Nora Lin
IST 772 Fall 2021
Final

Introduction:

To investigate how the state legislator's office should allocate financial assistance to the school districts, we use three data sets. First, the World Health Organization reporting vaccination rates in the U.S collected from 1980 to 2017. The Vaccines studied were Diphtheria/Pertussis/Tetanus (DTP1), Hepatitis B, Birth Dose (HepB_BD), Polio third dose (Pol3), Influenza third dose (Hib3), and Measles first dose (MCV1). Second, a list of California kindergartens and whether they reported vaccination data to the state in 2013. Third, a sample of California public school districts from the 2013 data collection, along with specific numbers and percentages for each district. Exploratory and predictive analysis was conducted in order to help answer the legislator's office's question.

Report Structure:

Report 1 : Examining the distributions of U.S vaccination rates for DTP1, Pol3, Hib3, MCV1 and HepB_BD

***Comparison of spread*:** we can see that the boxplot heights for DTP1, Pol3, Hib3 and MCV1 are significantly smaller than that of HepB_BD. This means that vaccination rates for HepB_BD has a wider spread and more variable than the other four vaccines.

***Comparison of quantiles*:** the upper and lower quantiles for DTP1, Pol3, Hib3 and MCV1 all are around 80 to 90, respectively, while the that for HepB_BD are 17 and 54.5, respectively.

Comparison of skew: the median for DTP1, Pol3, Hib3, and MCV1 are all greater than their respective means. This indicates that the distribution of vaccination rates for these four vaccines are negatively or left skewed. In contrast, the mean is greater than the median for HepB_BD. This indicates that the distribution of vaccination rate is positively or right skewed.

Summary: the distribution for vaccination rates HepB_BD across 1980-2017 does not follow the same distribution for vaccination rates of the other four vaccines: DTP1, Pol3, Hib3 and MCV1.

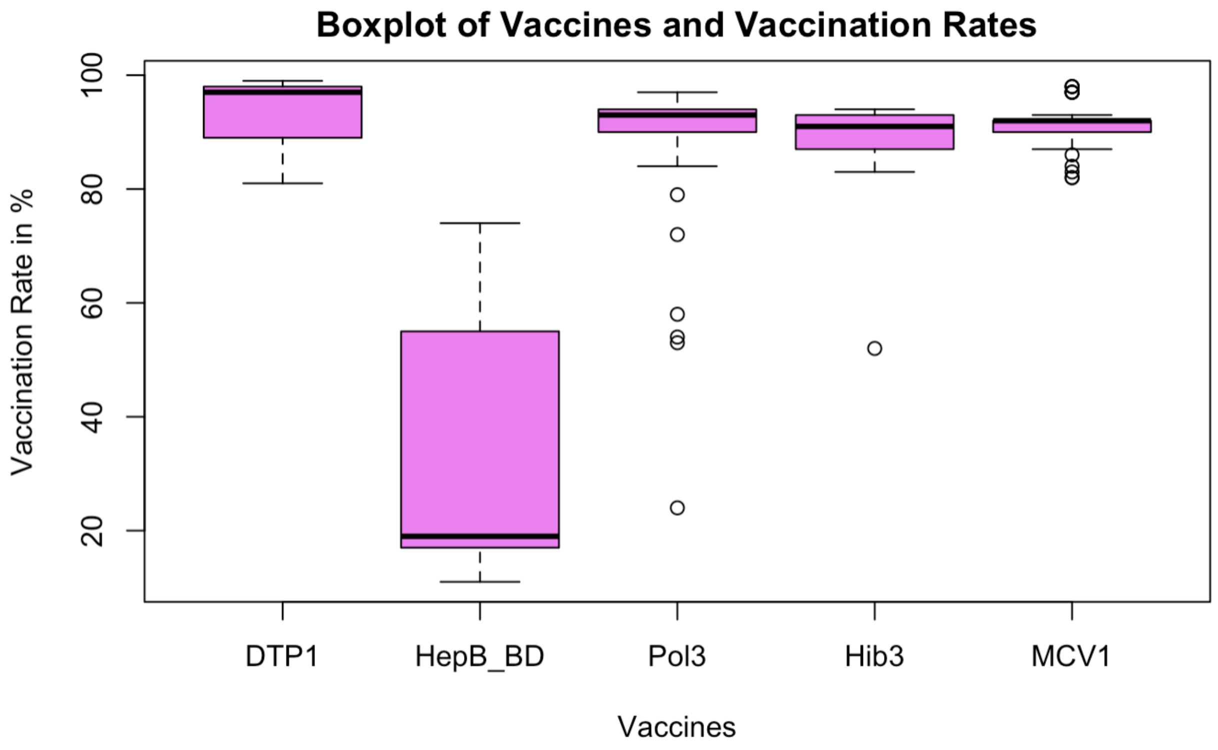


Figure 1: Boxplot of Vaccines and Vaccination Rate.

Report 2 : Examining how U.S vaccination rate varies over time

***Vaccination rate over time comparison*:** similar to our report of the distribution of vaccination rates, HepB_BD has the lowest vaccination rate in 1980. At the end of the time series data, in 2017, DTP1 and HepB_BD both start with low vaccination rates than follow a general upward trend, increasing from 1980 to 2017 by 15% and 48% respectively. Sin 1987, the overall trend for Hib3 vaccination rates from 1980 to 2017 are tightly bound between 85% and 94%. Similarly, sin 1987 where the vaccination rate for Pol3 drops down to 24%, the overall trend for Pol3 starts at 95% concaving between 1988 to 1991 then remains in the 90%s for the duration for the time series. In contrast, overall the vaccination rates for MCV1 was highest between 1982-1988, then experiences some volatility, than makes a gradual increase from 1990-2017. At the end of the time series, in 2017, HepB_BD had the lowest vaccination rate at 64% and DTP1 had the highest vaccination rate 98%.

