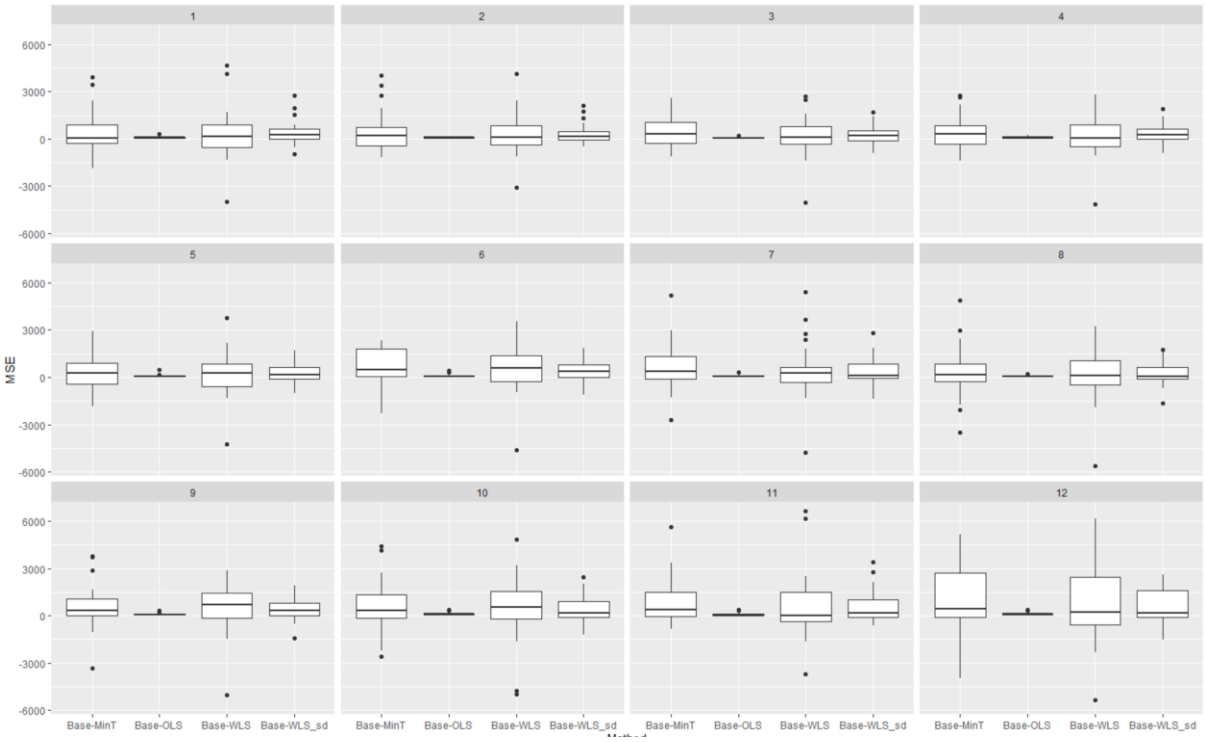
George 30/8/2019

Some differences with the past (let's concentrate on the JASA paper - the data and evaluation in the 2009 IJF paper is even more different).

- Evaluation in JASA is done by level not across the whole hierarchy. Evaluating across the whole hierarchy using MSE the errors on the top level will dominate due to the scale. Notice that in the JASA paper at the top-level OLS did better than MinT and MinT did worse than base as h increased.
- The training samples in JASA started from L=96 (end of 2005) and ended by the end of 2016 (you will see below why I am mentioning this).
- In JASA we did an expanding window – not sure that this help our case in any way and whether it makes a difference.

I am running two experiments one with L=100 and one with L=200 (training length). They both stopped when my computer fell asleep around iteration 26.

