#SASGF

VIRTUAL

SAS® GLOBAL FORUM 2021

AMERICAS | MAY 18 – 20   ASIA PACIFIC | MAY 19 – 20   EMEA | MAY 25 – 26

SAS® Time Series Analysis & Forecasting (TSAF) at the Canada Revenue Agency (CRA), with COVID impacts

Jason A. Oliver, MBA, CAAP, with the Canada Revenue Agency (CRA)

Jason Oliver is a Project Leader, Senior Compliance Analyst and Data Scientist with the Canada Revenue Agency, who manages a team of data scientists in the pursuit of predictive analytics for tax related data. He is SAS certified and has used SAS extensively, as well as R and Python.

SAS® Time Series Analysis & Forecasting (TSAF) at the Canada Revenue Agency (CRA), with COVID impacts

Jason A. Oliver, MBA, CAAP (with the Canada Revenue Agency – CRA)

Jason Oliver is a Project Leader, Senior Compliance Analyst and Data Scientist with the Canada Revenue Agency, who manages a team of data scientists in the pursuit of predictive analytics for tax related data. He is SAS certified and has used SAS extensively, as well as R and Python.

The Canada Revenue Agency (CRA)

Overview

- The Canada Revenue Agency (CRA) is Canada's federal tax administration.
- As with all tax jurisdictions, the CRA has been challenged to keep pace with COVID-19 shocks and manifestations, which began in March 2020 (the last month of our fiscal year).
- Fortunately, SAS® Enterprise Miner™ has been an invaluable aid in gauging these impacts.
- We will begin with a **Glossary of Terms** to explain some of the key concepts.

GLOSSARY

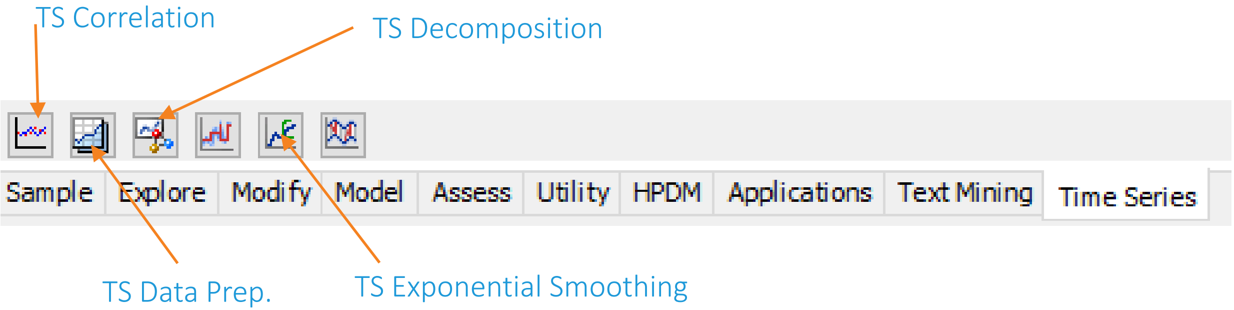
Of key terms at the CRA

- **TSAF:** Time Series Analysis & Forecasting.
- **TEBA:** *tax earned by audit,* which is the amount of tax collectible that is agreed upon in the course of a taxpayer audit.
- **TAR:** the *tax-at-risk*, which is the amount that CRA risk assessors arrive at as the precursor to auditing activity.
- **C/AR ratio:** the ratio of [audit] cases completed, to action requests [submitted] for assistance. It is a tentative measure of auditor productivity.
- **Integras:** the tool used by CRA auditors to process cases.

Time Series Functional Nodes

In SAS Enterprise Miner

- In SAS® Enterprise Miner™, you have six TSAF nodes in the "Time Series" bar; but we're just going to use four of them.
- To begin, we're going to use the TS Data Prep. & TS Decomp. nodes.

