Paper 1489-2017

# Timing is Everything: Detecting Important Behavior Triggers

Daymond Ling, Professor, School of Marketing, Faculty of Business, Seneca College

## ABSTRACT

Predictive analytics is powerful in its ability to predict likely future behavior and is used widely in marketing. However, the adage “timing is everything” continues to hold true—the right customer and the right offer at the wrong time is less than optimal. In life, timing matters a great deal, but predictive analytics seldom take timing into account explicitly, we should be able to do better. Financial service consumption changes often have precursor changes in behavior, and a behavior change can lead to multiple subsequent consumption changes. One way to improve our awareness of the customer situation is to detect significant events that have meaningful consequences that warrant proactive outreach. This paper presents the use of Change Point Analysis in event detection that has proven to work successfully. Real case studies are discussed to illustrate the approach and implementation. Adoption of this practice can augment and enhance predictive analytics practice to elevate our game to the next level.

## 1. INTRODUCTION

If you give people a sequence of numbers, for example, monthly spend on a credit card, and ask “did something change”, most people would look for periods where the spend appear to be different, and if the mean difference is notionally big, they would say yes. If you then ask what is the decision rule, they would say “difference in average spend is large”. When implemented as is, this common-sense decision rule does not take variability into account, hence it would generate false positives where the difference is just natural variability.

In contrast, statistical decision rules generally look at the mean difference in relation to variability, and if the change is out of the ordinary, it is declared as a statistically significant difference. This approach won’t generate false positives, but it may detect small changes that, while true, are uninteresting. In practice we combine both statistical significance (is it really there) and business interest (is it big enough to pay attention to) to identify relevant changes.

Change Point Analysis is a general method capable of statistically detecting whether a regularly spaced time series data has changed in some fashion, this paper discusses its use in detecting change in the mean of a time series. The paper is structured as follows. A short description of change point analysis of mean shift is given in section 2. Section 3 shows sample code and discusses implementation. Three financial services case studies are presented in Section 4. Relationship to predictive modelling is discussed in section 5.

## 2. CHANGE POINT ANALYSIS

Given a time series, change point analysis (CPA) is the problem of estimating the point at which its statistical properties change. It answers two basic questions:

1. Is there a significant change, and if so,
2. When did the change occur

If a change exists, the time series can be divided into two smaller timer series, each of them can then be examined further to detect more changes. By recursively checking the two smaller sections, a time series can be partitioned into a set of small sections until no more change can be detected.

While many statistical properties can be analyzed, this paper will focus on change in the mean of a time series.

Prior to CPA, time series data may need to be treated with other techniques such as seasonality removal or differencing.  Performing CPA on the first difference of a time series is equivalent to looking for slope change in the original series.

## 2.1  DID THE MEAN SHIFT?

Instead of the original series, we analyze the centered series where the overall mean has been removed. We then calculate the cumulative sum of the successive deviations, this is called the CUSUM, it ends on zero as the sum of all deviations of a centered series is zero.

If there is no structure to the time ordering of the data points, the deviations should be randomly ordered, the CUSUM would then not deviate far from zero as successive deviations cancel each other out. However, if there are systemic structures such as long sections where the deviations are all negative (below the overall mean) or positive (above the overall mean), then successive deviations would accumulate causing the CUSUM to move far from zero.  Therefore, the range of the CUSUM may reveal whether there is significant mean change or not.  The two graphs below illustrate the idea that large CUSUM range indicates a mean shift.

Figure 1 is the CUSUM of 15 normally distributed data with standard deviation of 1 in random order.  It shows that the deviations negate each other frequently, the CUSUM does not systemically build in any direction thus it does not move far from 0, it has a range of 1.8 from a min of -0.5 to a max of 1.3.  Figure 2 is the CUSUM of the same data sorted ascending.  As the graph shows, this sorted data arrangement has a CUCUM range of 3.2 which is larger than the range of 1.8 of the original series.