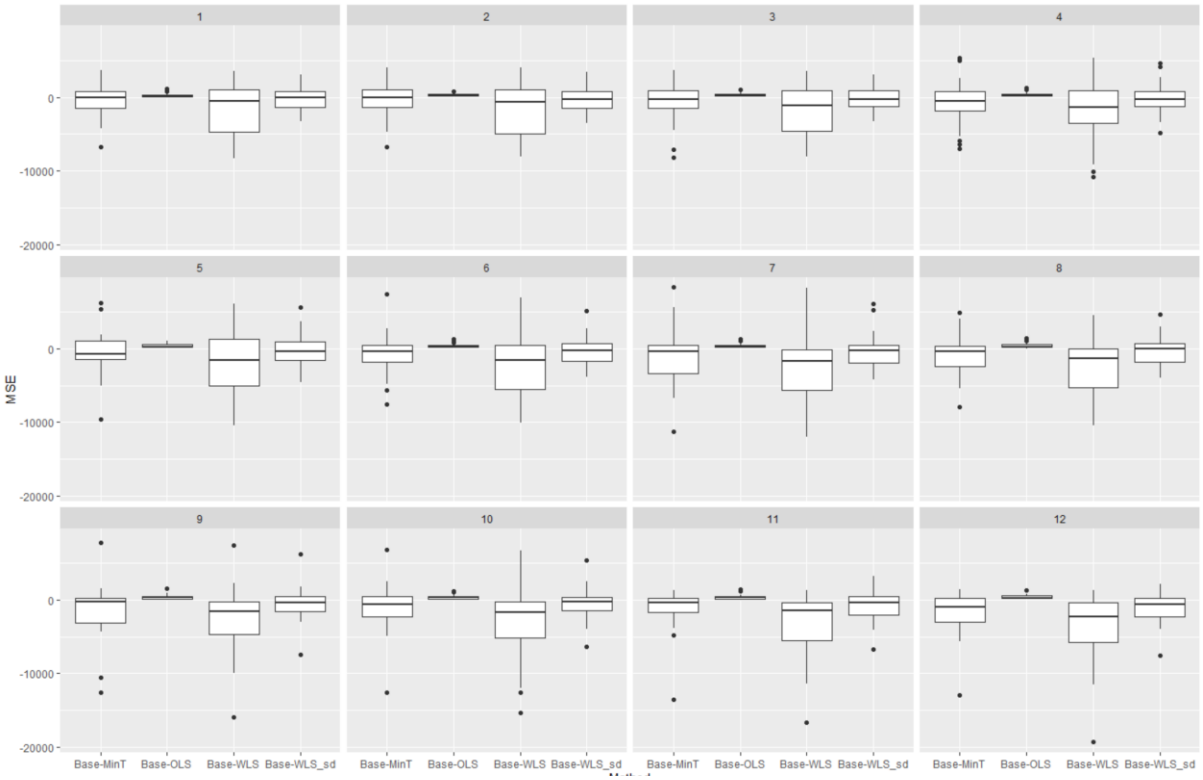
For L=100 (26 replications ran) - below are the results. This is the type of thing I expect to see. Ok MinT does have higher variation than OLS. With n=100 I don't think this is too bad. WLS_sd as Tas suspected/explained dampens this variation compared to normal WLS variance scaling. (WLS_sd is the fourth column labelled and instead of using variance I am using standard deviation as W).




The averages below show that MinT on average outperforms all methods.

```
> DF_MSE %>% group_by(`R-method`) %>%
+   summarise(MSE = mean(MSE)) %>%
+   spread(key = `R-method`, value = MSE)
# A tibble: 1 x 6
   Base `Bottom-up` `MinT(Shrink)`   OLS   WLS WLS_sd
  <dbl>       <dbl>          <dbl> <dbl> <dbl>  <dbl>
1 4888.       5049.          4283. 4797. 4469.  4484.
> |
```

For L=200 (27 replications ran). The results look sooooo different. I should be getting a better estimate of the variance for all methods that involve W but these are now much worse and…




… the only average that beats Base now is OLS.