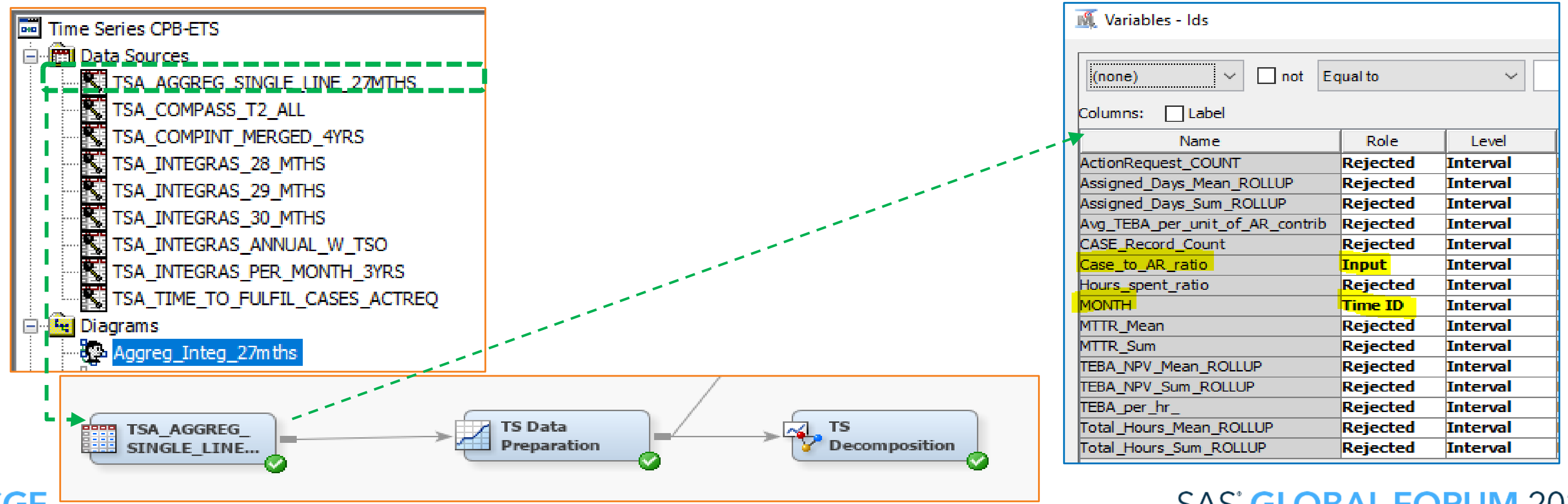


NOTE: the role of your data source must be "Transaction" for these nodes to work.

Train	
Output Type	View
Role	Transaction
Rerun	No
Summarize	No
Drop Map Variables	Yes

TSA Initial Setup

- We can first scrutinize on the **C/AR ratio** as a tentative measure of auditor performance.
- Our diagram is called "Aggreg_Integras_27mths", which runs from Jan. 2018 to March 2020.
- The dataset name is "TSA_AGGREG_SINGLE_LINE_27MTHS".
- So, on the initial node for Data Source, we only use the C/AR variable.

