

An event-state depiction algorithm using CPA methods with continuous feed data

Philip Sallis

Geoinformatics Research Centre
Auckland University of Technology
Auckland, New Zealand
psallis@aut.ac.nz

Sergio Hernández

Lab. de proc.de inf. geoespacial
Universidad Catolica del Maule
Talca, Chile
shernandez@ucm.cl

Abstract— Change Point Analysis (CPA) is a formal method used for determining whether a change-of-state has taken place in a set of observed events. There are claims that this method is capable of detecting subtle changes missed by for example, control charts. By providing confidence levels and confidence intervals it claims to better characterise the changes detected. This is a bonus when for instance, tracking continuous wind velocity data in order to detect when perturbations occur in the oscillation patterns because these are potential predictors of wind gusts, which can have devastating physical effects on land based objects such as buildings and crops. This early-stage research paper describes some previous work in predicting wind gusts using a branch-and-bound algorithm. It considers the need for precision and early detection in wind pattern state-change and examines how fit-for-purpose a change-point analysis method could be for the early detection of velocity oscillation perturbations in a mixed variable analysis of condition change. Wind velocity data is sampled in real-time. The continuous data feed is processed by a change-point analysis algorithm, which has been derived for this purpose. The results are depicted and described.

Keywords-change-point analysis; continuous data; real-time systems; algorithms design

I. INTRODUCTION

The literature relating to Change Point Analysis (CPA) mostly relates to its use as a method in financial modeling and more widely in economic dynamics and trend forecasting. In a recent article describing this application area Taylor [1] claims that although not fully appropriate for use with continuous data through say an on-line real-time feed, the additional confidence level information provided by the change-point analysis (CPA) method when used together in a hybrid approach, produces more reliable state-change models than previously developed. He goes on to say that when analyzing historical data, especially when dealing with large data sets, change-point analysis is in fact, preferable to control charting. He provides a useful

illustration of control charting where in Figure 1 below, the red horizontal lines indicate the upper and lower bounds of an individual chart where it is assumed over time no change has occurred. It can be seen here that at the point shown to be in the time period October 1987 that in fact, a change point did occur because it exceeded the upper limit of the normalised parameter.

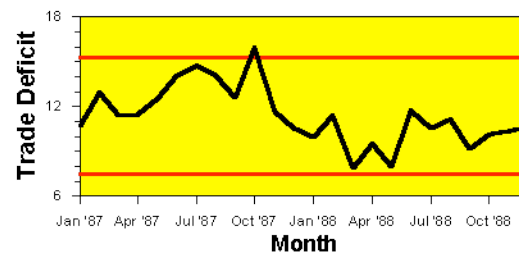


Figure1 Control Chart with assumed fluctuation limits [1 (op cit)]

This is argued to be the case because CPA is a so-called *more powerful* analyser in that it better characterizes the changes, controls the overall error rate, is robust to outliers, is more flexible and is more simple to use. These observations are attractive to those working in areas other than finance or economics because it may be that CPA is a tool that in one form or another could be used to observe state change considered to be statistically significant. This approach would for example, enhance our understanding of so-called *step-change* events, more precisely quantified using such non-parametric methods such as McNemar's Test for the Significance of Changes [2]. This long-standing method originating from early (1947) psychometric research is used on nominal data. It is applied to contingency tables of a dichotomous trait with 2×2 dimension, which have matched pairs of subjects. It is used to determine whether the row and column marginal frequencies are equal. McNemar [2 (op cit)] calls this derived value *marginal homogeneity*. It is determined by testing the null hypothesis of marginal homogeneity, which states that the two marginal probabilities for each outcome are the same.

So the marginal probabilities are compared thus,

$$p_a + p_b = p_a + p_c \text{ and } p_c + p_d = p_b + p_d$$

and the null hypothesis is expressed as $p_b = p_c$. The McNemar statistic equation below uses a chi-square derivation in order to produce a binomial distribution of matrix values thus,

$$\chi^2 = \frac{(b - c - 0.5)^2}{b + c}$$

If we apply this in an example where the same subjects in a sample are included in a before-and-after measurement matrix (so that they are matched pairs) we can see as in Table 1 below that we have the potential for observing a binomial distribution. The data depicted here relates to patients who were diagnosed with a particular disease and then treated with a prescribed drug. The effects of the drug before-and-after for each patient were observed [3].

