

# accrued: Tools for visualizing data quality of partially accrued data

Julie Eaton  
University of Washington Tacoma  
jreaton@uw.edu

Ian Painter  
University of Washington  
ipainter@u.washington.edu

August 30, 2013

---

## Abstract

A time series is *partially accruing* if data for each time point are accrued piecemeal and become more complete over time. The R package **accrued** provides tools for summarizing and visualizing data quality for partially accruing data. This vignette describes the data classes defined in **accrued** and the summary and visualization methods associated with each, with examples.

*Keywords:* Partially accruing data, times series

---

## 1 Introduction

The **accrued** package contains methods for visualizing data quality for partially accruing data. A time series is *partially accruing* if data for each time point are accrued piecemeal and become more complete over time. Data including the number of emergency visits per day (total counts) and the number of daily visits due to ILI were sent from various emergency departments to a single source. Data were received at different times, giving the time series a partially accrued structure.

The primary feature of partially accruing data we note is the effect of *accrual lag*—the difference between the *event time* (emergency department visit date) versus *record time* (also referred to as upload date), which is the time when the data become available to the system. Our summary and visualization methods primarily revolve around visualizing aspects of the data quality affected by accrual lag (referred to here as *lag*, though this is fundamentally different from the traditional definition of lag in time series analysis).

Due to the piecemeal accrual of data, the value of an indicator for a particular event date changes over time. The following definitions help clarify different ways of characterizing the indicator value. The *current value* of an indicator refers to the value of an indicator for a particular event date as of the current date. The *complete data value* of the indicator refers to the value of an indicator for a particular event date once all data for the event date has been received. Finally, the *lagged value* of the indicator refers to the value of an indicator for a particular event date a fixed number of days after that date. The number of days lagged is specified so that, for example, the five-day lagged value for a particular event date is the current value five days after the event date.

The methods provided in **accrued** were developed for understanding the data quality issues surrounding the Distribute project for real time influenza-like-illness (ILI) surveillance (Distribute: [1], methods: [2] and [3]).

## 2 Classes defined in accrued

The main class used in **accrued** is the “**accrued**” class. This class is a wrapper for a matrix containing partially accrued data. The function `data.accrued` creates an object of the “**accrued**” class. The only required argument, `x`, is a matrix of time series with rows as consecutive event dates (encounter dates) and columns as accrual lag. Accrual lag must be nonnegative integers and are assumed to increase one unit

with each column. Optional arguments are **start** (start date) and **final** (the vector of complete data). If **start** is omitted, the start date defaults to “1” and if **final** is omitted, then the right-most column of data is used in its stead. The matrix value in the  $i$ th row and  $j$ th column of **data** represent the value for the event date  $i$ , known  $j - 1$  time units after the event date.

To create the “accrued” object:

```
> library(accrued)
> data(accruedDataExample)
> accrued_data = data.accrued(accruedDataExample)
> names(accrued_data)
```

```
[1] "final" "data"  "start"
```

The accrued data in the “accrued” class can be extracted using “data”

```
> accrued_data[["data"]][201:215,1:10]
```

	0	1	2	3	4	5	6	7	8	9
201	2	2709	6226	6233	6234	6234	6234	6234	6234	6234
202	0	2352	6246	6287	6287	6287	6287	6288	6288	6288
203	NA	2224	5689	5845	5845	5845	5845	5846	5846	5846
204	0	2180	5810	5820	5820	5822	5823	5823	5823	5823
205	1	1885	5038	5175	5578	5579	5579	5579	5580	5580
206	1	1972	5056	5566	5568	5568	5568	5569	5569	5569
207	NA	2234	5695	5695	5695	5696	5696	5696	5696	5696
208	0	2523	6067	6164	6164	6164	6164	6164	6164	6164
209	1	1974	5628	5630	5781	5781	5781	5781	5781	5781
210	NA	2054	5305	5475	5475	5476	5476	5476	5476	5527
211	1	1939	5336	5337	5618	5618	5619	5619	5679	5680
212	0	1866	4962	5325	5327	5327	5327	5387	5387	5387
213	1	1721	4954	5047	5047	5224	5273	5273	5273	5273
214	1	2275	5085	5086	5281	5341	5341	5341	5341	5450
215	1	2133	5815	5990	6050	6050	6050	6050	6077	6077

The **summary** function for the “accrued” class provides a succinct summary of the data. Below is an example summary. The first column represents the percentage of dates on which a positive number of counts lagged by 0, 1, 2, ..., 20 days were received. The second column represents the mean percentage of counts relative to the final counts received, for each lag. The third column represents the mean counts received for each lag.

```
> summary(accrued_data)
```

Summary of accrued data object with time points.

