George 30/8/2019

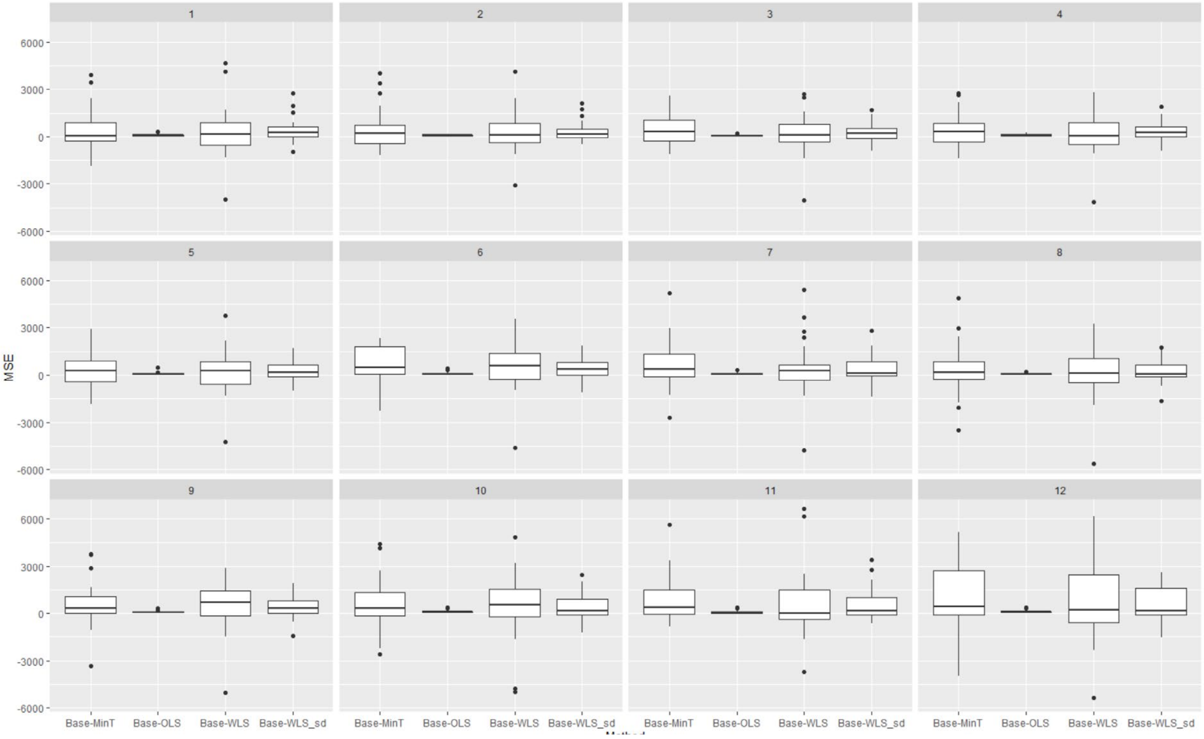
The boxplots are giving us not so good results for both MinT and WLS so we start to investigate where the problem is by starting with some debugging. Why did these methods forecast the tourism data better than OLS in the past?

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).

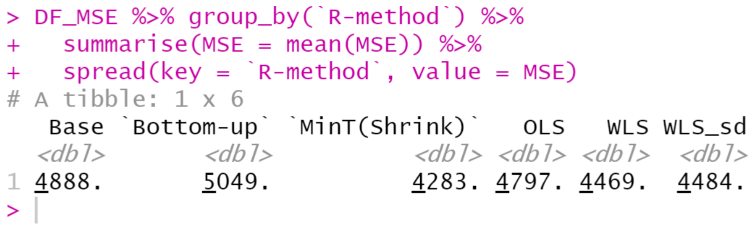
- Evaluation in JASA is done by level not across the whole hierarchy. Evaluating across the whole hierarchy using MSE the errors on the top level will dominate due to the scale. Note that in the JASA paper at the top-level OLS did better than MinT and MinT did worse than base as h increased.
- The training 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 and not the rolling window we do in this paper – 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. This was by accident as I didn’t switch off my auto sleep mode. In any case this has helped bring light to why the results are what they are.

Below are the results for L=100 (26 replications/rolling windows ran). These are (base-Reconciled) MSEs across the 26 windows across the whole hierarchy, i.e., a positive value is an improvement over base. 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 dampens this variation compared to normal WLS variance scaling (hence this was the reason this had creeped in the forecast package). (WLS_sd is the fourth column labelled and instead of using variance I am using standard deviation as W).