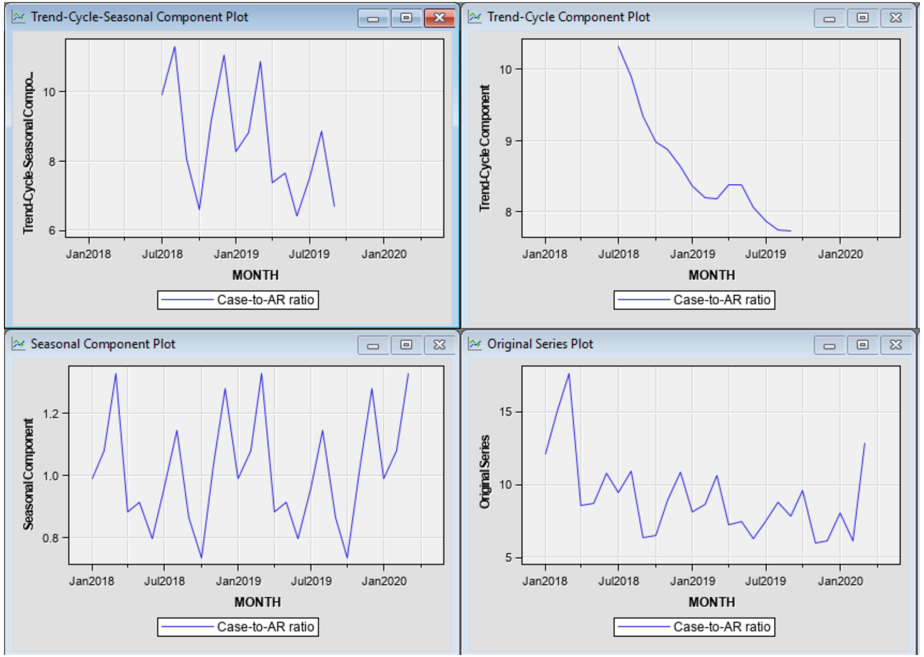


Name	Role	Level
ActionRequest_COUNT	Rejected	Interval
Assigned_Days_Mean_ROLLUP	Rejected	Interval
Assigned_Days_Sum_ROLLUP	Rejected	Interval
Avg_TEBA_per_unit_of_AR_contrib	Rejected	Interval
CASE_Record_Count	Rejected	Interval
Case_to_AR_ratio	Input	Interval
Hours_spent_ratio	Rejected	Interval
MONTH	Time ID	Interval
MTTR_Mean	Rejected	Interval
MTTR_Sum	Rejected	Interval
TEBA_NPV_Mean_ROLLUP	Rejected	Interval
TEBA_NPV_Sum_ROLLUP	Rejected	Interval
TEBA_per_hr_	Rejected	Interval
Total_Hours_Mean_ROLLUP	Rejected	Interval
Total_Hours_Sum_ROLLUP	Rejected	Interval

TSA Components:

C/AR ratio

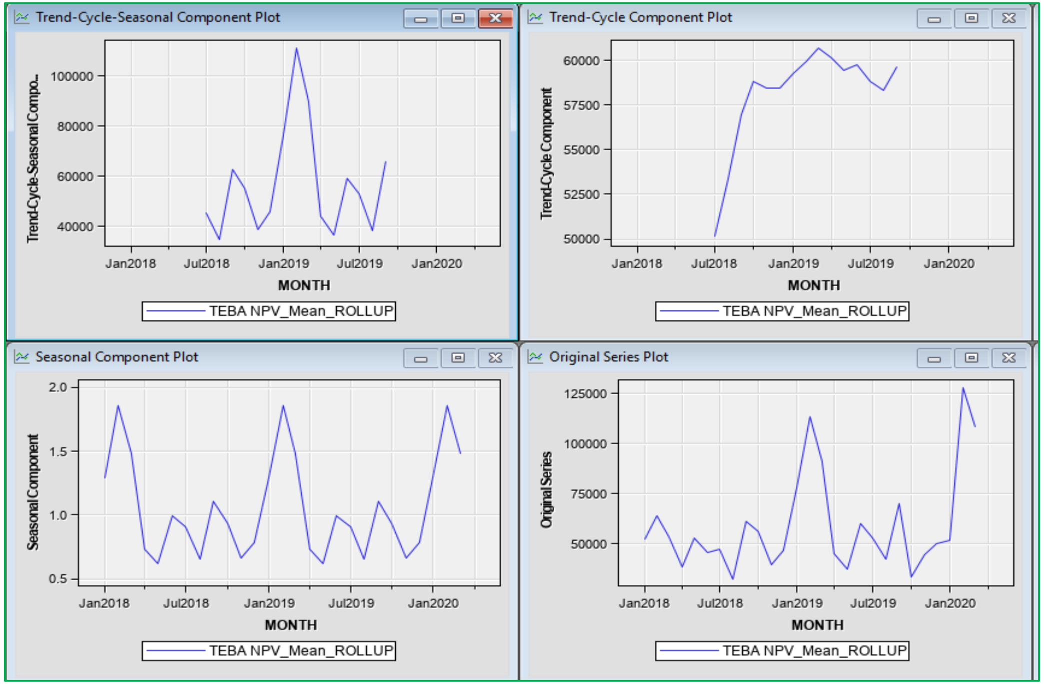
- If we run the TS Decomp. Node, then we can see the graphs for trend, seasonality, & cycle components, either in isolation or combined.



TSA Components:

Average TEBA

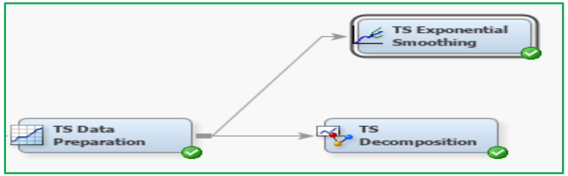
- Now, let’s substitute **Avg. TEBA** in place of C/AR ratio, to see how the components appear.



