants.

Davenport, T.H. and J.E. Short, 1990. The New Industrial Engineering: Information Technology and Business Process Redesign, *Sloan Management Review*, pp. 11-27, Summer.

Doan, T.A., Littleman R.B., and C.A. Sims. 1984. Forecasting and Conditional Projections Using Realistic Prior Distributions. *Econometric Reviews* 1, 1-100.

Dzeng S.C. 1994. A Comparison of Analytical Procedures Expectation Models Using Both Aggregate and Disaggregate Data. *Auditing: A Journal of Practice and Theory* 13 (Fall), 1-24.

Elliot, R.K. 2002. Twenty-First Century Assurance. *Auditing: A Journal of Practice and Theory* 21 (Spring), 129-146.

Enders, W. 2004. *Applied Econometric Time Series. Second Edition.* New York, NY: John Wiley and Sons.

Engle, R. 2001. GARCH 101: The Use of ARCH/GARCH model in Applied Econometrics. *Journal of Economic Perspectives* 15 (Fall): 157-168.

Felix, R.M. and L.C. Nunes. 2003. Forecasting Euro Area Aggregates with Bayesian VAR and VECM Models. Working Paper. Banco De Portugal. Economic Research Department.

Groomer, S.M. and U.S. Murthy. 1989. Continuous auditing of database applications: An embedded audit module approach. *Journal of Information Systems 3* (2), 53-69.

Hammer, M. 1990. Reengineering Work: Don't Automate, Obliterate! *Harvard Business Review*, Vol. 68, Issue 4 (July-August), 104-112.

Hentschel, L. 1995. All in the family Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics* 39 (1): 71-104.

Heston, S.L. and S. Nandi. 2000. A closed-form GARCH option valuation model. *Review of Financial Studies* 13: 585-625.

Hirst, E. and L. Koonce. 1996. Audit Analytical Procedures: A Field Investigation. *Contemporary Accounting Research.*, Vol 13, No 2, Fall.

Hoitash, R. A. Kogan, and M.A. Vasarhelyi. 2006. Peer-Based Approach for Analytical Procedures. *Auditing: A Journal of Practice and Theory* Vol. 25, No.2 (November), 53–84.

Hylas, R. and R. Ashton. 1982. Audit detection of financial statement errors, *The Accounting Review*, Vol. 57 No.4, 751-65.

Kinney, W.R. 1978. ARIMA and Regression in Analytical Review: an Empirical Test. *The Accounting Review*, Vol. 17, No. 1, 148-165.

Kinney, W.R. and G.L. Salamon. 1982. Regression Analysis in Auditing: A Comparison of Alternative Investigation Rules. *Journal of Accounting Research*. Vol. 20. No. 2, 350-366.

Kinney, W.R. 1987. Attention-Directing Analytical Review Using Accounting Ratios: A Case Study. *Auditing: A Journal of Practice and Theory* Vol. 6, No.2 (Spring), 59-73.

Knechel, W. R. 1985 (a). An analysis of alternative error assumptions in modeling the reliability of accounting systems. *Journal of Accounting Research* Vol. 23, No. 1, 194-212.

Knechel, W. R. 1985 (b). A Stochastic Model of the Error Generation Process in Accounting Systems. *Accounting & Business Research* Vol. 15, Issue 59 (Summer), 211-221.

Knechel, W. R. 1986. Applications and implementation: a simulation study of the relative effectiveness of alternative analytical review procedures. *Decision Sciences* Vol. 17, No.3, 376-394.

Knechel, W. R. 1988. The effectiveness of statistical analytical review as a substantive auditing procedure: A simulation analysis. *The Accounting Review* (January), 74-95.

Kogan, A. E.F. Sudit, and M.A. Vasarhelyi. 1999. Continuous Online Auditing: A Program of Research. *Journal of Information Systems* 13 (Fall), 87–103.

