Kaggle Rain Training Set - Radar Distributions

Scott Mark

October 10, 2015

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:data.table':
##
##     between, last
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## loading train from Rdata fileloading test from Rdata file
```

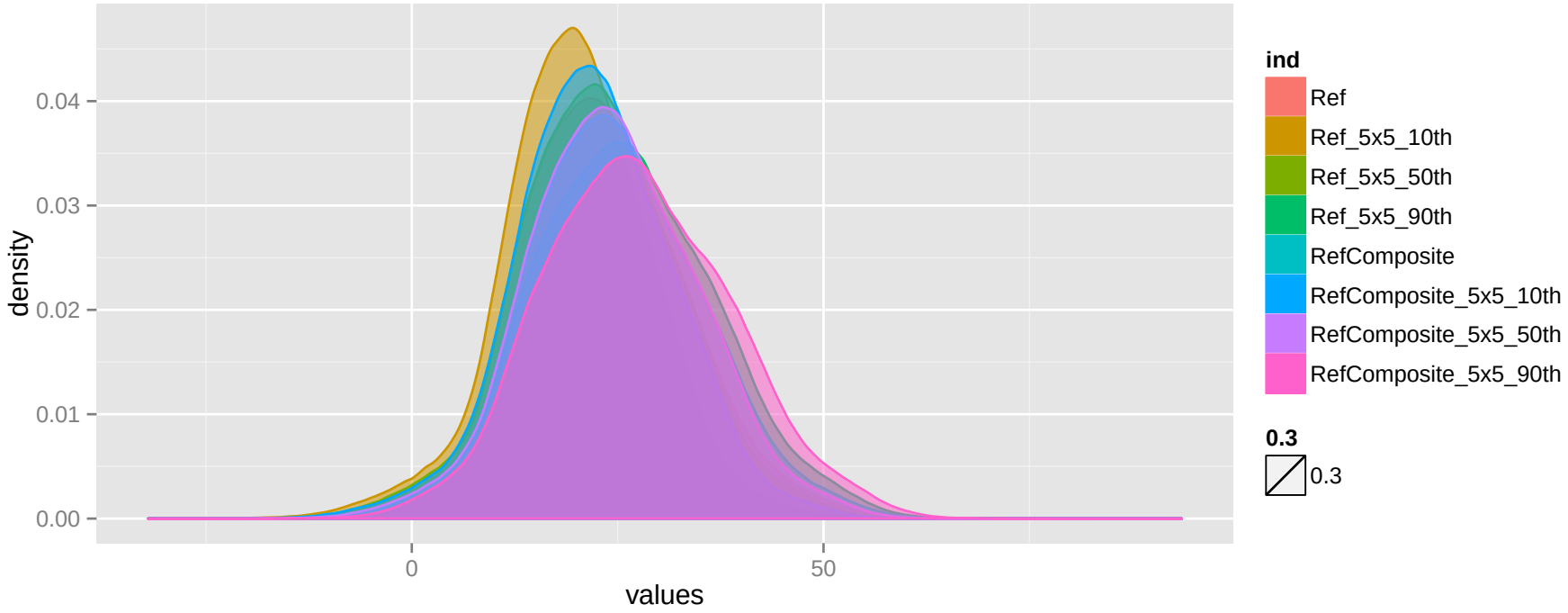
Densities by Measurement Group

Ref

```
train.refstack <- stack(as.data.frame(train)[,vars.refcols])
ggplot(train.refstack, aes(x=values)) +
  geom_density(aes(group=ind, colour=ind, fill=ind, alpha=0.3))
```

```
## Warning: Removed 2127591 rows containing non-finite values (stat_density).

## Warning: Removed 3192978 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2120484 rows containing non-finite values (stat_density).

## Warning: Removed 925685 rows containing non-finite values (stat_density).

## Warning: Removed 1760623 rows containing non-finite values (stat_density).

## Warning: Removed 2721293 rows containing non-finite values (stat_density).

## Warning: Removed 1765303 rows containing non-finite values (stat_density).

## Warning: Removed 647763 rows containing non-finite values (stat_density).
```



Rho

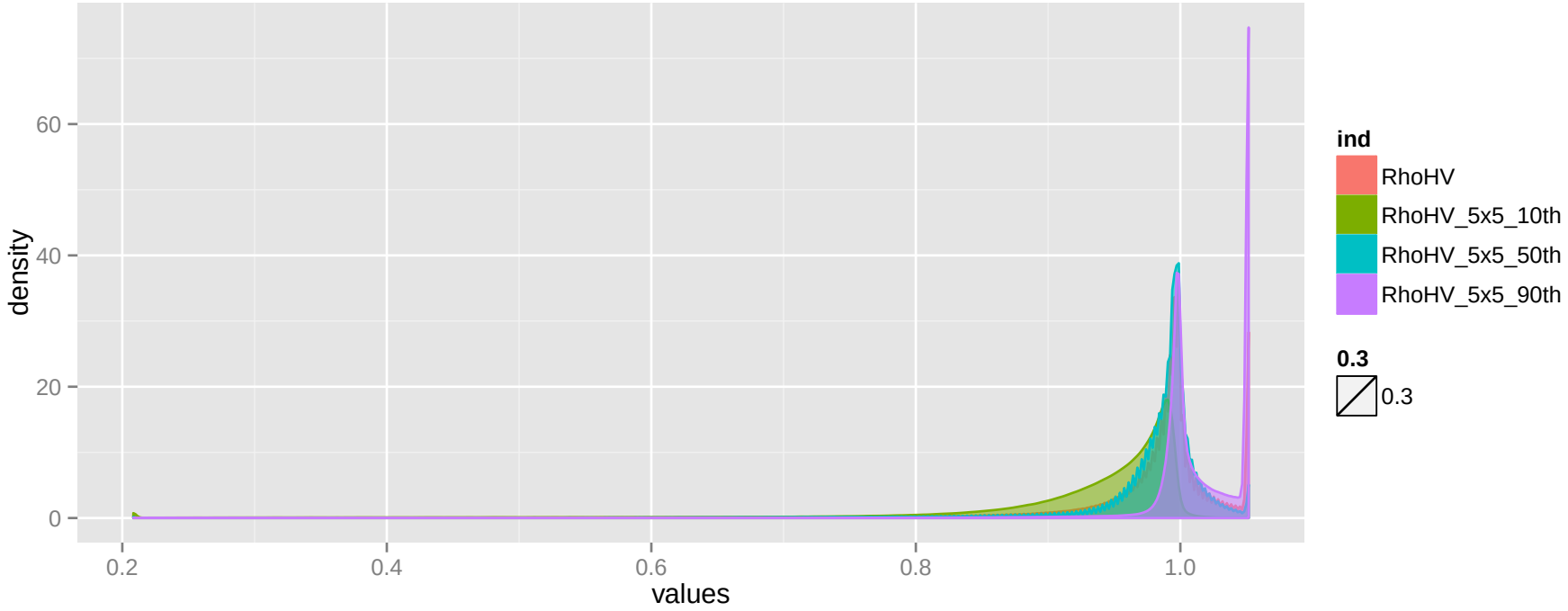
```
train.rhostack <- stack(as.data.frame(train)[,vars.rhocols])
ggplot(train.rhostack, aes(x=values)) +
  geom_density(aes(group=ind, colour=ind, fill=ind, alpha=0.3))
```

```
## Warning: Removed 3542050 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 4343812 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 3540398 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2571382 rows containing non-finite values (stat_density).
```



Zdr

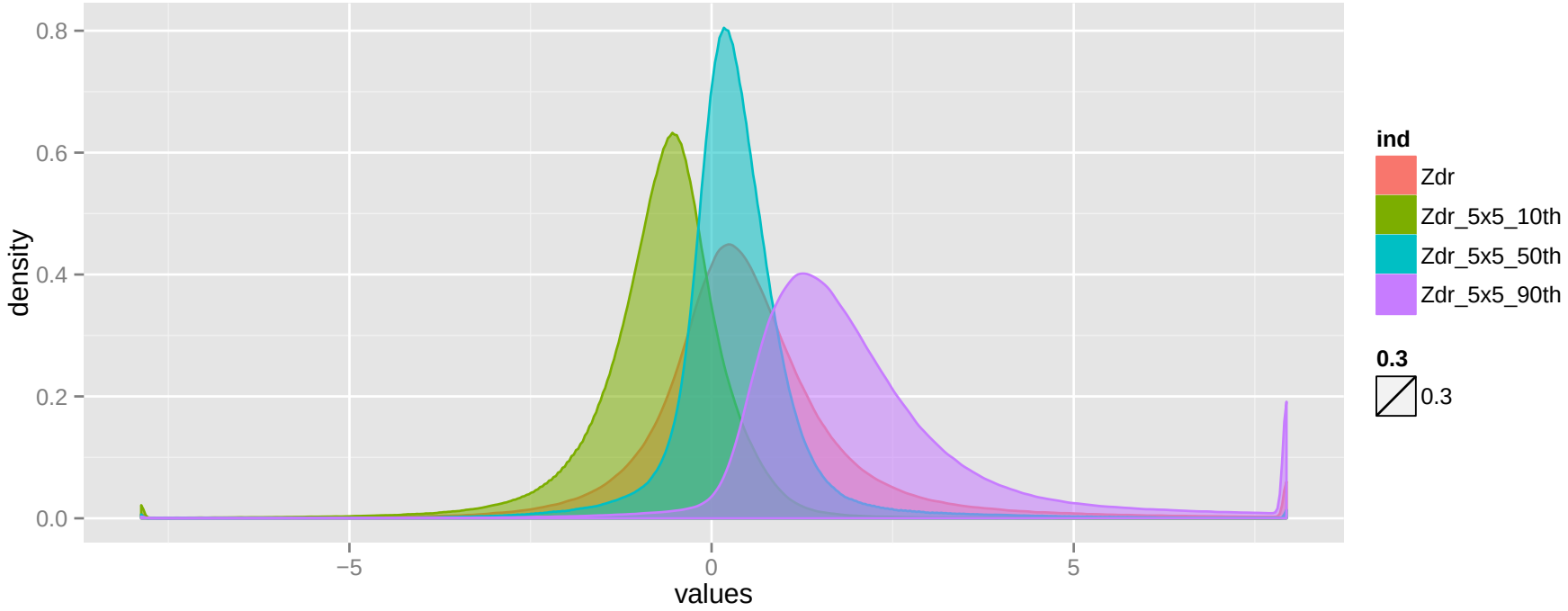
```
train.zdrstack <- stack(as.data.frame(train)[,vars.zdrcols])
ggplot(train.zdrstack, aes(x=values)) +
  geom_density(aes(group=ind, colour=ind, fill=ind, alpha=0.3))
```