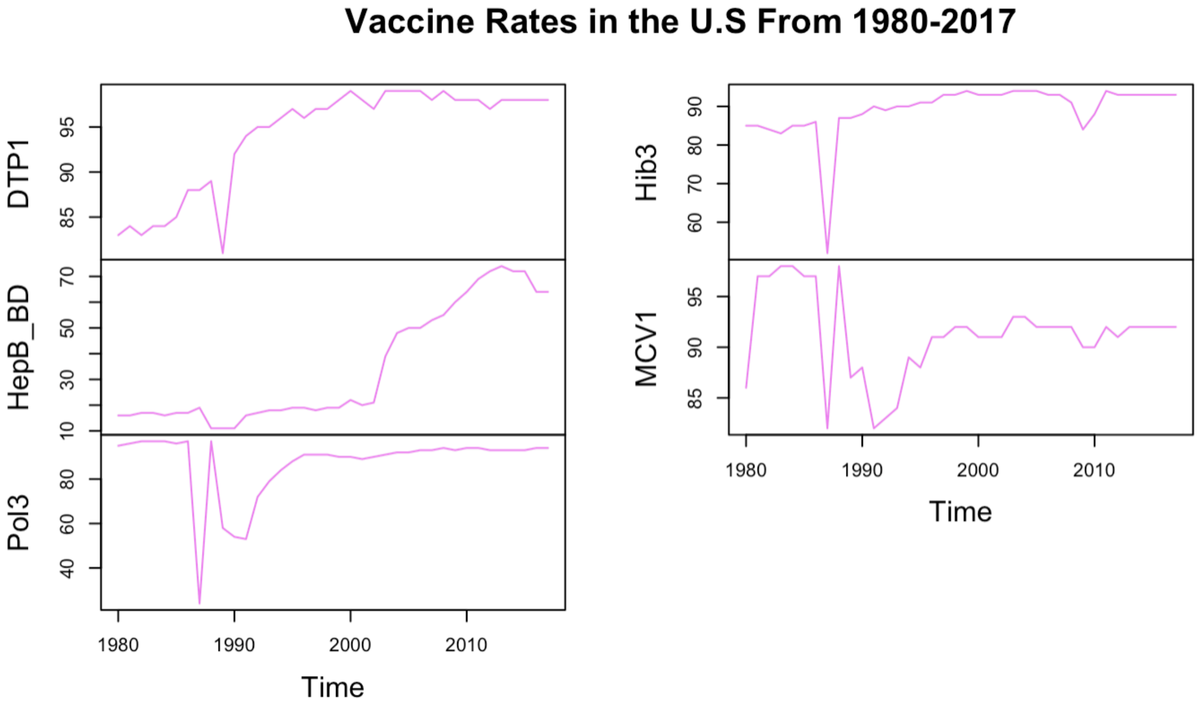


Figure 2: Vaccine rates in the U.S from 1980-2017

***Volatility comparison*:** Figure 3 shows that across all five vaccines, there's a period of volatility between 1985 and 1990. DTP1, Pol3, Hib3 and MCV1 all level out around 0 while HepB_BD has another peak around 2000 and 2005.

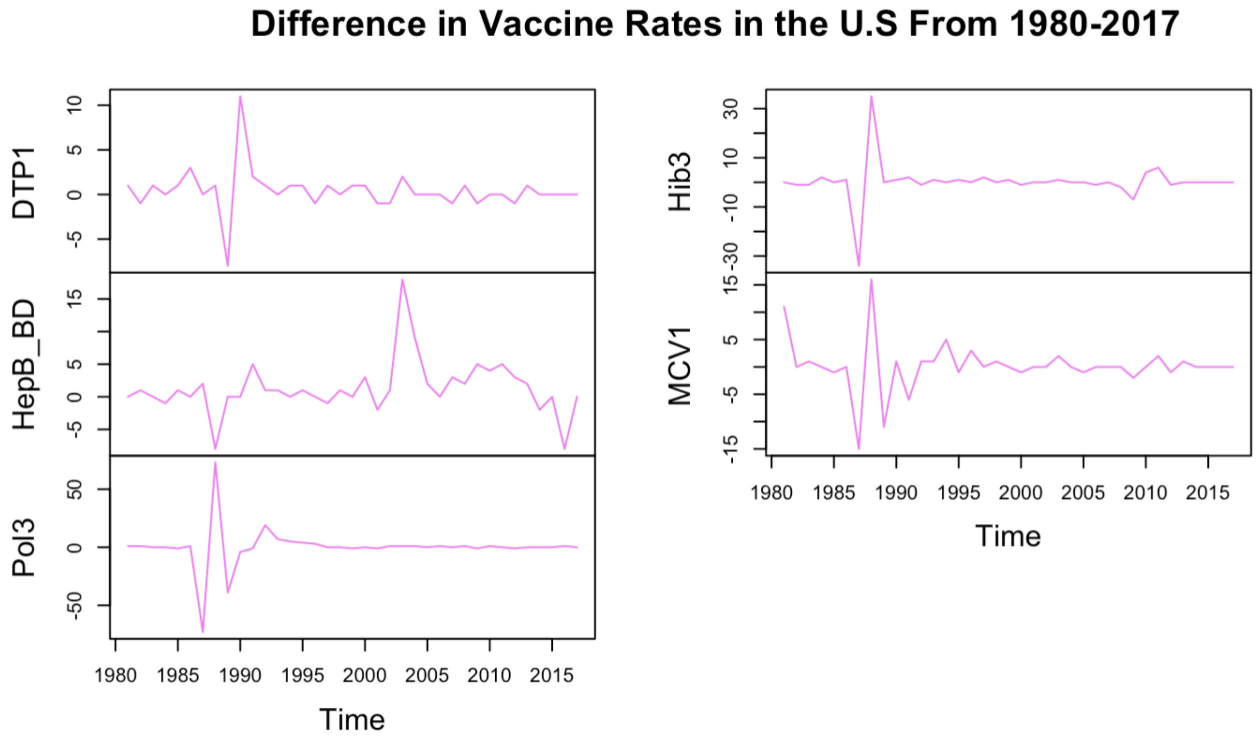


Figure 3: Difference in Vaccine Rates in the U.S From 1980-2017

Augmented Dickey-Fuller Test Results: the adf.test( ) procedure tests for stationarity. The resulting tests for adf.test(diff(VaccineRates[,"DTP1"])) and adf.test(diff(VaccineRates[,"Hib3"])) are significant so we reject the null hypothesis that the time series is non-stationary and accept the alternative hypothesis that the time series for DTP1 and Hib are stationary. However, the adf.test( ) for that of HepB_BD, Pol3, and MCV1 showed p-values greater than our alpha level and so fail to reject the null hypothesis that the time series is non-stationary.