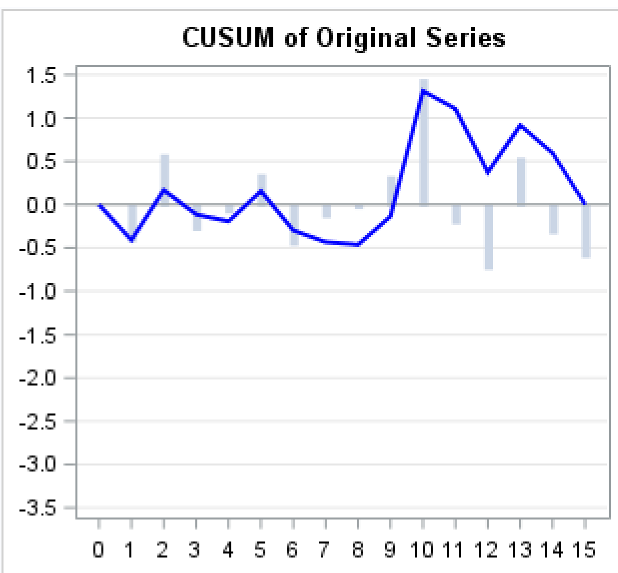



**Figure 1 CUSUM of Original Series**

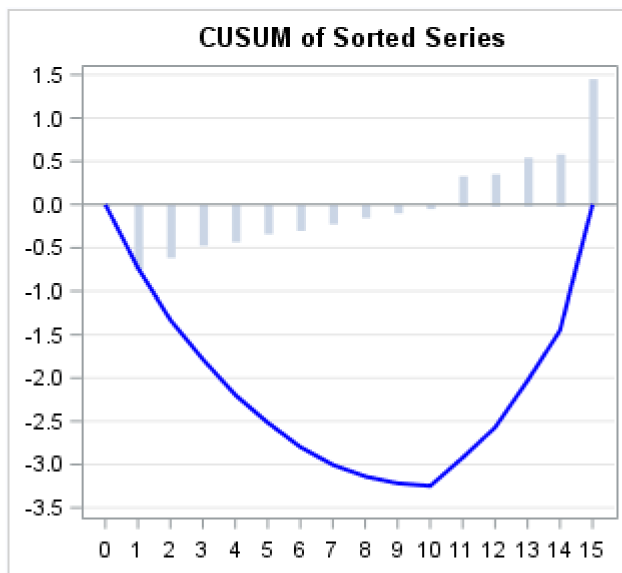



**Figure 2  CUSUM of Sorted Original Series**

Large CUSUM range provides indication of mean shift, but how large is large?  To answer this question, we use a randomization test to tell us the probability of the actual range.  The algorithm is as follows:

1. Calculate CUSUM range of the empirical time series (CSRE)
2. Shuffle the data randomly a large number of times.  Each time calculate CUSUM Range (CSR) of the shuffled data, count the number of times CSRE is greater than CSR, denote this as N.
3. Probability of significant change is N / number of shuffles

If probability is high, say 0.95 which corresponds to a p-value of 0.05, meaning it is difficult to obtain CSRE by shuffling the data alone, we conclude there is structure in the original series beyond data variability.

### 2.2 WHEN IS THE SHIFT?

Having determined there is a significant change in the mean of the time series, we would like to know when did it happen. A simple estimator of when the change occurred is the point that's furthest from zero in the CUSUM sequence, it marks the last point before the change, the next point is the first point after the change.

Another choice for when the change occurred is to pick the point that minimizes the variance after accounting for the mean of the two sections of the time series. The algorithm is:

1. Vary the cut-point from the first point to the second last point, each time:
    a. Calculate variance for the points up to the cut-point
    b. Calculate variance for the points after the cut-point
    c. Total variance for this cut-point is the sum of the two variances
2. Find the cut-point with minimum total variance. (If interest is in the most current change, pick the last cut-point when there are tied values.)

These two estimators are usually close to each other but there may be differences at times. The minimum variance estimator is usually better in the sense that it cuts where we would separate the series ourselves, but it comes at the cost of computational complexity.

## 3. SAMPLE CODE

This code block will calculate the range of CUSUM from the array _d[*] of length &dim:

```
calc_csrng:                                                * CUSUM Range  ;
   _cusum  = 0;                                            * CUSUM        ;
   _csmax  = 0;                                            * Max(CUSUM)   ;
   _csmin  = 0;                                            * Min(CUSUM)   ;
   do _ci = 1 to &dim;
      _cusum = _cusum + _d[_ci];
      _csmax = max( _csmax, _cusum );
      _csmin = min( _csmin, _cusum );
   end;
   _csrng = _csmax - _csmin;                               * Range(CUSUM) ;
   return;
```

This code block will shuffle the _d[*] array with length &dim:

```
shuffle:                                                   * Shuffle      ;
   do _si = 1 to %eval(&dim-1);                            * swap this    ;
      _sx=_si+floor(rand('uniform')*(%eval(&dim+1)-_si));  * with this    ;
      _st=_d[_si];  _d[_si]=_d[_sx];  _d[_sx]=_st;         * do swap      ;
    end;
    return;
```

