MNAR Analysis

Nicholas Link

11/20/2023

Overview

Here are results analyzing why the MNAR missing assumption yields the same results as MCAR in our simulations. The results show:

1) The WF fit on MNAR data and on the full data set.
2) The same plot but under quasipoisson data generation with $\theta = 9$.
3) The same plot but under quasipoisson data generation with $\theta = 100$.
4) Simulation metric results with different β parameters for data generation.
5) Discussion of other simulation results.

The long story short is this:

1) The quasipoisson data generation does create MNAR fits that we would expect, but these are only visibly noticeable at high over-dispersion values. And they are more pronounced when there is not a downward yearly trend.
2) The WF and CAR performance are unaffected by the choice of WF β parameters to create the data. However, freqGLM is very affected, with the performance decreasing as β_0 increases.

Proposed next steps:

1) Run full set simulations with WF DGP under quasipoisson with $\theta = 100$ and with MNAR missingness.

(1) The WF fit on MNAR data and on the full data set.

In the plot below, there is virtually no difference between the fits of WF on the full dataset and on the incomplete dataset.

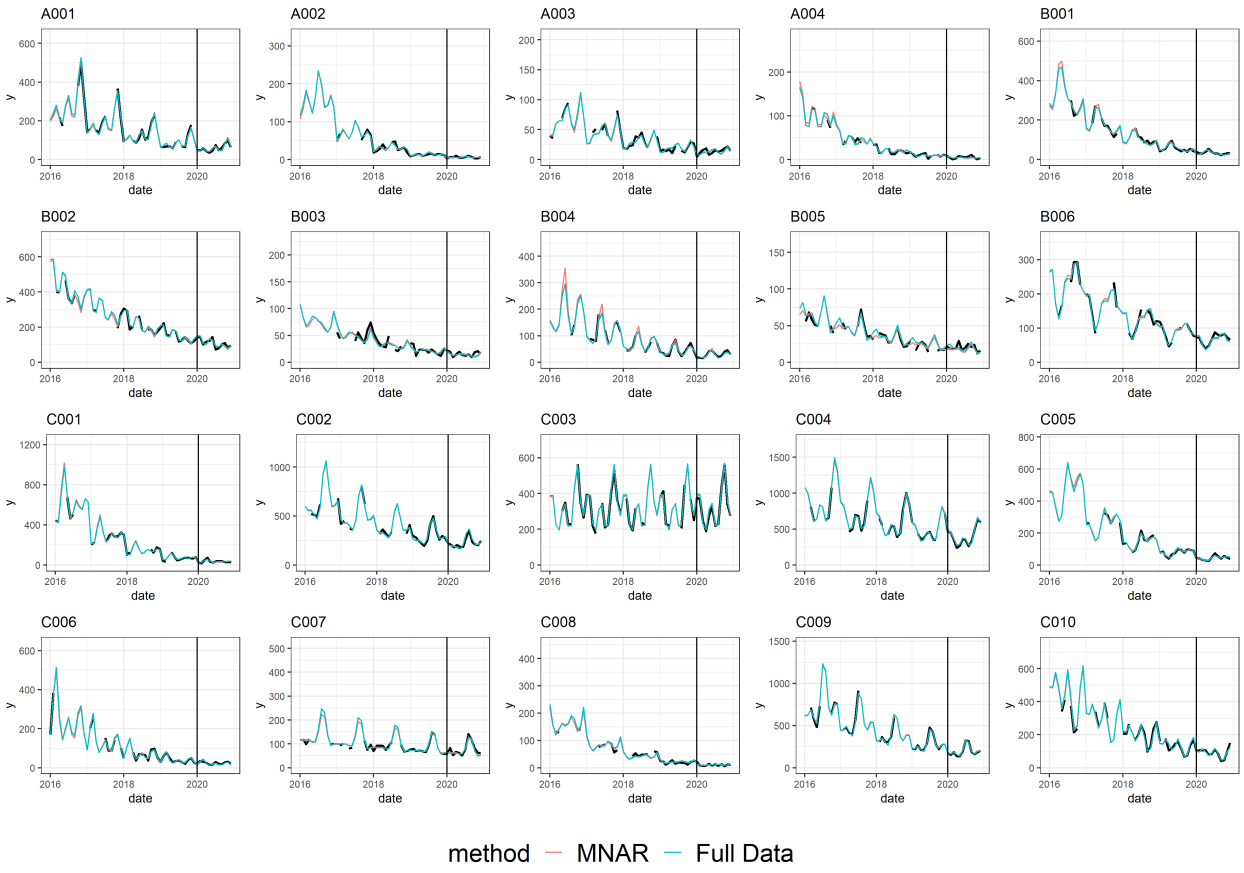


Figure 1: One simulation run with data generated by WF with E[B0] = 6 and E[B1] = -0.25. There is 30% MNAR missingness. The results show model fits from WF on the incomplete data and on the full data.