Change-point analysis: the change point analysis shows the times where series mean values shifted. The changepoint for DTP1, Hib3, and MCV1 were all before 1990 while that of Pol3 were between 1990-2000 and HepB_BD was after 2000.Figure 1-5 in the Appendix show the change points for the series mean values shifts.

The change point analysis also shows the times where difference variance shifted. The change points where difference variance shifted for DTP1 and Hib3 occurred around 1990, while that of Pol3 and MCV1 was around 1996. Figures 6-9 in the Appendix show these change points. The figure below shows the change point for HepB_BD. Note there is no change point. This show the volatility of the difference in variance of the vaccination rates for HepB_BD.

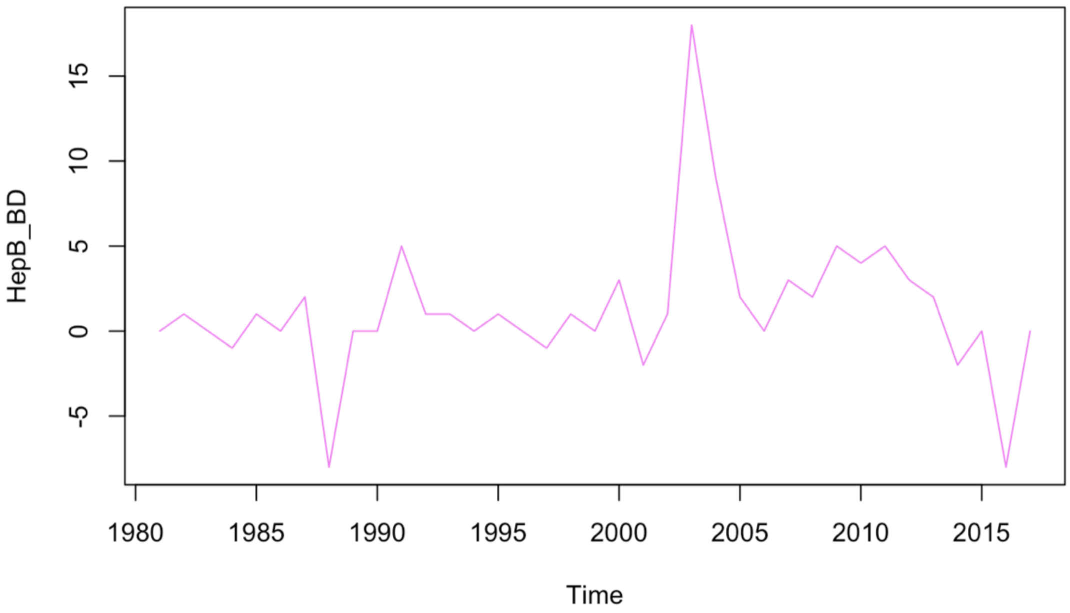


Figure 4: No change points showing difference variance shift for HepB_BD

Report 3: Examining a list of California kindergarteners and whether they reported vaccination data in 2013

All Schools Analysis: Data was acquired on 7,381 kindergarten classrooms in the state of California. Out of 7,381 schools, 6,981 school or 94.58% of total schools reported their vaccination data.

Public Schools Analysis: There were 5,584 public schools in the data acquired. 5,732 schools or 97.42% of those schools reported their vaccination data.

Private Schools Analysis: There were1,649 private schools in the data acquired. 1,397 or 84.71% of those schools reported their vaccination data.

Public vs. Private schools: to investigate whether there is a credible difference in overall reporting proportions between public and private schools, we use a chi square test for independence. The chi-squared returned a x-value of 402.49 with 4 degree of freedom. The p-value was 2.2e-16. Since our p-value is smaller than our alpha we reject the null hypothesis that school type (public or private) is independent from reporting status.

Since we used the frequentist approach, we will also use the Bayesian test on our variables. We used a Markov Curve/Monte Carlo (MCMC) simulation to examine if public/private school type is independent from the reporting status of the school. The mean Y/N proportion for public school was 37.4701 and median of 37.5724 with a 95% HDI of 32.098 to 44.404. The mean Y/N proportion for private school was 5.5510 and median of 5.534 with a 95% HDI of 4.8573 to 6.3304. Since the HDI intervals for public and private schools do not overlap, this further supports rejecting the null hypothesis that the public/private school types is independent of the reporting status for the school. Figure5 and 6 below show the distributions of the proportions for public and private school, respectively.