This code block will find the CUSUM peak estimator and store it in _cp1:

```
find_cspeak:
   _cusum  = 0;                              * CUSUM Peak estimator;
   _cspeak = 0;
   _cp1    = 0;
   do _i = 1 to &dim;
      _cusum = _cusum + _d[_i];
      if abs(_cusum) >= _cspeak-1e-12 then do;
         _cp1 = _i;
         _cspeak = abs(_cusum);
      end;
   end;
   return;
```

This code block will find the minimum variance estimator and store it in _cp2:

```
find_minvar:                                 * Minimum variance estimator;
   _cp2 = . ;
   _min_var = . ;
   _s1sum   = 0;
   _s1sumsq = 0;
   do _mi = 1 to %eval(&dim-1);              * _mi marks end of group;
      *--- group 1 ---;
      _s1sum   + t[_mi];
      _s1sumsq + t[_mi]**2;
      _s1mean  = _s1sum / _mi;
      _s1var   = max(_s1sumsq-(_mi*_s1mean**2),0);
      *--- group 2 ---;
      _s2sum   = _tsum - _s1sum;
      _s2sumsq = _tsumsq - _s1sumsq;
      _s2mean  = _s2sum / (&dim - _mi);
      _s2var   = max(_s2sumsq-((&dim-_mi)*_s2mean**2),0);

      _var = _s1var + _s2var;                * total variance         ;

      *--- Find largest index of min ---;
      if _var <= min(_min_var,_var)+1e-12 then do;
         _min_var = _var;
         _cp2 = _mi;                         * Min variance cut-point ;
         _u1  = _s1mean;                     * Mean of group 1        ;
         _s1  = sqrt(_s1var/_mi);            * Stdev of group 1       ;
         _u2  = _s2mean;                     * Mean of group 2        ;
         _s2  = sqrt(_s2var/(&dim-_mi));     * Stdev of group 2       ;
      end;
   end;
   return;
```

The variables _u1 and _u2 allow further processing such as selection of meaningful difference. The comparison fuzz tolerance of 1e-12 is added to pick the most recent change point when there are ties. With these four code blocks in place, the main driver routine is as follows:

```
data cpa;
   if _n_=1 then call streaminit(888);      * Initialize rand()      ;

   set test;                                * Read time series       ;

   array t[&dim];                           * Actual time series     ;
   array _d[&dim] _temporary_;              * Deviation array        ;

   _tsum    = 0;
   _tsumsq  = 0;
   do _i = 1 to &dim;
      _tsum   + t[_i];                      * sum(t)                 ;
      _tsumsq + t[_i]**2;                   * sum(t**2)              ;
   end;
   _tmean = _tsum / &dim;                   * mean(t)                ;

   do _i = 1 to &dim;
      _d[_i] = t[_i] - _tmean;              * Deviation array        ;
   end;

   link calc_csrng;
   d_csrng = _csrng;                        * CUSUM Range Actual     ;
   link find_cspeak;                        * CUSUM Peak estimator   ;

   prob = 0;
   do _i = 1 to &shuffle;                   * Shuffle iterator       ;
      link shuffle;                         * Do the shuffle         ;
      link calc_csrng;                      * Calculate CUSUM Range  ;
      if d_csrng > _csrng then prob+1;      * # Actual > Shuffle     ;
   end;
   prob = prob / &shuffle;                  * Probability            ;

   if prob >= 0.95 then link find_minvar;   * Find min var estimator ;

   return;                                  * Data step loop         ;

<place four code blocks here>

run;
```

Using LINK-RETURN allows clearer expression of the processing logic, beginning variable names with an underscore mitigates name collision with input dataset variables as DATA step variables have global scope.

This DATA step will read time series data and compute the probability of significant mean shift, the two cut-point estimators, the mean (_u1, _u2) and standard deviation (_s1, _s2) of the two segments for the minimum variance estimator.

## 3.1 DOES IT WORK?

The ability of CPA to identify the presence and location of mean shift for a number of time series is shown below in Table 1.