```
## Warning: Removed 3542050 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 4343812 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 3540398 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 2571382 rows containing non-finite values (stat_density).
```



Kdp

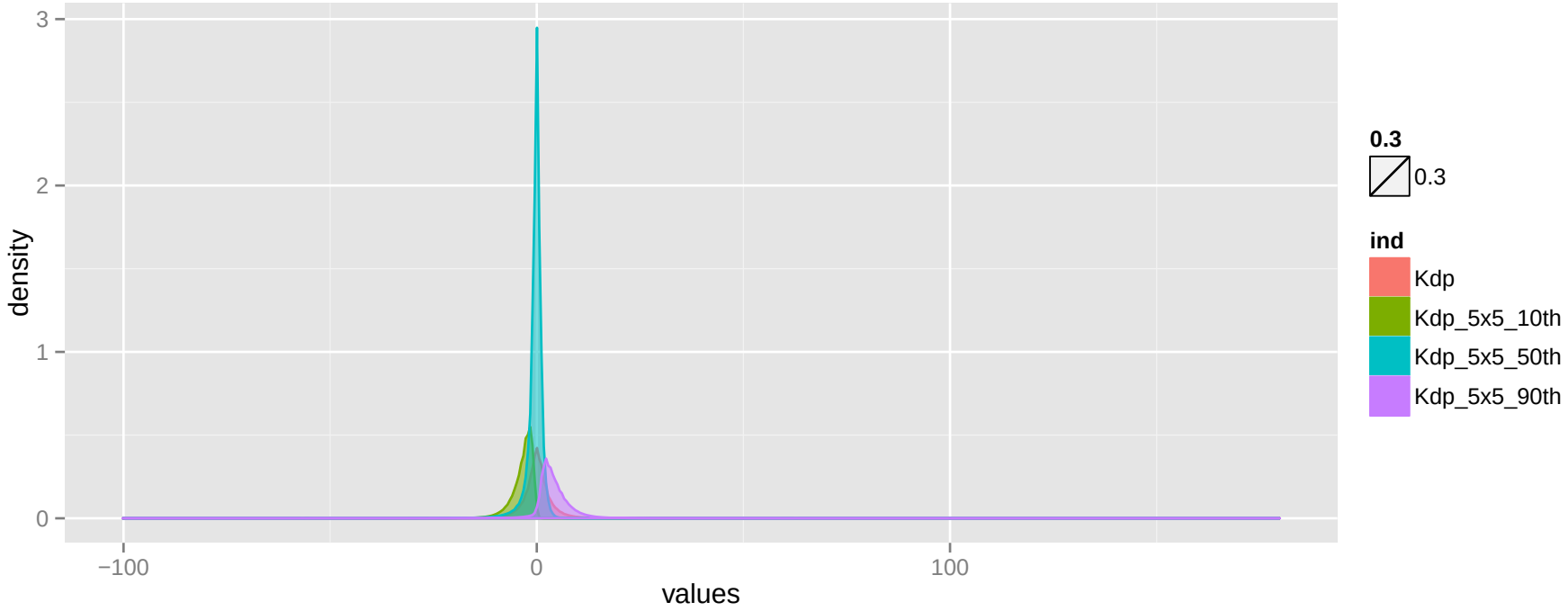
```
train.kdpstack <- stack(as.data.frame(train)[,vars.kdpcols])
ggplot(train.kdpstack, aes(x=values)) +
  geom_density(aes(group=ind, colour=ind, fill=ind, alpha=0.3))
```

```
## Warning: Removed 4294331 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 5048184 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 4289685 rows containing non-finite values (stat_density).
```

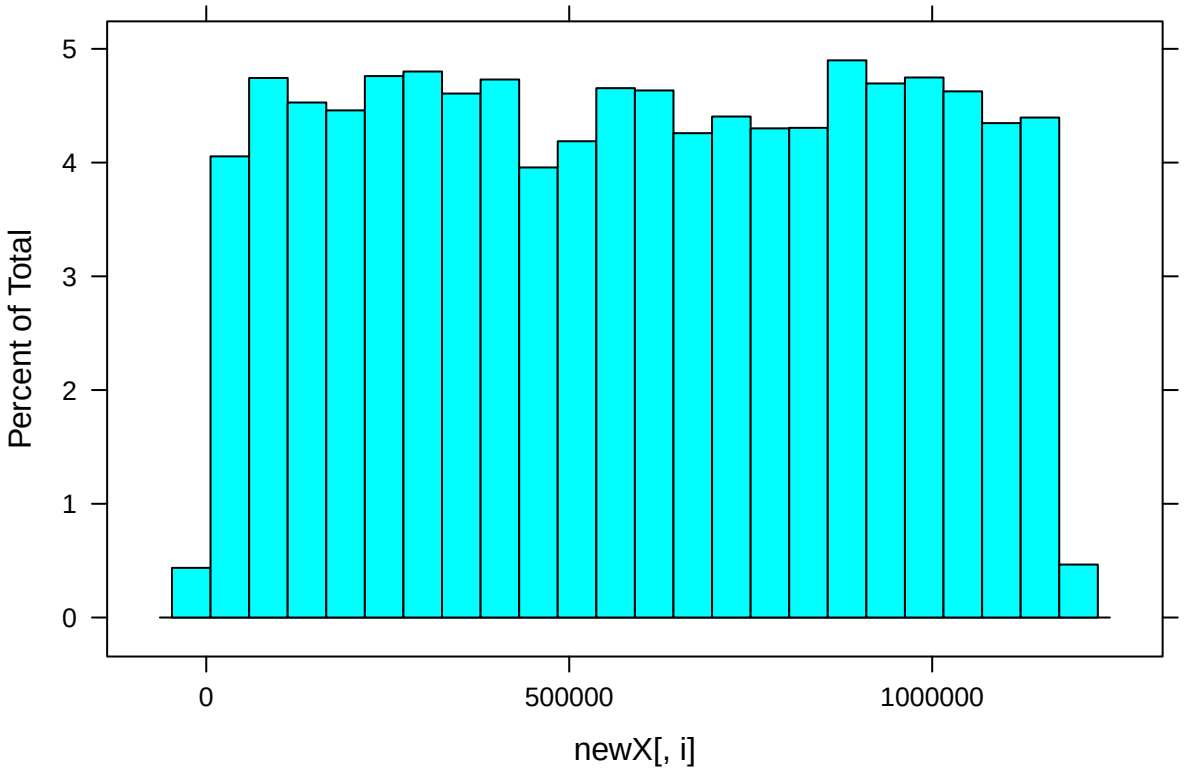
```
## Warning: Removed 3424190 rows containing non-finite values (stat_density).
```



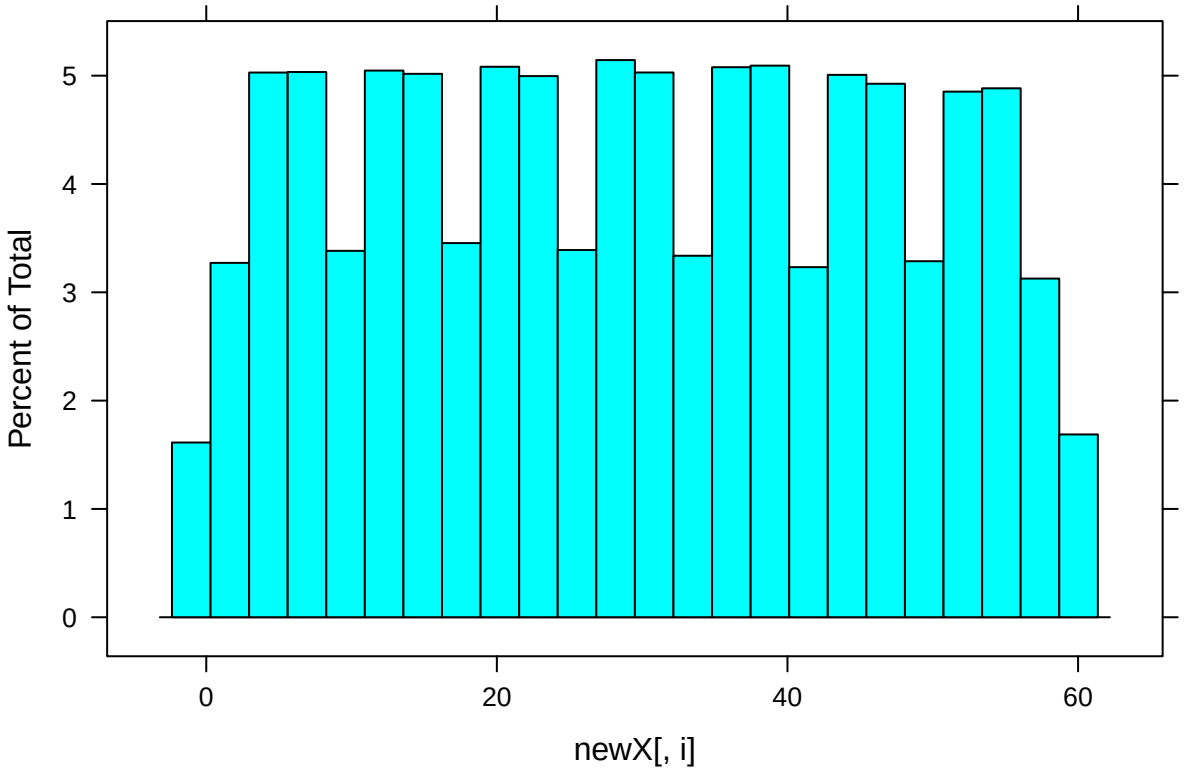
Histograms by Variable

```
histos <- apply(train,2,histogram)

