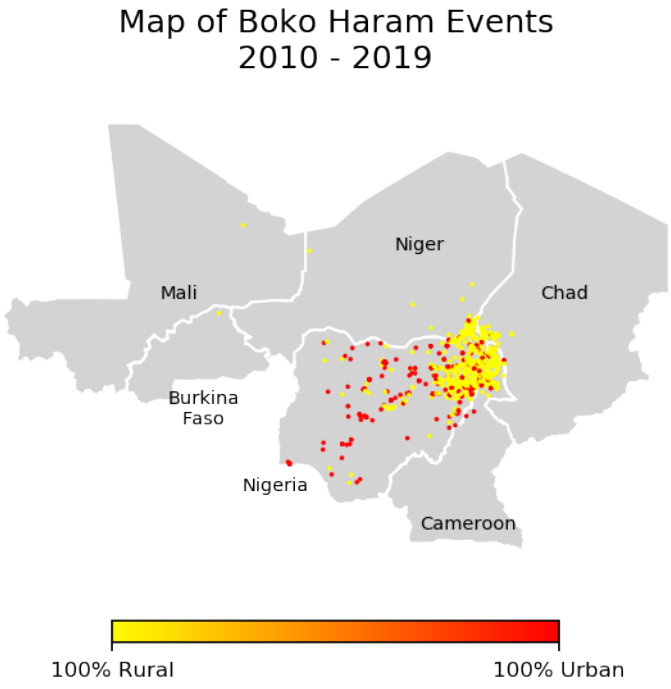
Boko_Haram_Analysis

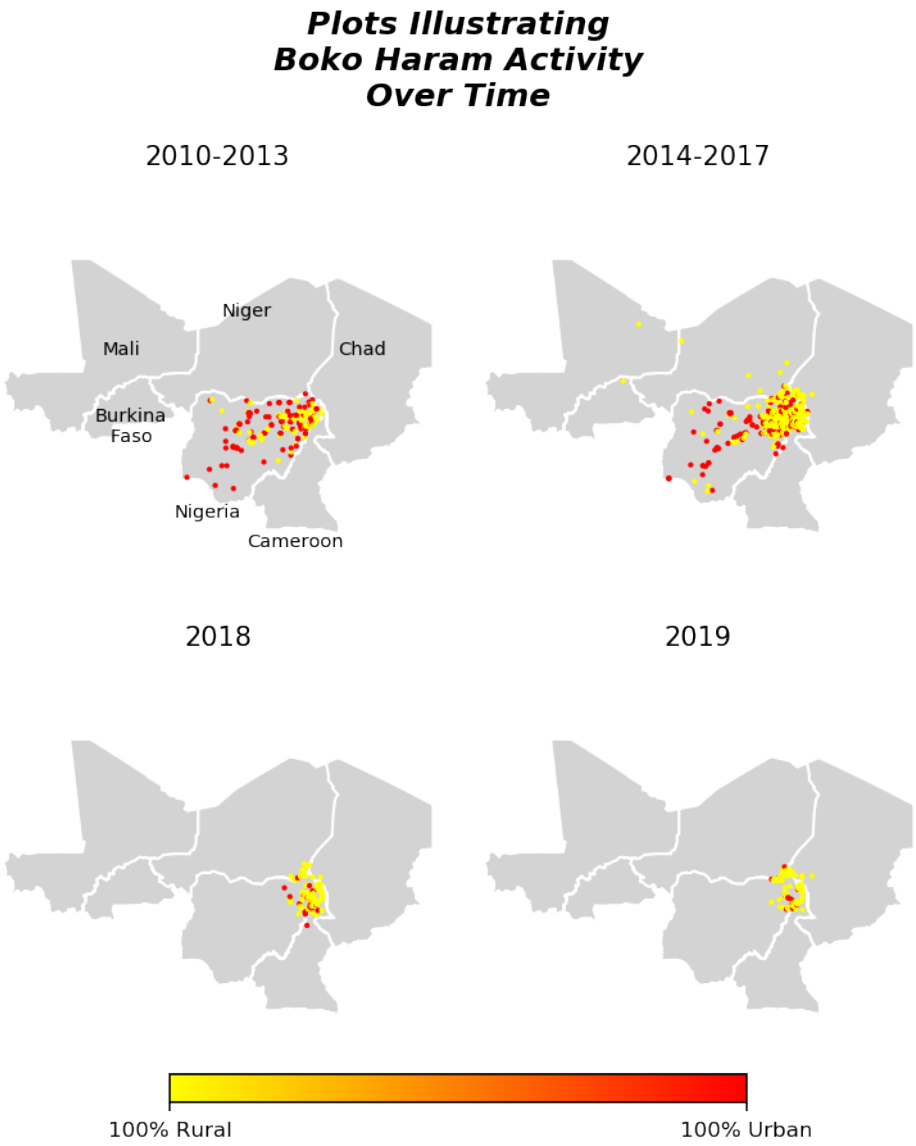
September 10, 2019

The purpose of this research is to investigate how the behaviour of Boko Haram has changed over time, specifically regarding their preference for rural or urban targets. For each event we determine whether it is urban or rural by checking whether its coordinates lie inside any city polygon in Africa, with the city polygon data provided by Africapolis.