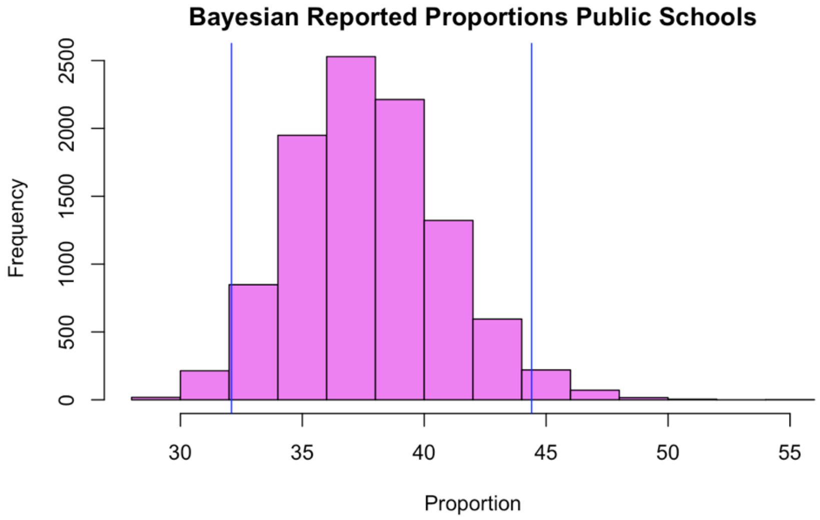


Figure 5: Distribution of public School Y/N proportion

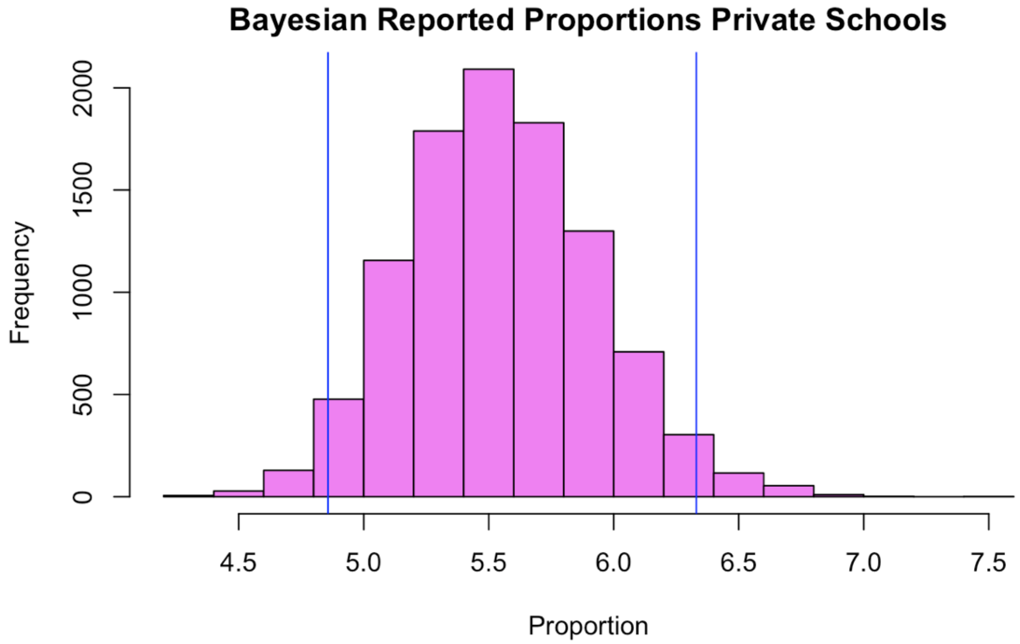
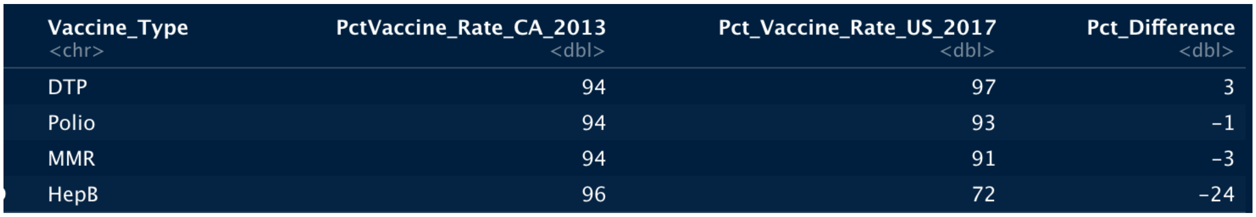


Figure 6: Distribution of private school Y/N proportion

Report 4: Examining California Public School Vaccine Rates for each vaccine in 2013

Vaccination Rates: The vaccination rate in 2013 for California public school for DTP1 was 93.59%, that of MMR was 93.61%, that of Pol3 was 94.06%, and that of HepB_BD was 95.93%.

Comparison against US 2017 vaccination rates: The vaccination rates for Pol3, MMR, and HepB were lower than the vaccination rates in the US in 2017, the last year of the given time series, by 1%, 3%, and 24%, respectively. The vaccination rate in California public schools in 2013 was higher than the vaccination rate in the US in 2017 by 3% for the DTP1 vaccine. The table below shows these results.



Vaccine_Type <chr>	PctVaccine_Rate_CA_2013 <dbl>	Pct_Vaccine_Rate_US_2017 <dbl>	Pct_Difference <dbl>
DTP	94	97	3
Polio	94	93	-1
MMR	94	91	-3
HepB	96	72	-24

Table 1: Comparison between Vaccination Rate in California public schools in 2013 and Vaccination Rates in the US in 2017

Report 5: Examining vaccination rates among districts are related

District 13: The chart below shows that all vaccines are related to each other. A student who is missing one vaccine is positively correlated with them missing another vaccine. A student who did not have the DTP vaccine had 0.983 correlation to not having the Pol3 vaccine. These correlations can be seen in the table below. The highest correlation is between Pol3 and MMR at 0.968. The lowest correlation is between HepB and DTP at 0.895.

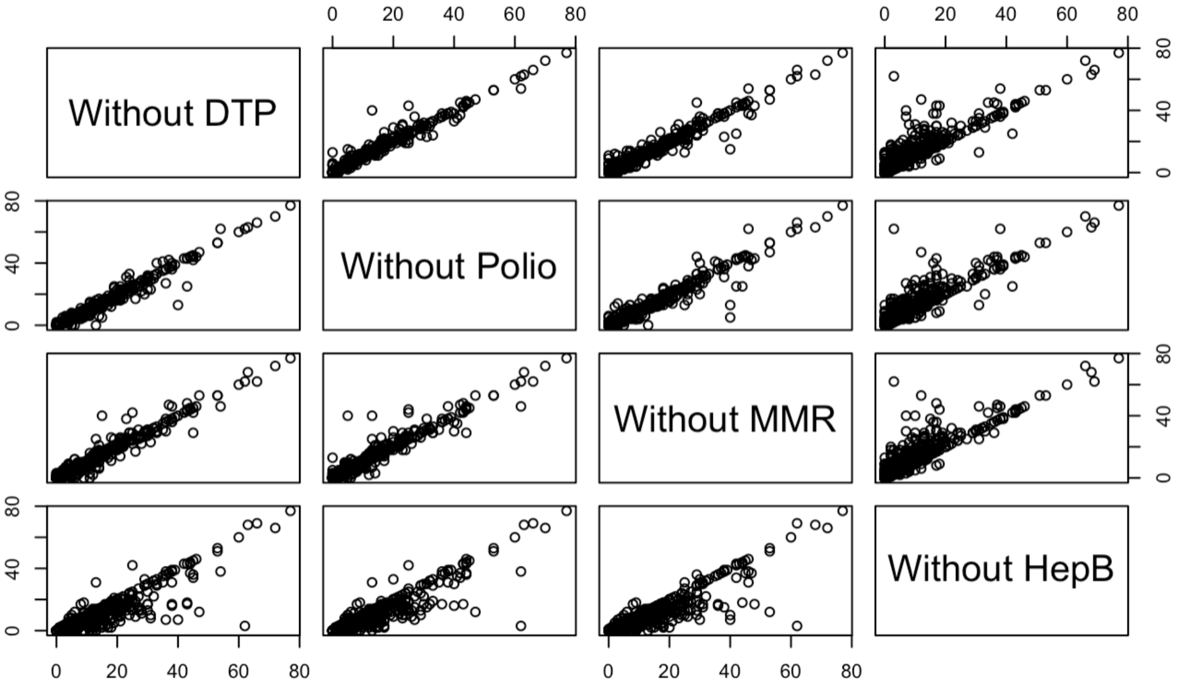
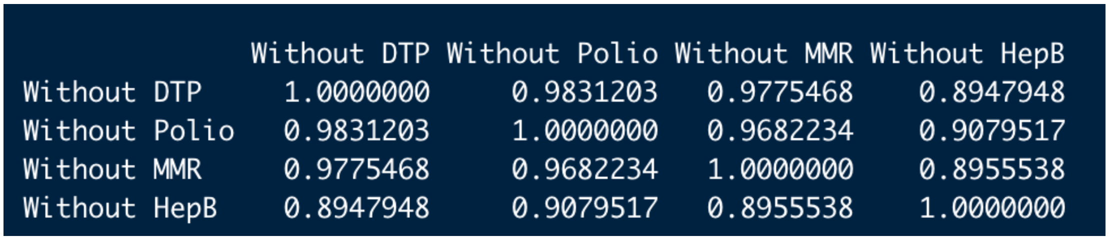