#TODO:  ideally would set xlab on the plots as well
for (i in 1:(length(histos))) {
  cat(paste("## ", colnames(train)[i], " ## \n"))
  plot(histos[[i]])
  cat("\n\n")
}
```

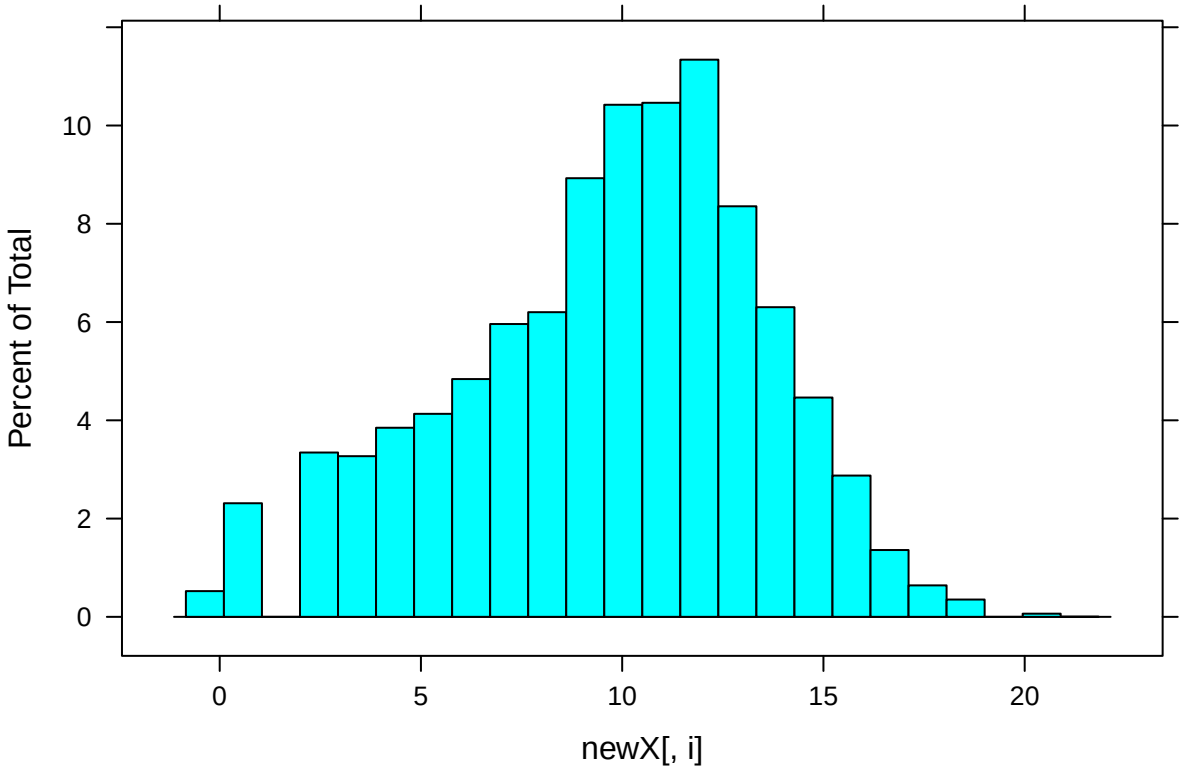
```
## ##  Id  ##
```



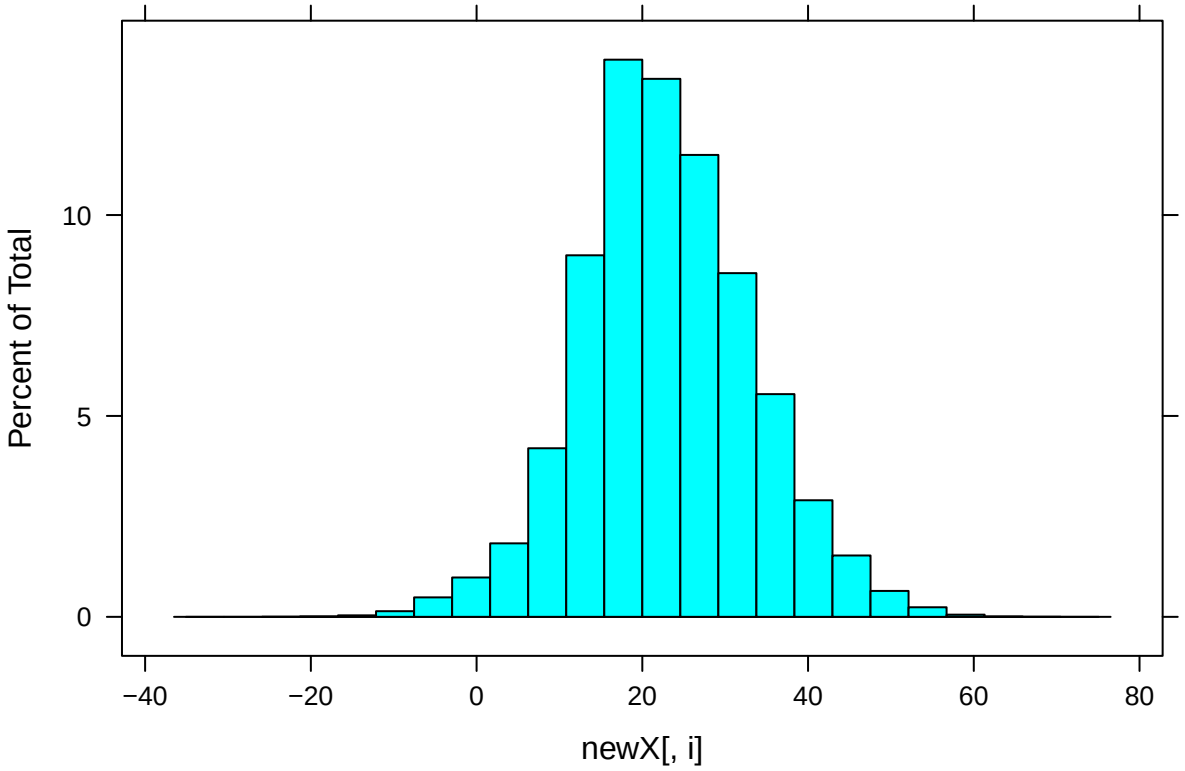
```
##
##
## ##  minutes_past  ##
```



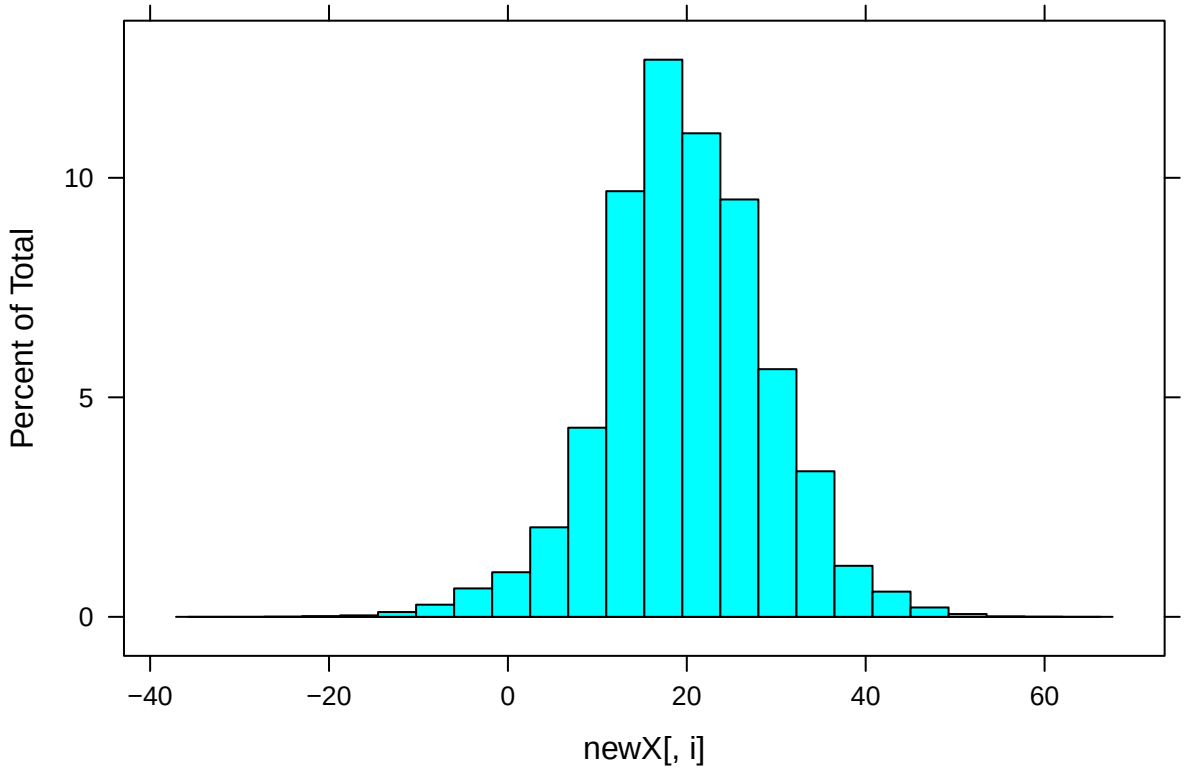

```
##
##
## ##  radardist_km  ##
```



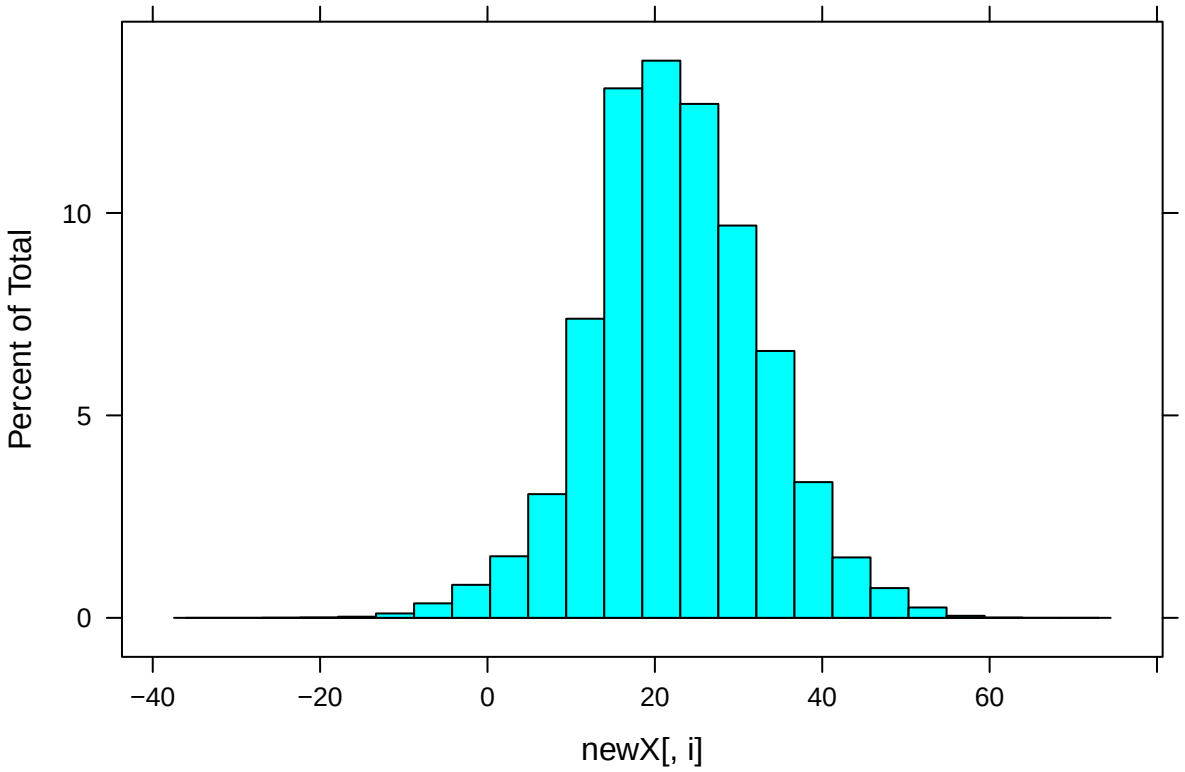