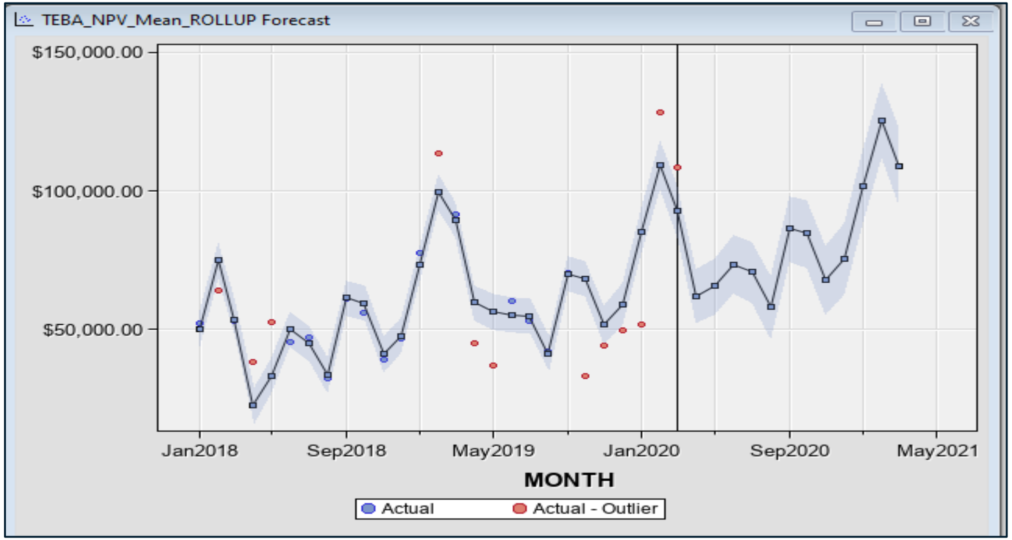
Forecasting Average TEBA

TS Exponential Smoothing node

- When we do forecasting, we use the TS Exponential Smoothing node. We let SAS® pick the best forecasting method, *and* selection criterion (forecast measure).
- Below, we see the forecast continues on a slight upward trajectory, despite the March disruption – because of *series momentum*.

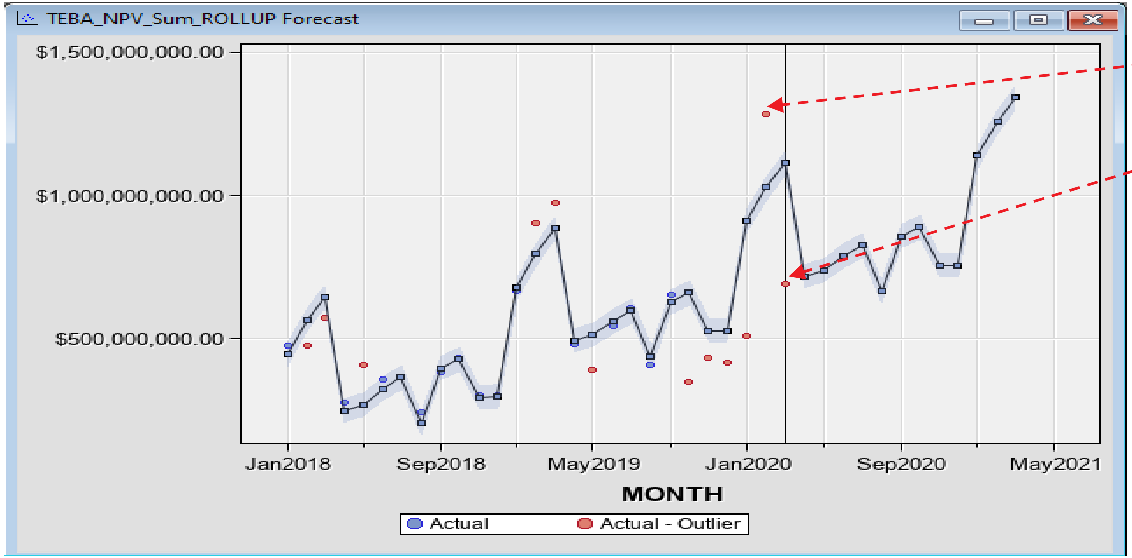


Train	
Variables	
Specify an Interval	Month
Accumulation	Total
Seasonality	Default
Forecasting Method	Best
Forecast Lead	18
Forecast Back	6
Forecast Sum Start	1
Significance Level	0.5
Input Time Series	
Forecast Input Time Series	Yes
Extended Value	Predicted Value
Best Model Selection	
Selection Criterion	Mean Square Error