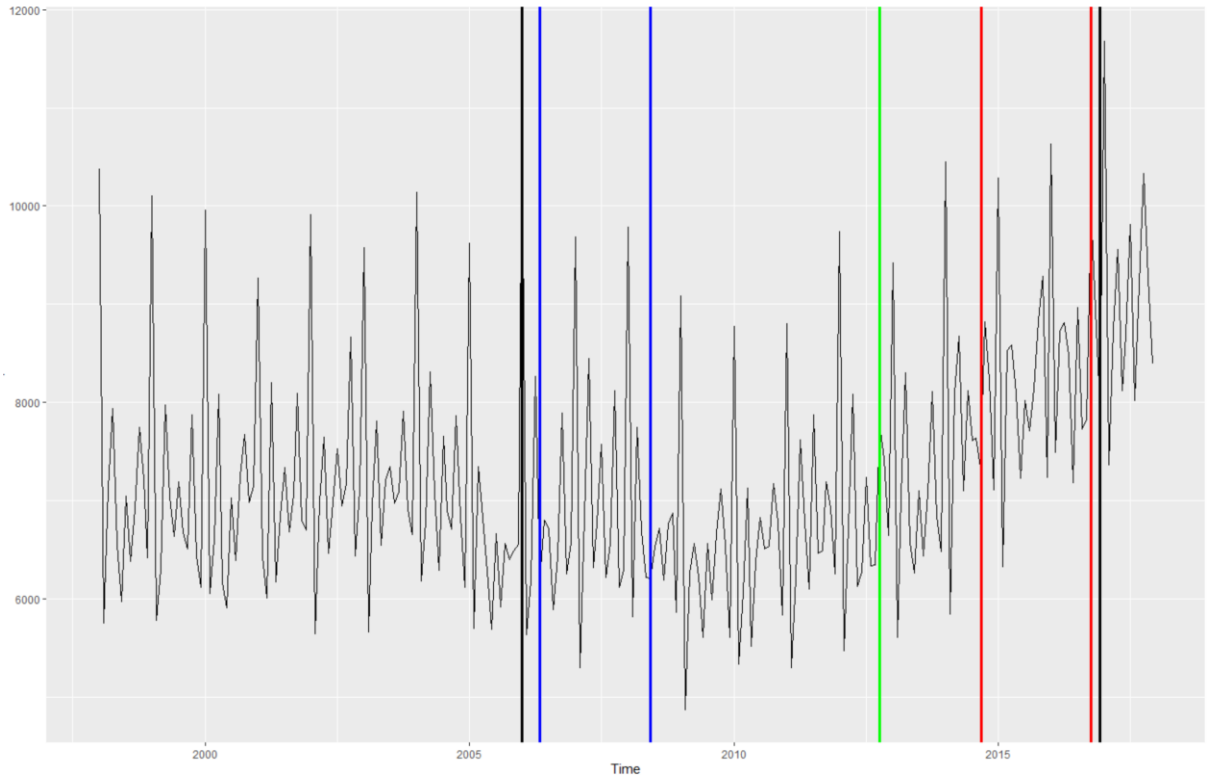
```
> DF_MSE %>% group_by(`R-method`) %>%
+   summarise(MSE = mean(MSE)) %>%
+   spread(key = `R-method`, value = MSE)
# A tibble: 1 x 6
   Base `Bottom-up` `MinT(Shrink)`   OLS   WLS WLS_sd
  <dbl>       <dbl>          <dbl> <dbl> <dbl>  <dbl>
1 5285.       9497.          6201. 4891. 7713.  5709.
```

Ok why is this happening?

Below are the test sets for the different settings for the top-level series.

- Black lines show the JASA period evaluation (remember MinT did worse than OLS for the top level).
- Blue are for L=100
- Red are for L=200.