```
##
##
## ##  Ref  ##
```



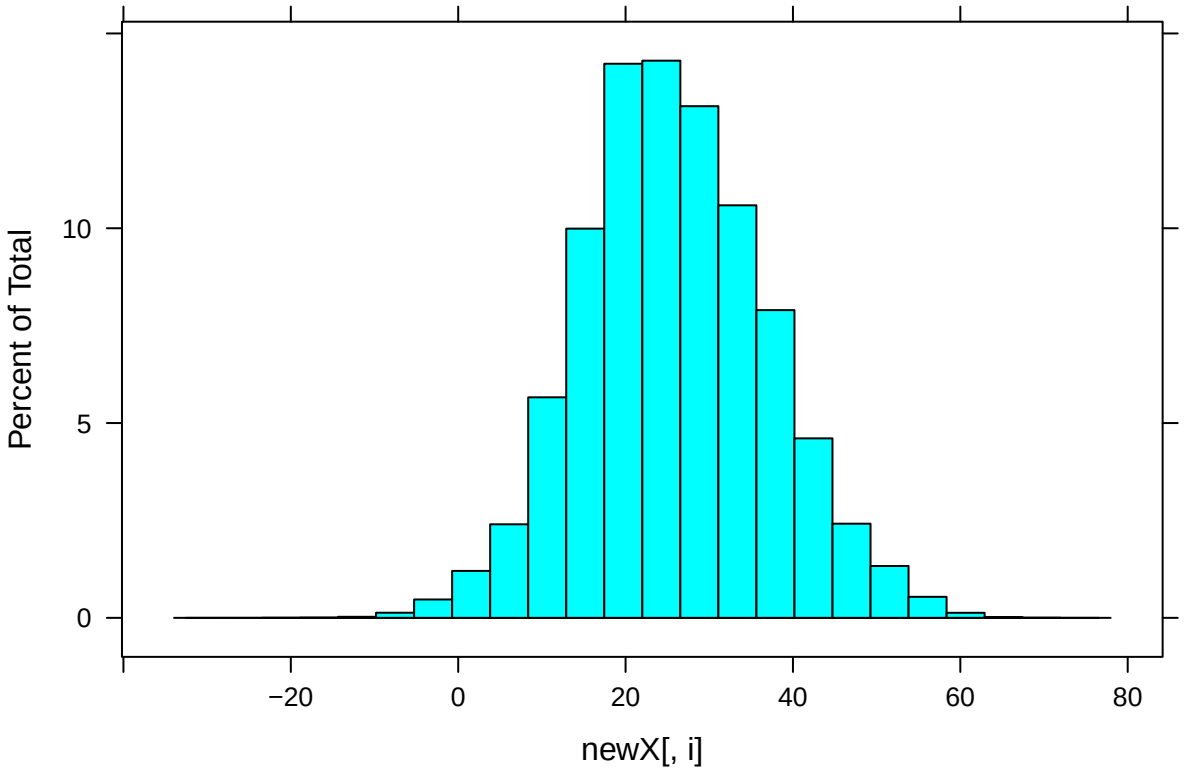

```
##
##
## ##  Ref_5x5_10th  ##
```



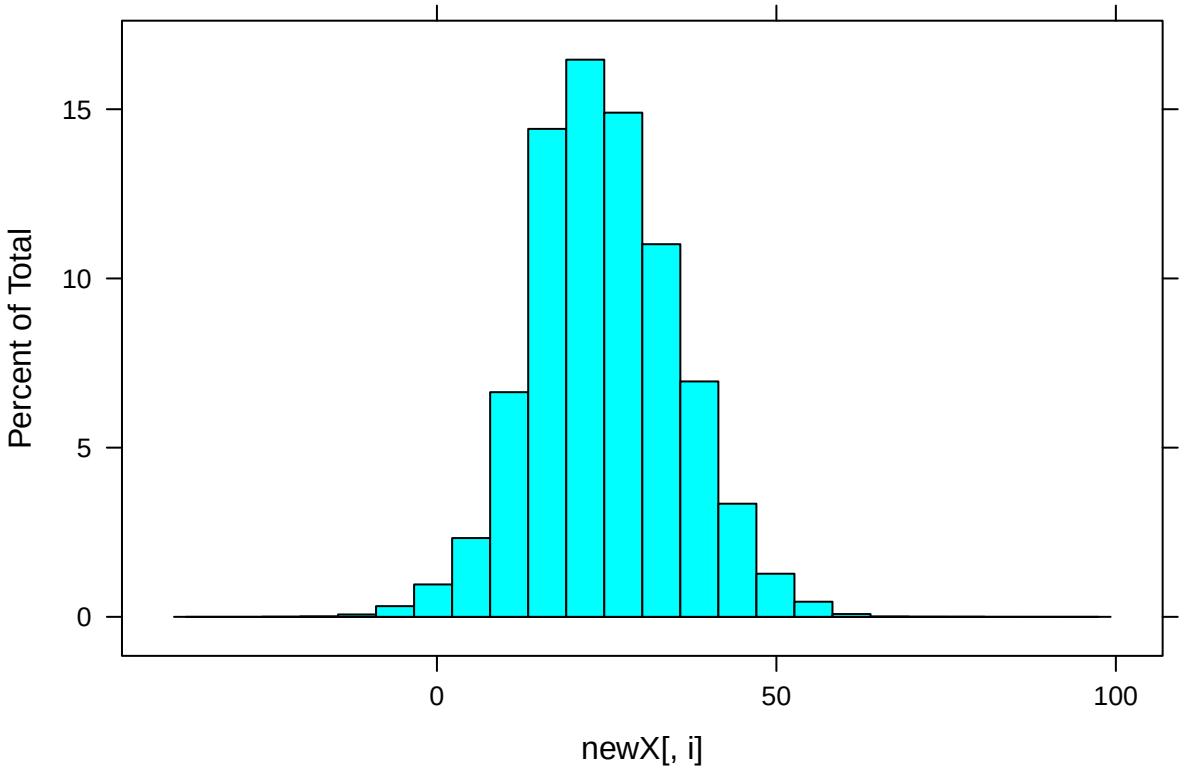

```
##
##
## ##  Ref_5x5_50th  ##
```



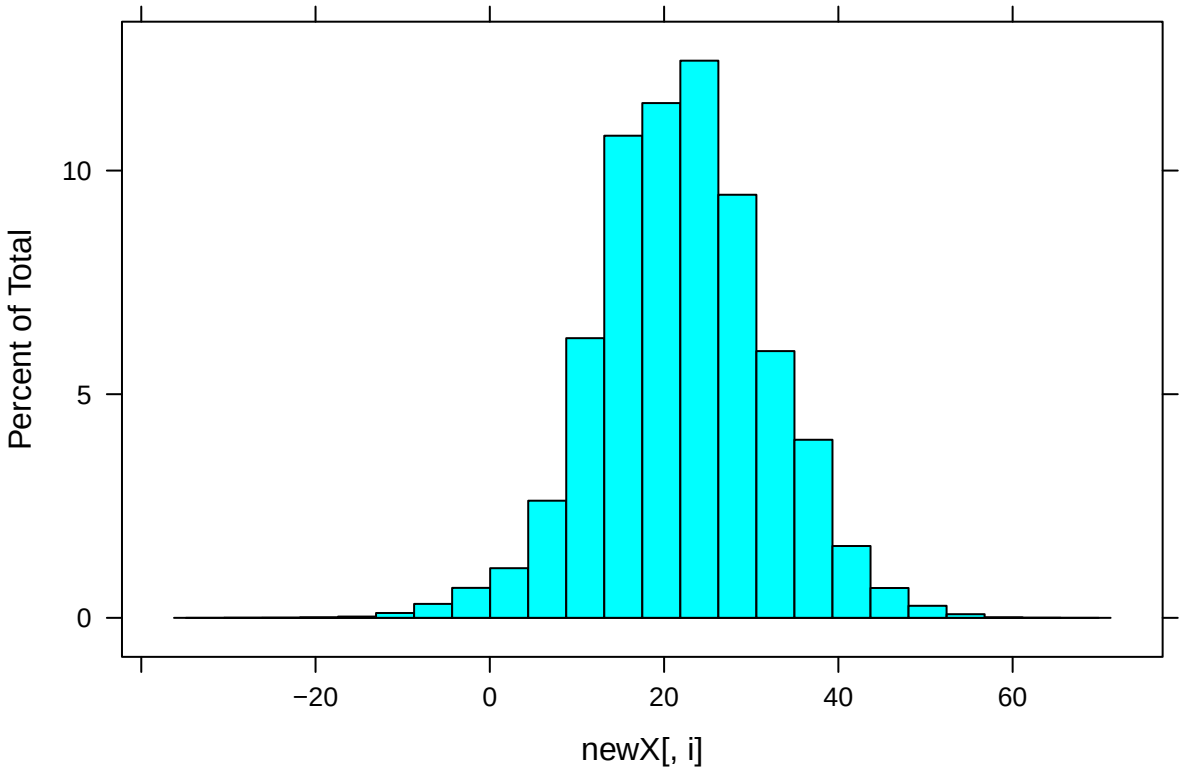

```
##
##
## ##  Ref_5x5_90th  ##
```



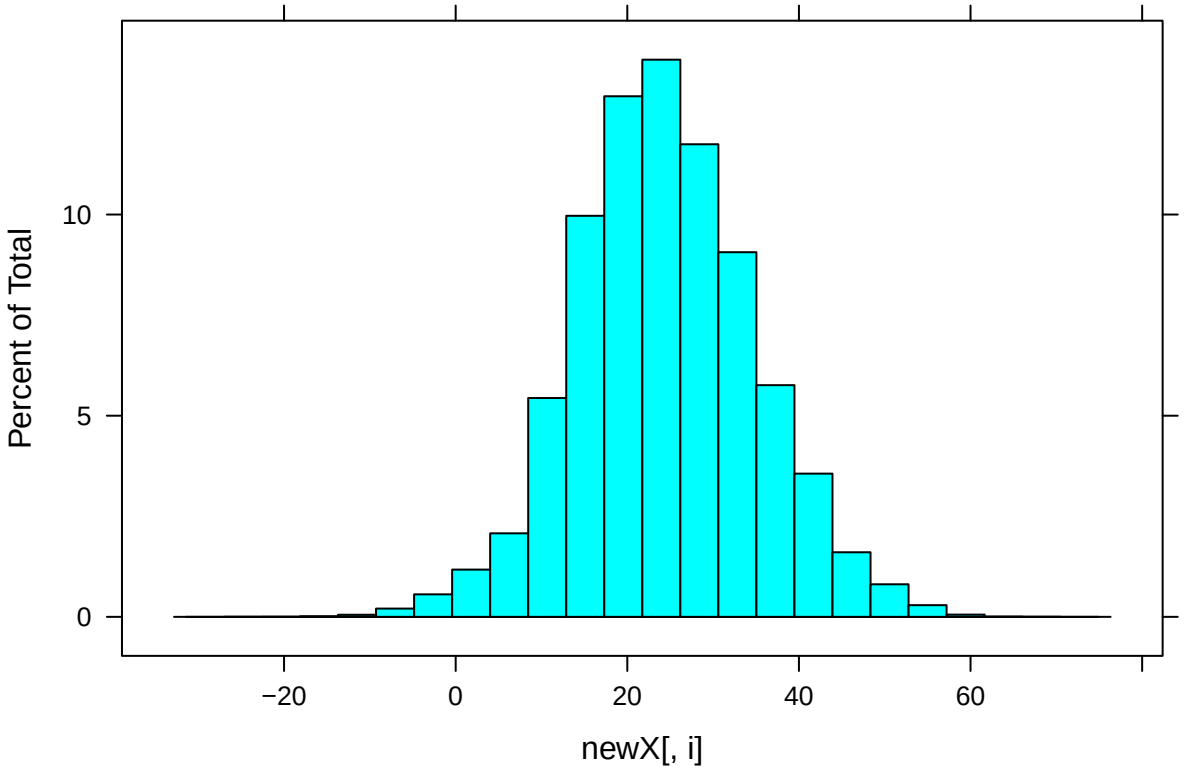