	After: present	After: absent	Row total
Before: present	101	121	222
Before: absent	59	33	92
Column total	160	154	314

Table1 The before-and-after sample observations matrix [3 (op cit)]

Populating the McNemar equation with this data results in it appearing thus,

$$\chi^2 = \frac{(121 - 59 - 0.5)^2}{121 + 59}$$

In this example, the null hypothesis of *marginal homogeneity* would show that the drug treatment had no positive effect. Placing the values in the McNemar equation results in it generating a value of 21.01, which is not the expected outcome from the distribution of values implied by the null hypothesis. The test therefore, illustrates strong evidence to reach the conclusion that null hypothesis of no treatment effect should be rejected. This test, given here as an example of how non parametric statistics deals with populations of sample data where difference is measured for significance, illustrates the potential for using alternative methods when considering state change over time, rather than for a snapshot of available discrete value data.

II. THE CPA APPROACH

Taylor [1 (op cit)] sets out the CPA equations and ascribes the conventional analytical component terms used with this approach. He defines cumulative sum charts (CUSUM) and so-called *bootstrapping* [4] as the primary inputs to operate on the data. A useful further reference to this concept is described by Hinkley et al [5] where the authors work through numerous examples of what they describe as a *mean-shift model*, which has the effect of bringing even greater precision to the state change observations by moving the average to compare with every realization of the event sample. To supplement this approach the CUSUM method can be used. This technique from statistical quality control is a sequential analysis technique [6] where it is typically used for monitoring state changes detections. The inventor of CUSUM, one E.S. Page [6 (op cit)], described a *quality number* θ , as being a parameter of the probability distribution and in particular he used it to refer to the *mean*. He devised CUSUM as a method to determine changes in (the *mean*), and proposed a criterion for deciding when to take corrective action. CUSUM relates to a sequence of events; it involves the calculation of a cumulative sum [7]. Samples from a process x_n are assigned weights ω_n , and summed as follows:

$$S_0 = 0$$

$$S_{n+1} = \max(0, S_n + x_n - \omega_n)$$

When the value of S exceeds a certain threshold value then a change in value has been found. The above formula only detects changes in the positive direction. When negative changes need to be found as well, the *min* operation should be used instead of the *max* operation. When this occurs a change has been found when the value of S is *below* the (negative) value of the threshold value. Figure 2 illustrates a sequence of inputs over time with the change-points mapped according to the threshold values used as parameters for the analysis [8].

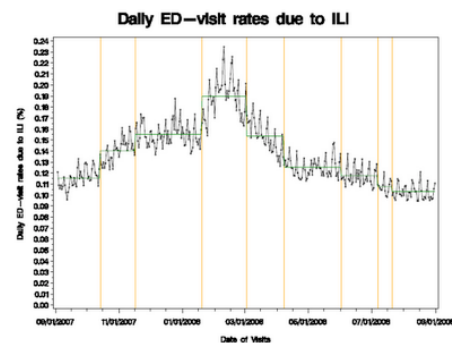


Figure2 Change Points in a sequence over time

The illustration shown in Figure 2 above is the result of calculating the CPA. Zhiheng [8 op cit] describes the procedure for calculating the CPA in a number of steps thus,

- Determine the Series Mean
- Accumulate Running Sum of differences between Mean and individual values
- Plot CUSUM series
- The point farthest from 0 denotes a Change-Point (CP)
- Break into two sections at CP:
- Analyze each subseries for additional significant CPs

Bootstrapping provides us with a measure of the CP's significance. Taking a statistical perspective on bootstrapping we could assign measures of accuracy to sample estimates using an algorithm designed for the purpose of measuring the properties of the estimator (say its variance) and generate a distribution for re-sampling purposes thereby, *bootstrapping* the CP with a matrix of covariance values. This is demonstrated later in the paper using a computer program designed to implement the CPA algorithm we will continue now to describe. If we are to model changes-of-state occurrences in a set of continuous data we first have to establish that a change occurred. This requires that we have a single value (or range of values) and the presence or absence of a condition that we can test the data stream against. We also need to know when the change occurred and to what extent it occurred. Cumulatively we need to know how many changes occurred in a given time series because this knowledge helps determine the pattern and severity (significance) of the changes. If we are to use the state-change information as input to another process or decision-making framework we need to know precisely with what confidence can we say that these changes have occurred in the set. Taylor [1 op cit] demonstrates how the confidence level is calculated by performing a large number of bootstraps and counting the number of bootstraps for which, S_{diff}^0 is less than S_{diff} .