Lag	Upload Percent	Proportion	Mean Count
0	0.40	0.00	0.9
1	0.92	0.35	1944.0
2	0.97	0.93	5204.1
3	0.99	0.97	5409.8
4	1.00	0.98	5488.0
5	1.00	0.99	5532.1
6	1.00	0.99	5555.6
7	1.00	1.00	5569.8
8	1.00	1.00	5579.5
9	1.00	1.00	5582.5

10	1.00	1.00	5585.1
11	1.00	1.00	5587.2
12	1.00	1.00	5587.2
13	1.00	1.00	5587.2
14	1.00	1.00	5587.2
15	1.00	1.00	5587.2
16	1.00	1.00	5587.2
17	1.00	1.00	5587.2
18	1.00	1.00	5587.2
19	1.00	1.00	5587.2
20	1.00	1.00	5587.2
final	1.00	1.00	5587.2

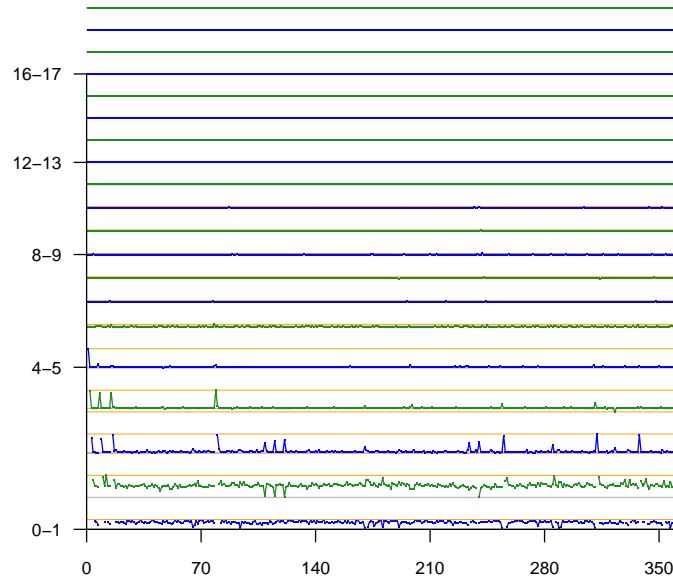
### 3 Stacklag difference plot

The default plot for an object of the “**accrued**” class is the stacklag difference plot. An example of this plot is shown in Figure 1 and is produced using the following function call.

```
> plot(accrued_data)
```

This plot shows changes in count values from one lag to the next by stacking scaled time series of differences between values. The  $x$ -axis shows the event date, and the  $(Z + 1)$ th layer shows the difference in count indicator values between lag  $Z$  and lag  $(Z - 1)$ , for  $Z \geq 1$ . For layer 1, the counts received at lag 0 are shown, and at any date if no counts have been received for the current or any previous lags no value is shown. The height of layer  $Z$  is proportional to the maximum over range of the difference in counts between lag  $Z$  and lag  $(Z - 1)$ , and a fixed gap is placed between each layer. Thus the lowest layer shows the counts for each date received on that date, and the subsequent layers show the additional counts received for a date a day later than the previous layer (note that negative differences are possible if counts are corrected downwards). The relative heights of lines within different layers are proportional to the size of the differences. The number of counts accrued decreases as lag increases, which gives the plots a “logarithmic decay” appearance.

Figure 1: Stacklag difference plot for the example data, the default plot for an “accrued” object.