```
##
##
## ##  RefComposite  ##
```



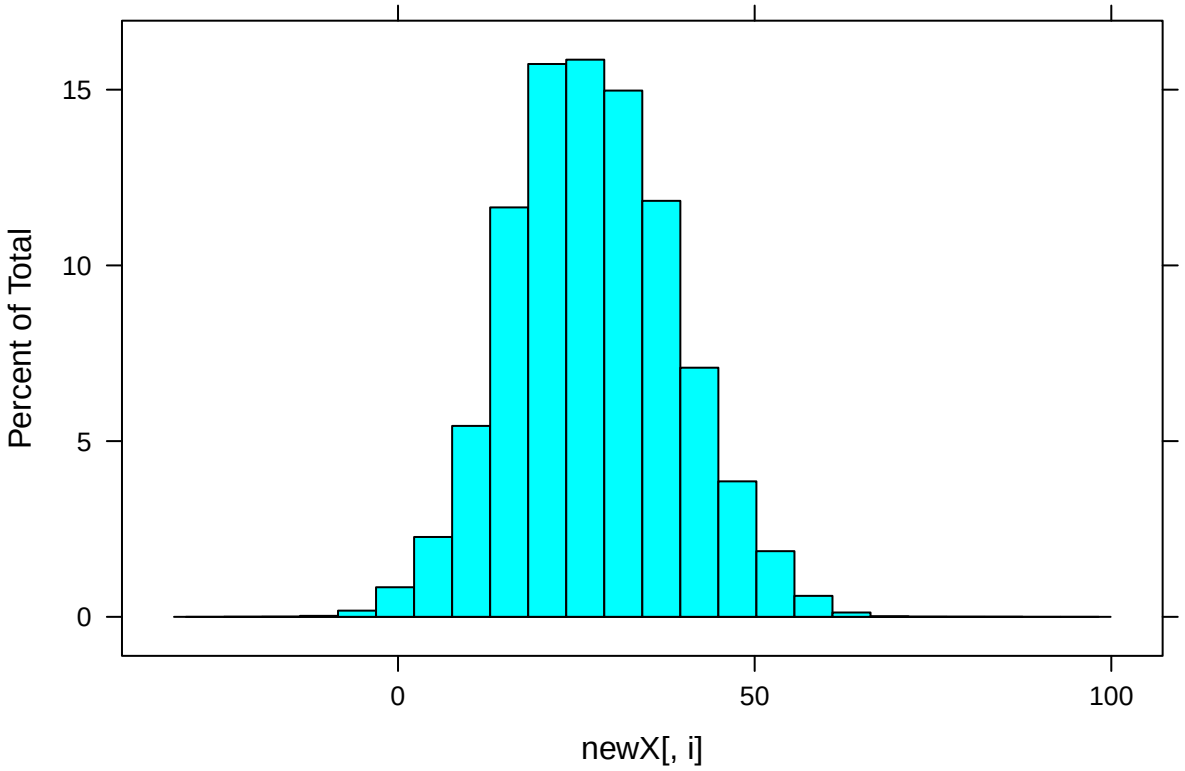

```
##
##
## ##  RefComposite_5x5_10th  ##
```



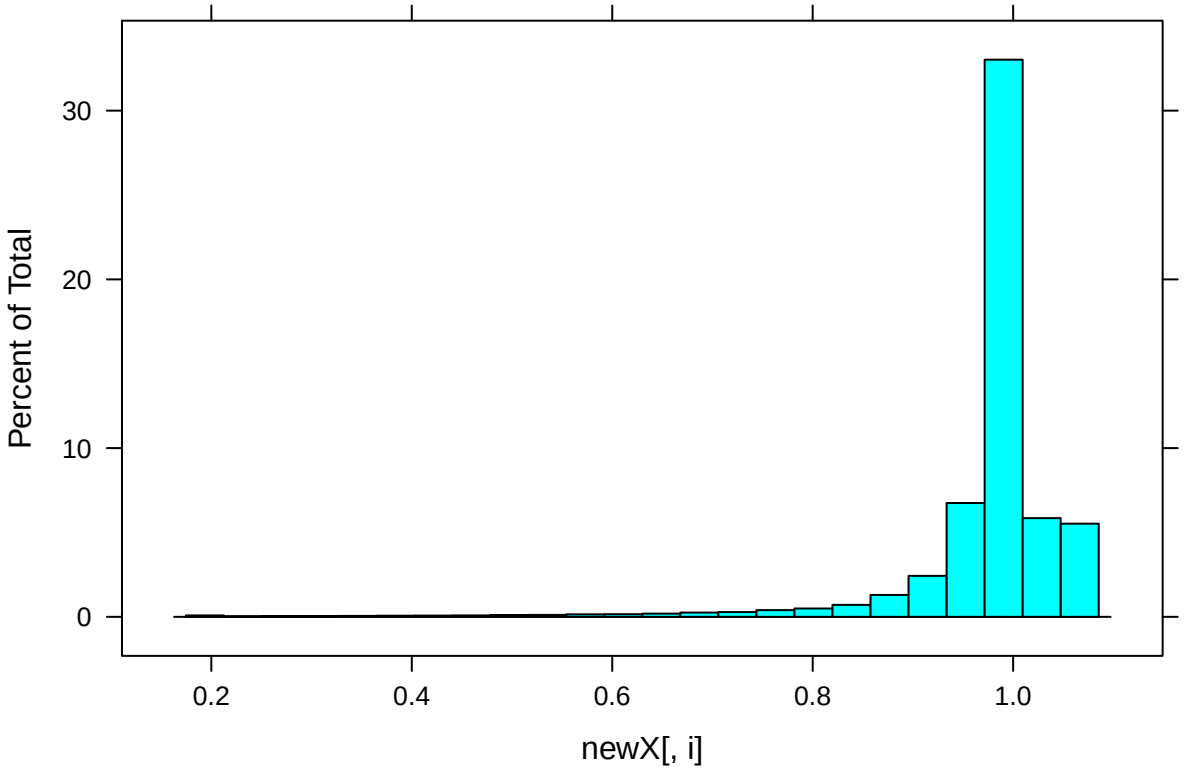

```
##
##
## ##  RefComposite_5x5_50th  ##
```



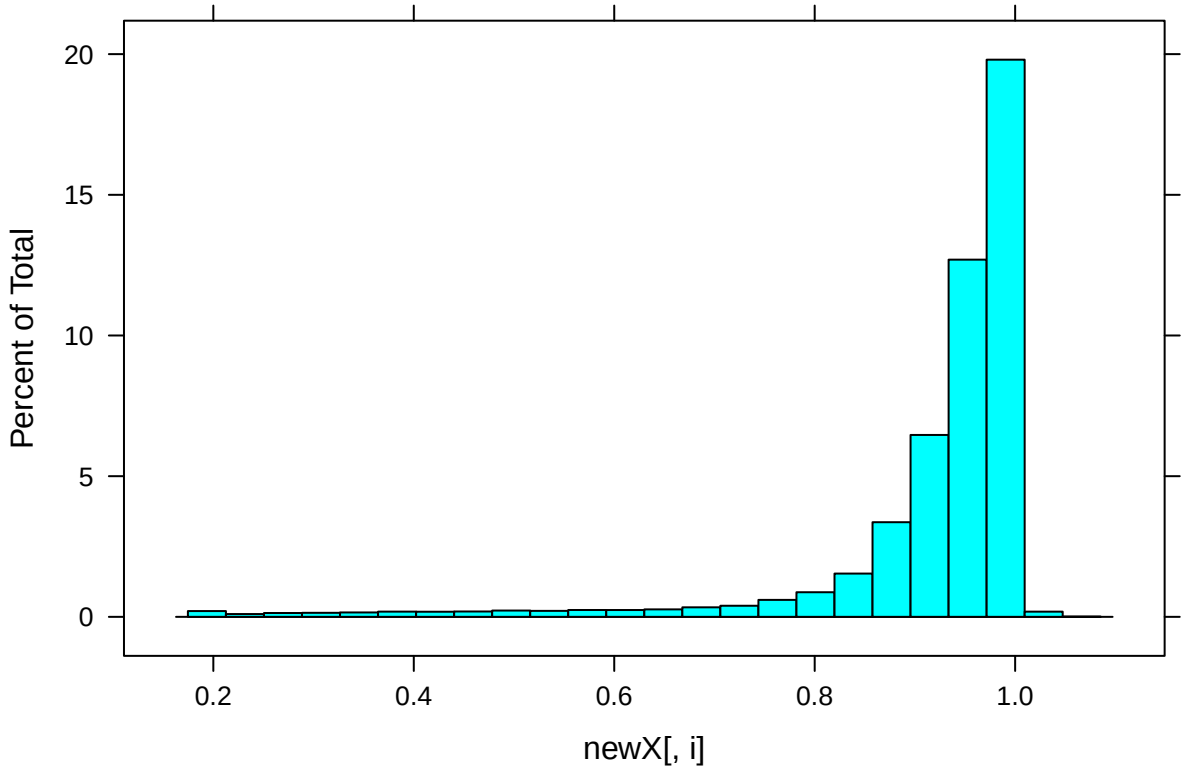