(2) The WF fit with quasipoisson dispersion, $\theta = 9$

Here everything is the same as the previous section except that the data is generated with over-dispersion.

$$Var(Y) = 9E[Y]$$

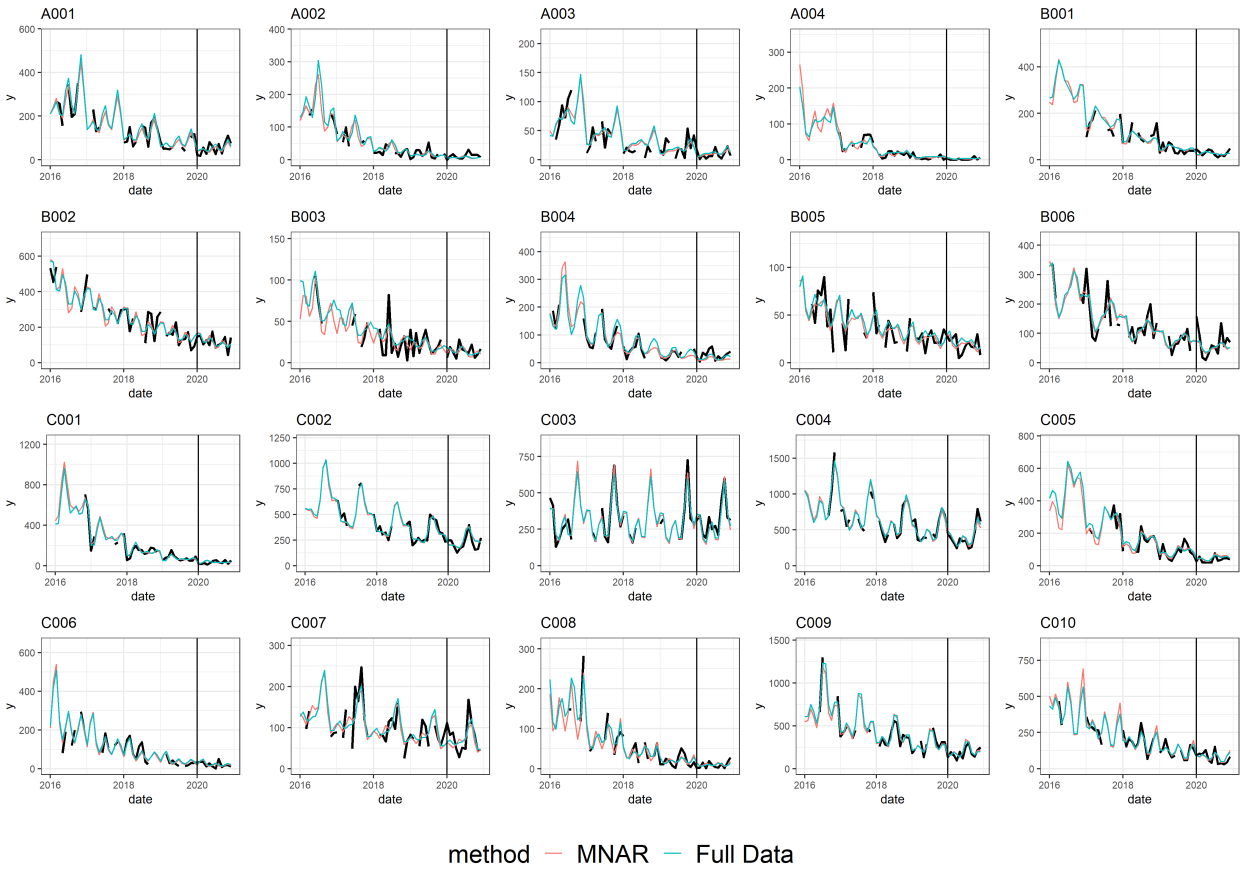


Figure 2: One simulation run with data generated by WF with E[B0] = 6 and E[B1] = -0.25 with quasipoisson overdispersion $\theta = 9$. There is 30% MNAR missingness. The results show model fits from WF on the incomplete data and on the full data.