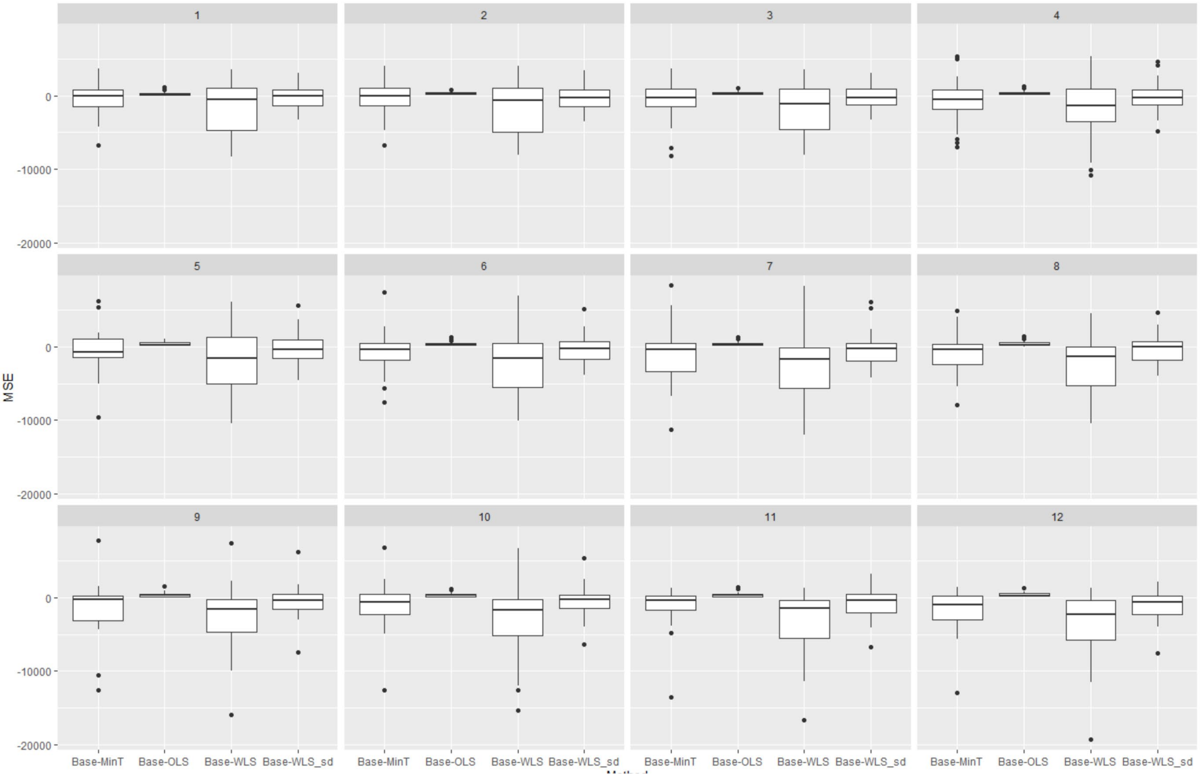


The averages below show that MinT on average outperforms all methods. So this is across all the repliations/windows.

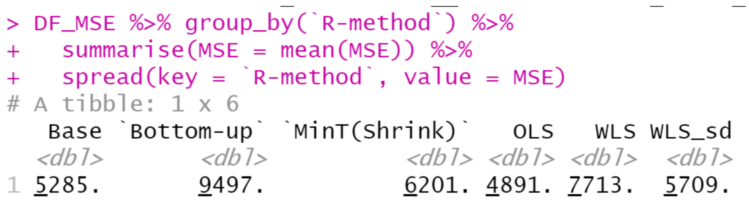


```
> DF_MSE %>% group_by(`R-method`) %>%
+   summarise(MSE = mean(MSE)) %>%
+   spread(key = `R-method`, value = MSE)
# A tibble: 1 x 6
   Base `Bottom-up` `MinT(Shrink)`   OLS   WLS WLS_sd
  <dbl>       <dbl>          <dbl> <dbl> <dbl>  <dbl>
1 4888.       5049.          4283. 4797. 4469.  4484.
>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

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



… the only average that beats Base now is OLS.



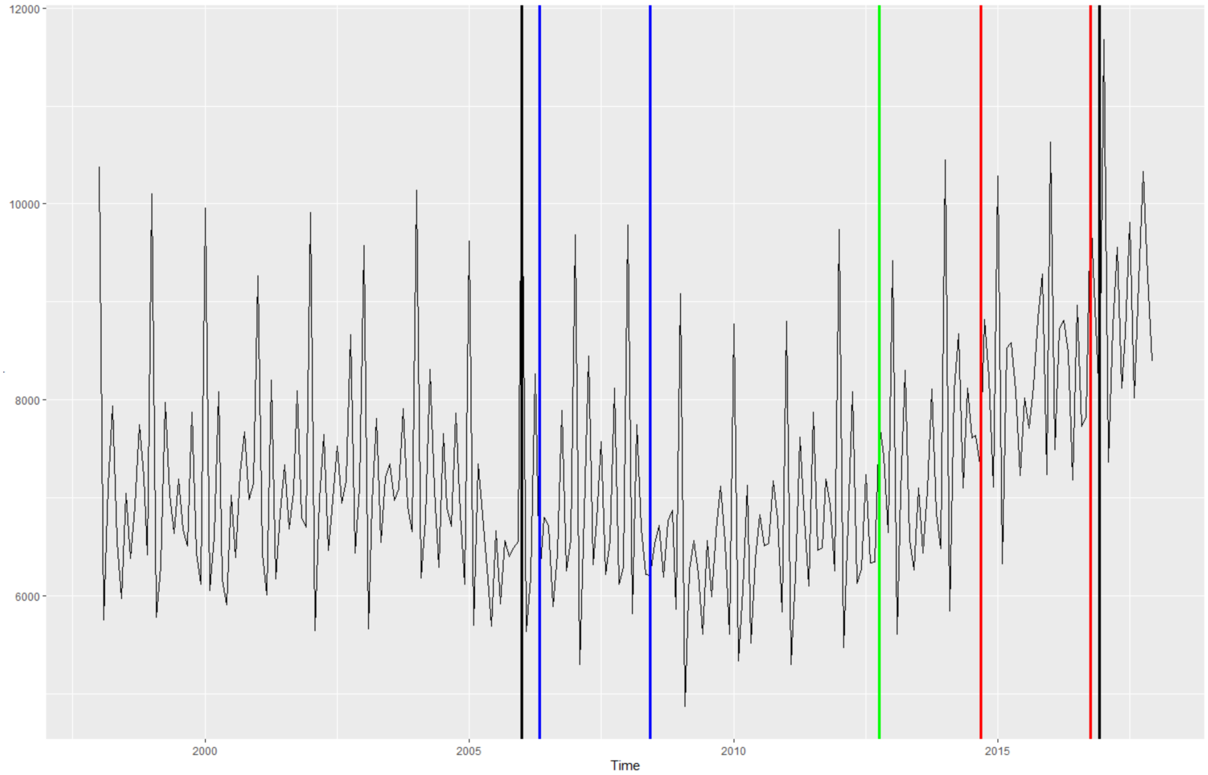
```
> DF_MSE %>% group_by(`R-method`) %>%
+   summarise(MSE = mean(MSE)) %>%
+   spread(key = `R-method`, value = MSE)
# A tibble: 1 x 6
   Base `Bottom-up` `MinT(Shrink)`   OLS   WLS WLS_sd
  <dbl>       <dbl>          <dbl> <dbl> <dbl>  <dbl>
1 5285.       9497.          6201. 4891. 7713.  5709.