```
##
##
## ##  RefComposite_5x5_90th  ##
```



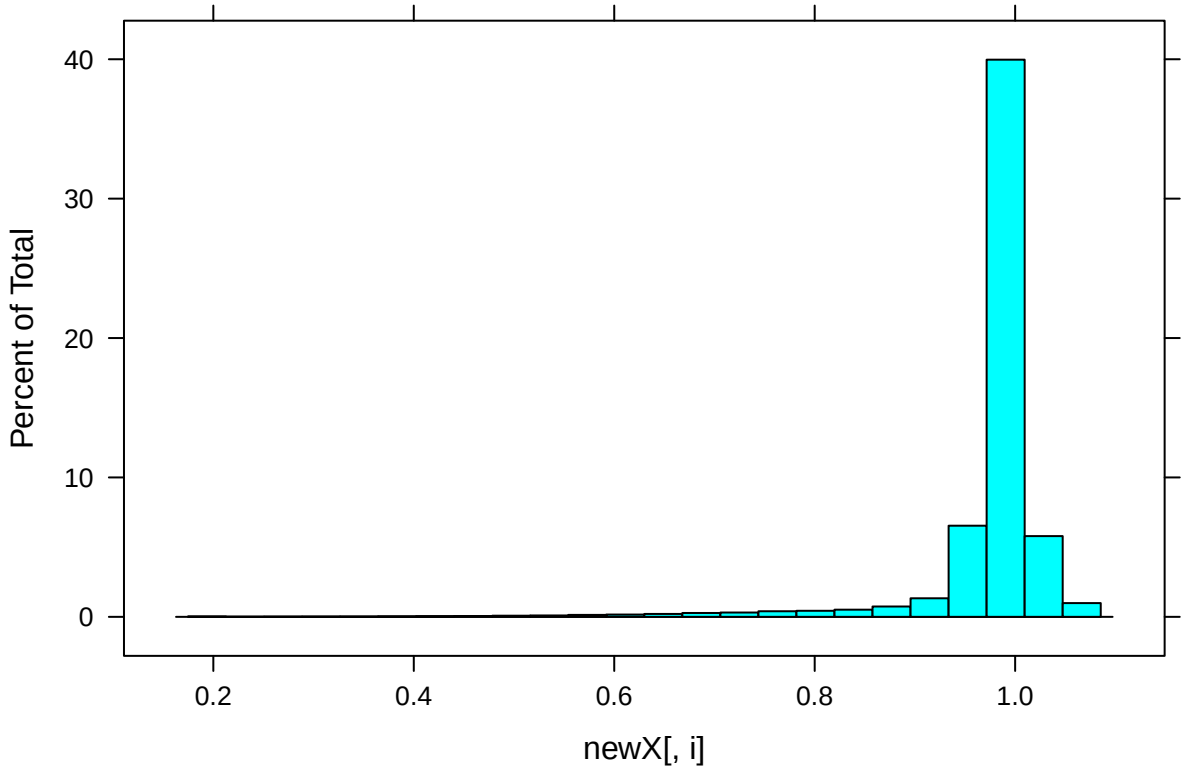

```
##
##
## ##  RhoHV  ##
```



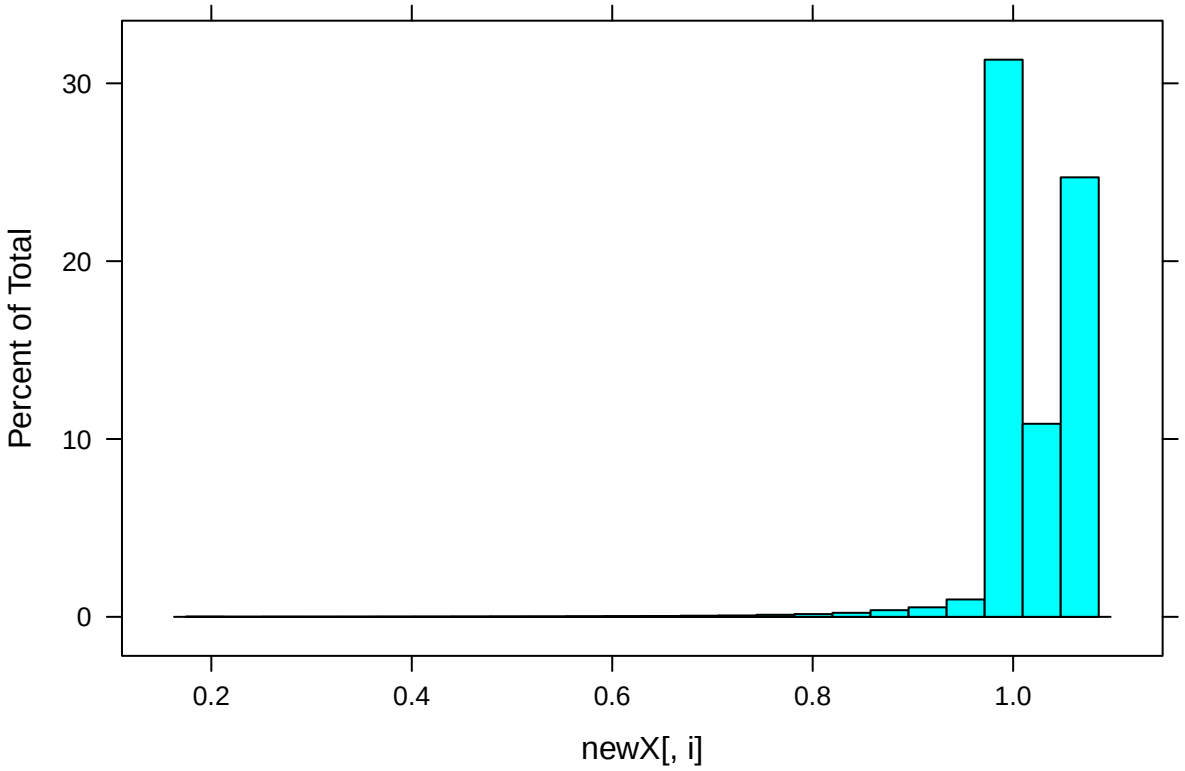

```
##
##
## ##  RhoHV_5x5_10th  ##
```



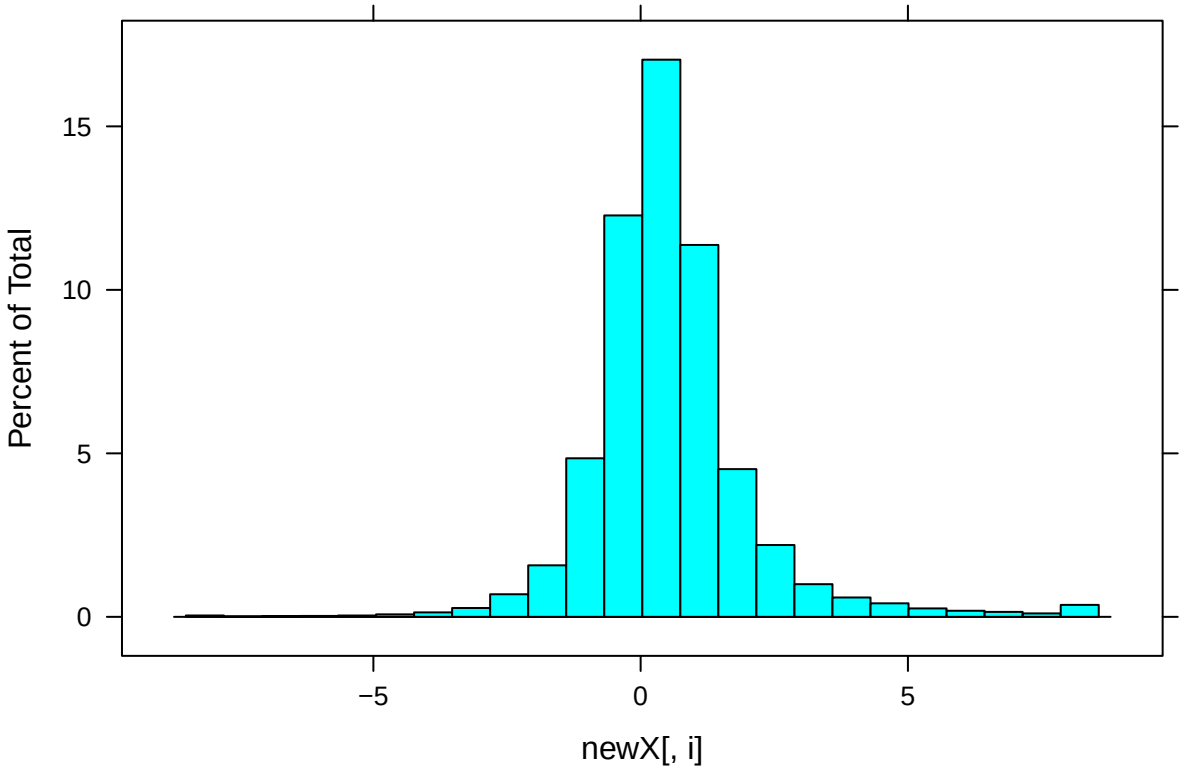

```
##
##
## ##  RhoHV_5x5_50th  ##
```



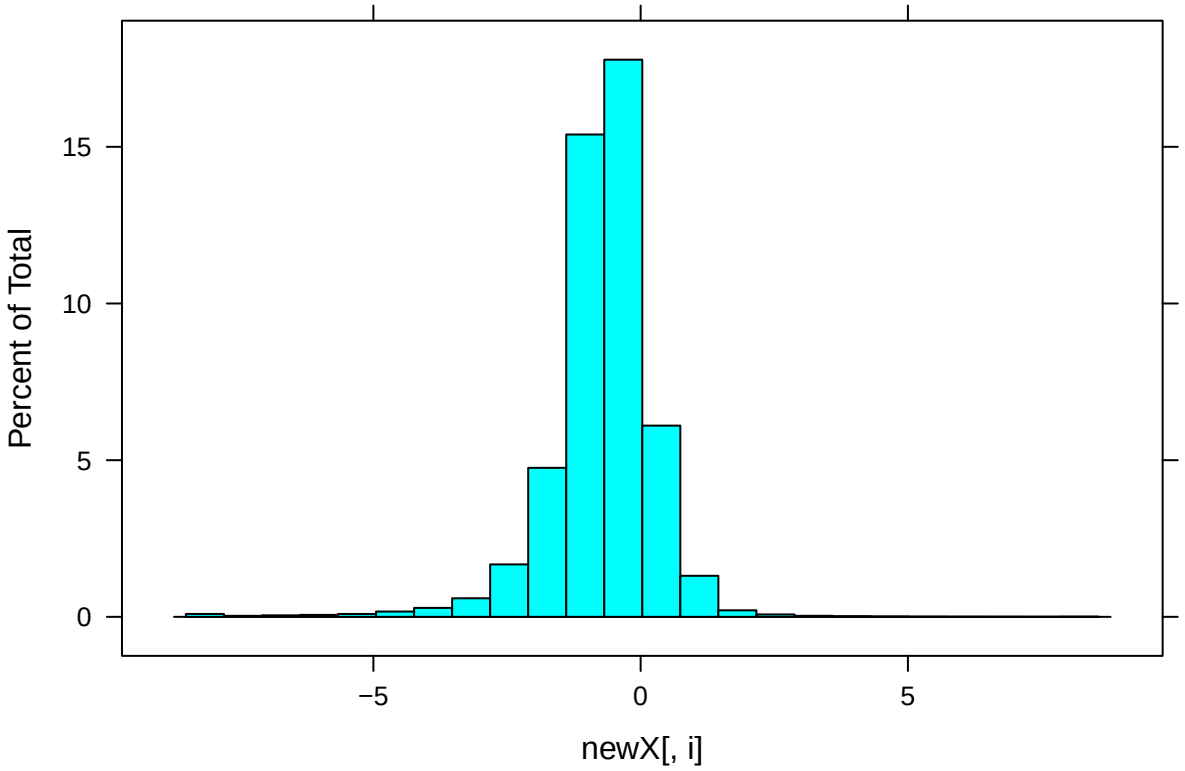