## 4 Summary completion curve

The “`summary.accrued`” function returns an object of type “`summary.accrued`”, which has a default plot function, called the *summary completion curve*, which graphs the second column of the `print.summary.accrued` output against lag, that is, mean percentage versus lag. The summary completion curve for the example data is shown in Figure 2. It is produced using the function call

```
> plot(summary(accrued_data))
```

## 5 Bar code plot

The bar code plot is a sparkline graph showing a vertical line for each date on which an record was received. This function requires the data be in the form of counts. The example data contain 385 rows of data, so we will view a window of only around 100 days instead, as demonstrated below. The reduced data set is created with the code below.

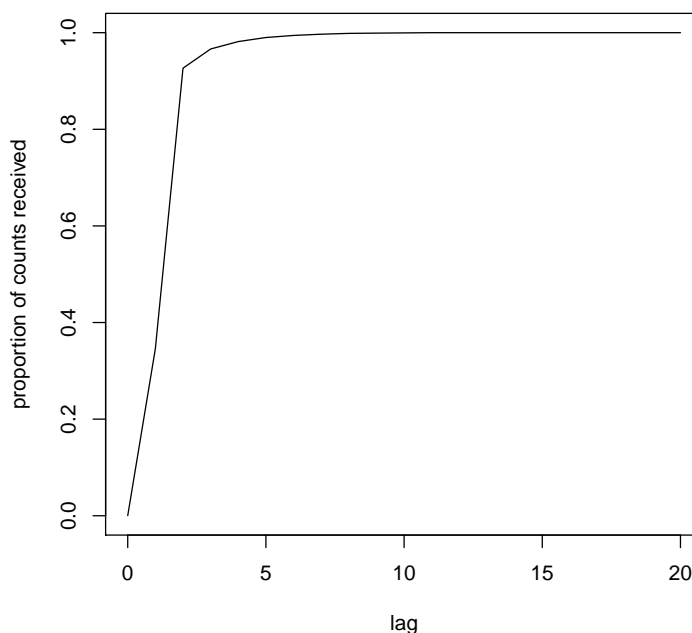
```
> accrued_data_reduced = data.accrued(accruedDataExample[100:200,])
```

The bar code plot call is made here

```
> barcode(accrued_data_reduced)
```

The bar code plot for the example reduced data is shown in Figure 3.

Figure 2: Summary completion curve for the example data, the default plot for a `summary.accrued` object.



## 6 Upload pattern plot

The upload pattern plot displays indicators of whether non-zero counts were received on a particular day. This function has two display options. In both cases count data are required. In the first way, record dates (upload dates) are represented on the horizontal axis and event dates (encounter dates) are represented on the vertical axis. This is the default plot (the parameter `horizontal` is set to `FALSE`). The point with coordinates  $(i, j)$  is plotted if a positive count was received on that record date for that particular event date. Since event dates must always occur on or before record dates, this produces a set of points with integer coordinates that lie on or below the line  $y = x$ . The upload pattern plots can be difficult to view if too many upload dates are included. The example data contain 385 rows of data, so we will view a window of only around 100 days instead. The upload plot (of the reduced data) is produced using the code

```
> uploadPattern(accrued_data_reduced)
```

and the plot is shown in Figure 4.

A more compact way of representing upload information is by replacing the vertical axis with lag instead of event date. Recall that for a particular record-event combination, the associated accrual lag is

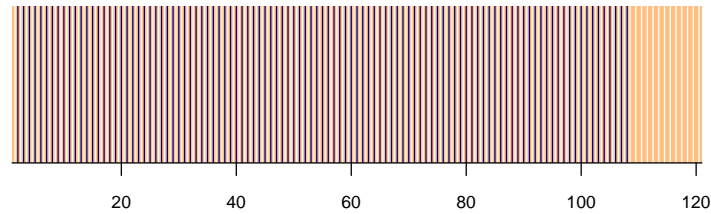
$$(\text{accrual lag}) = (\text{record date}) - (\text{event date}).$$

To obtain this plot, the optional argument `horizontal` must be set to `TRUE` (its default value is `FALSE`). The plot is produced using the code

```
> uploadPattern(accrued_data_reduced, horizontal=TRUE)
```

and resulting plot appears in Figure 5.