Forecasting SUM of TEBA

- Now we can see a drastic difference in using the sum total of TEBA as an aggregate.
- Note that SAS®, in auto-selecting the best forecast method (Multiplicative Winters), has graphed a "line of best fit" (blue points) around *known data* (the red points)

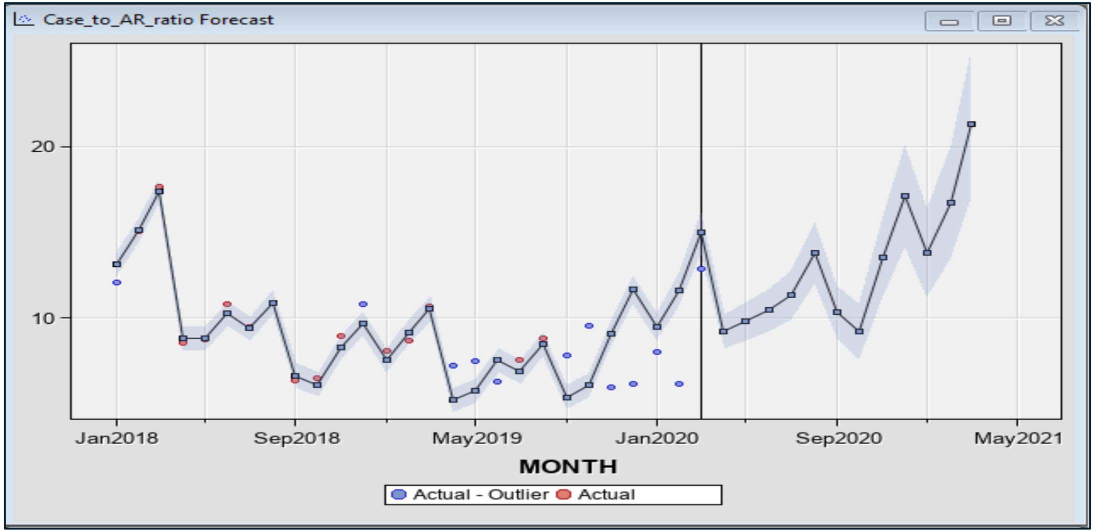


The SUM TEBA for Feb. 2020 is nearly double what it was for March 2020 (red dots).

Yet SAS® "thinks" that the trend will continue positively as it is "COVID-agnostic".

Forecasting C/AR ratio

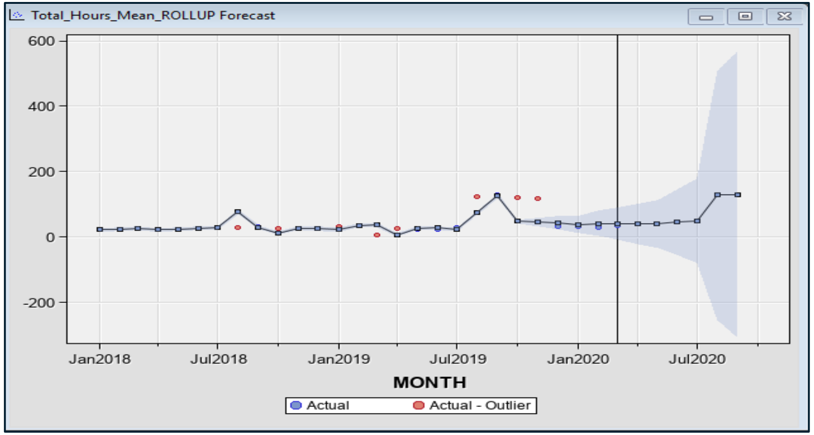
- In forecasting a fairly low continuous ratio variable such as **C/AR**, the prediction interval can be less reliable. We have to examine the midpoint distribution.
- While the midpoint post-March 2020 tends to be at or above the 10.0 line, this is rare for 2019 datapoints.



Forecasting Avg. Hrs. / case

- We also want to see how Avg. Hrs/case is forecasted.
- For this, I determined that the more ideal Selection Criterion is “Median Rel. Abs. Error”.
- **The midpoint** then goes very subtly upwards for the first few forecasted points, then sharply for summer.
- But with a lower scale, the prediction interval becomes spurious; you can’t have negative hours.