Koreisha, S. and Y. Fang. 2004. Updating ARMA Predictions for Temporal Aggregates. *Journal of Forecasting* 23, 275-396.

Lamoureux, C.G. and W. D. Lastrapes. 1990. Heteroskedasticity in Stock Return Data: Volume versus GARCH Effects. *The Journal of Finance* 45 (1), 221-229.

Leitch and Y. Chen. 2003. The Effectiveness of Expectation Models In Recognizing Error Patterns and Eliminating Hypotheses While Conducting Analytical Procedures. *Auditing: A Journal of Practice and Theory* 22 (Fall), 147-206.

Loebbecke, J. and P., Steinbart. 1987. An investigation of the use of preliminary analytical review to provide substantive audit evidence, *Auditing: A Journal of Practice and Theory*, Vol. 6 No.2, 74-89.

Lorek, K.S., B.C. Branson, and R.C. Icerman. 1992. On the use of time-series models as analytical procedures. *Auditing: A Journal of Practice & Theory*, Vol. 11, No. 2 (Fall), 66-88.

Murthy, U.S. 2004. An Analysis of the Effects of Continuous Monitoring Controls on e-Commerce System Performance. *Journal of Information Systems*. 18 (Fall), 29–47.

_____and M.S. Groomer. 2004. A continuous auditing web services model for XML-based accounting systems. *International Journal of Accounting Information Systems* 5, 139-163.

Murthy, U.S. and J.A. Swanson. 1992. Integrating Expert Systems and Database Technologies: An Intelligent Decision Support System For Investigating Cost Variances. *Journal of Information Systems*, 6(2) (Fall), 18-40.

Public Company Accounting Oversight Board. 2010. Auditing Standard No. 15: Audit Evidence. PCAOB Release No. 2010-004. Washington, DC.

Pandher G.S. 2002. Forecasting Multivariate Time Series with Linear Restrictions Using Unconstrained Structural State-space Models. *Journal of Forecasting* 21: 281-300.

Pesaran, M. H. and A. Timmermann. 2007. Selection of estimation window in the presence of breaks. *Journal of Econometrics*, Vol. 137, Issue 1 (March), 134–161.

Porter, M. E. 1996. What Is Strategy? *Harvard Business Review*, Vol. 74, #6, S. 61-78.

PricewaterhouseCoopers. 2006. State of the internal audit profession study: Continuous auditing gains momentum.
http://www.pwcglobal.com/images/gx/eng/about/svcs/grms/06_IAState_Profession_Study.pdf

Rezaee, Z., A. Sharbatoghlie, R. Elam, and P.L. McMickle. 2002. Continuous Auditing: Building Automated Auditing Capability. *Auditing: A Journal of Practice and Theory* 21 (Spring), 147-163.

Searcy, D. L., Woodroof, J. B., and Behn, B. 2003. Continuous Audit: The Motivations, Benefits, Problems, and Challenges Identified by Partners of a Big 4 Accounting Firm. *Proceedings of the 36th Hawaii International Conference on System Sciences*: 1-10.

Stringer, K. and T. Stewart. 1986. *Statistical techniques for analytical review in auditing.* Wiley Publishing. New York.

Swanson, N., E. Ghysels, and M. Callan. 1999. A Multivariate Time Series Analysis of the Data Revision Process for Industrial Production and the Composite Leading Indicator. Book chapter of *Cointegration, Causality, and Forecasting: Festchrift in Honour of Clive W.J. Granger.* Eds. R. Engle and H. White. Oxford: Oxford University Press.

Vasarhelyi, M.A and F.B. Halper. 1991. The Continuous Audit of Online Systems. *Auditing: A Journal of Practice and Theory* 10 (Spring), 110–125.

_____ 2002. Concepts in Continuous Assurance. Chapter 5 in *Researching Accounting as an Information Systems Discipline*, Edited by S. Sutton and V. Arnold. Sarasota, FL: AAA.