```

Ok why is this happening?

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

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

What is the big difference here? The trend in the red period does not exist in the blue period. My gut tells me that, MinT does not give enough weight to the top level forecasts and in this case these have a strong trend in them. In contrast OLS puts more weight on the top-level. 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 previously compared the reconciliation weights between OLS and WLS (structural scaling) and if my memory serves me right OLS gives more weight to the top. I suspect this is the case compared to MinT and the WLS_var here as well.

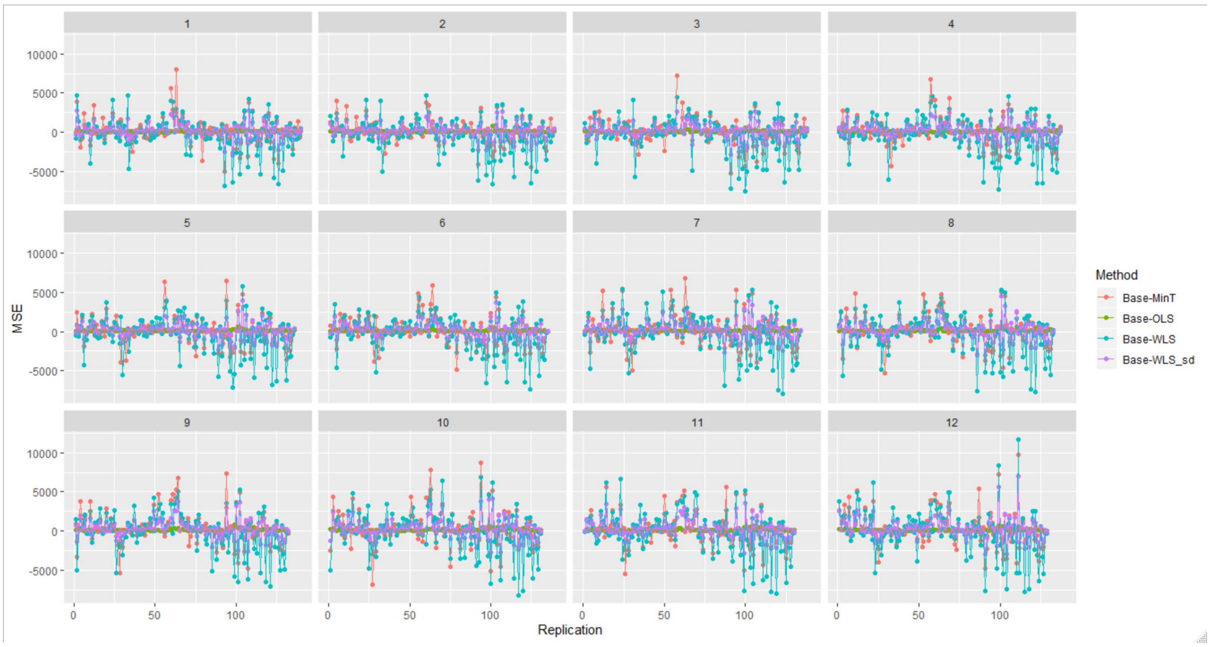


I suspect if we do the evaluation and include up to about the green period (or maybe a little before that) MinT will give us similar results as Figure 1 above.

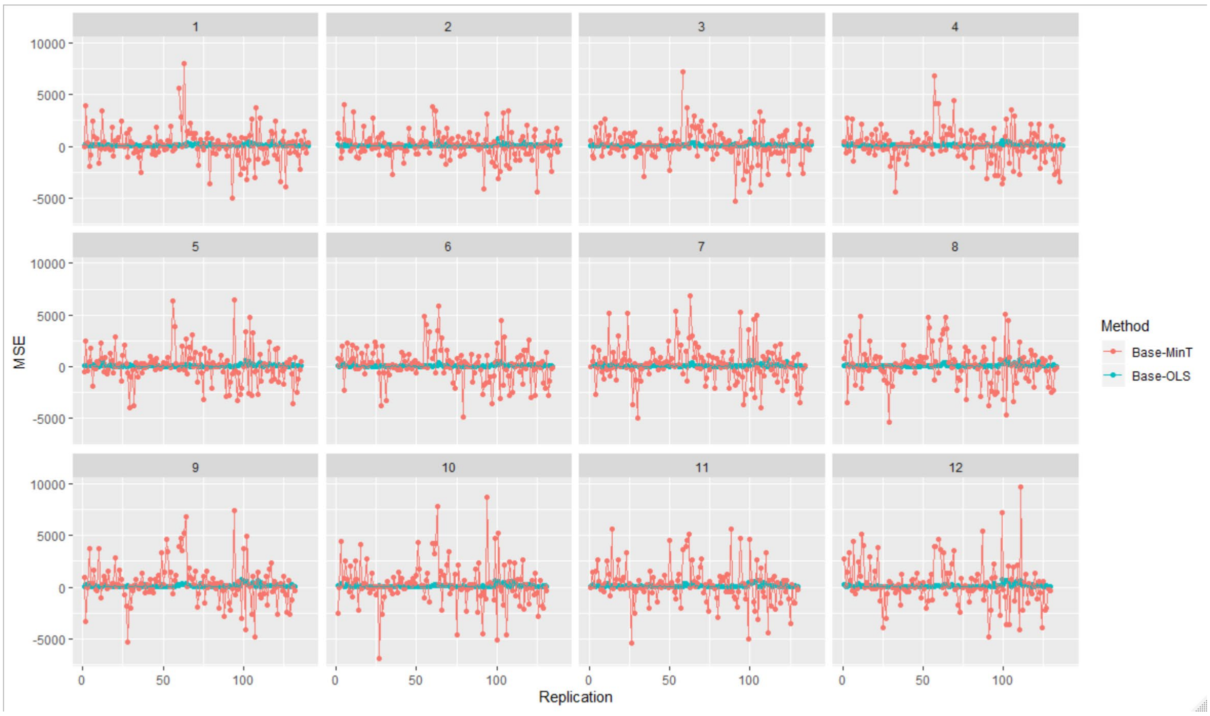
Ok so the question now is what is next? Of course we are waiting for Puwasala's results which I suspect will be something like an average between the two above.