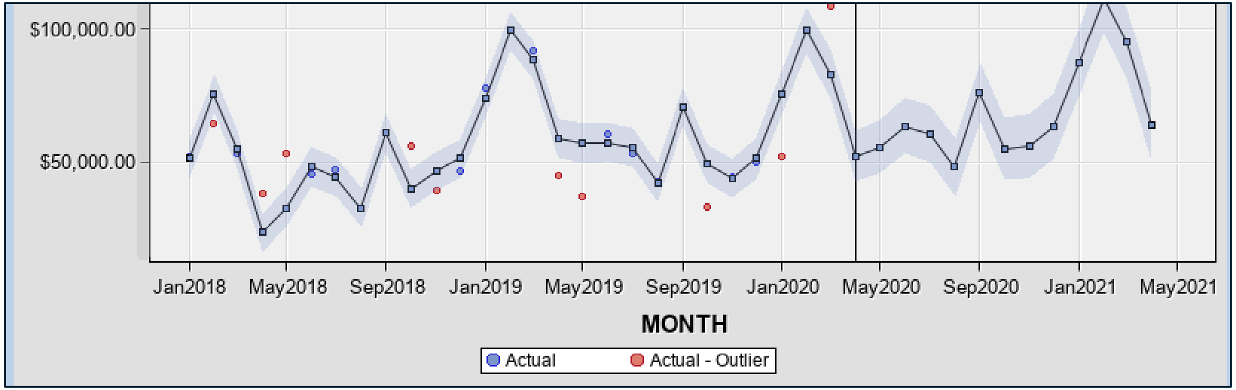
Train	
Variables	
Specify an Interval	Month
Accumulation	Total
Seasonality	Default
Forecasting Method	Best
Forecast Lead	18
Forecast Back	6
Forecast Sum Start	1
Significance Level	0.5
Input Time Series	
Forecast Input Time Series	Yes
Extended Value	Predicted Value
Best Model Selection	
Selection Criterion	Median Relative Abs. Error



Incremental alignment:

April 2020, known values

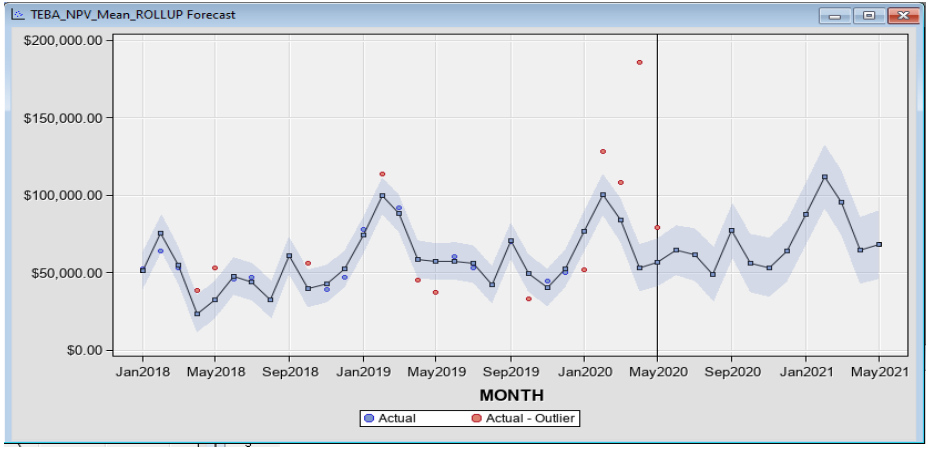
- Now when we add the month of April 2020 to our data (making it 28mths total), we would expect the **Avg. TEBA *actuals*** for subsequent months to become closer to / within forecast range.
- Example: the forecast for Sept., Oct., and Dec. becomes more within range of later-known actuals, once we add April 2020 data.
- However, the July 2020 actual ($122,000) is *still* above the forecast band for this incremental dataset's forecast.