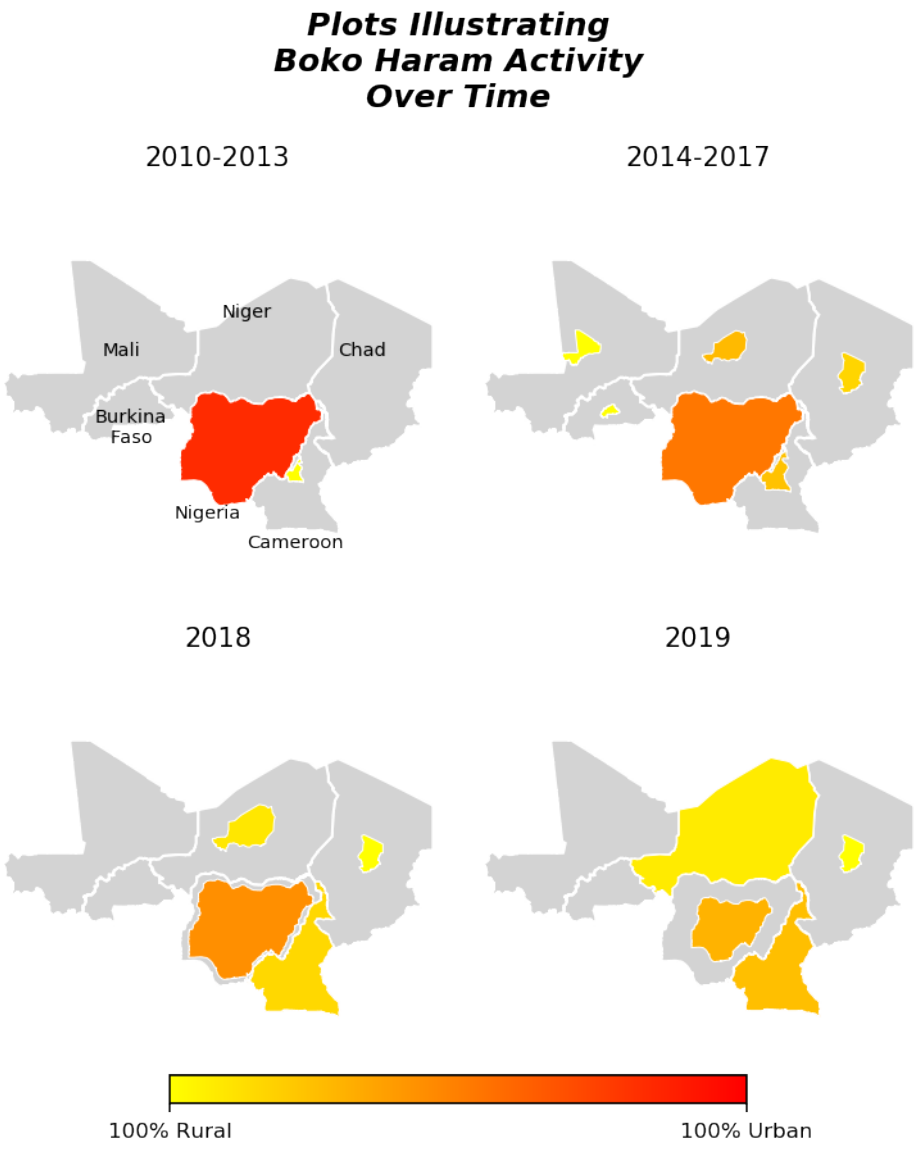
In the paper Strategic Risk of Terrorist Targets in Urban vs. Rural Locations, Hinkkainen and Pickering demonstrate the importance of understanding the differences of terrorism in urban and rural areas from the perspective of risk assessment for different target types. For example, identifying the risk to civilian, governmental and police targets in urban and rural areas differently. It follows that any statistical model used in the evaluation of risk must differentiate between urban and rural areas effectively. For this reason it is important to have effective and practical methods of identifying locations as urban or rural, such as the approach outlined in this article.

Let's take a look at the geographical distribution of terror events attributed to Boko Haram and see how things have developed over the years. Points range in colour from yellow to red, yellow meaning a rural event, and red meaning urban. The data is aggregated into specified time intervals and spatial pixels, with a pixel being roughly a square mile so that each point may represent multiple events with the colour representing the proportion of events that were urban. We see a clear shift to rural areas and a movement to North-East Nigeria and the bordering areas.

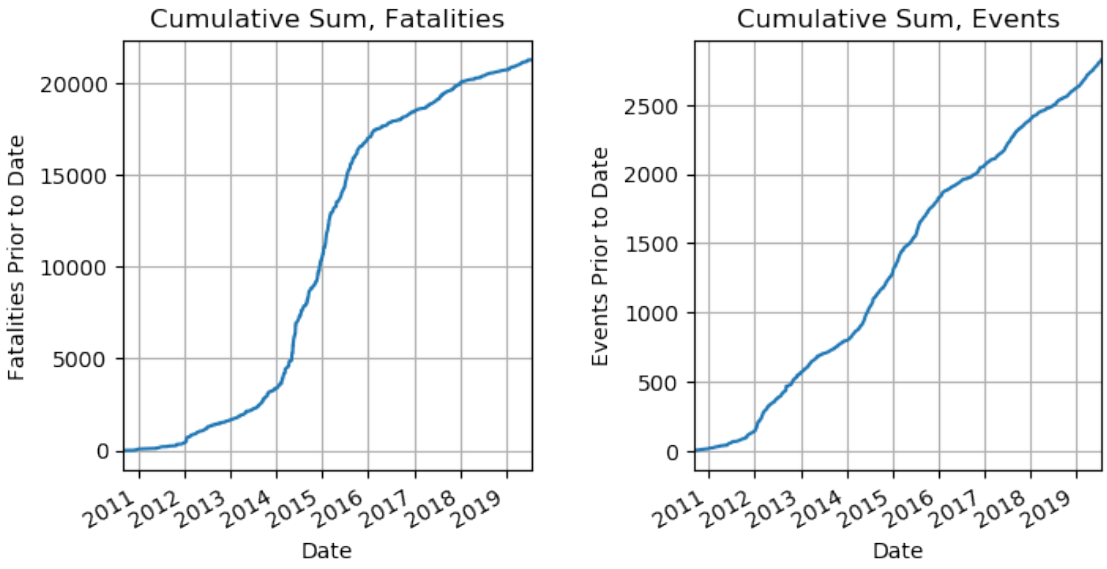




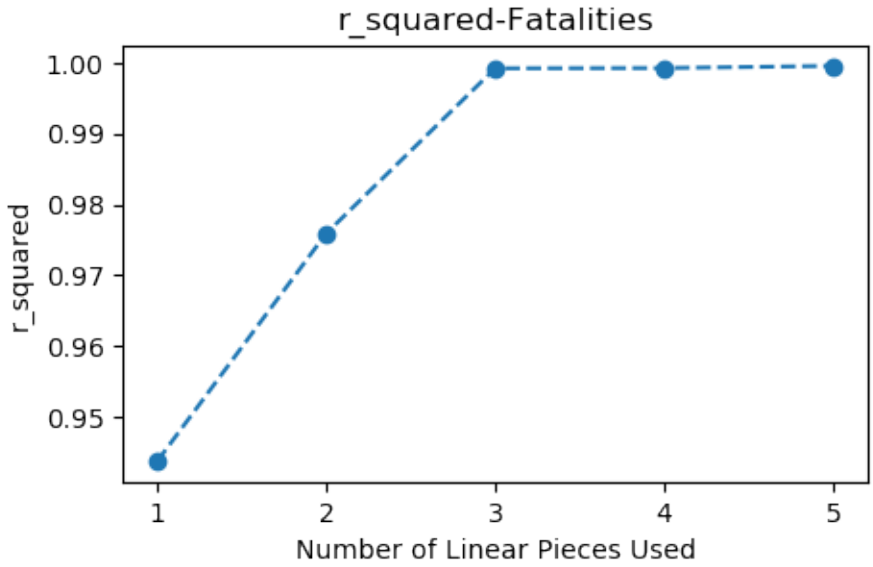
If we aggregate by country and plot a cartogram of events we get the following. Here the size scaling of a country indicates the proportion of events occurring in that country relative to the other countries.