What is the big difference here? The trend in the red period does not exist in the blue period. Unfortunately MinT down-weighs the top level forecasts and in this case these have a strong trend in them. In contrast OLS puts more weight to the top. I have also seen this with the prison data as exactly the same increase/trend in the number of prisoners happens during the 5-6 years towards the end of the sample and MinT is outperformed by OLS. I have compared the reconciliation weights between OLS and WLS before and OLS gives more weight to the top. I suspect this is the case with MinT as well.

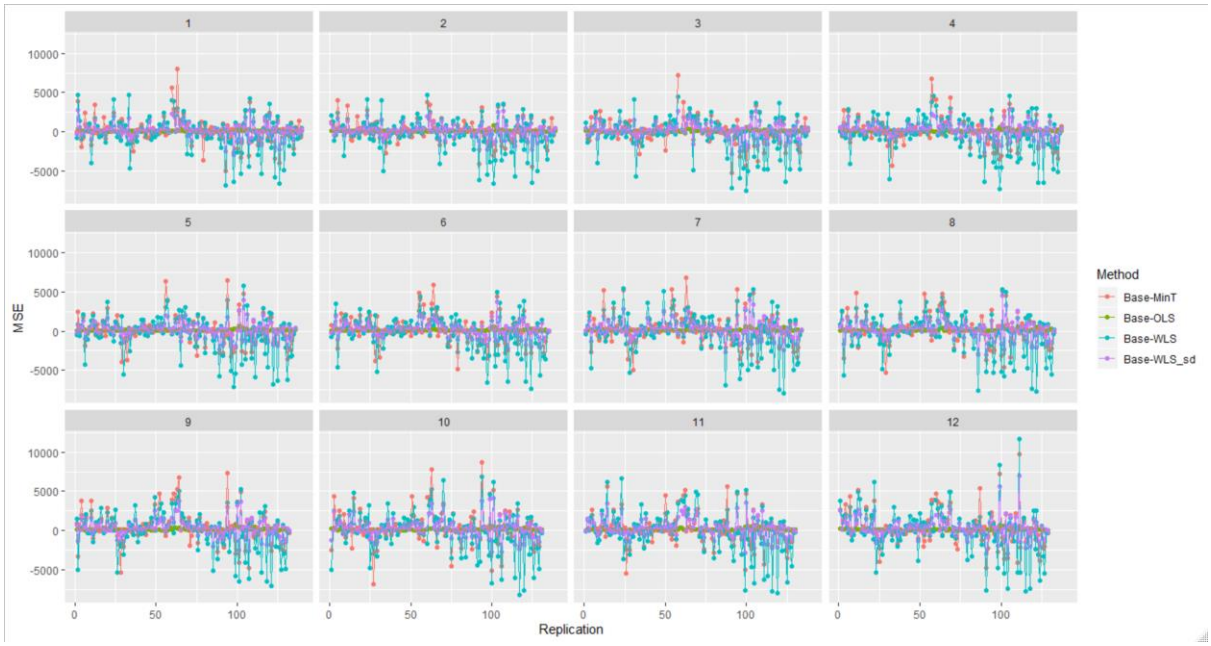



I suspect if we do the evaluation and include up to about the green period (or maybe a little before that) MinT will give us similar results as Figure 1 above.