Figure 3: Bar code plot for the example data.



## 7 Arrayed lag time series

The function `laggedTSarray` generates time series of indicator values, arrayed by lag. Its only required argument is an object of the “`accrued`” class. A running median line using the previous `daysOfHistory` days of data is drawn to show the average pattern of each time series. Lines at plus and minus two median absolute deviations (MAD) from the median to show the variability of each time series, both using a `daysOfHistory` window. If `daysOfHistory` is not specified, it defaults to 30.

The following code produces arrayed lag time series for all lags. This produces as many pages of plots as there are lags.

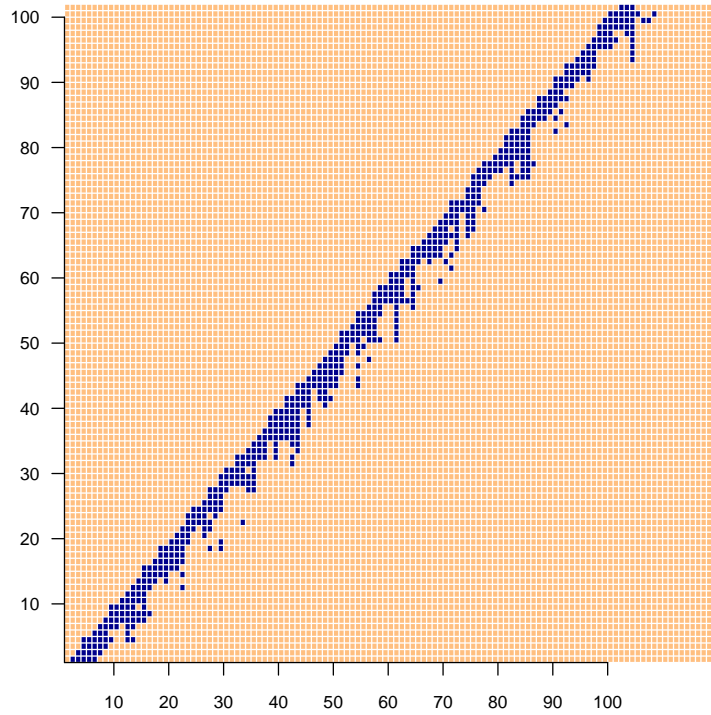
```
> laggedTSarray(accrued_data, daysOfHistory=20)
```

The following code produces a time series for a lag of 1 only.

```
> laggedTSarray(accrued_data, daysOfHistory=20, lags=c(1))
```

The output of three separate calls is shown in Figure 6.

Figure 4: Upload plot for the example data.



## 8 Lag histograms

The completion distribution refers to the percentage of total counts received for each lag. The lag histograms show the completion distribution for lags ranging from 0 to 14 days. For a fixed lag  $Z$  and for each event date, the proportion of counts available relative to the final counts present for that date by that lag is computed. That is,

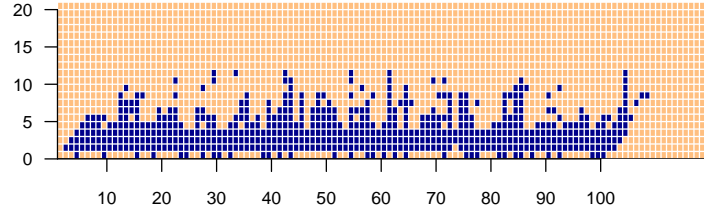
$$(\text{percent complete, } Z \text{ days lagged}) = \frac{(\text{counts, } Z \text{ days lagged})}{(\text{final counts})}.$$

Percentages above 100 are rounded down. Repeating this calculation over all event dates yields a completion distribution. The function `lagHistogram` graphs the completion distribution. Its only required argument is an object of the “`accrued`” class. The function call for producing the lag histograms is shown here:

```
> lagHistogram(accrued_data)
```

The lag histograms for the example data appear in Figure 7. Note that only lags 0 to 14 are shown, even though the data contain more lags.

Figure 5: Horiztonal upload plot for the example data.