George 31/08

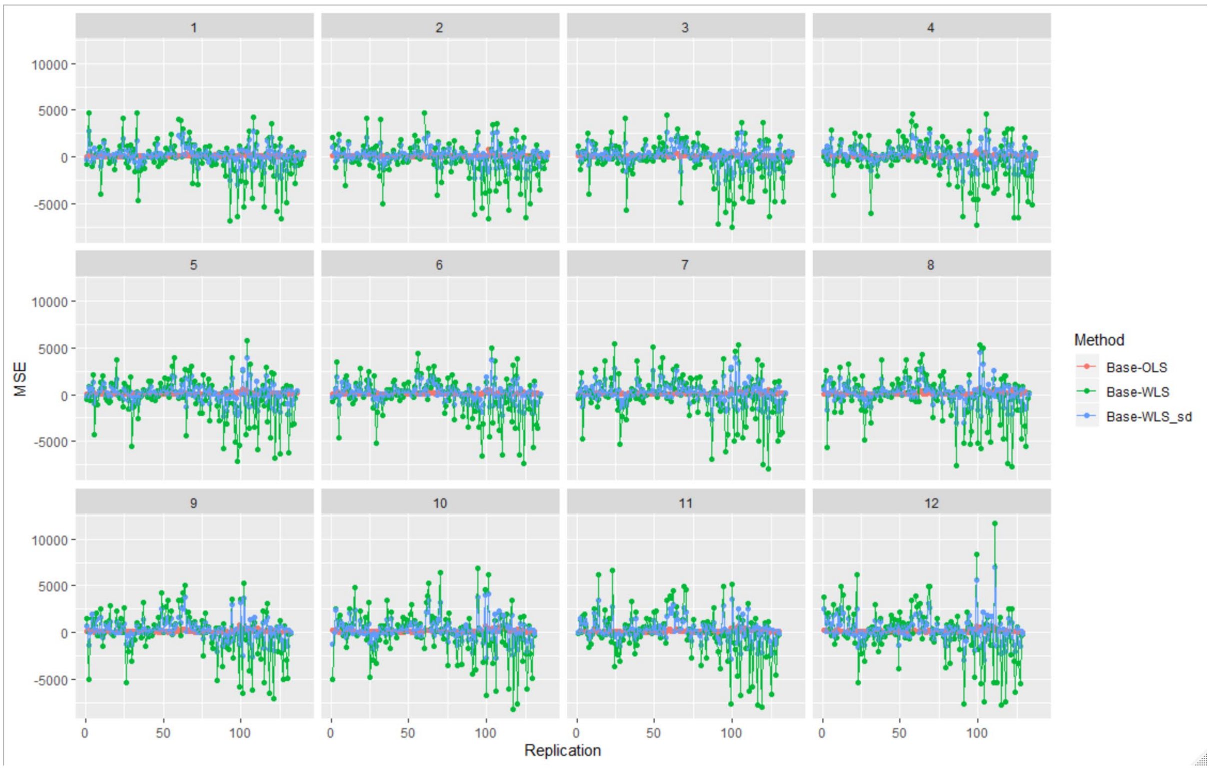
Some further analysis/investigation to see where the problem lies. The plots below show a times series plot of the MSE across the hierarchy over the rolling window. We would like to verify/identify when things start going wrong with MinT and WLS. To my eyes it is the trend in the second part of the top series that is causing MinT and WLS trouble. Notice lots more variability and many dots below the axis towards the second part of the sample.