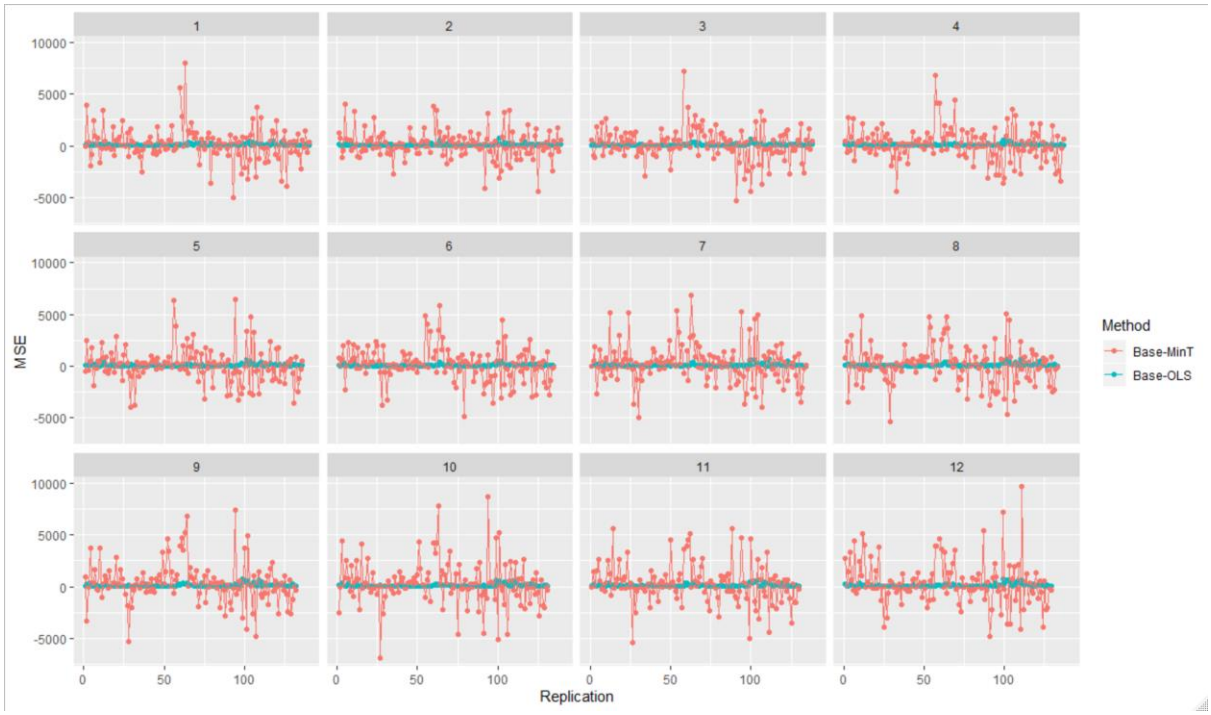
Ok so the question now is what is next? Of course we are waiting for Puwasala's results which I suspect will be something like an average between the two above.

