Alternatives to the bayesian change point biology essay

Science, Biology

Chapter 3

A time series is a sequence of values of a particular measure taken at regularly spaced intervals over time. Segments (split time series data) in a time series are defined when the sequence of measures is divided into two or more portions at change points. Change points are specific points in time where the values of the time series may exhibit a change from the previously established pattern. A change may be due to an ongoing stable process becoming unstable or a change due to an intervention measure that has been put in place. An algorithm for detecting change points using CUSUM CPA has been presented with an aim of showing how to detect changes due to epidemic and changes due to intervention. The outcome measure is the daily reported malaria cases at the health facilities. The epidemics were signaled if the change point detected in time series of malaria cases were outside the control limits while the detected change point due to intervention was felt if the cases of malaria were within the control limits. R statistical software version R 2. 14. 1 was used in the data analysis and to generate the graphs. In literature review we identified the following methods for change point detection; SPC methods for change point detection, frequentists methods which uses the likelihood ratio statistic tests and least square regression, Bayesian change point detection and non parametric change point detection methods. SPC methods for change point detection include CUSUM and Shewhart X chart. For the Statistical change point detection we have single change point detection for single hypothesis testing where it involves testing the hypothesis using likelihood ratio test. Multiple change point detection has multiple hypotheses testing using

likelihood ratio test. There are three search methods for multiple change point detection; Binary Segmentation, Segmented Neighborhood and Pruned Exact Linear Time. Instead of testing the hypothesis for change point, there exists a Bayesian change point detection method which uses the probability distributions - the probability of a change point at each location in a sequenceAlternatives to the Bayesian change point include the frequentist and the non-parametric CPA. For the frequentists they estimate specific locations of change points and they include Recursive Circular Binary Segmentation which uses likelihood ratio statistics test for hypothesis of no change point versus alternative hypothesis of change point. Test is applied recursively to splits data into different segments where there are change points. The other frequentist method is the Dynamic programming (Structural change model) which is an intercept regression which minimizes residual sum of squares) identifies optimal partitions with varying numbers of segments using both the minimum segment length, , and the maximum number of breaks, . Location of change point is estimated using least square regressionFinally we have the non-parametric CPA which is the CUSUM CPA. But in this particular study we used non parametric CPA (CUSUM CPA)

3. 2 Related works

In a study on change point detection problem, Taylor present CUSUM CPA for the mean of a process with abrupt changes (Taylor, 2010). Like in our study, they start with CUSM then combines it with bootstrapping to make inferences on the change point detected (Gavit et al., 2009; Kass-Hout et al., 2012; Sanghvi, 2008; Chang and Byun, 2011; Sallis and Hernandez, 2011).

The CUSUM CPA algorithm makes an assumption of normal and identical distribution. There are other varieties of change point detection techniques Bayesian and non-Bayesian approaches. There are other techniques for detection change points in variance too. Product partition model was used Bayesian change point detection technique by Barry and Hartigan (1993). The algorithm start with an estimator that assumes there is at most one change, and then use it to generate an approximate estimator in the general case (Erdman and Emerson 2007; Yigiter 2012). The frequentists approach to change point detection uses likelihood ratio statistics test to test for the hypothesis of no change point against alternative of single or multiple change points (Hawkins et al., 2005; Eckley et al., 2011; Killick et al., 2011). The method usually starts with single change point detection which then is extended to multiple change point detection. Furthermore, the method assumes normality of the data. Multiple change point detection was conducted using Binary segmentation and Pruned Exact Linear Time (PELT). There are other algorithms for detecting changes in variance of a process. They include the works of Kellick et al (2010) who have used penalized likelihood change point algorithm to detect changes in variance in oceanographic time series data. Structural change point detection technique has been used bay Bai and Perron (2003) to detect structural changes in times series data. The method aims to detect and estimate the break point and locate the change point using least square regression. CUSM and control charts have also been used in some studies to detect changes. Vries and Reneau (2010) used control charts to detect unplanned changes in animal production systems. CUSUM detection techniques detected changes by

raising an alarm when a given threshold was exceeded (McKelvie et al., 2012; Hawkins and Zamba 2005). Hay et al. (2002) also examined three simple techniques proposed for malaria epidemic detection (WHO, Cullen, and c-sum methods)Apart from CP detection Wagner et al. (2002) has used regression analysis with segmented to detect changes due to interventions. Wangdi et al. (2011) used time series plots in their study to show the trends in malaria cases and from the plots they concluded that the decrease in the trend could be attributed to the successful implementation of the control measures (introduction of ITN with IRS). However they were not able to detect the points of change. In this study we used non parametric change point detection technique which uses CUSUM plus bootstrapping so as to detect changes in mean of malaria cases.