Chart 1: Correlation between each pair of vaccines in District 13



```
              Without DTP Without Polio Without MMR Without HepB
Without DTP     1.0000000     0.9831203   0.9775468   0.8947948
Without Polio   0.9831203     1.0000000   0.9682234   0.9079517
Without MMR     0.9775468     0.9682234   1.0000000   0.8955538
Without HepB    0.8947948     0.9079517   0.8955538   1.0000000
```

Table 2: Correlations between each pair of vaccines in District 13

Pearson's product-moment correlation test: the t-value varied between the different pairings of the vaccines, but the p-value remained the same for all six cor.test( ) of 2.2e-16. Since this value is smaller than alpha we reject the null hypothesis that there is no correlation between the vaccines.

Report 6: Predictive analysis on whether or not a district's report was complete

Chi-Square Omnibus test: There are 700 school sample, there are 657 schools or 94% of schools that have completed their reporting and 6% who did not. We will run a logistic regression model to analyze which attributes would be the best predictors for whether or not a district's report was complete. We use a chi-square omnibus test to see whether the test on predictor is statistically significant.

For percent child poverty, the test returned chi-square(1) = 3.2817, p-value= 0.07006
For percent free meal, the test returned chi-square(1) = 7.8829, p-value =0.00499**
For percent family poverty, the test returned chi-square(1) = 5.747, p-value=0.01652*
For enrolled, the test returned chi-square(1) = 13.695, p-value=0.000215 ***
For total schools, the test returned chi-square(1)=19.586, p-value=9.618e-6***

Wald's z-test: for enrolled students, the z-test was significant at z=3.64, p-value =0.001557. for school counts the z-test was also significant at z=-3.589, p-value =0.000332. These low p-values for the Wald's z-test also show that enrolled students and total students are strong predictors for whether or not a district's report was complete.

Log-Odds:
The total schools is the strongest predictor followed by enrolled, because their levels of significance reported by the chi-square omnibus test. The odds for total schools is 0.9615. This means that for each additional schools, the odds of completing reports drops by 0.9615:1. Similarly, the odds for enrolled is 0.9996. This means that for every additional enrolled student the odds of completing reports drops by 0.9996:1.

Bayesian MCMC simulation: The traces for both predictors look normal since there are no high or low spikes. The coefficient density charts also show normal distributions centered near coefficients generated by our logistic modeling. In addition, the odds distribution charts show HDIs that don't overlap .The Bayesian simulation supports rejecting the null hypothesis that changing the enrolled and/or total schools doesn't change the odds of a districting reporting complete vaccination reports.

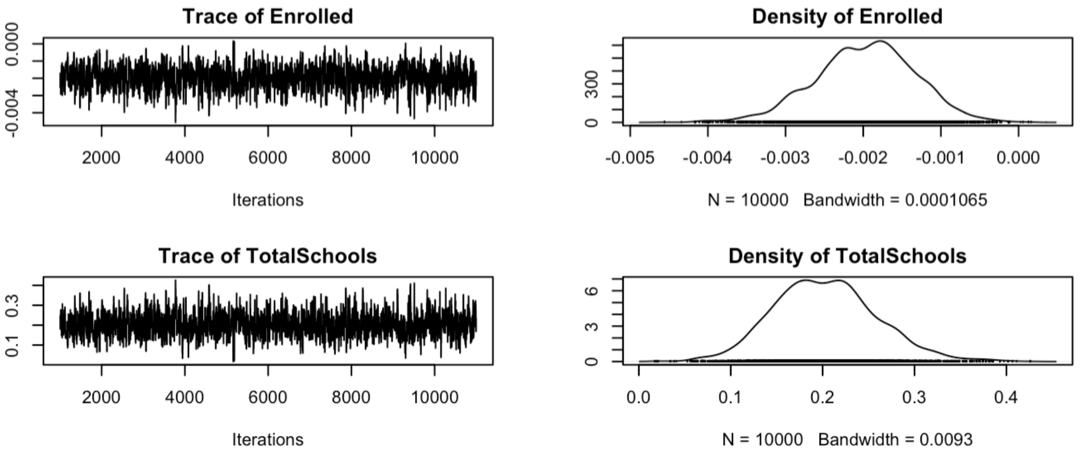


Chart 2: Bayesian MCMC simulation on enrolled and total schools as predictors for whether or not a district's report was complete