George 31/08

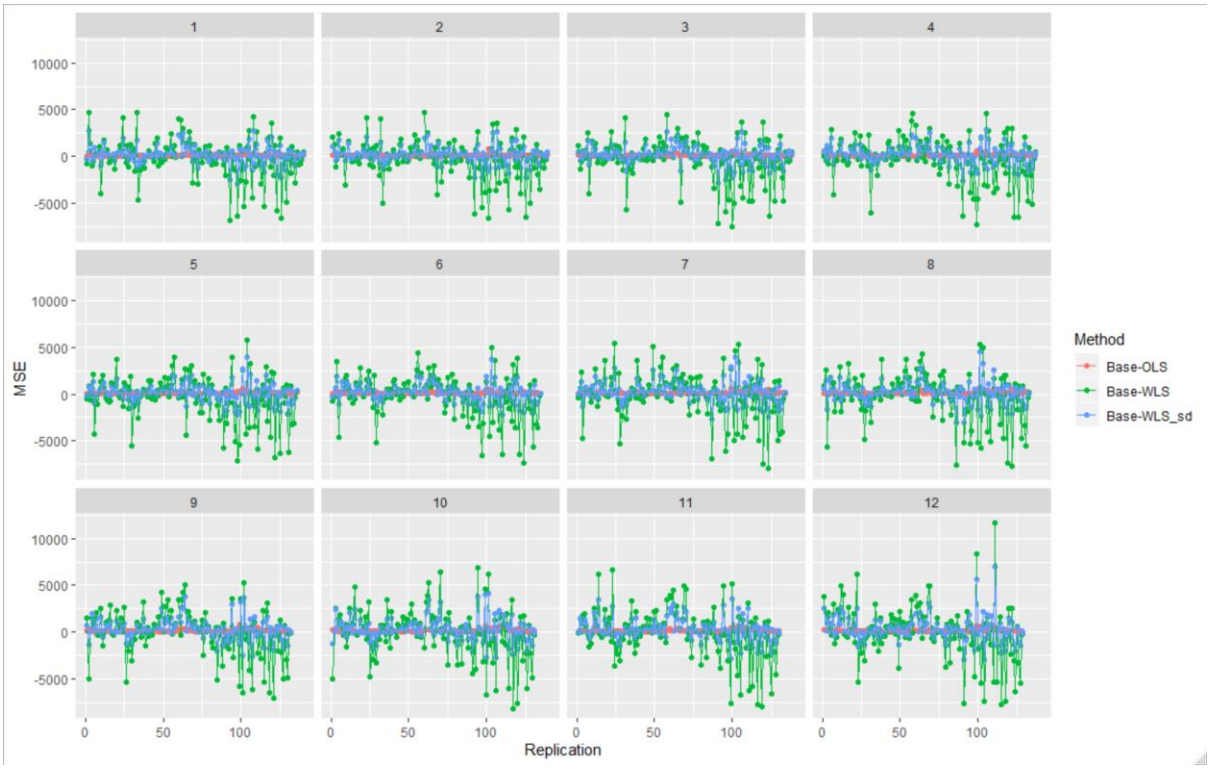
To my eyes it is the trend in the second part of the top series that is causing MinT and WLS trouble. Notice lots more variability and dots below the axis towards the end of the sample. (Replication 59 gave us trouble returned NaN not sure why).