Now let's look at the plots of the cumulative sums of number of fatalities and also the number of events due to Boko Haram between 2010 (this is to remove the outlier that occurs a year prior to anything else) and the end of the data, the end of July, 2019. When analysing these plots we're looking at two key features; break points in the curves and the gradients of the linear pieces. A break point is a point in time where there is a distinct change in the gradient of a curve. The gradient of a curve is the rate at which the curve is increasing, a measure of the steepness of the line. In our case this estimates the average number of fatalities/events per day. This will be done by fitting a piecewise linear function to the curve.



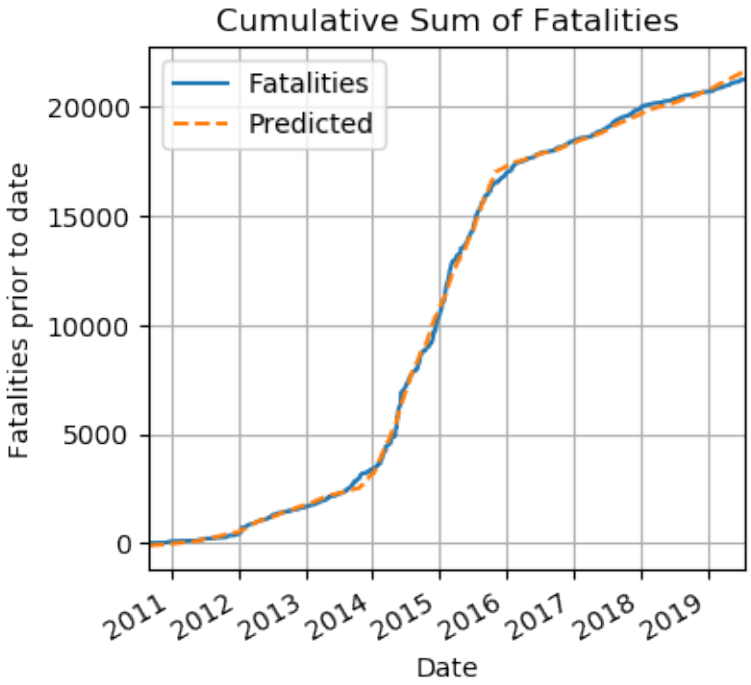
Above we see that the graph for fatalities seems to have two distinct break points, however the graph for events is very linear. I will fit a piecewise linear function to the curve for fatalities to see where the break points occur. For this I will implement the pwlf library which uses global optimisation techniques to minimise the mean squared error for a piecewise linear function with a specified number of break points. In order to find a the optimal number of break points we can measure how well the piecewise linear function fits the curve and see where we get little/no improvement. For this we will use the r_squared value provided by the library, a value close to 1 indicates a good fit.



So above we see that 3 linear pieces is indeed the correct number to fit this curve. Let's see where the break points are and plot the curve with the piecewise linear function fitted.

`Break points for 3 lines:`

	Break Points
0.0	2010-09-05
385.0	2013-10-24
824.0	2015-10-28
1466.0	2019-07-19



We can also test for the significance of the breakpoints found using a simple t-test on the parameters of our function. For each parameter (there is one parameter for each break point, including the end points) we perform the hypothesis test:

- H0: The parameter is 0
- H1: The parameter is not 0

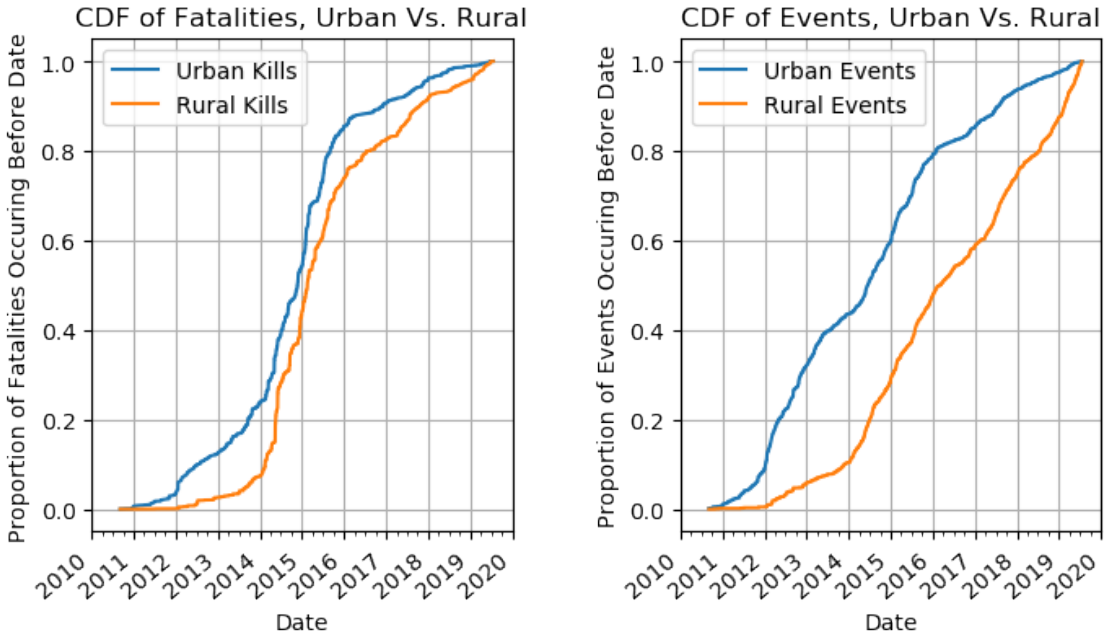
If the p-value is small we reject H0 and conclude that the parameter is not zero, and hence that the break point corresponding to that parameter is significant. We also have the estimate for the standard error for each parameter which tells us how certain we are in our choice of value for the parameter.