```
##
##
## ##  RhoHV_5x5_90th  ##
```



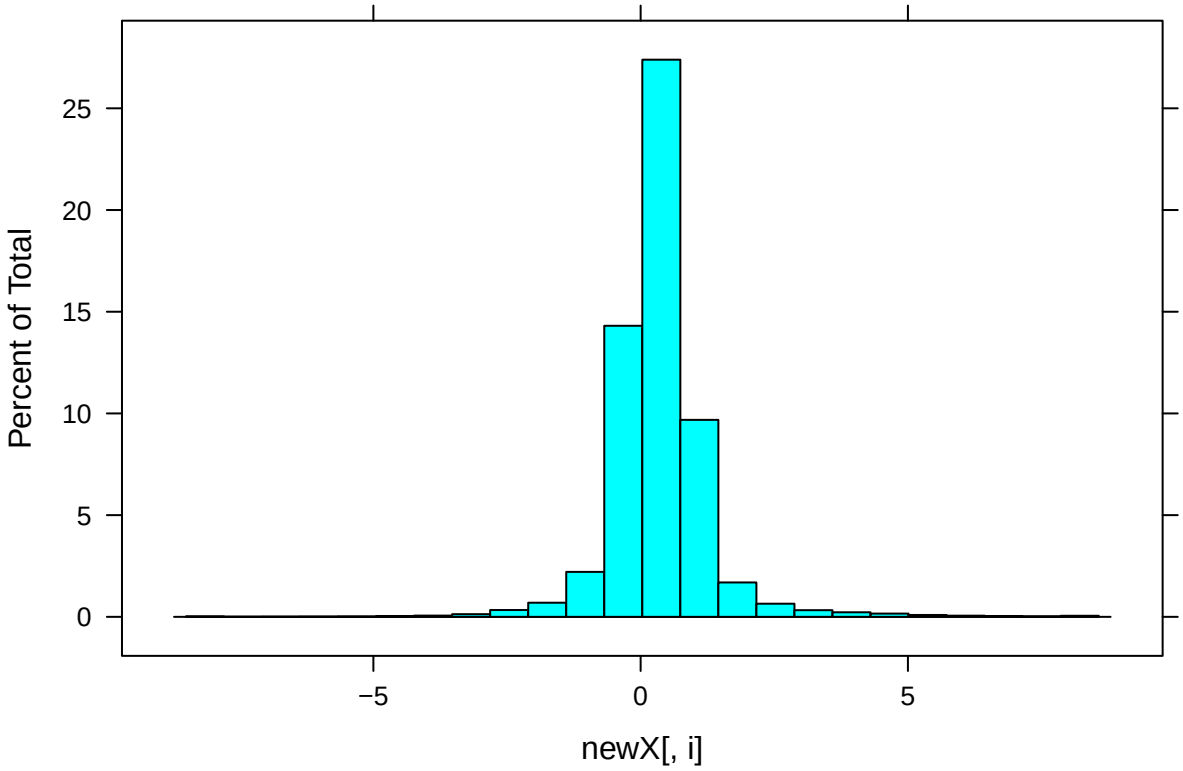

```
##
##
## ##  Zdr  ##
```



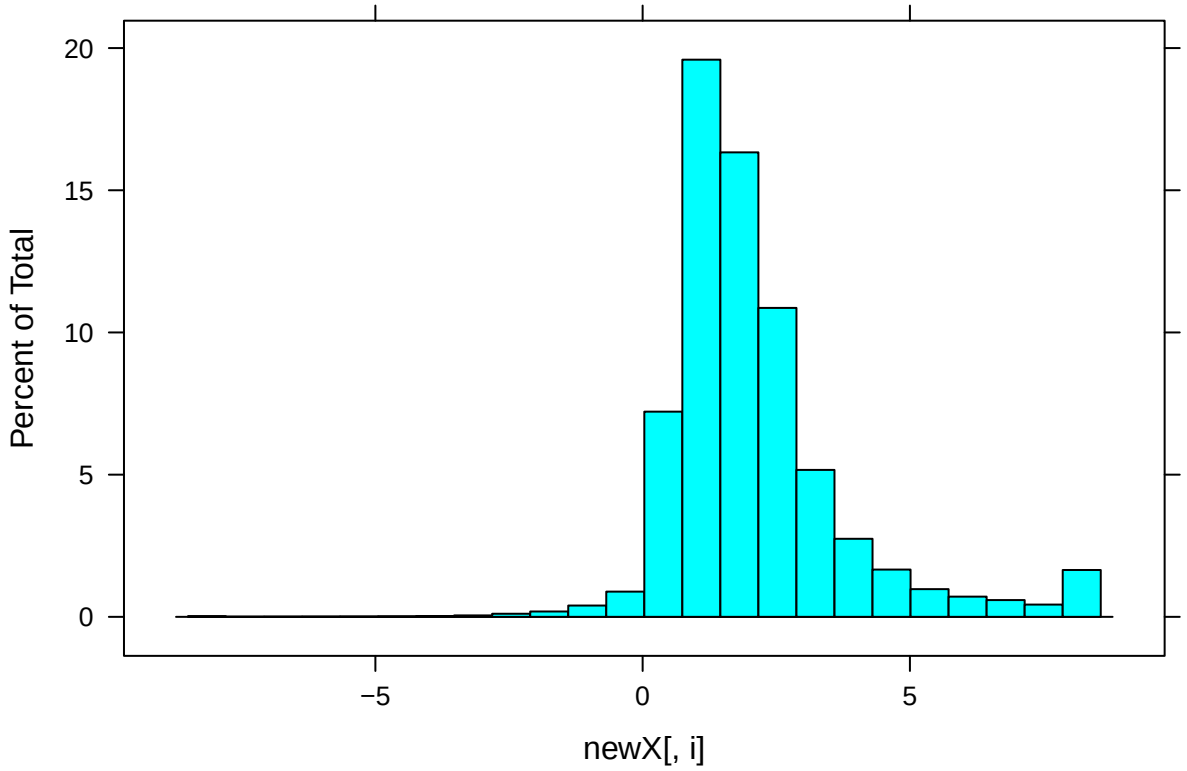

```
##
##
## ##  Zdr_5x5_10th  ##
```



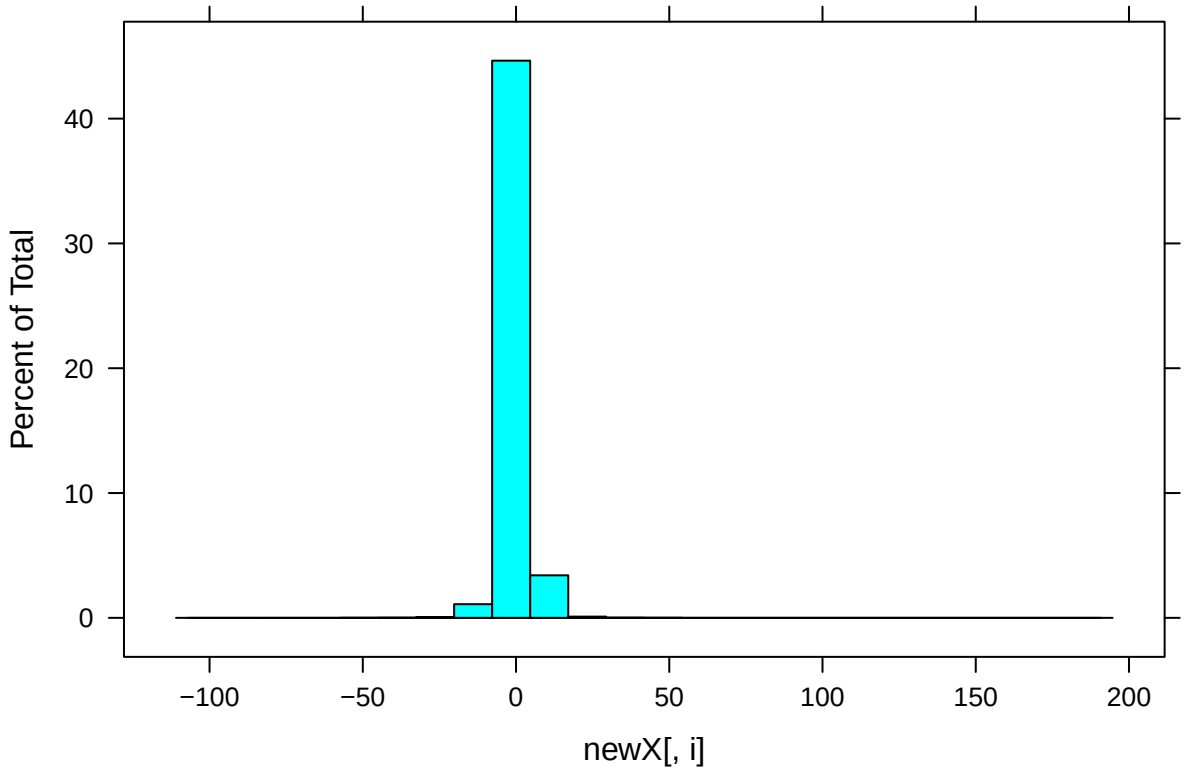