So,

let N be the number of bootstrap samples performed and
let X be the number of bootstraps,

where $S_{diff}^0 < S_{diff}$.

The confidence level that a change occurred can then be expressed as a percentage of the sample thus,

$$\text{sum}((100 * X / N) \%)$$

Typically 90%, or 95% confidence is required before one states that a significant change has been detected. For the five bootstraps derived in this example, the values of S_{diff}^0 are observed as being 7.0, 14.917, 7.975, 7.938 and 9.15 respectively. All of these values are below $S_{diff} = 17.74167$. Figure 3 shows a histogram of S_{diff}^0 based on 1000 bootstrap samples. Out of 1,000 bootstraps, 995 had $S_{diff}^0 < S_{diff}$. This gives a confidence level of 99.5%,

derived from the sum $((100 = 995/1000)\%)$. This is strong evidence that a change did in fact occur.

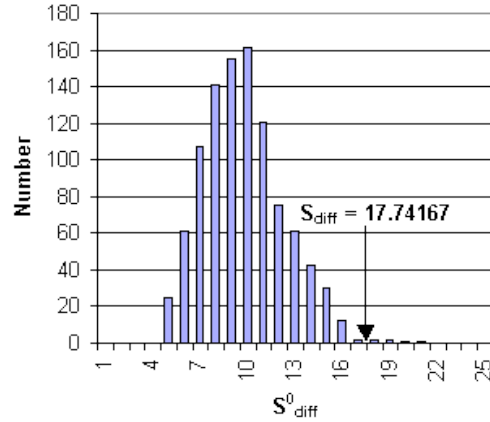


Figure3 Histogram of S_{diff}^0 for 1000 Bootstrap Samples

A feature of CPA that assists us is that the method features a multiple change detection capability. This is necessary for continuous stream data where the time interval between changes informs us about trends and change magnitudes. The CPA method generates a confidence level and the confidence interval associated with each change that is identified so we can use that information to determine the significance extent of the change and when it occurred.

It is generally held that change-point analysis analytical procedures are extremely flexible. In the examples of how CPA performs with numerous kinds of data, the article [1 (op cit)] demonstrates how it can be used as an analytical tool with various time ordered data types including attribute data, data from non-normal distributions, ill-behaved data such as particle counts and complaint data and data with outliers.

II. THE WIND VELOCITY OSCILLATION PROBLEM

Wind velocity is influential in numerous contexts and applications but is especially of concern where the preservation of property or crops is concerned. So-called wind gusts are perturbations in the wind oscillation patterns and they can be devastating in the damage they cause during a typically very short (generally no more than 20 seconds) duration. It has been observed [9] that erratic fluctuations in wind velocity followed by a rapid acceleration in velocity accompanied by a logarithmic increase in temperature and humidity (not strictly concurrent or equivalent orders of magnitude) and a drop in atmospheric pressure will accompany a wind gust event. This phenomenon is illustrated in Figure 4 below, where the red dots are wind gust events and the line plots show the climate factor variations. This category of event in Nature is often considered to be chaotic [10] and the prediction of occurrences is time-constrained such that any warning of a

wind gust event would be meaningless in terms of defense preparations. Identifying these perturbations is useful in order to observe the other climate variation factors relating to them and in real-time we can plot their occurrence over time but combining all the influence factors at a single moment in time in order to determine the range of values present at the change-of-state requires more analysis and some conditions that must be tested for a model to be built that leads us towards information concerning the severity of the event. CPA seems to provide a tool for analyzing this by providing information about the confidence level we can obtain for each event as it occurs.