3. 3 Description of Data and Data Sources

The data used in this study was extracted from daily routinely collected between the year 2010 and 2011 from an epidemic prone area for Eldoret East District which has five sentinel hospital sites. This data is usually collected on a daily basis from the health facilities consisting of confirmed malaria cases which are usually submitted to DOMC on a weekly basis for monitoring purposes. The use of IRS was launched between March 2010 and May 2010 while LLIN was launched between May 2011 and June 2011 as an intervention (vector control) to help reduce the malaria incidence. A separate data from one of the sentinel sites (Moiben Health Centre) for the year 2011 was used to demonstrate how to detect epidemics. Malaria epidemic detection threshold for Moiben Health centre was provided.

3. 4 Cumulative sum (CUSUM) and bootstrapping

Change-point analysis can be performed iteratively using a combination of cumulative sum charts (CUSUM) and bootstrapping to detect the changes (Taylor, 2010). It uses a recursive algorithm to detect multiple changes and the iterative algorithm decomposes the dataset into sub-datasets having different means. For each change point, CPA provides detailed information including a confidence level indicating the likelihood that the change occurred and a confidence interval indicating when the change occurred. In CPA, inferences is made via bootstrapping which is one of the methods of interval estimationChange point analysis is based on the mean-shift model. Let the user-defined values be as follows: (1) number of bootstrapping, (2) minimum confidence level, (3) level of confidence interval (CI) for CI (e. g. for 95% CI) and (4) is the significance level or p-value or the alpha level, CPA procedure is as shown below. To describe the mean-shift model, let represent the data in time order. The mean-shift model can be written as(3. 1)Where is the sample average at time as and is the residual term defined as for the observation. The cumulative sums of the residuals are calculated asCUSUM means cumulative sum, so we first calculate cumulative sum of the data, and then plot the sum. Let represent the n data points. The cumulative sums are calculated iteratively as shown in the following steps. Prepare the initial time series data and calculate the average as follows(3. 2)Start the cumulative sum at zero by settingCalculate the other cumulative sums recursively by adding the difference between current value and the average to the previous sum, i. e.(3. 3)forThe CUSUM series is then plotted. A sudden change in direction of the CUSUM either upward or negative

indicates a sudden shift or change in the average. This is the furthest point form 0 which also denotes the change point. A period where the CUSUM chart follows a relatively straight path indicates a period where the average did not change. The magnitude of change, , is defined as(3. 4)Where(3. 5)and(3. 6)Since we have already determined the estimator of the magnitude of the change, bootstrap analysis can be performed. For one to be sure that changes took place, a confidence level can be determined for the apparent change by performing a bootstrap analysis. Bootstrap analysis is performed as followsGenerate a bootstrap sample of n units, denoted as by randomly reordering the original values. This is called sampling without replacementBased on the bootstrap sample, calculate the bootstrap CUSUM, denoted byCalculate the maximum, the minimum and the difference of the bootstrap CUSUMDetermine whether the bootstrap difference is less than the original difference or notIterate the above procedure (1) - (4) N times and record the number of bootstrap samples which has less thanBootstrapping means, random reordering of the data. The benefit of using bootstrap is that, if no change has occurred in the data, the bootstrap samples will mimic the behavior of the CUSUM chart. Bootstrapping is a method for estimating the sampling distribution of an estimator by re-sampling with replacement from the original sample dataset. By performing large number of re ordering, we can estimate how much would vary if no change took place. Now we can compare this value with the of the original data to check the consistency. After generating for the original data and all the bootstrap samples, we can plot it on the same graph. If the CUSUM chart of the bootstrap samples is closer to zero than the CUSUM of the original samples then the change must