Incremental alignment:

May 2020, known values (Avg. TEBA)

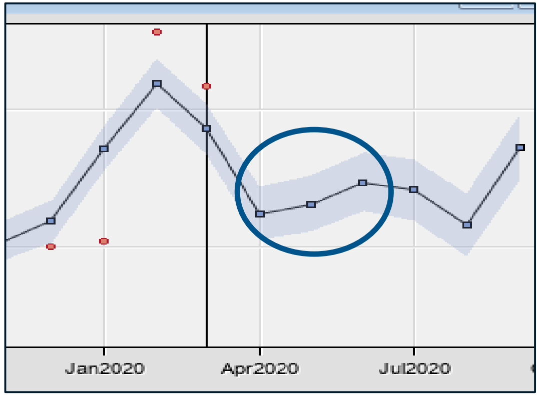
- Clearly, the addition of April wasn't enough to right the trajectory of the expanding "COVID window", so I added May 2020 AND I changed the forecast significance level from 0.5 to 0.25.
- But it makes no difference: July actual is *still* out of forecast range.
- We must simply accept that July 2020 is an irregular value (~$122K), since July 2018 had Avg. TEBA =~45K, and July 2019 Avg. TEBA = ~57K. *This is likely a COVID-adjustment spike.*

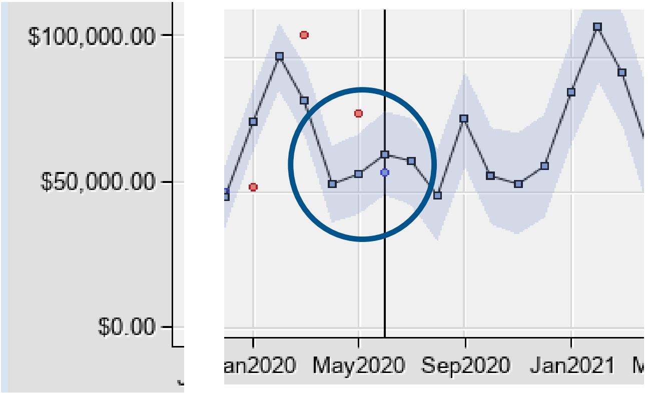


Incremental alignment:

June 2020, known values (Avg. TEBA)

- For the addition of June, it didn't improve the forecast band to include actual Avg. TEBA of July.
- So this strengthens the theory that July's value was a one-time event, or *pulse*, in the time series.
- It also strengthens the theory that Avg. TEBA was more resilient to initial COVID-19 transition measures.
- To wit: note that the April-May-June line for the original forecast (left) and actual (right) is just above the $50K line, and follows the same trajectory.





Fallacy: comparing SUM of TEBA shift to AVG. TEBA changes

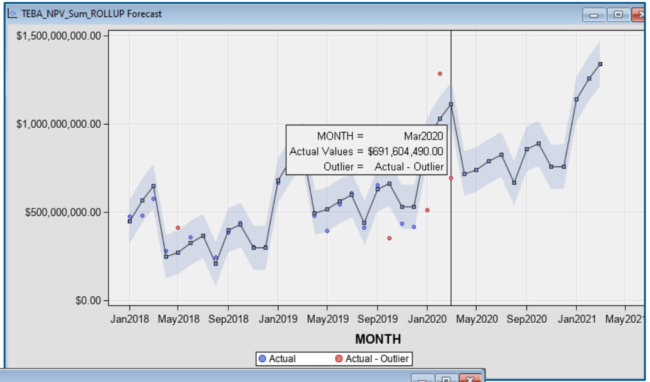
- TSA works best when you accumulate data records by ***average***, not by sum total.
- If we tried this exercise using SUM TEBA per month, it wouldn't work very well, since sum totals are immediately impacted by any severe transition, i.e. work re-arrangements in March 2020 due to COVID.
- Evaluating the March 2019-2020 comparison: the **TEBA_SUM** and **Case Count** have dropped significantly in March 2020, yet the **C/AR** ratio has gone up.
- However, as the staffing situation has attempted to stabilize in the intervening months (April-June 2020), the C/AR ratio has dropped dramatically. The same is true for the TEBA/AR pattern.

Mth / Var.	TEBA_SUM	TEBA_AVG	Case Count	C/AR	TEBA/AR	Avg. Case Hrs.
March 2019	$973,573,844	$91,561.54	10,633	10.65	$975,524.89	6.2526
March 2020	$691,604,490	$108,300.11	6,386	12.85	$1,391,558.33	35.44