## 9 Errors

For partially accrued data, we refer to count error (or uncertainty) as the difference between the final count for a day and the lagged count for a day. One expects that the error in the total counts decreases as lag increases. The function `accruedErrors` inputs an “accrued” object and outputs one of the “`accruedErrors`” class. `accruedErrors` computes the count errors using the error function `func`. The default error function is the difference between final and lagged observation. The data set is truncated to include only lags 0 to 15 so there is space for printing on this page.

```
> dat = data.accrued(accruedDataExample[,1:10])
> countErrors = accruedErrors(dat)

> summary(countErrors) # the summary function prints the 0.1, 0.5, and 0.9 quantiles.
```

	0	1	2	3	4	5	6	7	8	9
0.1	5179.4	3239.2	52.4	1.9	0.0	0.0	0	0.0	0	0
0.5	5605.0	3613.0	287.0	123.0	58.5	19.0	1	0.0	0	0
0.9	6111.4	4102.8	782.6	405.8	257.0	137.6	86	48.8	1	0

```
> errorQuantileSummary(countErrors, quantiles=c(0.1,0.2,0.5,0.8,0.9))
```

	0	1	2	3	4	5	6	7	8	9
0.1	5179.4	3239.2	52.4	1.9	0.0	0.0	0.0	0.0	0	0
0.2	5280.6	3369.2	103.0	29.8	1.0	0.0	0.0	0.0	0	0
0.5	5605.0	3613.0	287.0	123.0	58.5	19.0	1.0	0.0	0	0
0.8	5910.2	3933.0	579.4	309.0	153.0	86.0	47.4	1.0	0	0
0.9	6111.4	4102.8	782.6	405.8	257.0	137.6	86.0	48.8	1	0

```
> # errorQuantileSummary allows the user to specify a desired quantile vector
```

```
> plot(countErrors)
```

```
> # generic plot call does not allow user to specify quantiles plotted.
```

```
> # Quantiles plotted are 0.1, 0.5, and 0.9.
```

```
>
```

```
> plot.accruedErrors(countErrors, quantiles=c(0.2,0.5,0.8))
```

```
> # this version of the plot call allows the user to specify the quantiles.
```

In Figure 8 the count errors are shown.

The ILI ratio is defined by

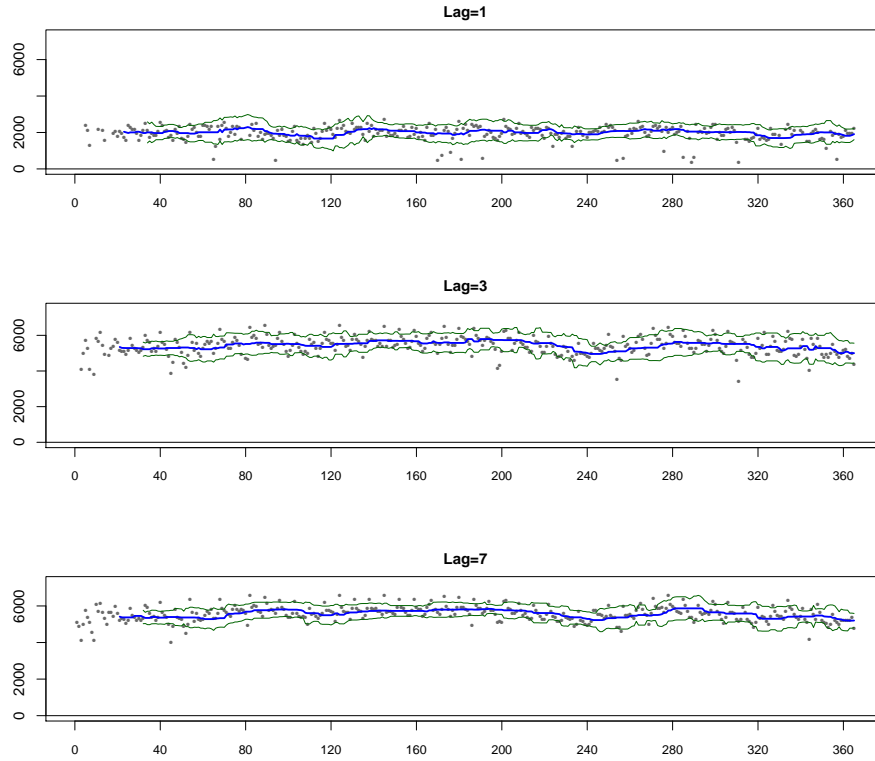
$$(\text{ILI ratio}) = \frac{(\text{ILI counts})}{(\text{total counts})}.$$