have occurred (Taylor, 2010). Confidence level is calculated by performing a large number of bootstraps and counting the number of bootstraps for which, . Let be the number of bootstrap samples performed and let be the number of bootstraps for which , then the confidence level (CL) at which a change point occurred as a percentage is(3. 7)Generally, at least 95% confidence is required before one concludes that a significant change has been detected. If the confidence level, it indicates that the detected change point is statistically significant and then we split the data into two subsets from this significant change point. If the confidence level, it indicates that detected change point is not statistically significant and we then stop the splitting. Let the initial datasetof sizeOnly when (), do the following steps: Get the time when a change occurred between the time and by change-point selection rule (CUSUM-maximizing). Calculate the associated bootstrapped CI of the time of change based on a selected bootstrapped CI formula (bootstrapped confidence interval, percentile interval and bias-correlated & accelerated interval). Brake the dataset into two segments (sub-datasets) and where one is each side of the change-point ; andCalculate the mean changes andThere are two change-Point selection rulesCUSUM-maximizing rule estimates the time of change as , if for a given sub-datasetMSE-minimizing rule estimates the time of changes as , if for a given sub-dataset where , and . Note that for. Once we are sure that change has occurred and we know the total number of changes, we can now go ahead and find out the estimate of when the change has occurred. One of such methods is the CUSUM estimator. To locate the location of change point, we define such that(3. 8)Is the point the point which is at the maximum distance (farthest) from zero in the CUSUM

chart. The point estimates the last point before the occurrence of the change point.

3. 4. 1 Confidence Level for Change

Each change detected by the change-point analysis has a confidence level associated with it. ?? The confidence level ranges from . ?? The higher the confidence level, the greater the certainty that a change took place and confidence is required to state that the change is significant. The confidence level is based on the significant level, alpha-level or p-value. ?? The p-value is the probability that the observed effect could have been due to the natural variation in the data assuming no change took place. Generally, the smaller the p-value the greater the evidence of changeThe confidence level is given by and in summary. ??

Confidence Level

p-value

99%0. 0195%0. 0590%0. 180%0. 250%0. 5Table 3. 1 Confidence level with associated p-value

Chapter 4

Results

4. 1 Sample description

The CUSUM CPA method was applied to a time series of malaria cases from Eldoret East and Moiben Health facility from 2010 to 2011. There were 7545 cases of malaria from Eldoret East district and 4417 cases from Moiben health facility. The total reported cases varied between the sentinel sitesThe

change points detected using CUSUM CPA for the cases of malaria and statistical inferences (i. e., confidence levels) are provided for each change point. In the CUSUM CPA procedure, the CUSUM of the different data points are calculated then plotted and the location and the magnitude of change is then determined. The location of the change point is the furthest CUSUM point from zero. This is a non-parametric method of locating and estimating the change point. Once the significant level of 95% has been determined, the data set is split into 2 segments at the points where the change has been located and the CUSUM CPA procedure is repeated until no more significant changes is detected.

4. 2 Detecting outbreak change point

The initial step of change point detection begins by calculating the CUSUM, then plotting the CUSUM and finally locating the furthest CUSUM point from zero was around week 30 as show in the figure 4. 1 below. Figure 4. 1 CUSUM plot for Moiben malaria casesFrom the CUSUM plot the change point is the furthest point from zero indicating that a change had occurred and the magnitude of change was 561. 58. We go ahead and determine whether this change point is significant or not via bootstrapping. 1000 bootstrap samples were generated and it yielded a confidence level of 99. 8%, the location of the change point was week 30 since it had the highest CUSUM value (454. 73). After detecting the first significant change point the time series data for malaria cases is split into 2 segments; from week 1 to week 30 and from week 31 to week 52. Figure 4. 2 CUSUM plot for the split time series data 1The analysis was repeated on each of the two segments to determine their