(3) The WF fit with quasipoisson dispersion, $\theta = 100$

Now there is a lot of overdispersion. Specifically the standard deviation here is 10 times bigger than in the original Poisson data generation. The first plot shows these results with an average downward yearly trend and the second plot shows the results with an average neutral yearly trend.

From these results, the model fit on the incomplete data is generally lower than the model fit on the full data set, which is what we would expect. However, this seems to only happen at very high levels of over-dispersion.

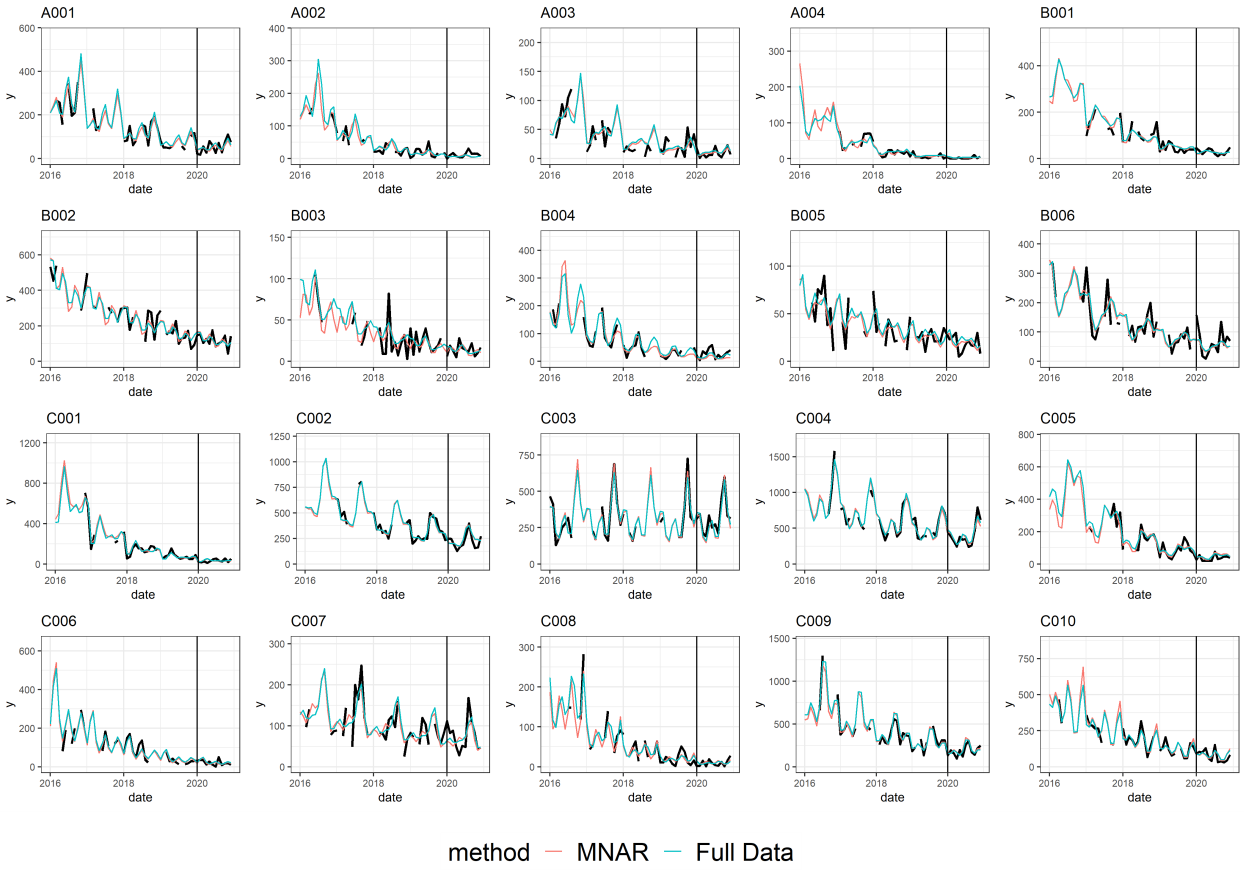


Figure 3: One simulation run with data generated by WF with E[B0] = 6 and E[B1] = -0.25 with quasipoisson overdispersion $\theta = 100$. There is 30% MNAR missingness. The results show model fits from WF on the incomplete data and on the full data.