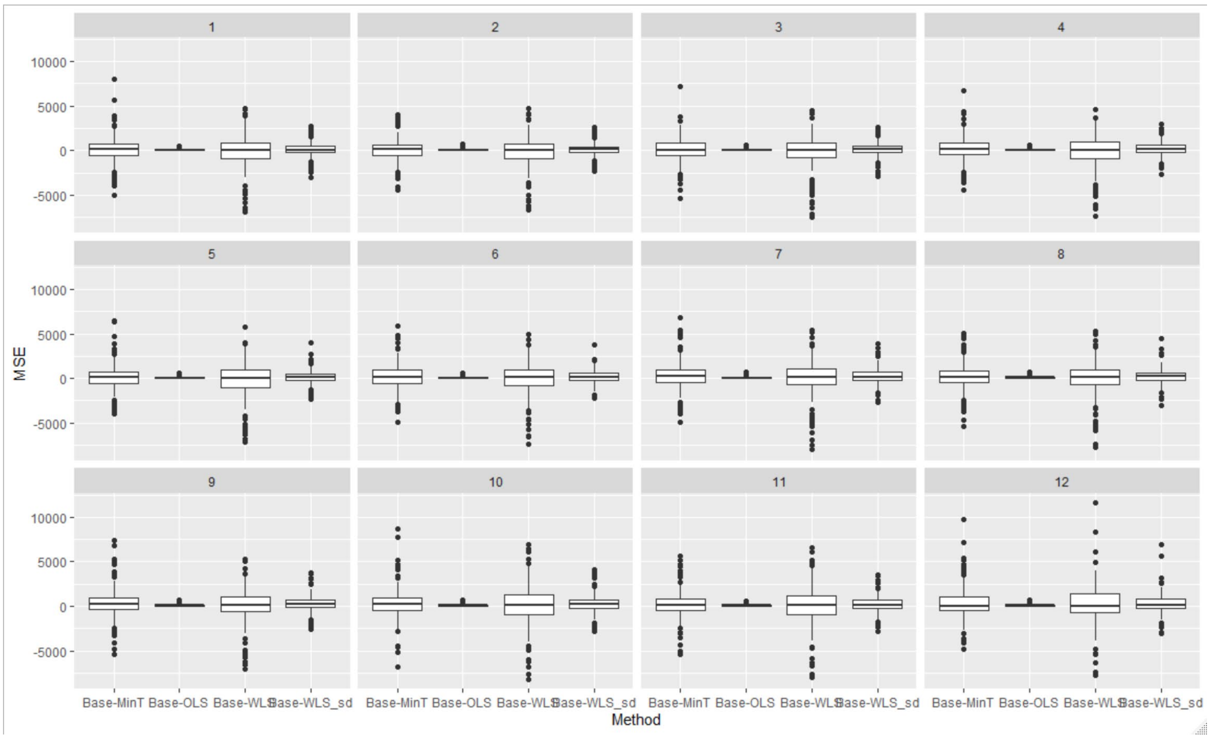
Only MinT



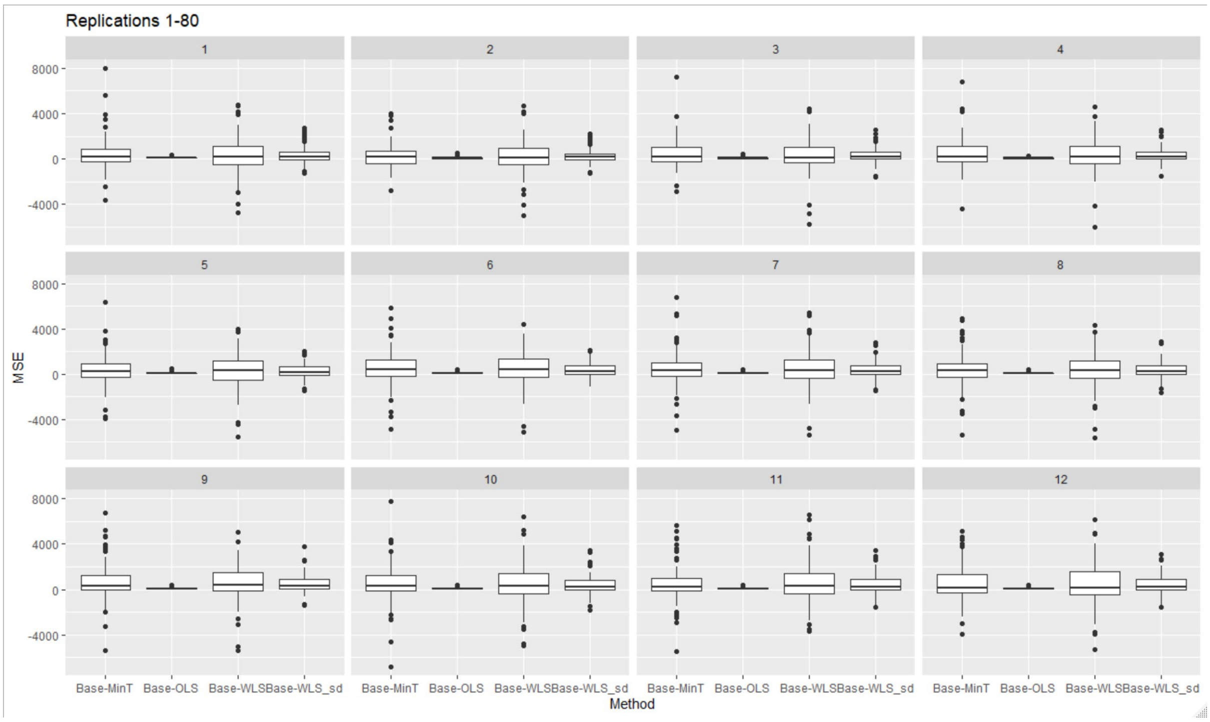
WLS_sd does a better job than WLS.

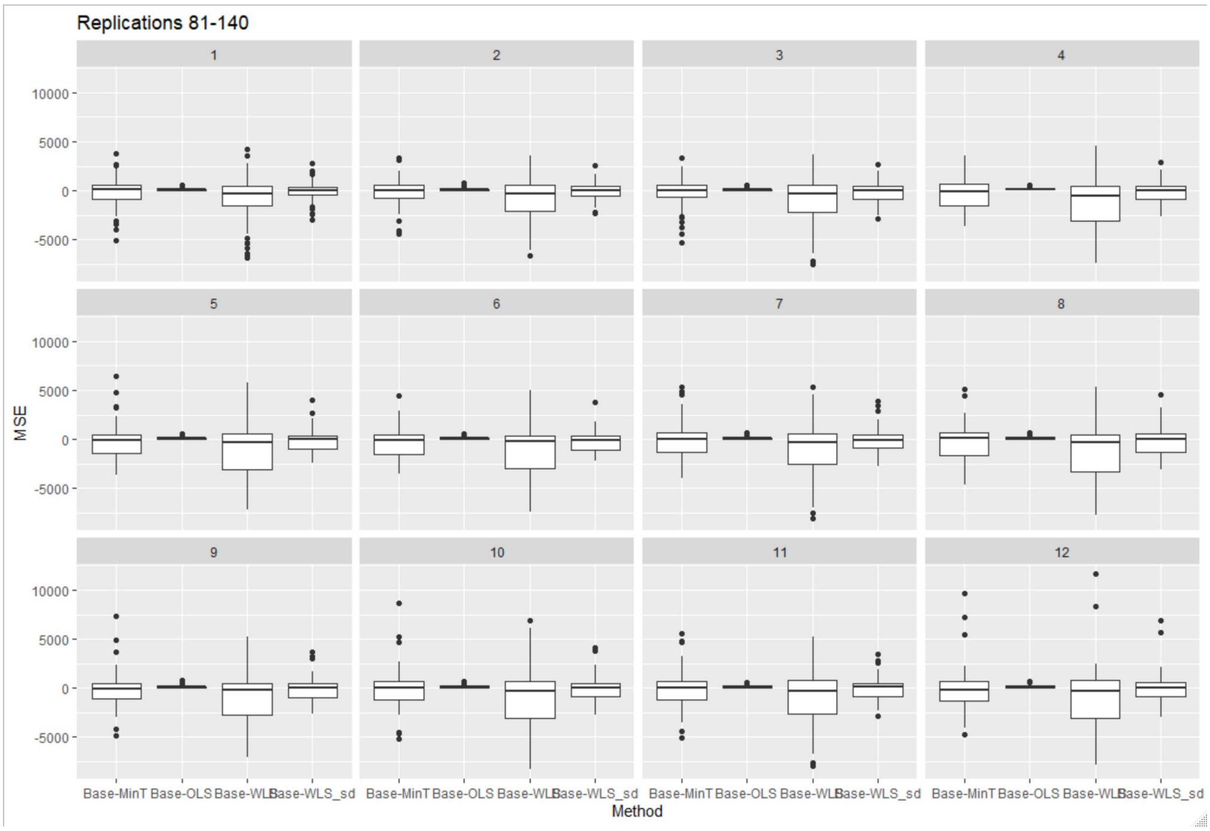


I think these now look fine for the current paper (they are over the entire 140 replications for h=1 of the rolling window) and they are what they are. I think for the paper we can comment on the variability of MinT and WLS and stop there.