change points. Performing CPA on the first split data between week 1 and week 30 gave the magnitude of change as 424. 2, and a confidence level of 99%. The location of change point was at week 21 with a CUSUM value of 402. 1. This result showed that there was a negative change point as show in the split CUSUM plots below and since the change was significant we split the data set in two from week 1 to week 21 and from week 22 to week 30 and perform the CUSUM CPA. Figure 4. 3CUSUM plot for the split time series data 2Performing CPA on the second segment of the time series data between week 31 and week 52 gave the magnitude of the change as 159. 82 and a confidence level of 98%. The location of the change point was week 45 with CUSUM value of 136. 91. This therefore means that the second change was at week 45 and the data is split into two segments between week 31 to week 45 and between week 46 to week 52 and CUSUM CPA is then conducted. Figure 4. 4 CUSUM plot for the split time series data 3Performing CUSUM CPA on the time series data set between week 31 and week 45 gave 87. 6 as the magnitude of change. The change produced a confidence level of 93% and the location of change point was at week 37 with CUSUM value of 86. 8. Since the change point was not significant, we therefore stop splitting the data setThe table 4. 2 below summarizes the magnitude, confidence level and the location of the change points for Moiben Health Centre using the CUSUM CPA

Year

Magnitude

Confidence Level

CP Location

2011561. 5899. 8%Week 302011159. 8298. 6%Week 452011424. 299%Week 21Table 4. 2 Change points for the cases of malaria from Moiben health centerIn summary, CUSUM CPA detected only three change points within the data set at week 30, week 45, and week 21 which had confidence level above 95%. Those are the points in the week of the year where the cases of malaria were significantly out of control (epidemics). Trend chart/ run chart is similar to control charts but do not show the control limits of the process. The trend chart showed the basic trends of weekly malaria cases for the year 2011 generally in a decreasing and increasing trends. The same data set was subjected to X-bar control charts and one was able to detect four points which were out of the control limits that had been set. The Upper Control Limit was set at 153. 4133 and the Lower Control Limit was 16. 4713. The points out of control were at week 22, week 24, week 25 and week 28. The significant level and the magnitude of change for those points were however not given. The control charts generally give the idea of the change point. Figure 4. 5 Control chart for Moiben malaria casesUsing the CUSUM control charts 28 point were detected to be out of control with upper decision interval being and lower decision interval being and shift detection (standard error) being equal to one. CUSUM chart could detect points missed by the Shewart (X-bar) control chart could not detect. In the CUSUM control

chart, each point plotted represents the difference between an individual data point and the mean, which is added to or subtracted from the previous point on the graph (depending on whether the difference between the individual data point is positive or negative). Control charts only shows you the change but not the exact place and the estimate of the change, it is therefore hard to interpreting the change and hence control charts are not the best option. We therefore use change point analysis to detect, locate and estimate the significant changes together with their corresponding confidence levels. Change-point analysis basically builds on a CUSUM chart by determining a confidence level for each change. The same data set was exposed to the conventional method of detecting malaria epidemic using the 3rd quartile (WHO) and mean1. 5SD (Cullen). Figure 4. 6 Malaria epidemic detection system for Moiben health centreFrom the analysis there were 18 cases above the action threshold and 34 cases above the alert threshold. However, one was not able to know whether the changes (exceeding the threshold) were significant or not since the procedure depends on previous five-year data. The 3rd quartile was used as an alert point and mean1. 5standard deviation was the action point. This is the conventional method that DOMC uses to detect the epidemic. The epidemic detection method uses the five year previous data to set the alert and the action thresholds which at times may not be reliable. Figure 4. 7 Time series plot for Moiben malaria cases with the location of change pointsThe figure 4. 7 above shows the time series plot of together with the detected significant change points

4. 3 Detecting Intervention (Impact) change point

Our study provides the first, to our knowledge, observation of a reduction in malaria incidences (cases) following introduction of IRS and LLITNs solely in a stable malaria-endemic setting. There were several change points detected before and after the launch of IRS in 2010 and LLITNs in the year 2011as shown in the figure 4. 8 below. Figure 4. 8 Trends of malaria cases for Eldoret East year 2010 and 2011Figure 4. 9 CUSUM plot for Eldoret East malaria cases year 2010 and 2011The above Figure 4. 9 shows the CUSUM plots for malaria trends from January 2010 to December 2011 for Eldoret west District. This is the period where there were the two interventions IRS and LLITNs. There are several change points in the trends for the malaria cases. From the CUSUM plot a change had occurred (the furthest point from zero ??? around week 28) and the magnitude of change is 2494. 06. We go ahead and determine whether this change point is significant or not via bootstrapping. 1000 bootstrap samples were generated and it yielded a confidence level of 99. 6%, the location of the change point was week 28 year 2011 since it had the highest CUSUM value (2453). After detecting the first significant change point the time series data for malaria cases is split into 2 segment; from week 1 year 2010 to week 28 year 2011 and from week 29 year 2011 to week 52 year 2011. The CUSUM CPA analysis was repeated on each of the two segments to determine their change points. Figure 4. 10 CUSUM plot for the split time series data APerforming CPA on the first split data between week 1 and week 28 year 2011 gave 1665. 8 as the magnitude of change, and a confidence level of 99%. The location of change point was at week 42 year 2010 with a CUSUM value of 1030. The data set is