**Table 1 CPA output for a variety of time series**

| Obs | Type | t1 | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | t10 | t11 | t12 | CSRE | prob | _cp1 | _cp2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Flat line | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 0.00 | 0.000 | 12 | . |
| 2 | Step change | 10 | 10 | 10 | 10 | 10 | 10 | 11 | 11 | 11 | 11 | 11 | 11 | 3.00 | 0.980 | 6 | 6 |
| 3 | Bump 1 up | 10 | 10 | 10 | 10 | 10 | 11 | 10 | 10 | 10 | 10 | 10 | 10 | 0.92 | 0.000 | 6 | . |
| 4 | Bump 2 up | 10 | 10 | 10 | 10 | 10 | 11 | 11 | 10 | 10 | 10 | 10 | 10 | 1.67 | 0.842 | 7 | . |
| 5 | Bump 2 up high | 10 | 10 | 10 | 10 | 10 | 20 | 20 | 10 | 10 | 10 | 10 | 10 | 16.67 | 0.843 | 7 | . |
| 6 | Bump 3 up | 10 | 10 | 10 | 10 | 10 | 11 | 11 | 11 | 10 | 10 | 10 | 10 | 2.25 | 0.940 | 5 | . |
| 7 | Bump 4 up | 10 | 10 | 10 | 10 | 11 | 11 | 11 | 11 | 10 | 10 | 10 | 10 | 2.67 | 0.975 | 8 | 8 |
| 8 | Change variation | 10 | 10 | 10 | 10 | 10 | 10 | 11 | 9 | 11 | 9 | 11 | 9 | 1.00 | 0.000 | 11 | . |
| 9 | Up down up down | 10 | 10 | 10 | 11 | 11 | 11 | 10 | 10 | 10 | 11 | 11 | 11 | 1.50 | 0.134 | 9 | . |
| 10 | Sloped line | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 18.00 | 0.982 | 6 | 6 |
| 11 | Tail 1 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 20 | 9.25 | 0.444 | 10 | . |
| 12 | Tail 2 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 9 | 20 | 20 | 16.67 | 0.845 | 10 | . |
| 13 | Tail 3 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 20 | 20 | 20 | 22.25 | 0.987 | 9 | 9 |

Row 1 is a flatline with no noise, CPA shows mean shift probability of 0. Row 2 is a step function, CPA shows mean shift probability of 0.98, meaning it is highly likely there is a mean shift. This shows CPA can distinguish the presence and absence of mean shift in a simple situation.

Rows 3-7 shift progressively more points up in the middle of the time series with no other noise. We can see shift probability increases from a probability of 0 for a single point to a probability of 0.98 for four points, this shows CPA can make more confident decisions as evidence mounts. We also see that CPA is sensitive to the length of the shift but not the magnitude when we compare rows 4 and 5.

Row 8 shows this CPA is not affected by heteroscedasticity. Row 9 shows CPA again making the correct assessment of no shift when the data has fluctuations. Row 10 shows it detecting mean shift for a sloped line which is technically correct.

Similar to rows 3-7, rows 11-13 show a noisy series with the shift occurring at the end of the series. CPA continue to demonstrate increasing probability of mean shift as more points are shifted, this shows the location of the shift doesn't matter, if it's there, CPA will identify it. Row 11 and row 3 also demonstrate that CPA is relatively robust to a single point outlier.

We can see that _cp2, the minimum variance estimator, and _cp1, the CUSUM peak estimator, both agree with visual inspection of when the change happened as they both mark the last point prior to change.

The demonstrated ability to correctly identify presence and location of mean shift under a variety of conditions give us confidence in using it in practice.

## 3.2 WHAT IS CPA'S REACTION SPEED?

How soon can CPA signal a change has happened? This is an important question to answer as it governs our reaction speed to a change event.

If we look at rows 4-7 and 11-13, we see that probability is greater than 0.94 at 3 consecutive shift out of 12 points. In a business setting, it means if we perform CPA on 12 monthly data, CPA needs to see three consecutive months of shift before it triggers, if we use a threshold probability of 0.94. In other words, at a threshold probability of 0.94, CPA has a reaction speed, or detection lag, of three months. Experimentation shows this speed doesn't change with time series longer than 12 points.

This speed makes intuitive sense. If we are monitoring a stable process, when we encounter the first shift, it could be a singular event, it is too early to raise an alarm, the CPA program calculates a small probability of change (row 11 shows 0.44). When we encounter two shifts in a row, the probability rises to 0.85, this is at the cusp of should we sound an alarm or not as there is still some doubt as to whether it will continue or not, sounding an alarm may run the risk of false positive. With three shifts in a row, CPA reaches probability of approximately 0.95 with row 13 showing 0.99, near certainty. It is interesting to note the Bayesian-like behavior of CPA probability increasing as it receives more evidence although the algorithm wasn't designed with Bayesian Statistics in mind.

To reduce time lag to two months requires a lower threshold probability of 0.8, CPA will trigger earlier but at a slightly higher risk of false positives. If the business requirement is earlier detection with minimal false positive consequences, this would be a good trade-off. However, if false positive consequences are serious, we can raise the threshold probability higher to be certain that a change has been occurred.

### 3.3 IMPLEMENTATION CONSIDERATIONS

The probability calculation depends on the number of shuffles. To obtain robust estimates, several hundred shuffles are required at a minimum, more than a thousand is usually not necessary. You should experiment with actual data to find the right balance between robust estimate and execution speed.