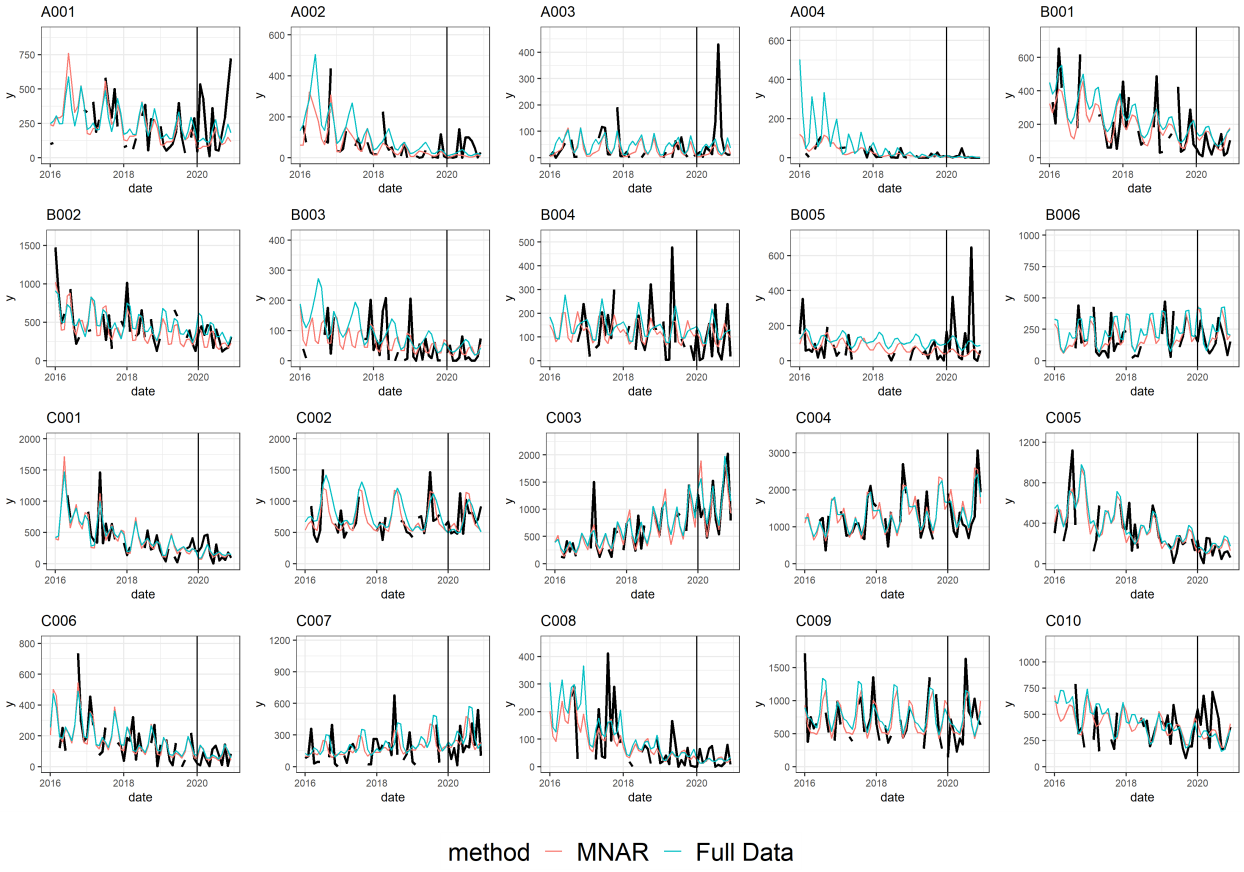
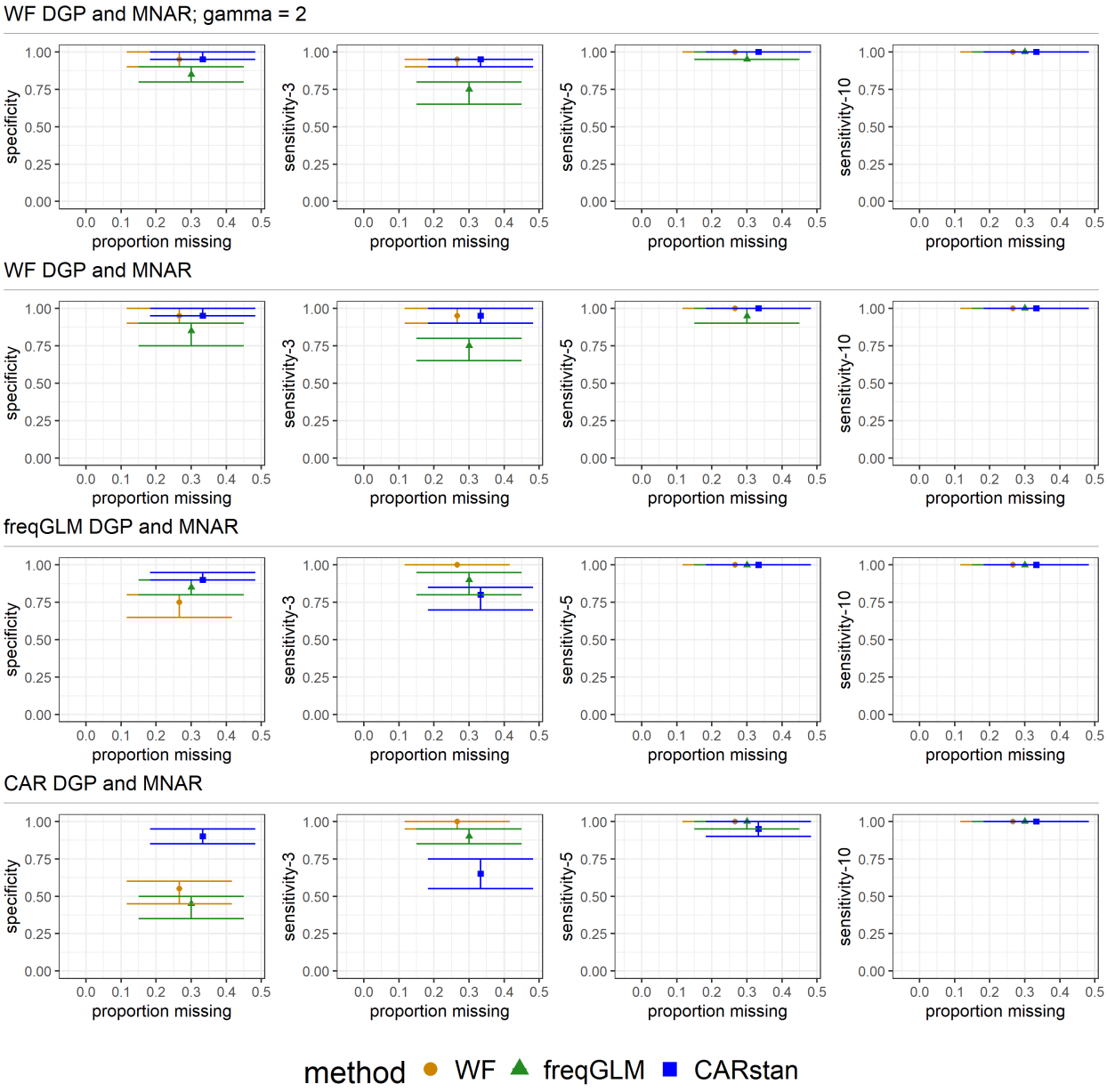


Figure 4: One simulation run with data generated by WF with E[B0] = 6 and E[B1] = 0 with quasipoisson overdispersion $\theta = 100$. There is 30% MNAR missingness. The results show model fits from WF on the incomplete data and on the full data.

(4) Simulation metric results with different β parameters for data generation.

In the results below, we can see that increasing the extremity of the MNAR assumption does not affect results. Interestingly, the β values don't have a noticeable effect on the WF and CAR results. However, the freqGLM model results are affected by the β values, specifically the specificity and sensitivity are lower for higher β_0 values. I am not sure what the cause of this is.



(5) Discussion.

The CAR DGP and freqGLM DGP results show no discernible difference between the MCAR missing and MNAR missing assumptions (I am not showing the results here).

These results show that the dispersion does have an effect on the model fits under MNAR. I am curious how the sensitivity and specificity metrics are with this level of over-dispersion. Does it make sense to compute the results with this over-dispersion to see how each of these models perform when they are incorrectly specified?