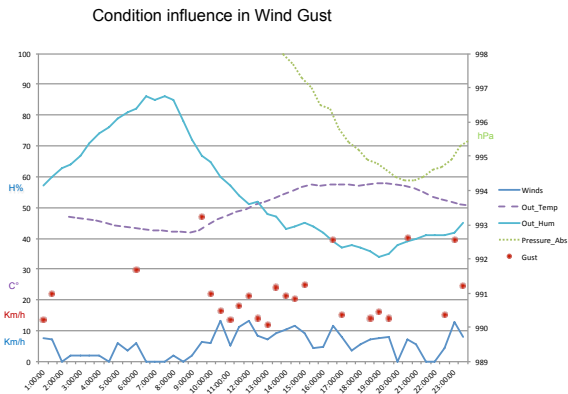


Figure4 Discrete variables as factors relating to wind gust events

III. APPLYING CPA TO THE WIND GUST PROBLEM

The first issue to address with respect to using CPA with the wind velocity data is that there are four potential predictor variables in the mix. In order to use one CPA operation we would need to normalize the values such that we have only one value to be concerned with in the algorithm. This is not appropriate due to the distinctly different measures used for each variable. We are faced then with developing an algorithm that will account for each variable simultaneously in the analysis. Software developed by Taylor [11] enables continuous data to be classified for event state change points over time. Visualisations of the results of analysis appear as groups with outlying values readily depicted as in Figure 5 below.

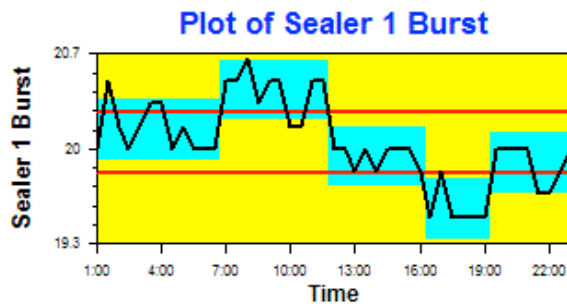


Figure5 CPA state change visualisation [11 (op cit)]

If we now consider the wind velocity measurement problem in light of this visualization, we could classify the events above and below the parameter limit (horizontal) lines as being perturbations in the oscillation distribution. These determinations would be based on the observed wind gust events as illustrated in Figure 4. Furthermore, we could describe the rules implied in this depiction for predicting wind gusts as being:

- Erratic velocity fluctuations over a brief time period
- Rapidly increasing temperature
- Rapidly decreasing humidity
- Rapidly decreasing atmospheric pressure

The CPA in this instance is both a state change identification tool and also a useful benchmarking device for comparing the event state change and the climatic conditions nominated as being precursors for perturbations in the oscillation distribution, in other words wind gusts.

An executable algorithm for processing this data is the subject of ongoing research for the work related to this paper but an early version of it appears below. It assigns values to a CPA partitioning process using the four variables in the gust recognition data plotted in Figure 4. It associates previously classified data that represents near-ground truth real-time values with CPA partition boundary estimates of state event change expectations. The four variables are assumed here to have the data point values from the real-time feed.

```

Begin
begin partition
  let upper-bound = p, lower-bound = q,
  bootstrap = 0
begin assign
  let vector = null,
  pert = 1 [assume event present],
  est = 0 [event_start_time],
  eet = 1 [event_end_time under 1 minute],
  (tv)n = temperature,
  (hv)n = humidity,
  (vv)n = wind velocity,
  (pv)n = atmospheric pressure
  vector S{((tv)n..(hv)n..(vv)n..(pv)n} [feed]
while feed do until exhaust
  for eetn (>0) and estn (<1) [minute]
    if (tv)n < (eeti+esti)
      and (dv)n < (eeti+esti)
      and (vv)n > (eeti+esti)
      and (pv)n < (eeti+esti)
    then compute S = pert+1;
      bootstrap=bootstrap+1;
    else S = 0, pert = 0; end;
end assign;
while S = 1 do
  if pert >= 1 and S0diff < Sdiff
    do compute ((100*X/N)%); p=1
    then plot partition [upper_bound]
  else
    p=0; plot partition [lower_bound]
end partition;
end.

```

Taking the form of a time series non-linear function based procedure, it is expected that this algorithm once proven, can add considerably to the proposition that change point analysis methods are more generally useful than where they are seen to have analytical success mostly in the fields of finance and economics.

Assuming this algorithm was imbedded in a process whereby parameters are set for the upper and lower limits of a CPA process, the data in Figure 4 would take the depicted form if Figure 5 below.