Even though computers are very fast, shuffling still takes time. In practice CPA is run in two steps:

1. CPA analysis to calculate probability, change point, means and standard deviations. This is the time intensive part for large datasets.
2. Filter output based on probabilities and magnitude and difference direction between _u1 (mean prior to change) and _u2 (mean after change) to isolate series that increased by some threshold and series that decreased by some threshold.

Step 1 deals with CPA calculations. Step 2 deals with selections of business relevance. The separation into two steps allows you to exercise a great deal of control in surfacing interesting business events quickly without re-running the compute-intensive step 1.

## 4. CASE STUDIES IN A FINANCIAL INSTITUTION

Financial institutions are data-rich, CPA has enormous potential to discover useful business insights. We will discuss three examples:

1. Credit card spend volume decline,
2. Pay increase, and
3. Interest rate sensitive customers

### 4.1 CREDIT CARD SPEND VOLUME DECLINE

The first example is of monthly spend of credit cards. The portfolio consists of approximately 4.5 million accounts belonging to 4 million customers, during planning it was determined there is an annual spend shortfall of around $700 million. Many ideas were suggested as the cause such as card acquisition slowdown, slowing economy causing broad based spend decline, increased market competition, increased card attrition, … etc.

Change Point Analysis was used to assess purchase volume decline of the existing customer base. The processing steps are outlined in Table 2 below:

**Table 2 CPA analysis of credit card spend**

| Step | Records | CPU Time | Elapsed Time |
|---|---|---|---|
| **1.** Extract three years of purchase volume | 178 Million | 6 minutes | 1 hour |
| **2.** Aggregation to customer | 126 Million | 15 minutes | 15 minutes |
| **3.** Eliminate low volume / closed | 48 Million | 6 minutes | 6 minutes |
| **4.** Change Point Detection | 48 Million | 4 hours | 4 hours |
| **5.** Change Point Selection | 8K | 1 minute | 1 minute |

The graphs below show examples of large purchase volume decline detected via CPA:

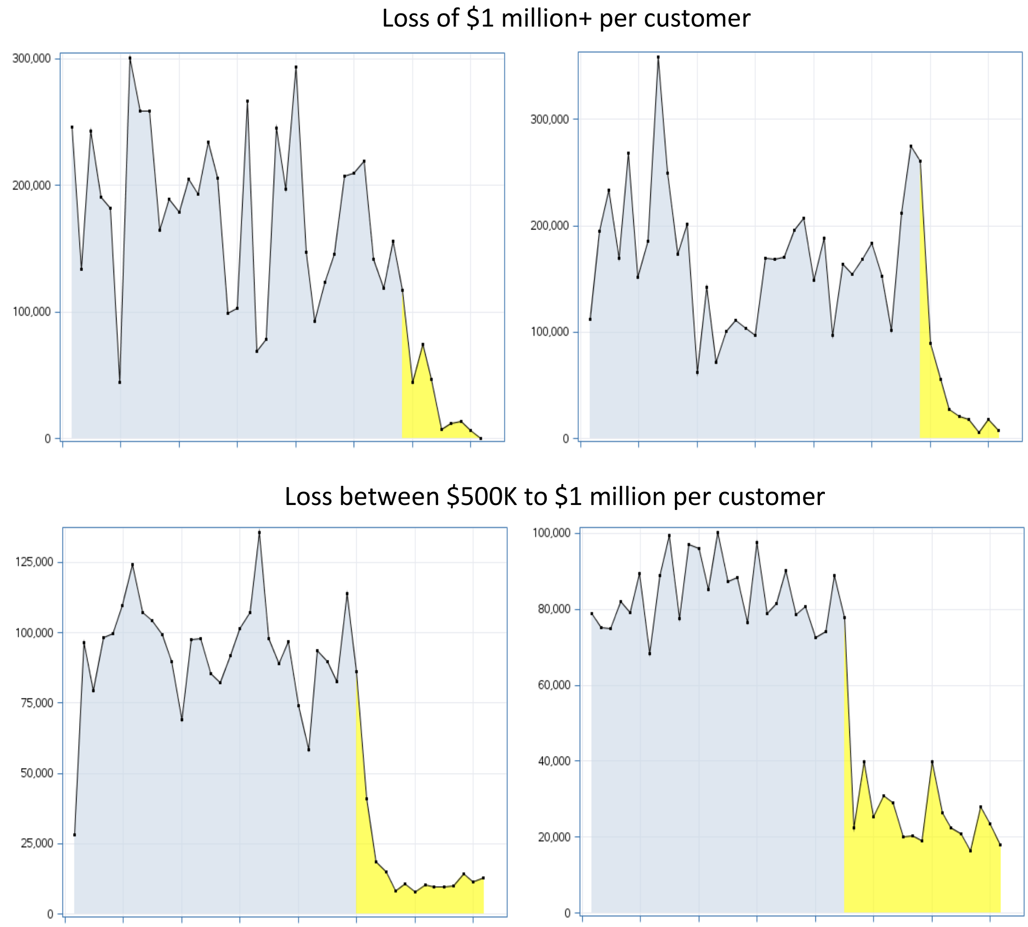



**Figure 3 CPA detection of large spend decline**

The annual purchase decline of the eight thousand customers identified via CPA amounted to $545 million, this means 78% of the annual spend shortfall is in 0.2% of the customer base, this potentially debunks the broad based spend decline argument.

Table 3 shows detailed breakdown of their spend decline and that pareto principle continue to hold:

**Table 3  Card spend decline identified by CPA**

| Group | Customer | % Customer | Average Annual Spend Loss | Annual Spend Loss | Index |
|---|---|---|---|---|---|
| 1 | 100 | 1% | $1 million+ | $109 million | 14.9 |
| 2 | 400 | 5% | $266K | $109 million | 3.9 |
| 3 | 1K | 13% | $110K | $109 million | 1.6 |
| 4 | 2.1K | 26% | $52K | $109 million | 0.8 |
| 5 | 4.4K | 55% | $25K | $109 million | 0.4 |
| Total | 8K | 100% | $68K | $545 million | 1.0 |

These are important customers that a financial institution should reach out as early as possible to understand what's driving the decline, and then propose better solutions to earn their business. Prior to CPA analysis, management did not know this was happening.

## 4.2 PAY INCREASE

The second example is examining a customer's payroll increase. We believe pay increase events are important occasions where a person's usage of financial products could change significantly. Two factors need to be dealt with prior to CPA:

**1.** There are four different pay frequencies: monthly, twice a month, bi-weekly, and weekly. If we aggregated to monthly pay, we would introduce artificial variability due to the varying number of pay events per month. This was unacceptable, so we used individual transactions. We also split the four pay frequencies into four streams and ran CPA on them separately. The different amount of data available meant we can set different control parameters for each of them, this was done mainly in the selection of interesting business events process.

**2.** There are unusual pay outliers when people receive bonuses or go on vacation. These onetime events are identified via Kalman Filter and removed for other purposes. The Kalman Filter was set to remove onetime outlier with minimal impact to the rest. CPA was done on the Kalman Filter filtered data.

The processing steps are shown below in Table 4:

**Table 4  CPA processing of payroll change**

| Step | Records | CPU Time | Elapsed Time |
|---|---|---|---|
| **1.** Extract four years of payroll | 155 Million records | 18 minutes | 25 minutes |
| **2.** Aggregation to customer | 145 Million records | 2 minutes | 2 minutes |
| **3.** Eliminate low and closed | 85 Million records | 5 minutes | 5 minutes |
| **4.** Pay frequency determination | 62 Million records | 2 minutes | 2 minutes |
| **5.** Remove irregular off cycle pay | 60 Million records | 5 minutes | 5 minutes |
| **6.** Kalman Filter removal of outlier | 60 Million records | 4 minutes | 4 minutes |
| **7.** Change Point Detection | 60 Million records | 7 hours | 7 hours |
| **8.** Selection of pay increase | 17K accounts | 1 minute | 1 minute |

You can see from the pay increase graphs below how noisy real data is.  There are also many onetime pay outliers.  A simple mean change rule results in many false positives that CPA correctly rejects.