_____, M. Alles, and A. Kogan. 2004. Principles of Analytic Monitoring for Continuous Assurance. *Journal of Emerging Technologies in Accounting*, Vol. 1, 1-21.

_____, and M. Greenstein. 2003. Underlying principles of the electronization of business: A research agenda. *International Journal of Accounting Information Systems* 4: 1-25.

Vasarhelyi, Miklos A., Alles, M.G., Williams, K.T. Continuous Assurance for the Now Economy, A Thought Leadership Paper for the Institute of Chartered Accountants in Australia, forthcoming May 2010.

Wright, A. and R.H. Ashton. 1989. Identifying audit adjustments with attention-directing procedures. The Accounting Review (October), 710-28.

Woodroof, J. and D. Searcy 2001. Continuous Audit Implications of Internet Technology: Triggering Agents over the Web in the Domain of Debt Covenant Compliance. *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Wu, J. Continuous Test of Details and Continuity Equations in Continuous Audit, Ph.D Dissertation, Rutgers University, 2006.

Figure 1: Architecture of Continuous Data Level Auditing System

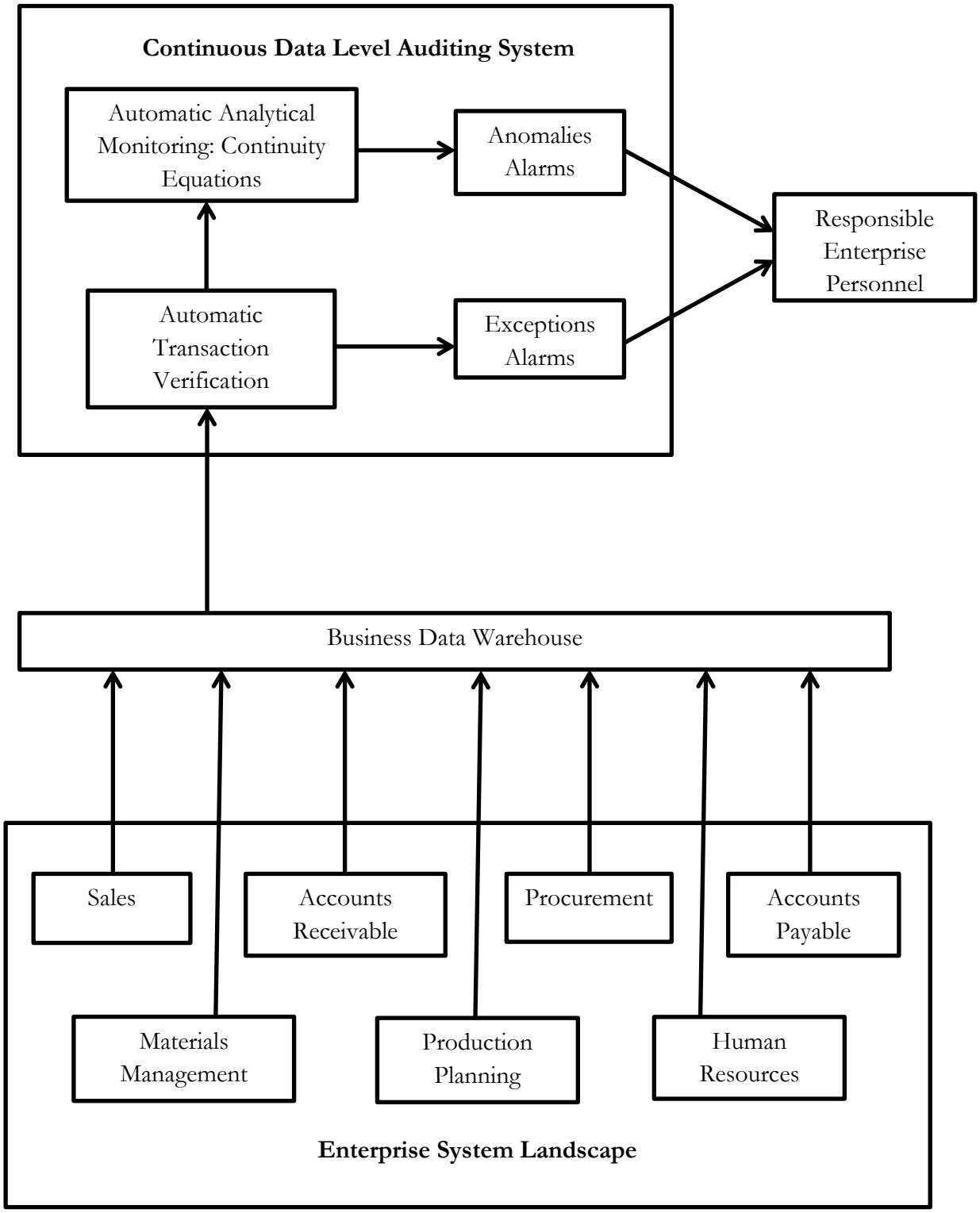

Figure 2: Model Updating Protocol

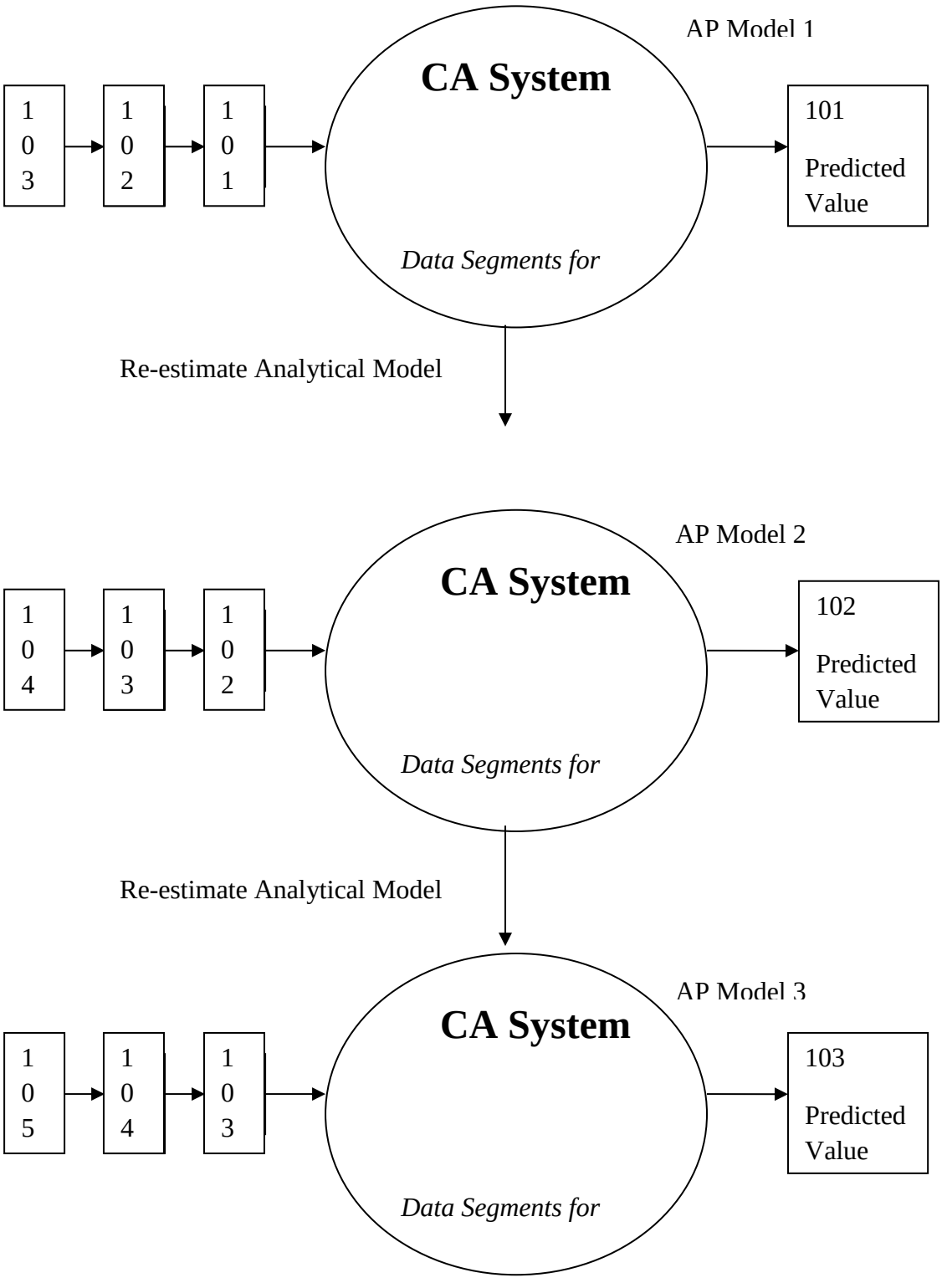

Figure 3: Multivariate Time Series Model Selection

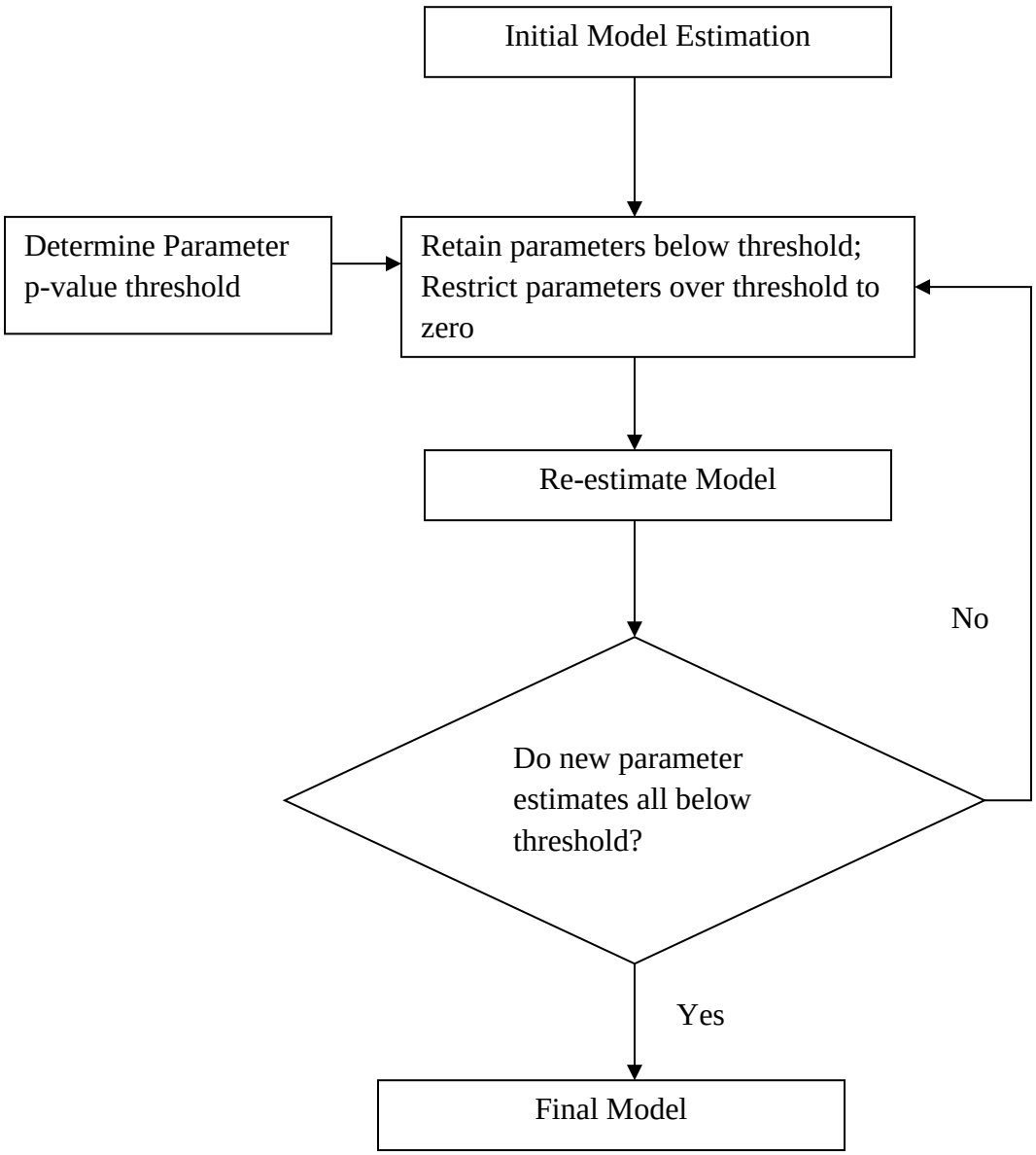

Table 1: Summary Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
Order	180	1076911.56	605393.82	342202.4	4549114.60
Receive	180	8876.26	3195.16	4430	31412.00
Voucher	180	869595.80	591950.79	12107.96	7026331.26

The table presents the summary statistics for the dollar amount daily aggregates for each Order and Voucher process and transaction quantity