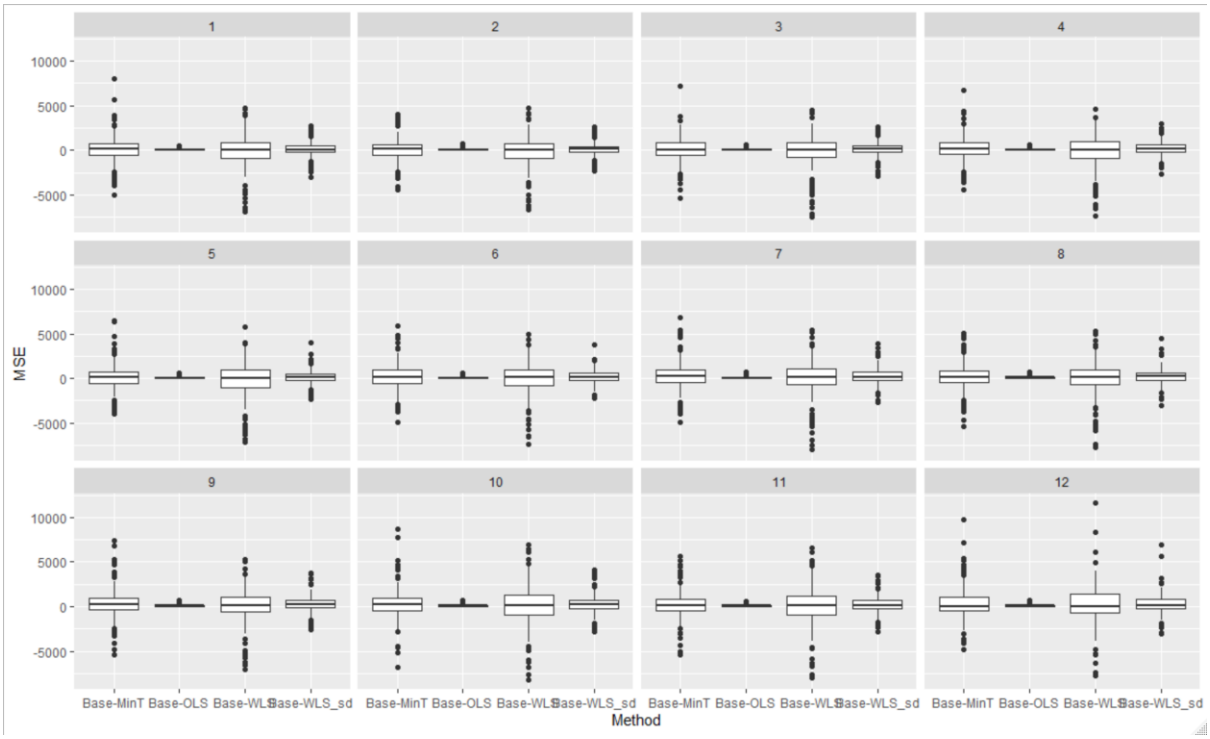

Only MinT



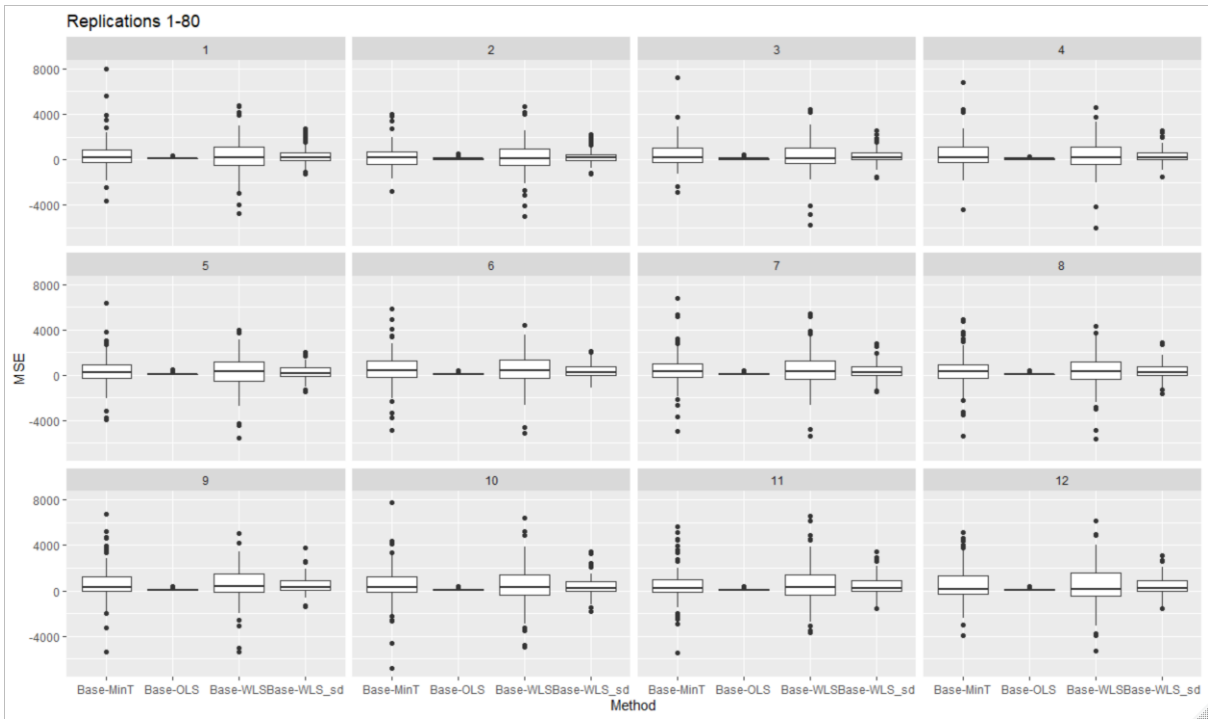

WLS_sd does a better job than WLS.