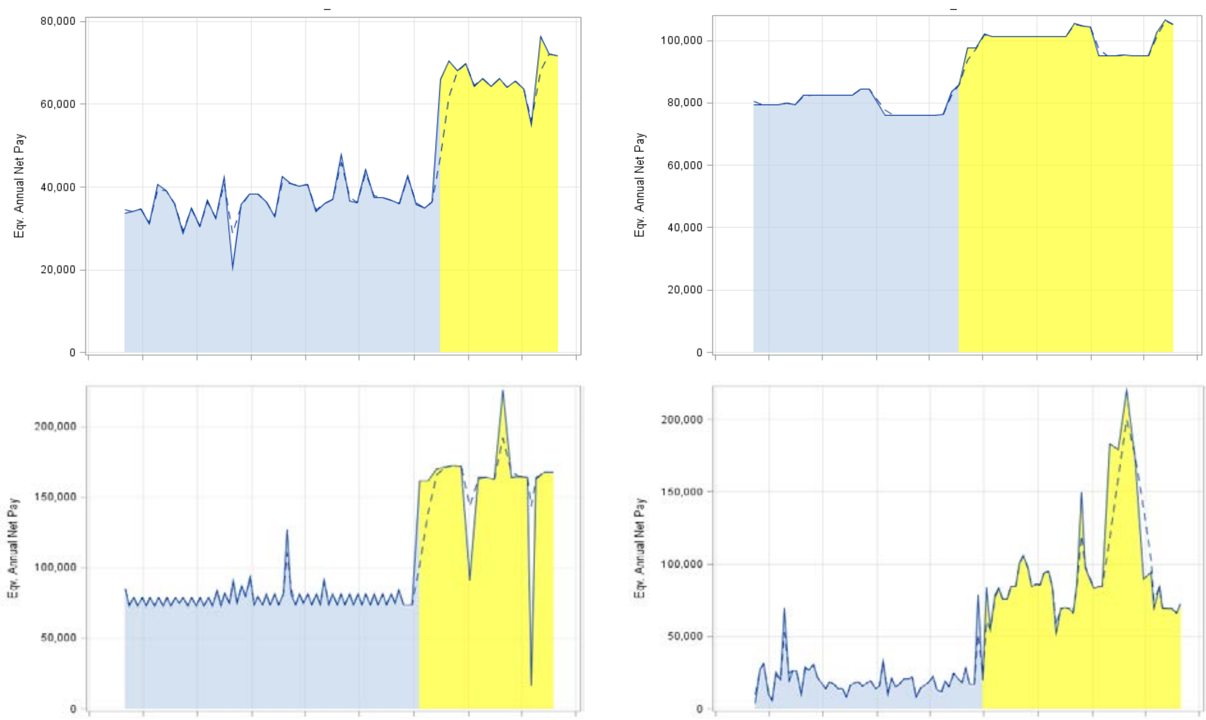



**Figure 4  CPA detection of payroll increase**

Further analysis of customer balance after pay increase showed significant financial changes, details are in table 5:

**Table 5  Pay increase impact identified via CPA**

| After Pay Increase | | | Per Customer Funds Increase | | | Total in $Million | |
|---|---|---|---|---|---|---|---|
| Age | Customer | Account Increase | Asset | Lending | Card Spend | Funds | Card Spend |
| 19 - 35 | 7,800 | 6.6% | $5,900 | $10,800 | $3,200 | $130 | $25 |
| 35 - 45 | 5,700 | 4.3% | $5,500 | $5,400 | $2,400 | $62 | $15 |
| 45 - 55 | 5,000 | 3.2% | $9,600 | $3,500 | $3,100 | $66 | $15 |
| 55 - 65 | 2,500 | 2.2% | $16,100 | $600 | -$1,900 | $42 | -$5 |
| All | 21,000 | 4.3% | $7,900 | $6,400 | $2,300 | $300 | $50 |

The results agree with intuition:

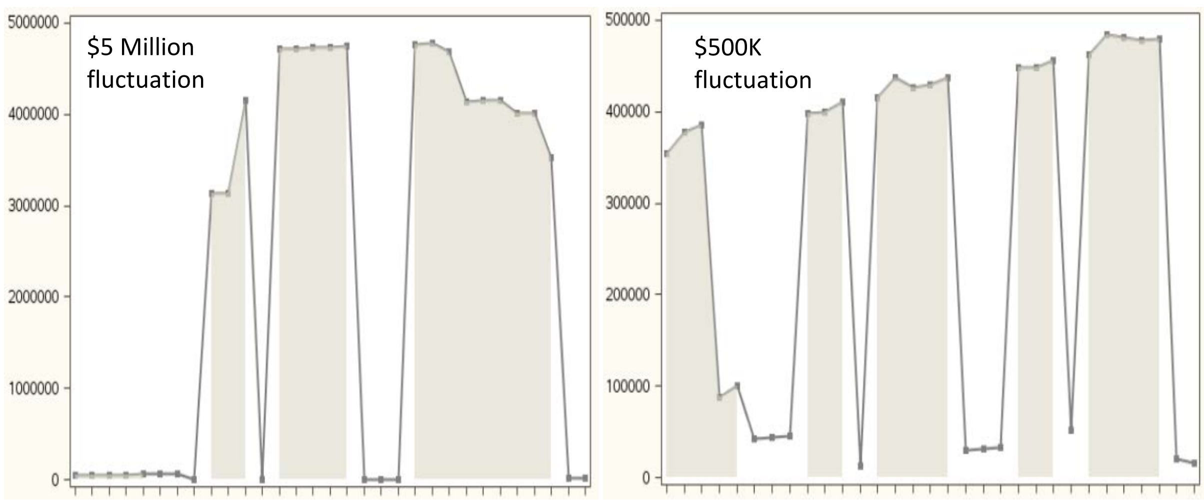
1. More young people have pay increase events, perhaps older people don't change jobs as often when they become settled
2. Young people open more new accounts as they expand their financial product usage
3. Older people tend to save their new wealth and not borrow more, perhaps they have moved beyond the borrowing life stage
4. Young people will save, borrow (new car, bigger house), and spend more when they have more money, i.e., they reward themselves when their pay increases