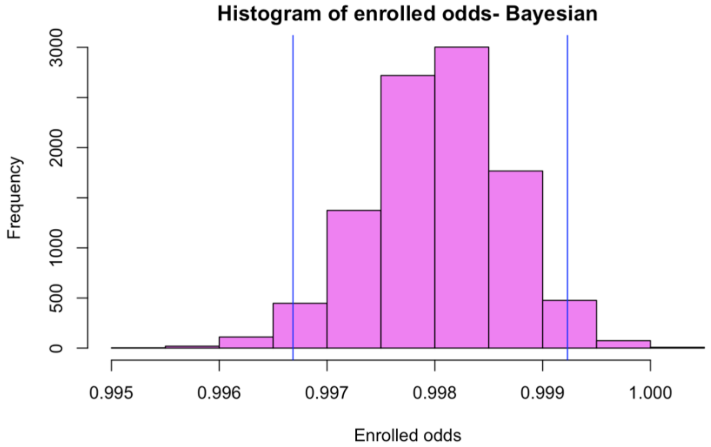


Figure 7: Histogram of enrolled odds

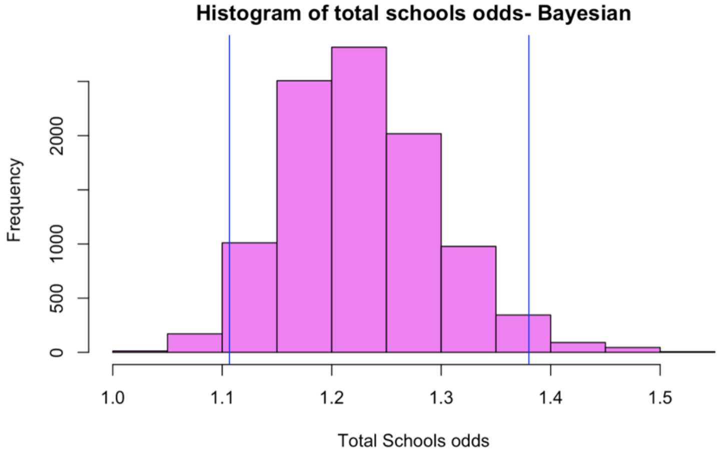


Figure 8: Histogram of total schools odds

Summary: the strongest predictors for whether or not a district's report was complete are total schools and enrolled.

Report 7: Predictive analysis on the percentage of all enrolled students with completely up-to-date vaccines

Linear regression modeling: We want to use a linear regression model to see which predictor would best predict the percentage of all enrolled students with completely up-to-date vaccines.

The modeling with just one predictor, PctFamilyPoverty as predictor was significant with a F-statistic of 45.3, t-value=6.731, p-value=3.529e-11, that using PctFreeMeal as predict was also significant with a F-statistic of 48.15, t-value=6.939,p-value=9.015e-12, and that using PctChildPoverty as a predictor was also significant with F-statistic of 31.15, t-value=5.582, p-value=3.413e-08.

The model using two predictors jointly with PctFamilyPoverty and PctFreeMeals as predictors was significant. The PctFamilyPoverty returned a coefficient of 0.2104, t-value of 2.508, and p-value of 0.01238. The PctFreeMeals returned a coefficient of 0.08128, t-value of 2.998, and p-value of 0.00281.

Odds: The odds for PctFreeeMeal is 1.14. This means that for each additional percent in free meals, the odds of completing reports drops by 1.14 :1. Similarly, the odds for PctFamilyPoverty is 1.48. This means that for each additional percent in family poverty, the odds of completing reports drops by 1.48 :1.

Bayesian MCMC simulation: we run a Bayesian simulation to confirm our findings. The HDI for PctFreeMeal ranged from 0.0265 to 0.1330. the HDI for PctFamilyPoverty ranged from 0.0456 to 0.366. Neither predictors' HDIs included zero. This means that there is a 95% chance that the coefficient is not 0. We can therefore, reject the null hypothesis that there is no relationship between PctFreeMeal and PctFamilyPoverty with the percent of students up to date with vaccination.

Report 8: Predictive analysis on the percentage of all enrolled students with belief exemptions

Linear regression modeling: after running a linear regression model for each of the predictors, PctFreeMeal was the strongest predictor. The test resulted in F-statistic of 69.47, t-value=-8.335, and p-value of 4.12e-16. This supports the coefficient is significant and we an reject the null hypothesis that there is no relationship between belief exemptions and percentage of free meal. As a note, PctChildPoverty was also a good predictor with F-statistic of 23.58, t-value of -4.856, and p-value of 1.48e-06. Lastly, PctFamilyPoverty was also a good predictor with F-statistic of 43.72, t-value of -6.612 and p-value of 7.53e-11.

Bayesian MCMC simulation: the HDI for PctFreeMeal ranged from -0.1340 to -0.08252. The HDI do not include zero, which means there is a 95% chance that the coefficient is not 0. We can therefore, reject the null hypothesis that there is no relationship between PctFreeMeal with the percentage of students with belief exemptions.

Recommendations to the state legislator's office with regards to how to allocate financial assistance to school districts to improve both their vaccination rates and their reporting compliance.

In comparing the data collected from the districts in 2013, there is room for improvement. The rates for individual vaccines in California lagged behind the overall US vaccination rates. The analysis showed public schools reporting was complete 97.42% of the time while private schools reporting was complete 84.71% of the time. That is a difference of 12.71% between the different types of schools. At a district level, our analysis showed that total number of schools in the district as well as total number of enrolled students in the district were predictors of whether a district's report was complete. For each additional schools, the odds of completing reports drops by 0.96:1. Similarly, for every additional enrolled student the odds of completing reports drops by 0.9996:1. The number of enrolled students per district is more predictive of the reporting completion.

Action-items:

- The district should focus resources to better understanding the gap between public and private school's vaccine reporting completion. The current study and collected data is not enough to understand the driving factors.
- The district should also focus financial resources to districts with higher number of enrolled students, especially if the number of total schools in that district is also high. The resources should be used specially for vaccine reporting and not other programs.

In examining factors to improve the districts vaccination rates, our analysis showed that percent family poverty and percent free meals were predictors for whether students were completely up to date on their vaccines. For each additional percent in free meals, the odds of completing reports drops by 1.14:1. For each additional percent in family poverty, the odds of completing reports drops by 1.48:1. The percentage of family poverty is more predictive of whether a student is completely up to date on their vaccines. This makes sense as our other finding showed that if a student is missing one vaccines that they are missing all vaccines. For districts with a higher percentage of families living under the poverty line, their children might miss one vaccine but that also means they probably are missing another vaccine.