	Dates	p_values	Standard Error
0	2010-09-05	9.646650e-09	20.450168
1	2013-10-24	0.000000e+00	0.075049
2	2015-10-28	0.000000e+00	0.112758
3	2019-07-19	0.000000e+00	0.078809

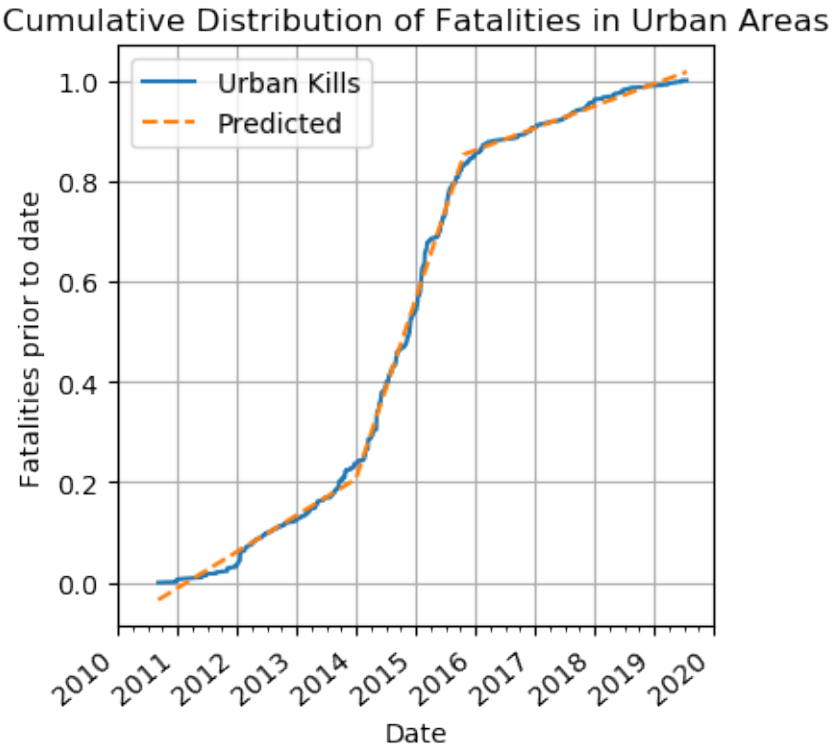
```
R_squared: 0.9992380920913639
```

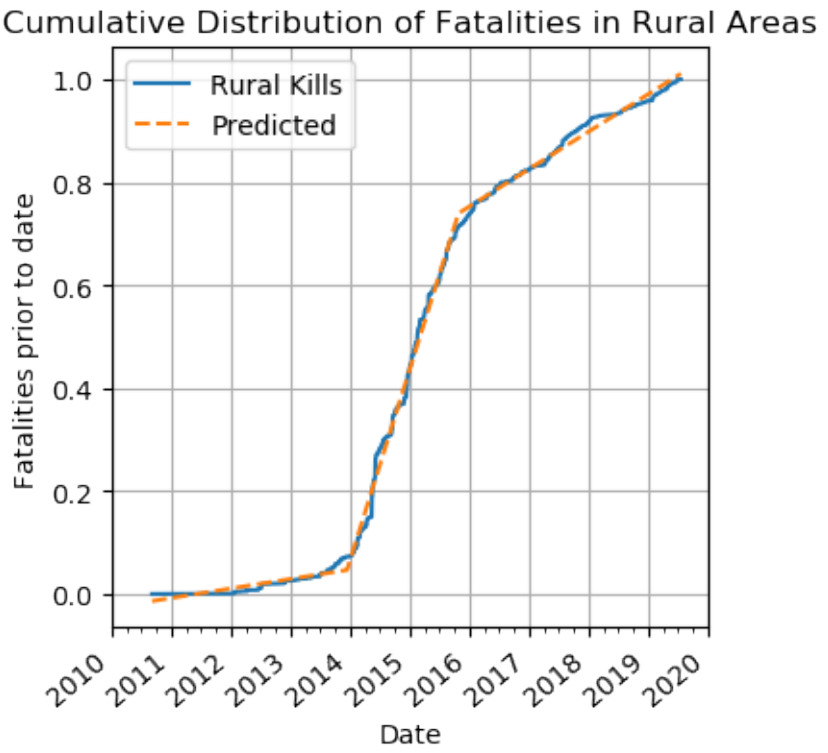
We see that we do indeed have significant break points and a very well-fitting curve.

Let's plot the cumulative distribution curves for urban and rural areas separately to compare Boko Haram's behaviour in these areas. Here I'm plotting the Cumulative Distribution Functions (CDFs) to accentuate the gradient changes for each curve and make it easier to spot breaking points.



We can note straight away that all curves show a marked increase in gradient around 2014, and all but 'Rural Events' show a decrease in gradient around 2016. 'Rural Events' is very linear post 2014. I will fit a piecewise linear function to these curves to determine breaking points which are apparent in the curves for fatalities in particular. It is clear for both fatalities curves the appropriate number of break points is 3 as with the overall curve, for the other two we'll do some experimentation again.

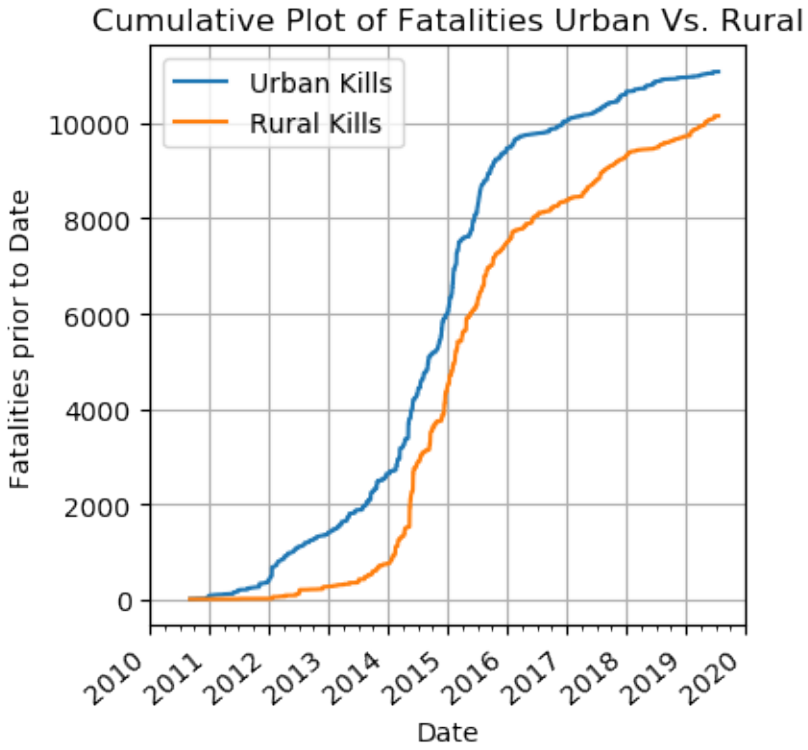

Cumulative Distribution of Fatalities in Urban Areas
Urban Kills
Predicted
Fatalities prior to date
1.0
0.8
0.6
0.4
0.2
0.0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Date


Cumulative Distribution of Fatalities in Rural Areas
Rural Kills
Predicted
Fatalities prior to date
1.0
0.8
0.6
0.4
0.2
0.0
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Date

Below we have the dates for the break points in the curves for number of fatalities in both urban and rural areas. We have found that the break points for 'Fatalities' for urban and rural areas occur within 3 weeks for the first and just a few days for the second which lends some weight to the probability that something occured across the board which caused Boko Haram to change their behaviour at these times.