the split into two segment (week 1 to week 42 and week 43 to week 28) since the change point is significant. Figure 4. 11 CUSUM plot for the split time series data BPerforming CPA on the second segment of the time series data between week 29 year 2010 and week 52 year 2011 gave the magnitude of the change as 805. 25 and a confidence level of 98%. The location of the change point was at week 48 year 2011 with CUSUM value of 412. This therefore means that the second change was at week 48 year 2011 and the data is split into two segments from week 29 to week 48 year 2011 and from week 49 to week 52 year 2011 and CUSUM CPA is performed. Figure 4. 12 CUSUM plot for the split time series data CPerforming CPA on the time series data set between 29 to week 48 year 2011 gave 639. 65 as the magnitude of change. The change produced a confidence level of 99% and the location of change point was at week 38 year 2011 with CUSUM value of 593. The data set was split into two segments and CUSUM CPA procedure is then repeated. CPA on time series data between week 49 to week 52 year 2011 gave 74. 75 as the magnitude of change and a confidence level of 55%. The location of change point was at week 51 year 2011 with a CUSUM value of 75. Since the point of change was not significant (CL <95%) we stop splitting the data set further. Figure 4. 13 CUSUM plot for the split time series data DPerforming CPA on the split data set from week 39 year 2011 to week 48 year 2011 gave a magnitude of change to be 219. 4 and a confidence level of change to be 95%. The location of change is at week 44 year 2011 with a CUSUM value of 219. The data set is the split into two segments from week 39 year 2011 to week 44 year 2011 and a second data set from week 45 year 2011 to week 48 year 2011 and CPA procedure

is repeated. For the data set for week 39 to week 44 year 2011, magnitude of change was 52. 67 and a confidence level of 76%. The location of change was week 43 year 2011 with CUSUM value 30. 3. The data set is then not splitted further since the change point detected is not significant. Figure 4. 14 CUSUM plot for the split time series data ECPA on data set for week 45 year 2011 to week 48 year 2011 gave a 33. 5 as magnitude of change, confidence level of 93% and the location of change was at week 46 year 2011 with CUSUM value being 34. The data is not split further since the change point is not significant. CPA for week 1 to week 42 year 2010 gave the location of change point to be at week 14 year 2010 with CUSUM value of 979, magnitude of change being equal to 1026. 57 and 98% confidence level. CPA for week 43year 2010 and week 28 year 2011 gave a confidence level of 72%, the location of change point was at week 4 year 2011 with a magnitude equal to 851. 16. For the data set for the period between week 15 and week 42 year 2010, the change point was located at week 26 year 2010 with a magnitude of 607. 75, CUSUM value of 351 and confidence level of 98%. The data set is the split into two segments and the procedure is then repeated until no significant change points are located. The table 4. 3 below summarizes the change points located or detected by the CPA

Year

Magnitude

Confidence Level

CP location (Week)

20112494. 0699%282011805. 2598%482011639. 6599%3820101665. 899%4220101026. 5798%142010607. 7598%26Table 4. 3 Change points for Eldoret East malaria cases year 2010 and 2011Subjecting the data set into control chart shows that the points within the control limits are in control while those outside the control limits are out of the process. From the control charts, 10 weeks are out of the control limits while 94 weeks are within the control limitsFigure 4. 15 Control chart for Eldoret East malaria cases year 2010 and 2011The figure 4. 15 above shows the control chart with CL= 499. 4423, LCL= 301. 8333, and UCL= 697. 0513. There were 10 points out of limit, while the remaining 94 points were within the control limits. Further research could be done to ascertain the changes in the trends of malaria cases. Figure 4. 16 Time series plot for Eldoret East malaria cases with the significant change pointsThe figure 4. 16 above shows malaria cases together with the detected change points