Action-items:

- Focus financial resources to districts with higher percentage of families under the poverty line and to educate and encourage them to receive vaccines. If a child has one vaccine, they are more likely to get the other vaccines. However, if a child is missing one vaccine, they are more likely to be missing all the other vaccines. Likewise, districts should also focus financial resources on districts with higher percentage of free meals.

APPENDIX:

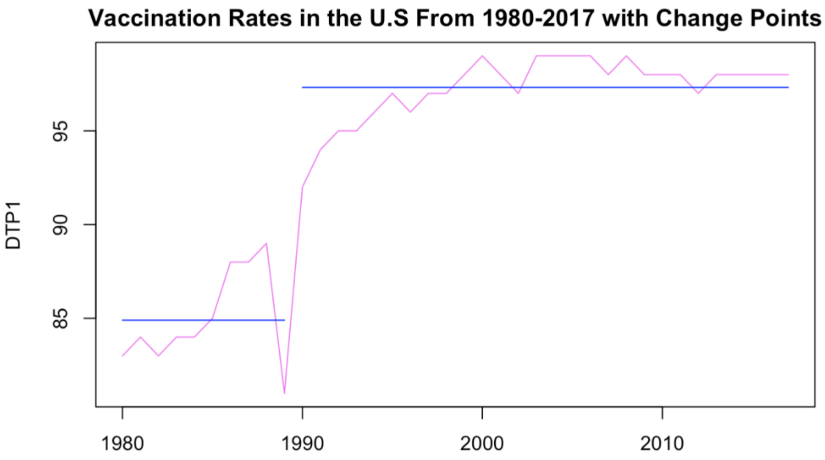


Figure 1: Change points showing times where series mean values shifted for DTP1

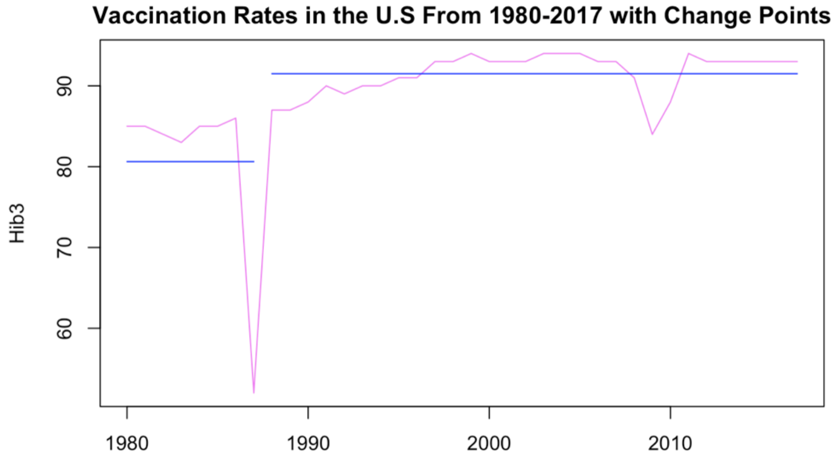


Figure 2: Change points showing times where series mean values shifted for Hib3

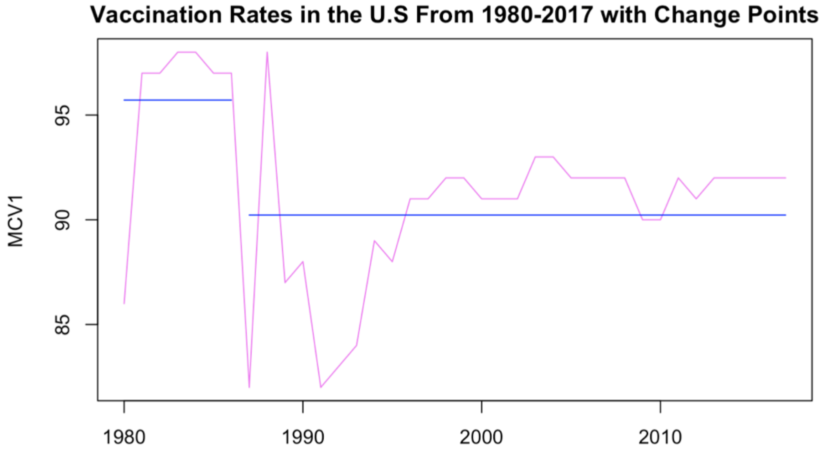


Figure 3: Change points showing times where series mean values shifted for MCV1

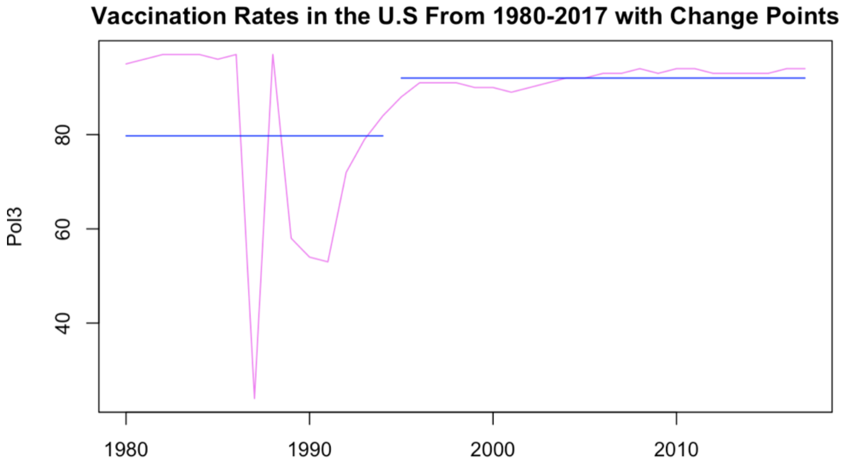


Figure 4: Change points showing times where series mean values shifted for Pol3

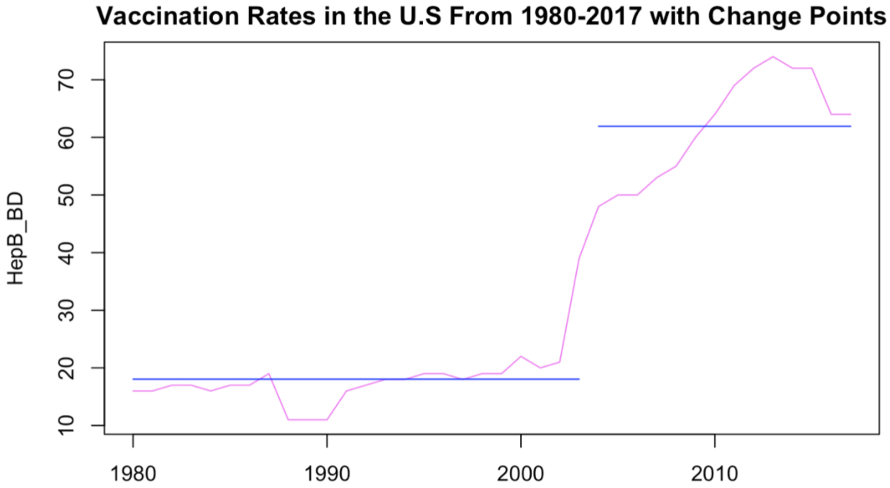


Figure 5: Change points showing times where series mean values shifted for HepB_BD