	Urban Breaks	Rural Breaks
0	2010-09-05	2010-09-05
1	2013-12-29	2013-12-12
2	2015-10-21	2015-10-25
3	2019-07-19	2019-07-19

In order to compare the behaviour of the curves for Urban and Rural in the regions pre-2014 and post-2016 we will look at the gradients of the cumulative curves in these regions. Here gradient refers to the steepness of the curve at a certain point. This is estimated by fitting the piecewise linear function to the curve and taking the gradients of the lines. In this case specifically, the gradient measures the number of fatalities/events per day. To compare the gradients of the curves for Urban and Rural areas we should use the cumulative sum, not the cumulative distribution, so let's do that.

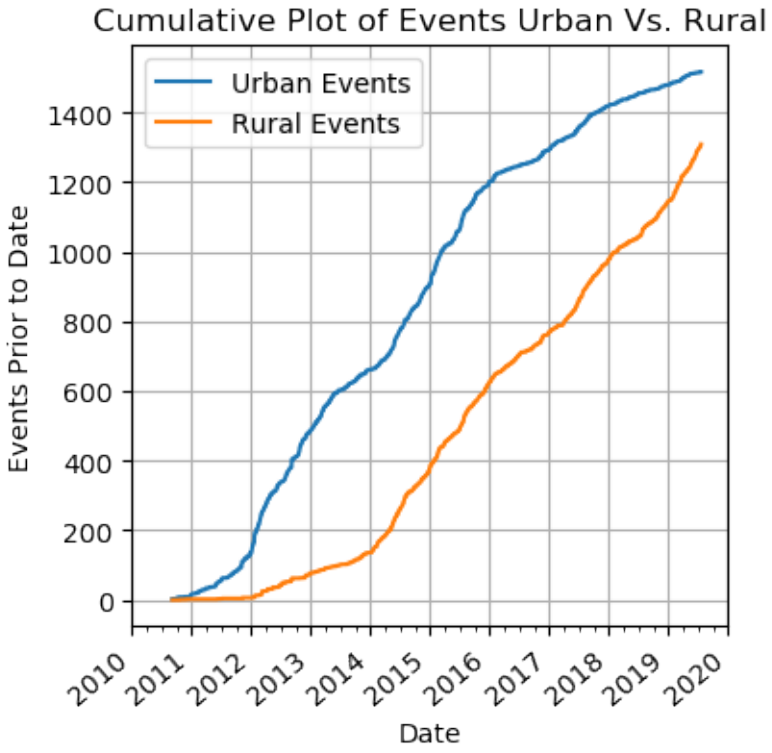


Note: Here Piece 1 refers roughly to the region before 2014, Piece 2 refers to 2014-2016 and Piece 3 is 2016 onwards.

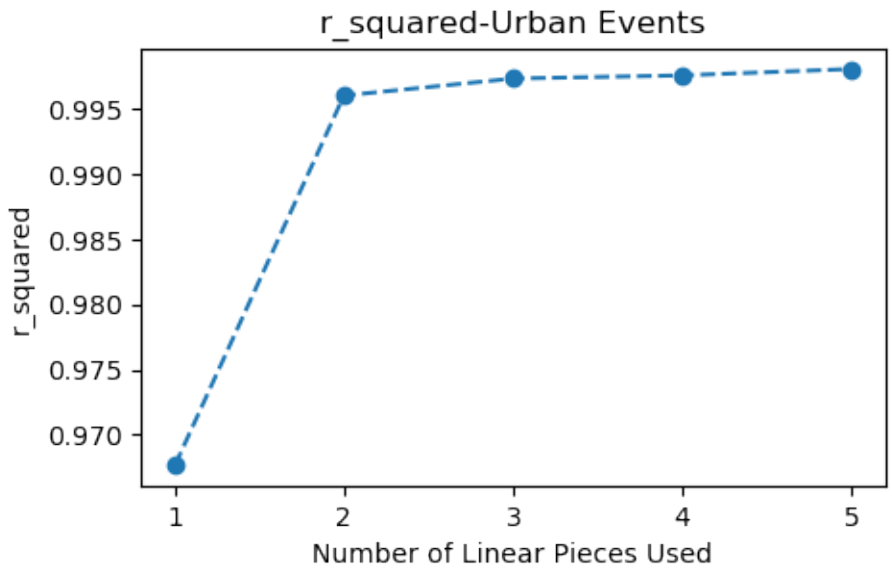
	Urban Gradients	Rural Gradients
Piece 1	2.199055	0.513796
Piece 2	10.829354	10.327389
Piece 3	1.335443	2.024063

The gradients of the piecewise linear function fitted to the curves give estimates for the average rate of increase in the number of fatalities. Here we see that there has been a significant shift in the rate of killing over time from urban to rural areas, though it is true that the rate in urban areas remains reasonably high, and the rural gradient on piece 3 is still less than the urban gradient on piece 1. The above analysis demonstrates that there have been two distinct breaking points and that since the first breaking point Boko Haram have been focussing on rural areas as much as urban ones, where before they were focussing on urban areas largely, and since 2016 they have been focussing significantly more on rural areas.

I now want to run a similar analysis for the number of events.



First we'll fit a function to the curve for urban events. Let's take a look at the r_squared values for different numbers of pieces first.



An r_squared value of over 0.99 for 2 linear pieces indicates that this is a good fit and there is almost no gain from increasing this value. Let's take a look at the break points for a few values anyway:

`Break points for 2 lines:`

	Break Points
0.0	2010-09-05
2033.0	2016-03-30
3239.0	2019-07-19

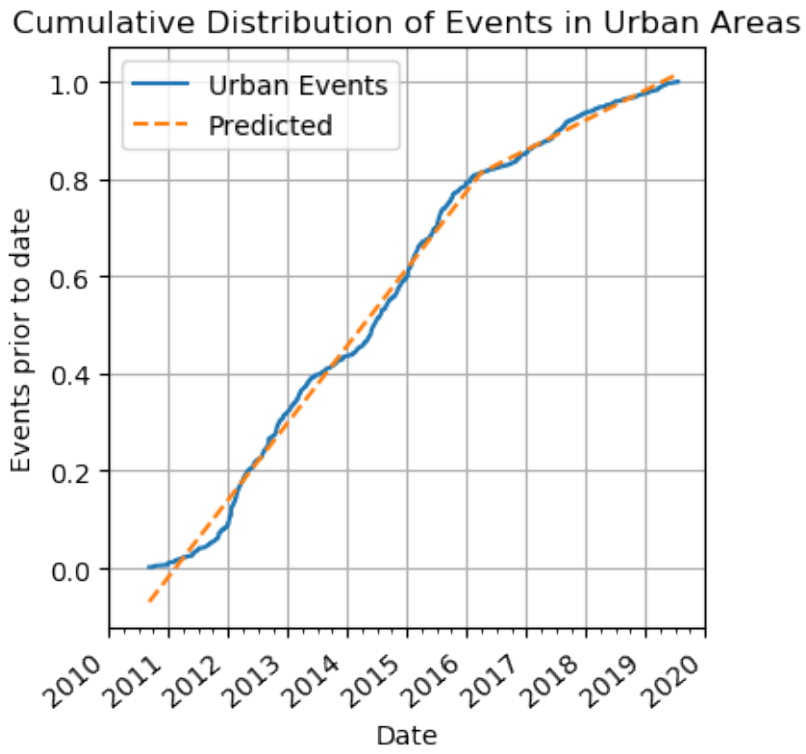
`Break points for 3 lines:`

	Break Points
0.0	2010-09-05
259.0	2011-05-22
1996.0	2016-02-22
3239.0	2019-07-19

`Break points for 4 lines:`

	Break Points
0.0	2010-09-05
428.0	2011-11-07
582.0	2012-04-09
2031.0	2016-03-28
3239.0	2019-07-19

We see that early 2016 is consistently identified as a breaking point. Let's have look at the fitted plot for 2 break points:



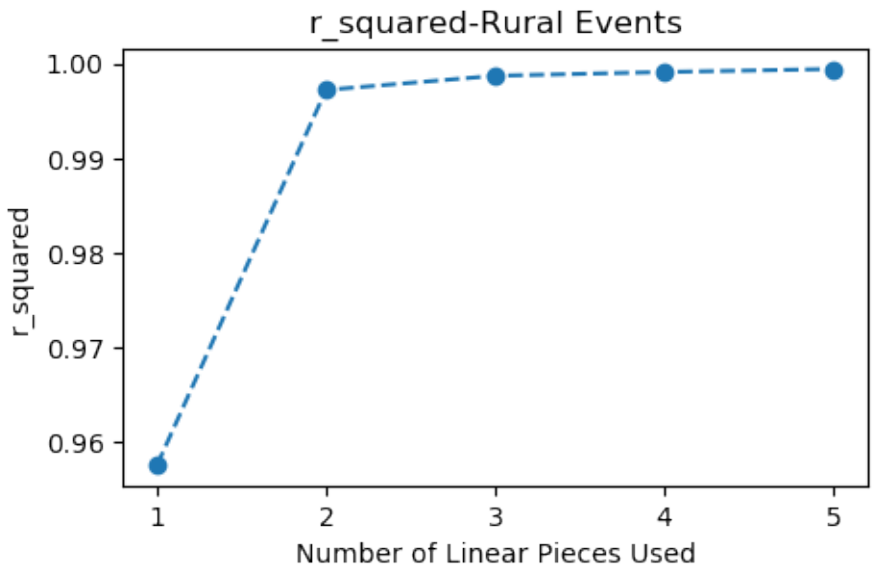
	Dates	p_values	Standard Error
0	2010-09-05	0.0	8.884614e-04
1	2016-03-29	0.0	6.750664e-07
2	2019-07-19	0.0	1.747982e-06

`R_squared: 0.9960281231335155`

(Plotted curve fitted with 2 lines)

We see here that there is just one particularly persistent break point which occurs in the fitting for every number of pieces, the dates for this break point are 2016-03-29, 2016-02-21, 2016-03-28 so I think it is reasonable to say that there was significant change in the region of late February/March 2016 where the frequency of events significantly slowed in Urban areas. Prior to this point in time the curve is very linear as we see a good fit with just two pieces. This roughly coincides with the marked reduction in the rate of increase in fatalities in both urban and rural areas at the end of 2015, though it is a few months later.

Now on to the curve for rural events.



Again a good fit with just 2 pieces and minimal improvement with more pieces added, but let's take a look at the breakpoints:

`Break points for 2 lines:`

	Break Points
0.0	2010-09-05
1048.0	2013-07-19
3239.0	2019-07-19

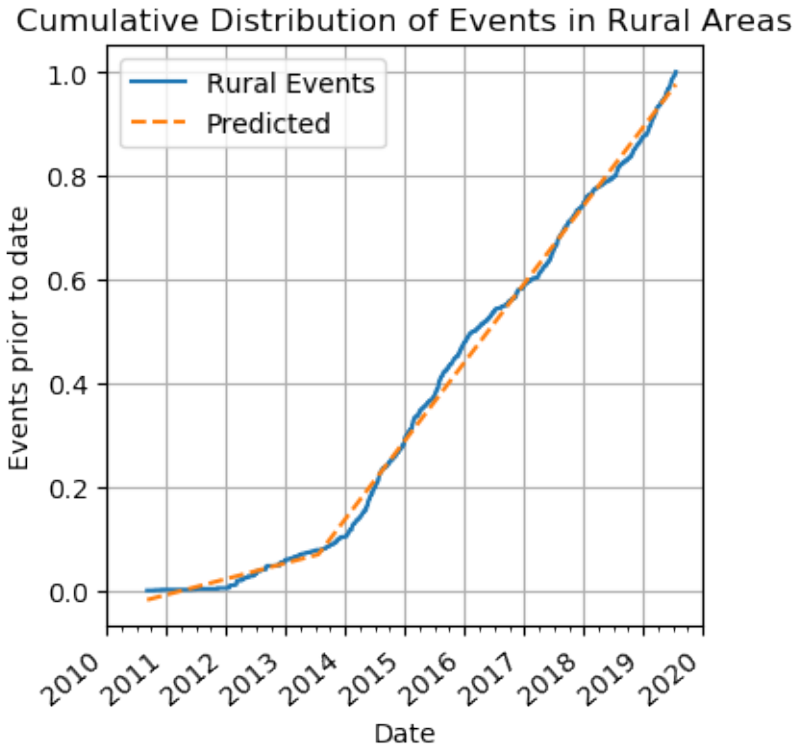
`Break points for 3 lines:`

	Break Points
0.0	2010-09-05
1176.0	2013-11-24
1802.0	2015-08-12
3239.0	2019-07-19

`Break points for 4 lines:`

	Break Points
0.0	2010-09-05
1174.0	2013-11-22
1965.0	2016-01-22
2338.0	2017-01-29
3239.0	2019-07-19

We see a breakpoint mid-late 2013.