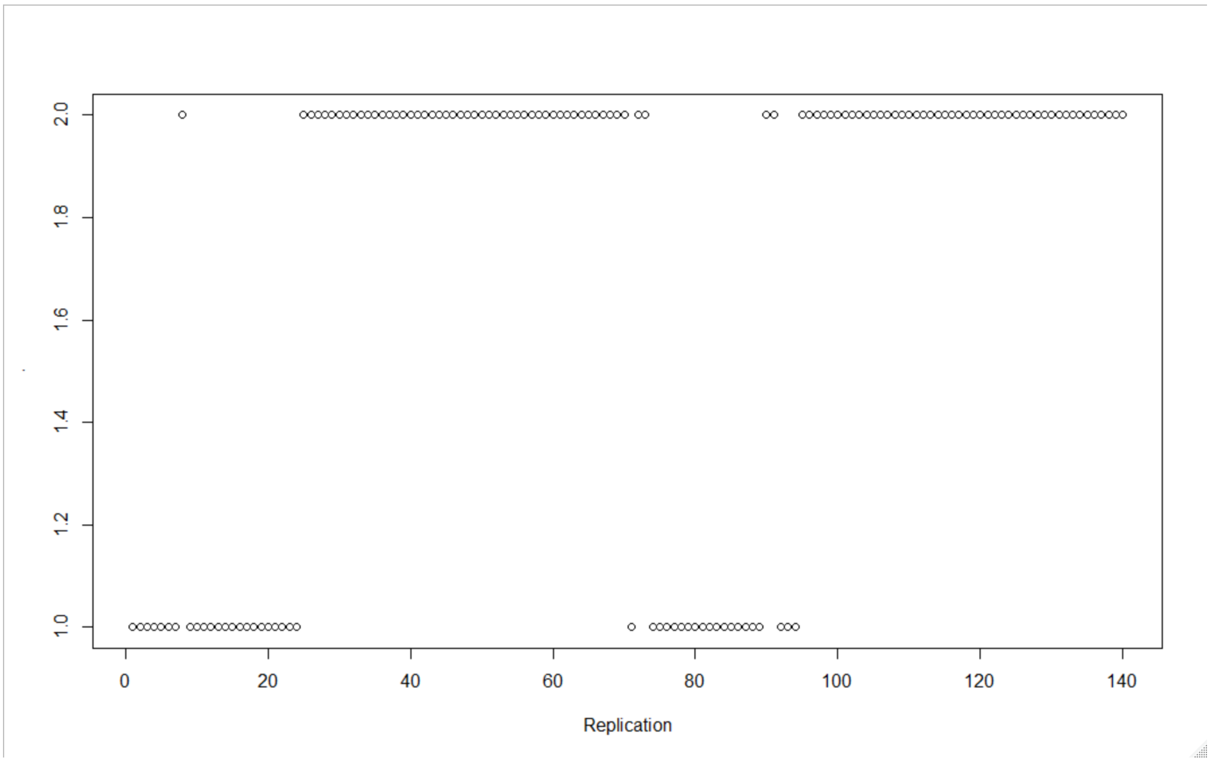


How about if I split the first and the second part of the sample. I think it shows again where the trouble is coming from. Again notice the good job sd is doing compared to variance scaling.