Together these customers, approximately 1,750 per month, add $300 million in funds and $50 million in credit card spend a year. These are good financial planning opportunities that a financial institution should leverage to proactively reach out to their clients at the right time.

## 4.3 INTEREST RATE SENSITIVE CUSTOMERS

The last example is the identification of interest rate sensitive customers. The financial institution periodically runs aggressive high interest rate promotion for saving accounts to attract deposit balance. Thus, the total balance of the approximately 400,000 accounts fluctuate with promotion periods, management would like to understand the nature of the rate sensitive hot money.

The analysis uses CPA to identify accounts whose balance has significant changes, followed by time series correlation with the promotion periods. We identified large balance accounts whose balance moved in near lockstep with promotion periods as shown in the graphs below:




Grey = high interest promotion    White = regular low interest

**Figure 5 CPA analysis of high interest promotion**

The portfolio has approximately 400,000 customers with aggregate balance of $20 Billion. We identified 8,800 customers that had $4.6 Billion hot money that moved with the promotion, when the promotion was on, money flowed in, when it came off, money flowed out. The hot money was highly concentrated rather than broad based, 2.2% of the customers hold 23% of the portfolio as interest seeking hot money, shown below in Table 6:

**Table 6 Interest seeking hot money identified via CPA**

| Hot Money | # of Customer | Total ($Million) |
|---|---:|---:|
| $ Million+ | 815 | $2,167 |
| $500k - $1Million | 1,313 | $904 |
| $250K - $500K | 2,381 | $838 |
| $100K - $250K | 4,361 | $699 |
| Total | 8,870 | $4,607 |

These are wealthy customers that are seeking return on their money in a safe instrument. The dialog for these customers is no longer just about interest rate on a saving account, it is about a proper investment strategy to balance the customer's risk/reward appetite, and part of the solution includes safe high interest flexible instruments that stops the yo-yoing of deposit balances. CPA allowed us to get to the core of the problem and change our focus to a larger and more important business issue altogether.

## 5. RELATIONSHIP TO PREDICTIVE MODELS

Predictive modelling is used to address many customer management issues such as upsell, cross-sell, acquisition and attrition, they provide tremendous value to marketing to uncover revenue opportunities. However, they run into difficulties when dealing with balance change problems, one of the major issues being target definition.

With acquisition, cross-sell, and attrition, the target definition is usually clean and unambiguous – is this a new customer, was a new product purchased, did the customer attrite. With balance change, however, targets are usually defined as the average balance of two time-window being different by some amount – the naïve decision rule that would include false positives. Good targets would also be missed due to the choice of the arbitrary time-window, CPA could scan the time series and find the correct change point windows. Thus, CPA can assist predictive model setup by providing an accurate view of who underwent what level of balance change with what level of probability at what time. Target definition can then be based on this improved understanding of behavior change rather than the problematic mean-change within arbitrary window rule.

Depending on the business situation, there are times when CPA results lead directly to business action, there is no need to refine the insight further with predictive models. And there are times where CPA output should be complemented and analyzed further by other techniques.

CPA output can also be used as input variables to standard predictive models. Indeed, if many events have been analyzed, it is possible to assemble an events history database that should prove very useful in discerning the relationship between behavior events.

## CONCLUSION

CPA is a valuable tool for detecting changes in behavior. It can be used in any industry. It can be used to monitor changes in sales by SKU, inventory levels, balances, purchases, transactions, inflow/outflow, claims, acquisitions, attritions, just about any key metric a business runs on. It can run on macroscopic aggregated data, or it can run on the most granular detailed level of data. The possibilities are endless. It's up to the imagination of the analytics professional to use this simple yet powerful tool to discover powerful business insights that alert the organization that something is changing.

## REFERENCES


Website Taylor, Wayne A. "Change-point analysis: a powerful new tool for detecting changes." *preprint, available as* http://www.variation.com/cpa/tech/changepoint.html (2000).

Book Chen, Jie, and Arjun K. Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2011.


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Daymond Ling
daymond.ling@gmail.com