```
##
##
## ##  Zdr_5x5_50th  ##
```



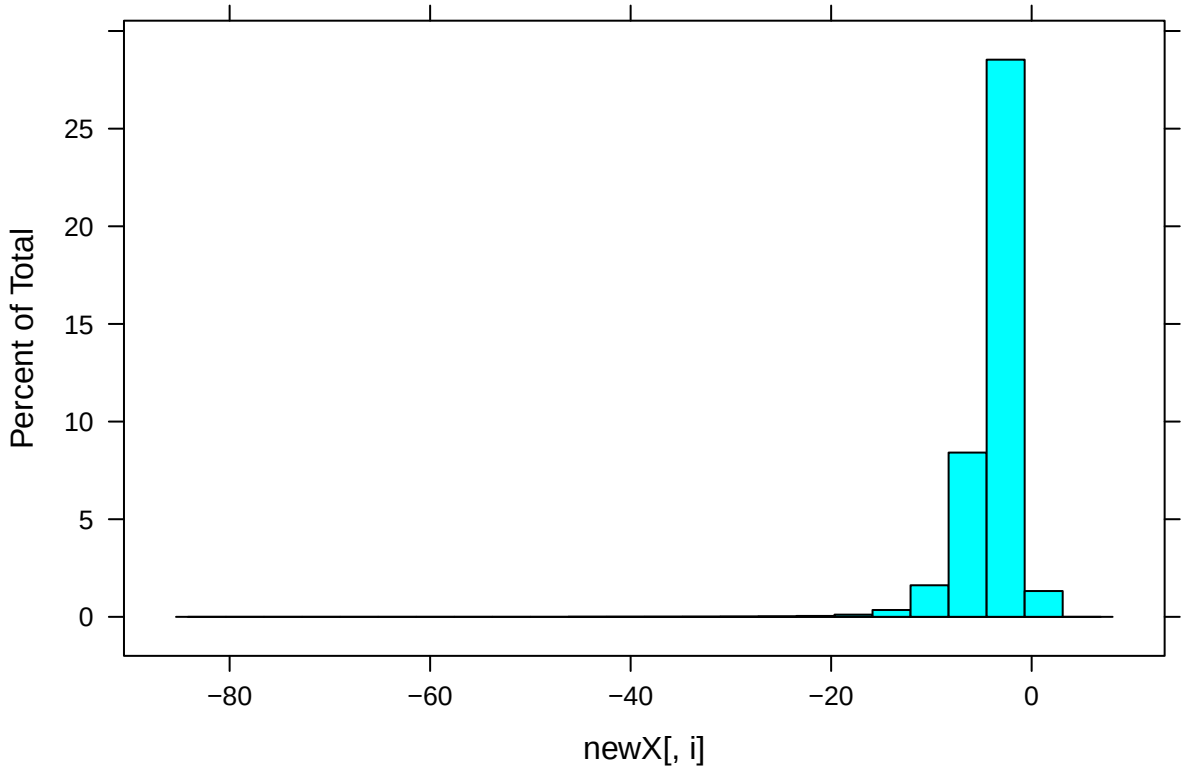

```
##
##
## ##  Zdr_5x5_90th  ##
```



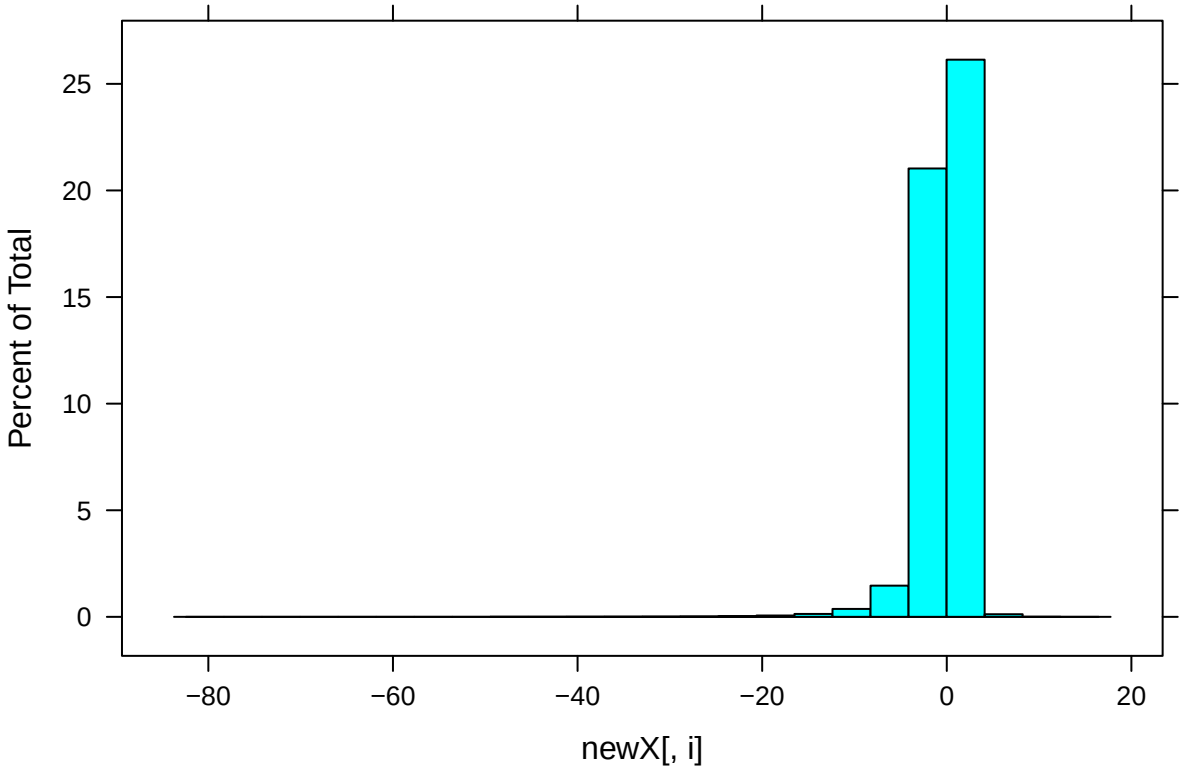

```
##
##
## ##  Kdp  ##
```



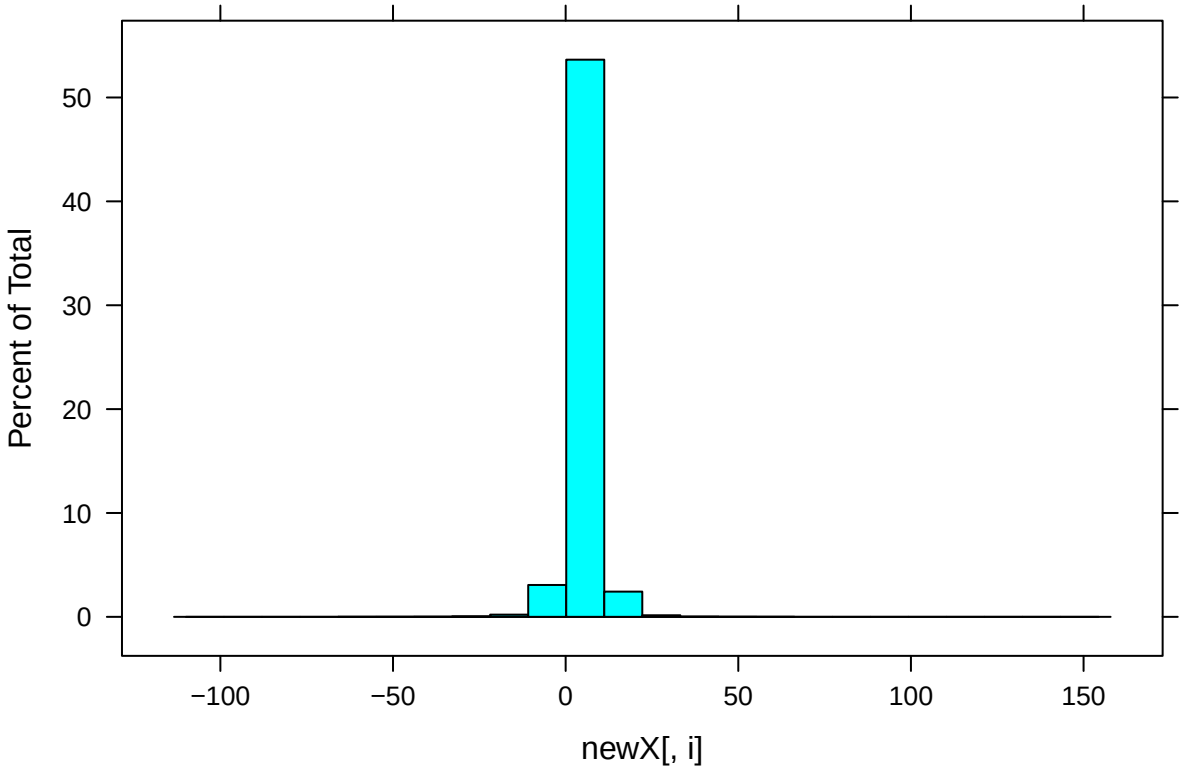

```
##
##
## ##  Kdp_5x5_10th  ##
```



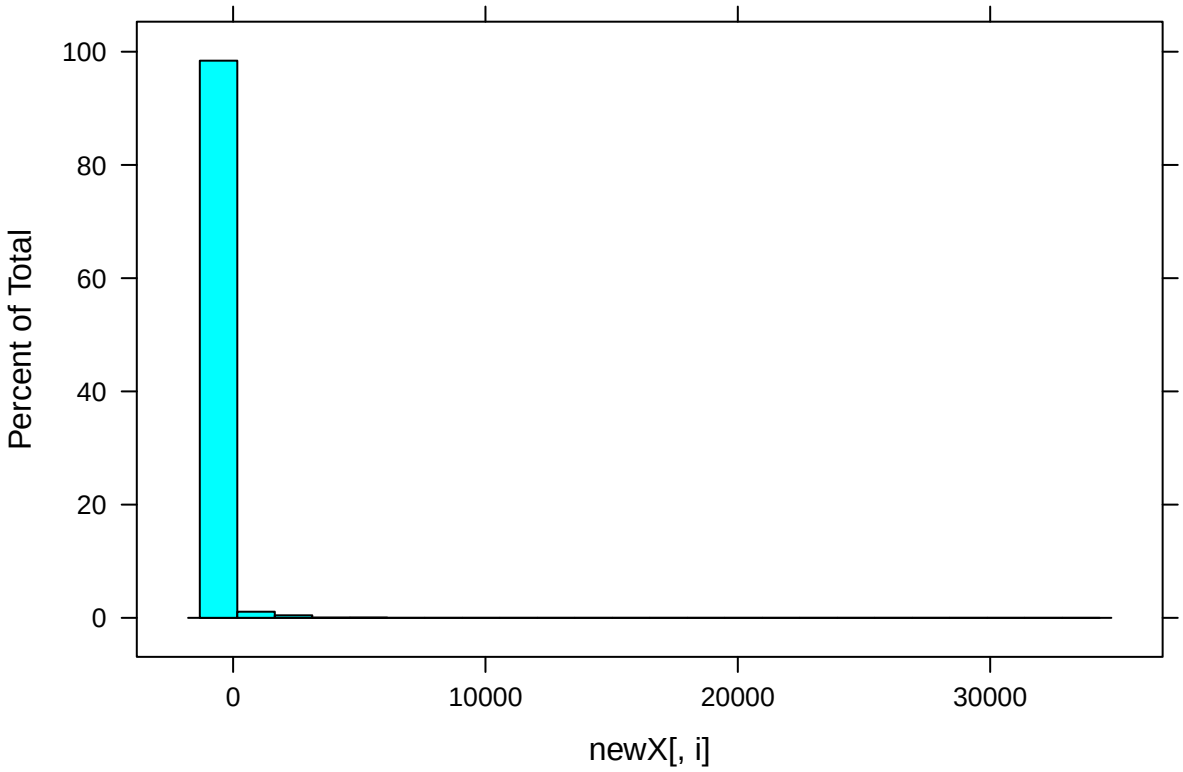

```
##
##
## ##  Kdp_5x5_50th  ##
```




```
##
##
## ##  Kdp_5x5_90th  ##
```




```
##
##
## ##  Expected  ##
```


Percent of Total
100
80
60
40
20
0
0
10000
20000
30000
newX[, i]