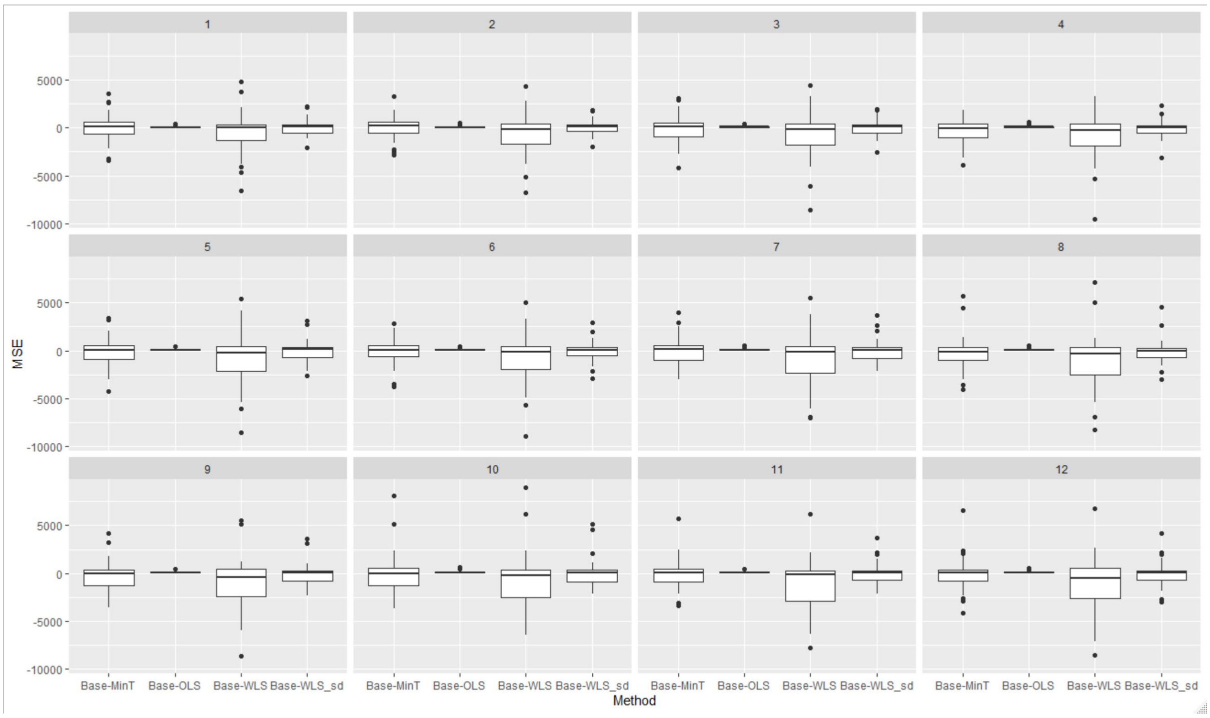


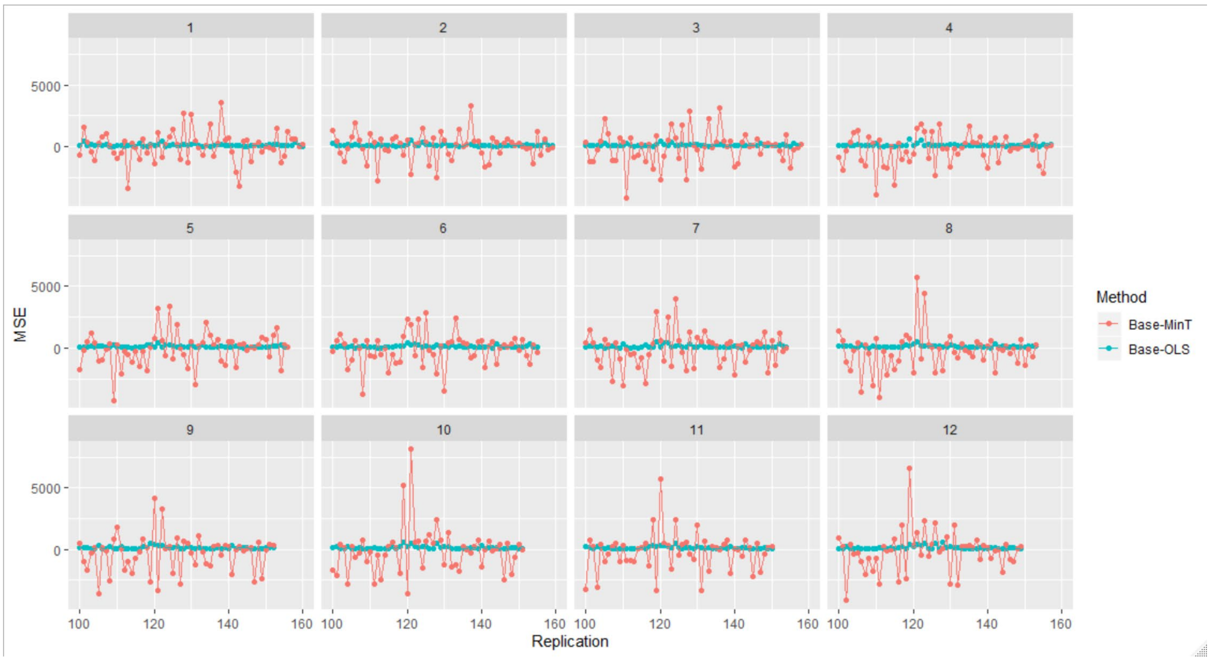
What model are we picking at the top level? Is this model misspecified and possibly we have non-stationary errors causing trouble to the covariance estimation? A 2 below indicates that there was a model with trend picked, i.e., d+D+intercept+drift =2 (a combination of these). Hence, for the latter part of the sample clearly a model with trend is picked. There were other instances that a model with trend was picked from about replication 25 to 60. I suspect this is the downwards trend in the middle of the sample. More on this below.