	Dates	p_values	Standard Error
0	2010-09-05	1.499304e-78	0.000938
1	2013-07-19	0.000000e+00	0.000001
2	2019-07-19	0.000000e+00	0.000002

`R_squared: 0.9972497874865895`

(Plotted curve fitted with 2 lines)

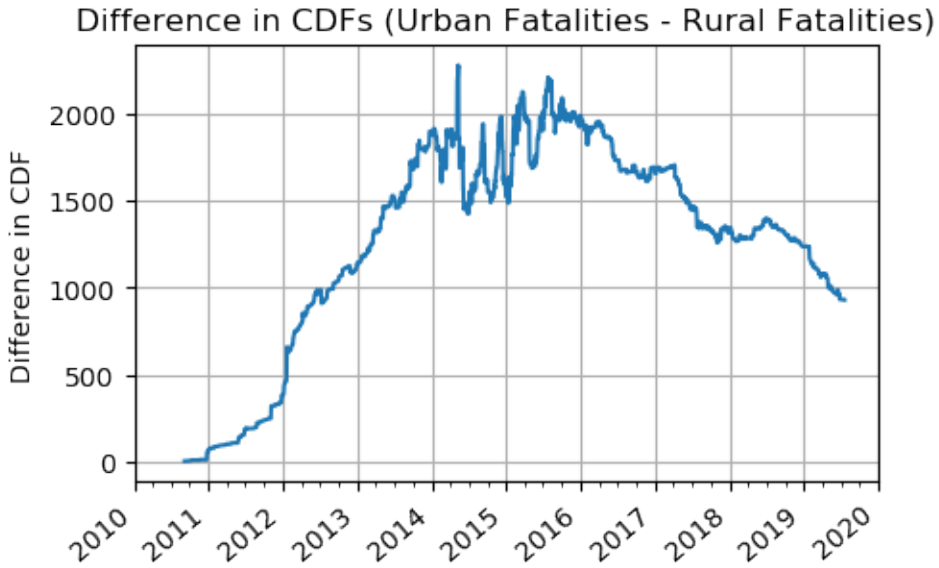
We see here that in rural areas there was a significant increase in the rate of events around late 2013, with a similar point in time appearing when fitting a curve with three and four pieces at the end of November 2013. After this point the curve is very linear, illustrated by the fact that we get very little gain by fitting more linear pieces to the function. Perhaps whatever caused the reduction in frequency of events in urban areas in 2016 did not affect rural areas in quite the same way.

Since the break points when using both 3 and 4 lines agree on a break point occuring in late November 2013 this is likely a significant point in time. This increase in the rate of events in rural areas occurs just one month before the marked increase in the number of fatalities in both urban and rural areas. This is not, however, a notable point in time for the cumulative distribution of the number of events in urban areas.

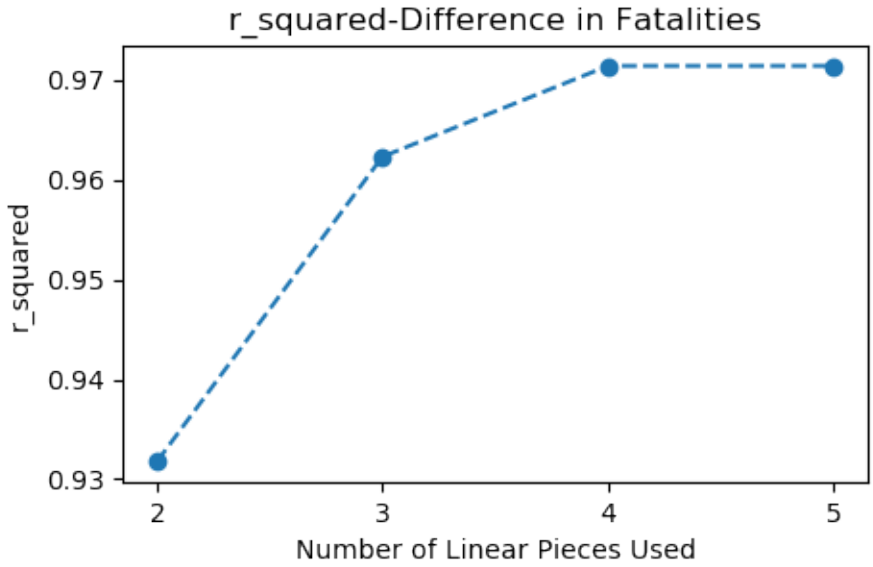
The only other thing to note is the appearance of 2015-08-12 and 2016-01-19 as break points for the number of events occuring in rural areas, at which point the occurences of events appear to slow slightly. This occurs shortly before the previously remarked decrease in frequency of events in Feb/March 2016 in urban areas. This also occurs shortly after the break point identifed in the number of fatalities in both urban and rural areas.

The fact that we find break points that lie so close to one another lends some weight to the hypothesis that something caused large change in the behaviour of Boko Haram.

To further illustrate the differences in both number of fatalities and number of events in urban and rural areas let's actually plot the differences.



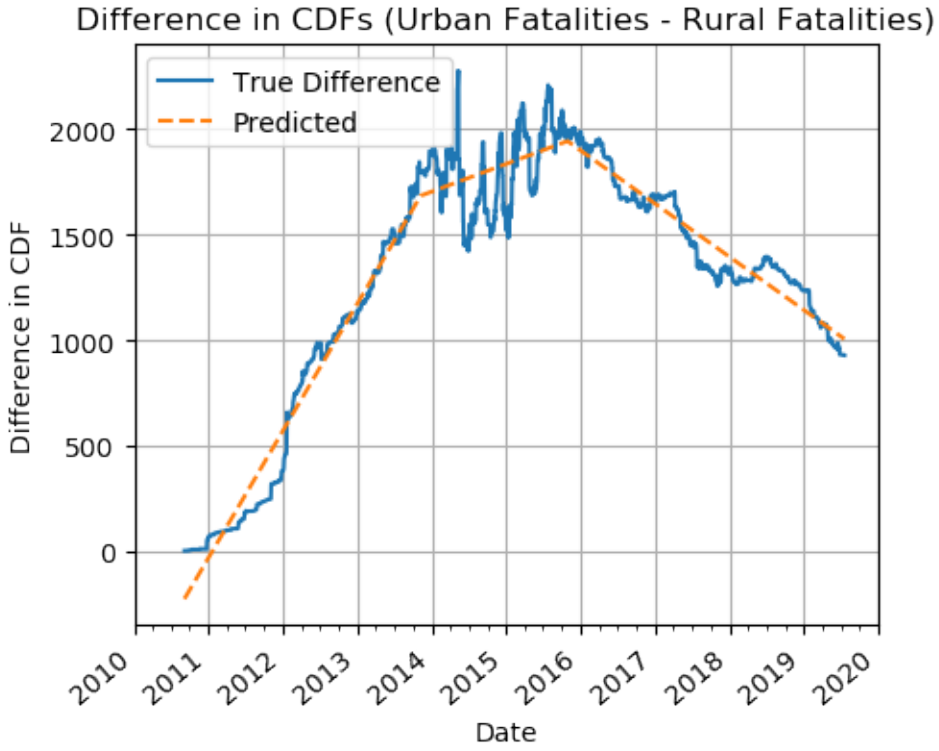
Let's fit a piecewise linear function to this curve to see the trends.