The ILI ratio error is given by

$$(\text{ILI ratio error, lagged by } Z \text{ days}) = \frac{\frac{(\text{ILI counts, lagged by } Z \text{ days})}{(\text{total counts, lagged by } Z \text{ days})}}{\frac{(\text{final ILI counts})}{(\text{final total counts})}}. \quad (1)$$



Figure 6: Arrayed lag time series using 20 days of previous data for lag=1, 3, and 7.



If two arguments are passed to the `accruedErrors` function instead of one (both must be counts), then a ratio is computed (the ILI ratio), and the log error ratio is computed as the default error function, that is, the log of (1) is returned. In the ratio computation, the first argument (`x`) is the denominator, and the second argument (`y`) is the numerator. The function call appears below, and the associated plot appears in Figure 9

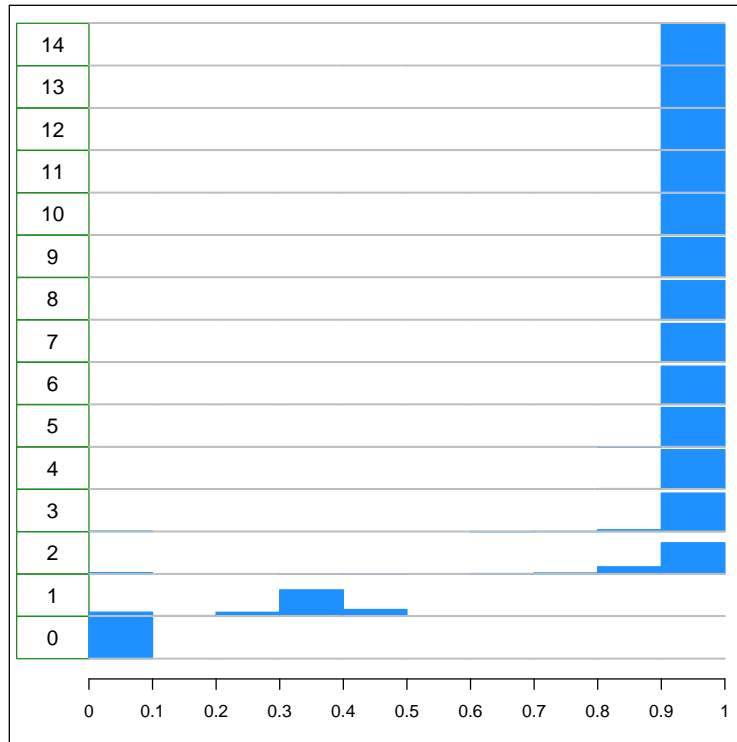
```
> data(accruedDataILIExample)
> dat2 = data.accrued(accruedDataILIExample[,1:10])
> combinedErrors = accruedErrors(x=dat, y=dat2)
> plot.accruedErrors(x=combinedErrors, quantiles=c(0.1, 0.2, 0.5, 0.8, 0.9))
```

The user can also specify the error function. In the following example, the square root function is used to compute the error of the counts contained in `accrued_data`. Note that the “`x`” in the argument of the function `func` is *different* from the first argument of `accruedErrors`, `x`. The plot of the errors appears in Figure 10.

```
> errors = accruedErrors(x=dat, func = function(x,y) {sqrt(x) - sqrt(y)} )
> round(summary(errors), 4)
```

	0	1	2	3	4	5	6	7	8	9
0.1	70.9427	26.9387	0.3480	0.0125	0.0000	0.0000	0.0000	0.0000	0.0000	0
0.5	74.0800	29.9978	1.9503	0.8479	0.3842	0.1316	0.0064	0.0000	0.0000	0
0.9	77.6084	35.2220	5.4921	2.7977	1.7922	0.9322	0.5735	0.3282	0.0065	0

Figure 7: Lag histograms for the example data.