for the Receive process. The data sets span from 10/01/03 to 06/30/04. Many related transactions for the Receive and Voucher for the first 2 days in the data sets may have happened before 10/01/03.

Table 2: MAPE Comparison among BVAR, Subset VAR, SEM and Linear Regression Model: Prediction Accuracy for Daily Dollar Amounts Aggregate of Entire Company's Vouchers

	MAPE	**Std. Dev**	**Min**	**Max**
BVAR	0.4208	0.4334	0.0119	2.6829
Subset VAR	0.3919	0.4271	0.0061	3.6514
SEM	0.4352	0.6135	0.0062	5.1651
LRM	0.3877	0.4594	0.0044	3.8854

MAPE: Mean Absolute Percentage Error
Std. Dev: Standard deviation for Absolute Percentage Error
Min: The minimum value for Absolute Percentage Error
Max: The maximum value for Absolute Percentage Error

Table 3: Error Rates Comparison between Error-correction and Non-correction Using Dollar Amounts, α=0.05, 0.1%, 0.5%, 1%, 2%, 4% of account bal.

	False Positive			**False Negative**		
	Error Magnitude	Error Correction	Non-Correction	Error Magnitude	Error-Correction	Non-Correction
Bayesian	0.1%	0.0139	0.0139	0.1%	1	1
	0.5%	0.0153	0.0153	0.5%	0.975	0.975
VAR	1%	0.0153	0.0153	1%	0.5625	0.6375
	2%	0.0153	0.0153	2%	0.0125	0.0375
	4%	0.0153	0.0167	4%	0.0125	0.0125
Subset	0.1%	0.0139	0.0139	0.1%	1	1
	0.5%	0.0139	0.0139	0.5%	0.9375	0.9375
VAR	1%	0.0139	0.0139	1%	0.5375	0.5375
	2%	0.0139	0.0139	2%	0	0.0125
	4%	0.0139	0.0139	4%	0	0
SEM	0.1%	0.09	0.09	0.1%	0.8625	0.8625
	0.5%	0.0843	0.06	0.5%	0.375	0.4125
	1%	0.09	0.0471	1%	0.025	0.025
	2%	0.0857	0.0329	2%	0.0125	0.0125
	4%	0.0829	0.0057	4%	0.0125	0.0125
LRM	0.1%	0	0	0.1%	0.975	0.975
	0.5%	0	0	0.5%	0.8875	0.8875
	1%	0	0	1%	0.2125	0.3125
	2%	0	0	2%	0	0
	4%	0	0	4%	0	0