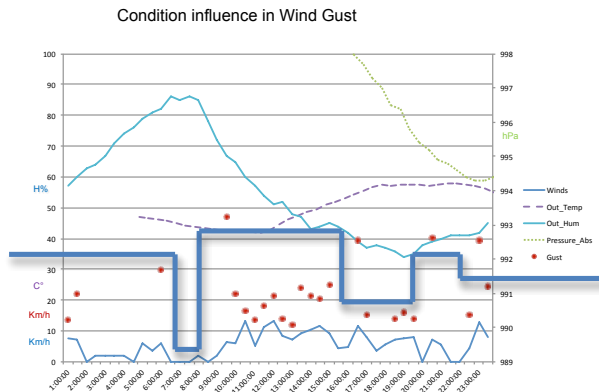


Figure5 Superimposing the CPA parameters

The change points are above and below the assumed (estimated) parameter line depicted here. It can be seen that the greatest number of actual state change events occurs in the partition where all four variables are behaving as predicted by the casual relationship of the combined conditions of fluctuating wind velocity, increasing temperature, decreasing humidity and decreasing air pressure. This partitioning depends on the upper and lower parameters being set for a CPA and in order to gain the benefit of the improved confidence levels provided by the method, the bootstrapping and difference calculations are necessary. The *a priori* parameters superimposed on the wind velocity data in Figure 5 while arbitrary are based on the kind of intuition often used for CPA applications. When implemented, we expect a considerably more precise result using the algorithm referred to in this paper.

IV. CONCLUSIONS

The early stage research to which this paper refers is motivated by the quest to find innovative methods for determining change of state in continuous sample data. In particular the domain under investigation is generally known as *agrometeorology*, where wind velocity modeling and prediction is an imperative for crop management situations. A non-parametric method, *McNemar's Test for*

the Significance of Changes is described as being useful for discrete value data with the observation that an alternative is needed for continuous data. Change Point Analysis (CPA) is considered to have potential in this regard and following a conceptual description of its applications in finance and economics, a proposed use of it in a modified form for the analysis of multivariate continuous data is outlined. This application relates to wind velocity event state changes and a conceptual implementation of the method with some recently sampled near-ground truth event state data is described. An algorithm has been proposed to implement the CPA method within the context of rules derived from empirical observations of climate conditions immediately preceding a wind gust event. The development of this algorithm and its testing is the subject of the ongoing work for the research being undertaken to which this paper refers and will be reported in due course. Although not yet fully developed, the title of this paper indicates the quest for the research to produce such an algorithm.

REFERENCES

- [1] Taylor, W (2011) Change-point analysis: a powerful new tool for detecting changes. Taylor Enterprises Inc. Libertyville, USA. On-line at <http://www.variation.com/cpa/tech/changepoint.html>
- [2] McNemar, Q. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2): 153–157.
- [3] A worked example using patient treatment data can be seen at http://en.wikipedia.org/wiki/McNemar's_test
- [4] Efron, Bradley and Tibshirani, Robert (1993), An introduction to the Bootstrap. Chapman & Hall, New York.
- [5] Hinkley, David and Schechtman, Edna (1987) Conditional bootstrap methods in the mean-shift model. *Biometrika*, 74 1, 85-93.
- [6] Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42, 523-527.
- [7] Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42, 523-527.
- [8] Zhiheng, X. et al. (2010) Change Point Analysis. On-line at <https://sites.google.com/site/changepointanalysis/>
- [9] Sallis, P., Claster, W., and Hernandez, S. (2011) An algorithm for predicting wind gust events. *Computers and the Geosciences*, Elsevier [in print for 2011] Online at http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V7D52FVDJ01&_user=3166782&_coverDate=03%2F24%2F2011&_rdoc=1&_fmt=high&_orig=gateway&_origin=gateway&_sort=d&_docanchor=&view=c&_acct=C000059927&_version=1&_urlVersion=0&_userid=3166782&md5=93dcb002d356addf7689b0c0b1c3be56&se archetype=a
- [10] Delyam, A.M. Chaotic Climate Dynamics. Lunivar Press, 2007. ISBN-13 978-1-905986-07-1
- [11] Taylor, Wayne (2000a), Change-Point Analyzer 2.0 shareware program, Taylor Enterprises, Libertyville, Illinois. Web: <http://www.variation.com/cpa>