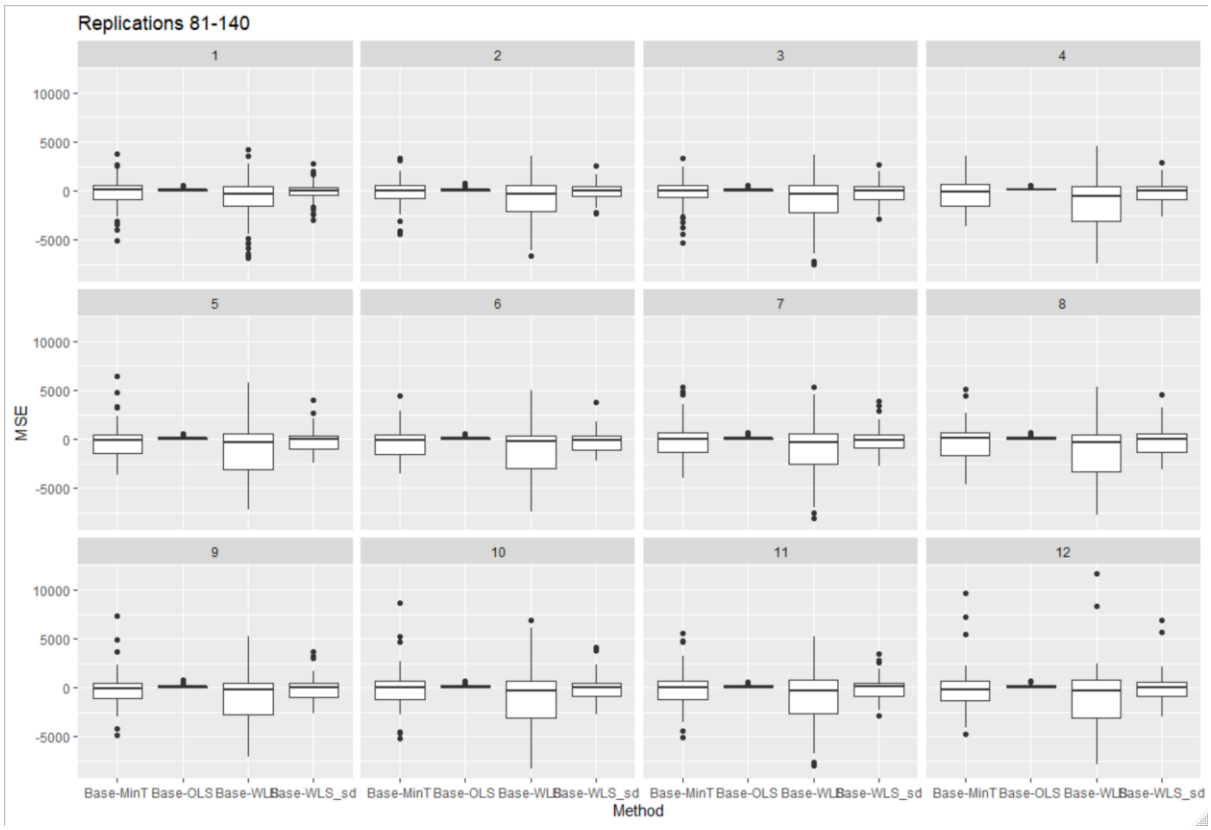



I think these look fine and they are what they are. I think for the paper we can comment on the variability of MinT



How about if I split the first and the second part of the sample. I think it shows again where the trouble is coming from. Again notice the good job sd is doing compared to variance scaling.



Averages

```
+   spread(key = `R-method`, value = MSE)
# A tibble: 1 x 6
   Base `Bottom-up` `MinT(Shrink)`   OLS   WLS WLS_sd
  <dbl>       <dbl>          <dbl> <dbl> <dbl>  <dbl>
1 5036.       6073.          4820. 4920. 5197.  4846.
```

The 2 shows there was a model with trend picked d+D+intercept+drift =2 (a combination of these).