We see a good improvement with three pieces and not too much increase after, so we'll go with 3. Let's see the break points and plot:

```
Break points for 3 Pieces:
```

	Break Points
0.0	2010-09-05
1158.0	2013-11-06
1880.0	2015-10-29
3239.0	2019-07-19



This picture tells us more or less what we could see from previous analysis:

- 2010-2014: Urban fatalities occuring much faster than Rural fatalities
- 2014-2016: The difference closes somewhat with urban still edging ahead
- 2016-2019: Rural fatalities start occuring faster than rural fatalities

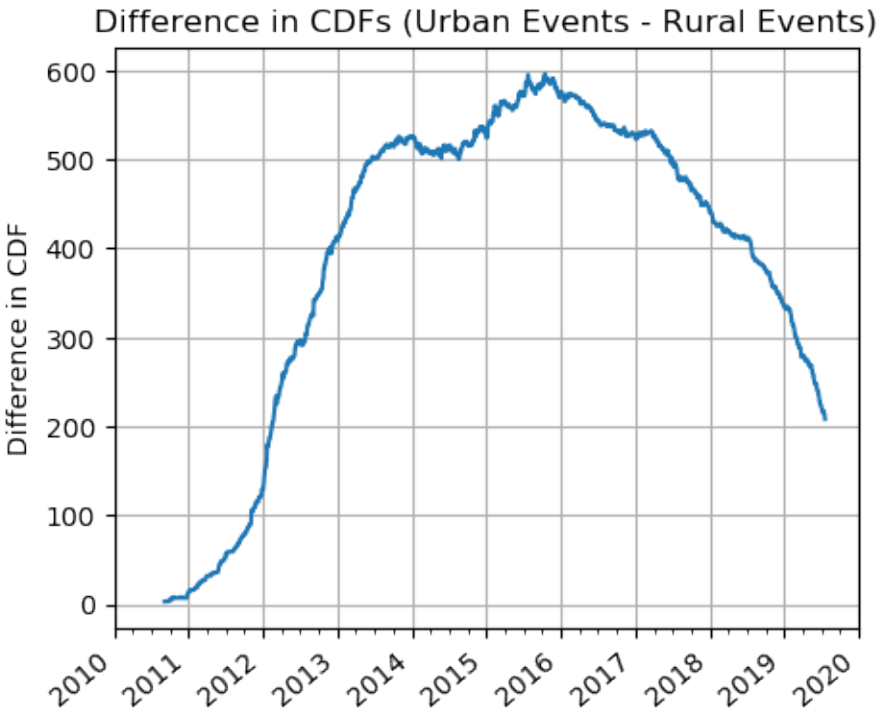
But we can also have a peak and see when this linear model predicts our curve will hit 0 again:

	Dates	p_values	Standard Error
0	2010-09-05	2.961174e-237	6.305952
1	2013-11-06	0.000000e+00	0.008230
2	2015-10-29	0.000000e+00	0.017099
3	2019-07-19	0.000000e+00	0.015663

```
R_squared: 0.9622938802004559
```

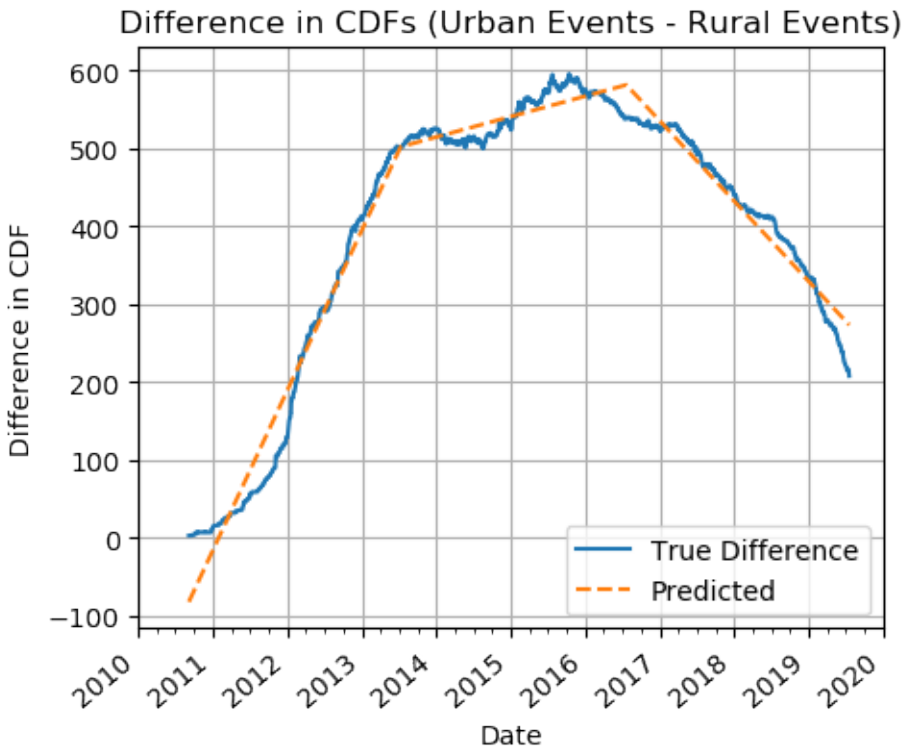
This linear model predicts that at this rate rural fatalities will overtake urban fatalities by 16/07/2023

Now let's see if the curve for the difference in the cumulative number of events has anything to say.



Break points for 3 Pieces:

	Break Points
0.0	2010-09-05
1039.0	2013-07-10
2146.0	2016-07-21
3239.0	2019-07-19



	Dates	p_values	Standard Error
0	2010-09-05	0.0	1.410457
1	2013-07-10	0.0	0.001942
2	2016-07-21	0.0	0.003071
3	2019-07-19	0.0	0.002955

```
R_squared: 0.9808813287622986

2022-03-14
```

It seems there isn't much to add. This model predicts that the number of events in rural areas will overtake those in urban areas by 14/03/2022.