Table 4: False Negative Error Rates Comparisons (Weekly, Weekly Best Case, Weekly Worst Case) Dollar Amounts, α=0.05, 0.1%, error sizes=0.5%, 1%, 2%, 4% of account balance

	Error Magnitude	Weekly Aggregation		Weekly Best Case		Weekly Worst Case	
		Error-Correction	Non-Correction	Error-Correction	Non-Correction	Error-Correction	Non-Correction
BVAR	0.1%	1	1	0.975	0.975	0.99	0.99
	0.5%	1	1	0.975	0.975	0.99	0.99
	1%	1	1	0.575	0.6	0.99	0.99
	2%	1	1	0.025	0.025	0.99	0.99
	4%	0.625	0.775	0.025	0.025	0.98	0.99
Subset VAR	0.1%	0.9	0.9	0.975	0.975	0.99	0.99
	0.5%	0.9	0.9	0.925	0.925	0.99	0.99
	1%	0.9	0.9	0.425	0.525	0.99	0.99
	2%	0.775	0.775	0	0	0.98	0.985
	4%	0.075	0.4	0	0	0.865	0.88
SEM	0.1%	0.85	0.85	0.9	0.9	0.985	0.985
	0.5%	0.775	0.8	0.475	0.675	0.975	0.975
	1%	0.75	0.775	0.05	0.05	0.945	0.945
	2%	0.175	0.25	0.025	0.025	0.86	0.88
	4%	0	0.05	0.025	0.025	0.37	0.53
LRM	0.1%	0.95	0.95	1	1	0.99	0.99
	0.5%	0.95	0.95	0.95	0.95	0.985	0.99
	1%	0.925	0.95	0.2	0.275	0.985	0.985
	2%	0.675	0.75	0	0	0.95	0.95
	4%	0	0.275	0	0	0.57	0.705

Table 5: False Positive Error Rates Comparison (Weekly, Weekly Best Case, Weekly Worst Case) Dollar Amounts, α=0.05, 0.1%, error sizes=0.5%, 1%, 2%, 4% of account balance

	Error Magnitude	Weekly Aggregation		Weekly Best Case		Weekly Worst Case	
		Error-Correction	Non-Correction	Error-Correction	Non-Correction	Error-Correction	Non-Correction
BVAR	0.1%	0	0	0.0118	0.0118	0.0133	0.0133
	0.5%	0	0	0.0132	0.0132	0.0133	0.0133
	1%	0	0	0.0132	0.0132	0.0133	0.0133
	2%	0	0	0.0132	0.0132	0.0133	0.0133
	4%	0	0	0.0132	0.0132	0.0133	0.0133
Subset VAR	0.1%	0.1231	0.1231	0.0118	0.0118	0.0133	0.0133
	0.5%	0.1385	0.1385	0.0118	0.0118	0.0133	0.0133
	1%	0.1385	0.1462	0.0118	0.0118	0.0133	0.0133
	2%	0.1769	0.1692	0.0118	0.0118	0.0133	0.0133
	4%	0.1692	0.2461	0.0118	0.0118	0.0133	0.0133
SEM	0.1%	0.1077	0.1077	0.0919	0.0919	0.0283	0.0283
	0.5%	0.1077	0.0923	0.0892	0.0689	0.0383	0.0383
	1%	0.0923	0.0769	0.0919	0.0581	0.0367	0.0383
	2%	0.1077	0.0769	0.0878	0.0392	0.0383	0.0383
	4%	0.1077	0.0769	0.0851	0.0135	0.0383	0.0383
LRM	0.1%	0.0615	0.0615	0	0	0.0133	0.0133
	0.5%	0.0615	0.0615	0	0	0.0133	0.0133
	1%	0.0615	0.0615	0	0	0.0133	0.0133
	2%	0.0615	0.0615	0	0	0.0133	0.0133
	4%	0.0615	0.0615	0	0	0.0133	0.0133