## References

- [1] D. Olson, M. Paladini, W. Lober, L. Buckeridge, and I. D. W. Group. Applying a new model for sharing population health data to national syndromic influenza surveillance: Distribute project proof of concept, 2006 to 2009. *PLOS Currents Influenza*, 3, 2011.
- [2] I. Painter, J. Eaton, D. Olson, D. Revere, and W. Lober. How good is your data. 2011 ISDS Conference abstract. *Emerging Health Threats Journal*, 4, 2011.
- [3] I. Painter, J. Eaton, D. Olson, D. Revere, and W. Lober. Visualizing data quality: tools and views. 2011 ISDS Conference abstract. *Emerging Health Threats Journal*, 4, 2011.

Figure 8: Count errors for the example data.

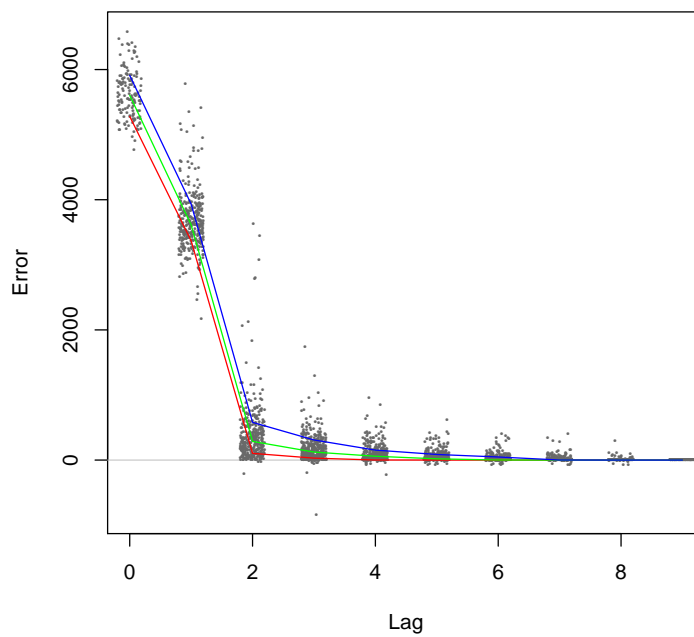


Figure 9: Log of the error in ILI ratio for the example data.

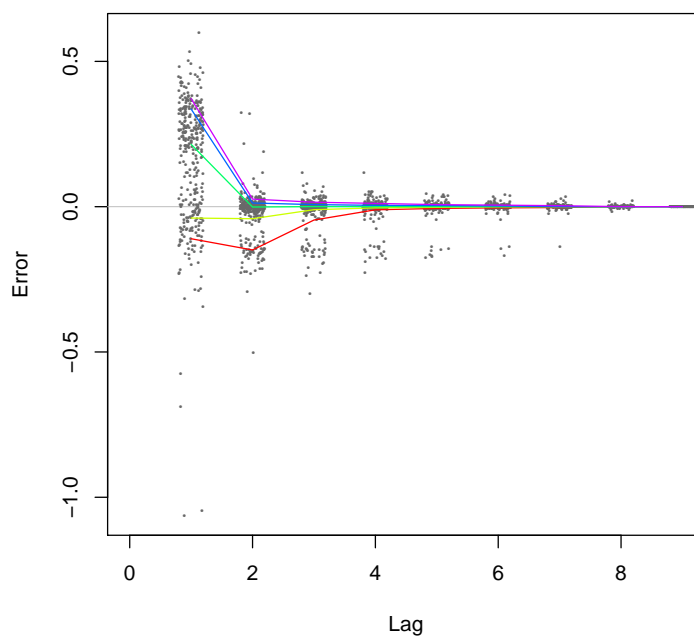


Figure 10: Graph of count errors with the user specified error function.