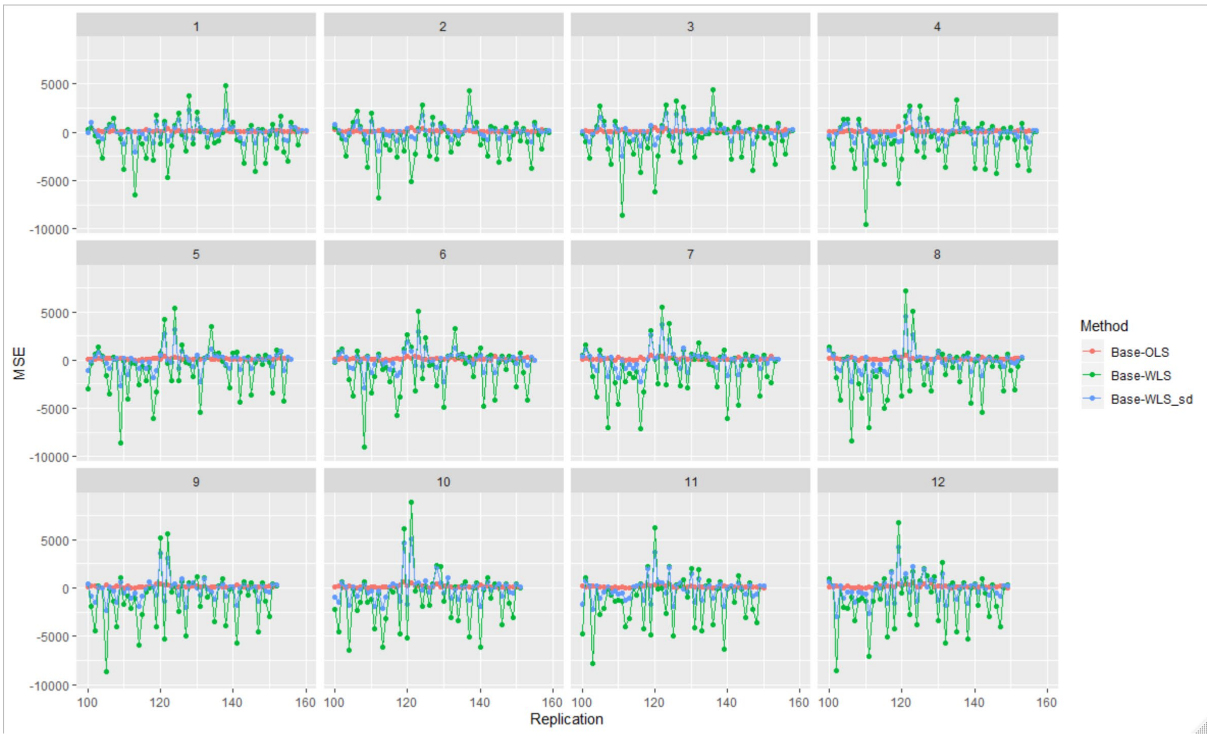
George 1/9

Is it the structural break in the trend that is causing us trouble? Well let's start the training set at replication 100 and beyond so that only clearly trended series at the top are considered in all windows and we do not have a structural break causing us trouble (could have started a little later but still ok). I am using a training set of 80 obs so all training data is strongly trended. We still have lots of problems with MinT and WLS so it is not just the structural break in the trend causing us trouble.

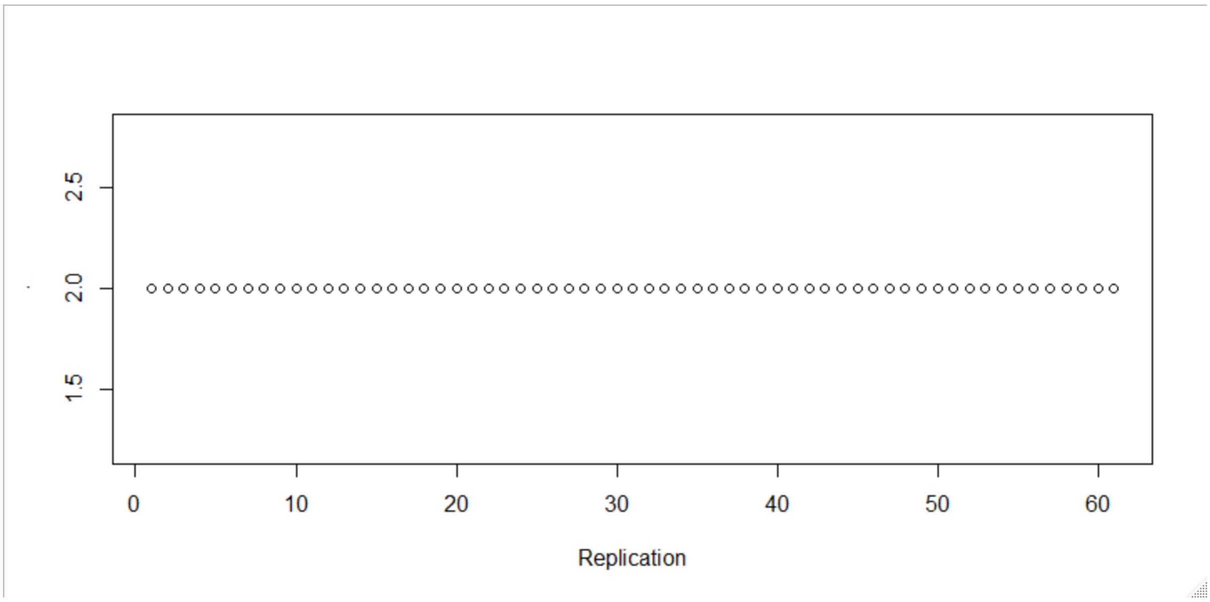




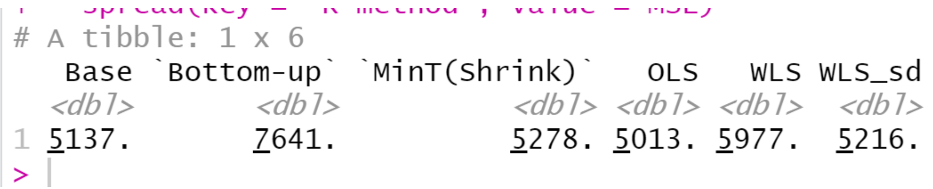
WLS has even more trouble



Yes we are always picking a trend at the top



And in this case only OLS beats Base



```
+   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 5137.       7641.          5278. 5013. 5977.  5216.
> 
```

Issues to investigate / random thoughts:

1. Where is our theory of MinT breaking down? The model at the top does not seem to be misspecified. Are the models in the levels below misspecified? However, if they are misspecified these should take even lower weights and therefore more weight should go to the top correctly specified model.
2. The investigation shows that we pick a model with trend at the top, so I suspect we have well behaved residuals there. In any case could it be non-stationarity in the residuals somewhere along the line that is causing a bad estimate of W? Is it the residuals for the below that are problematic?
3. If everything is fine is it purely the OLS reconciliation weights that are beating theory? Is it the MSE that may be exaggerating the aggregate results. How about if we used relative MSE?
4. In any case what more do we need to do when the data is so strongly trended? Do we need to check for residual stationarity?
5. WLS_sd. We have seen it over and over that it does better than variance scaling. Can we build a proper case (theory, simulation, etc.) so that this is used? Can we/should we write a note about somewhere (IJF?)? Everyone is using these methods.