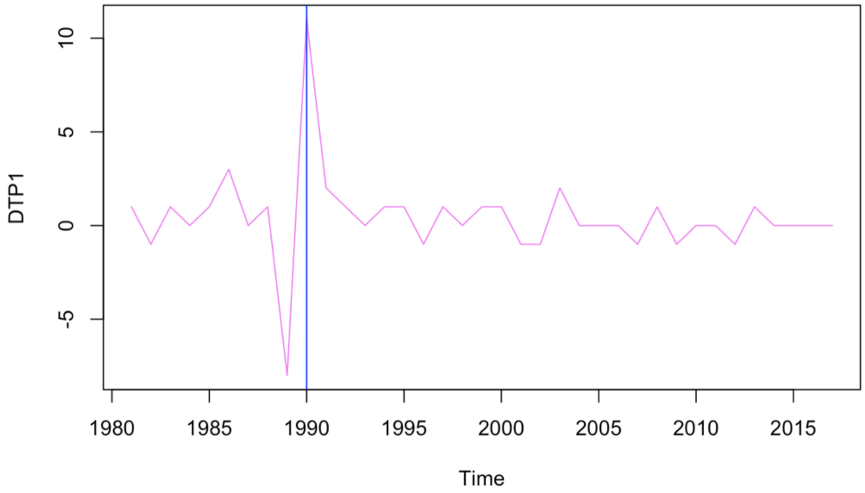


Figure 6: Change points showing difference variance shifts for DTP1

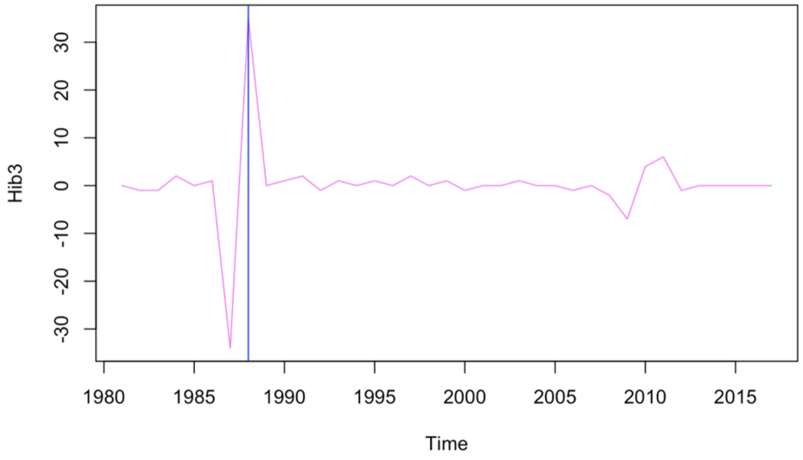


Figure 7: Change points showing difference variance shifts for DTP1

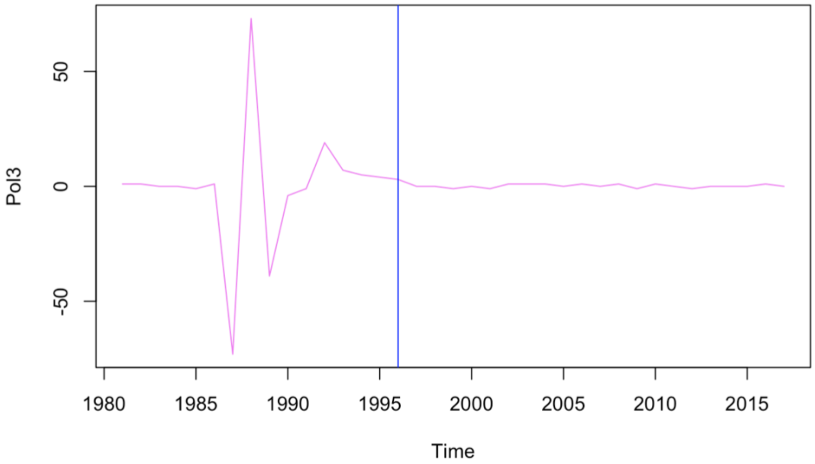


Figure 8: Change points showing difference variance shifts for DTP1

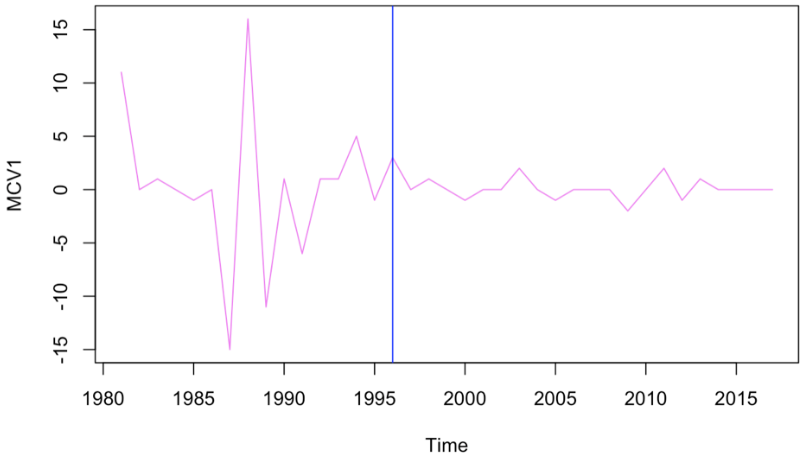


Figure 9: Change points showing difference variance shifts for DTP1