SUM of TEBA: drastic change

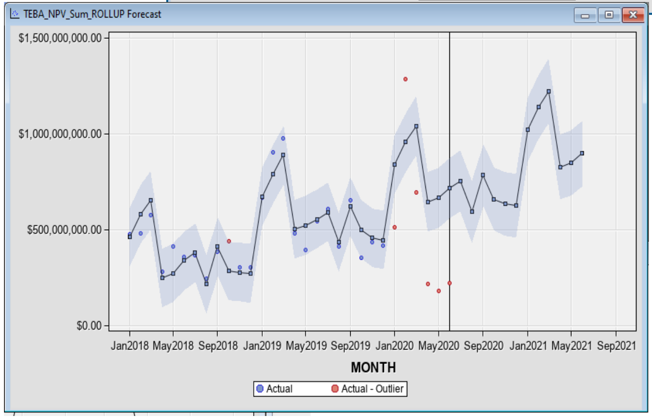
Last month of actuals: MARCH 2020

None of the actuals of the last six months of 2020 fall in the forecast band.



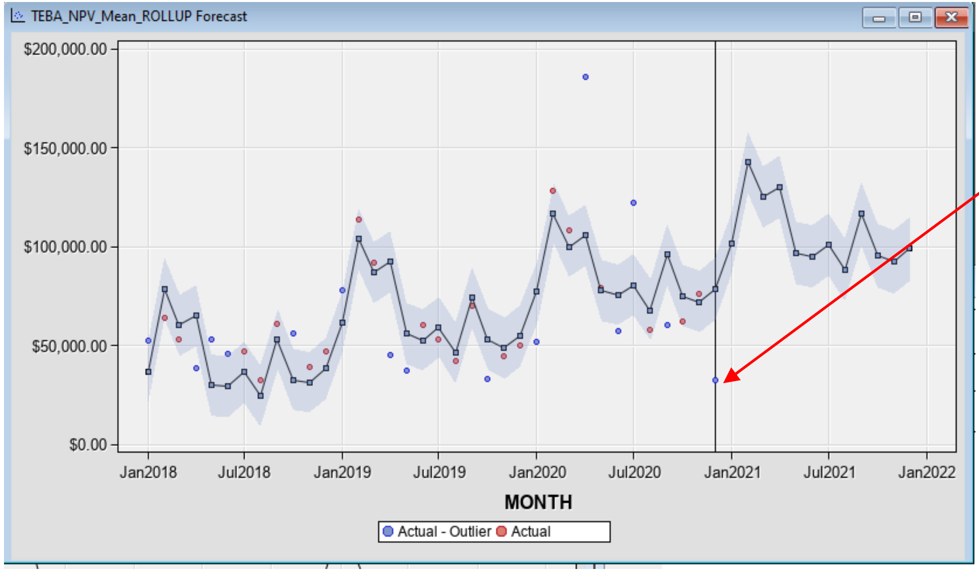
Last month of actuals: JUNE 2020

Two of the actuals of the last six months (Oct.,Nov.) of 2020 fall in the forecast band.



Latent Effects of Shocks

- We would also expect that lower Avg. TEBA wouldn't manifest until much later in the fiscal year 2020-21, due to most of 2020 consisting of *past year* audits.
- Given this, we would need to resort to the use of *interventions* in our time series.

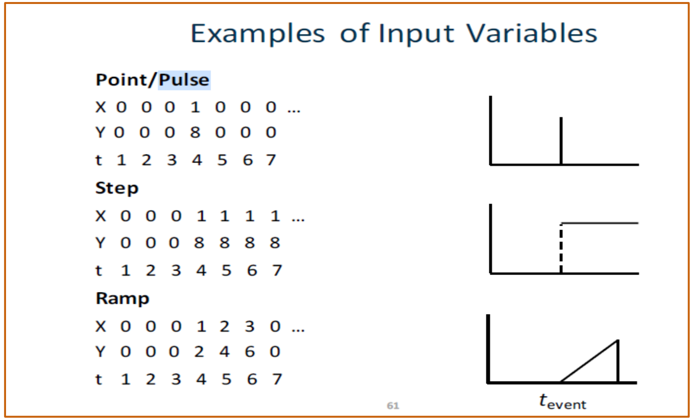


Lowest actual in 3 years;
Dec. 2020
Avg. TEBA of $32,404

Interventions

- A TSA may use ***interventions***, if the extreme or irregular event is known in advance.
- This is an adjustment to the time series, using a "dummy" variable for the period of observation.
- An intervention would be recommended for the SUM of TEBA as of March 2020, and for AVG TEBA as of Dec. 2020. Plus, a "pulse effect" for July 2020.
- Programming an intervention requires SAS® Studio™, which is out of scope for this presentation.

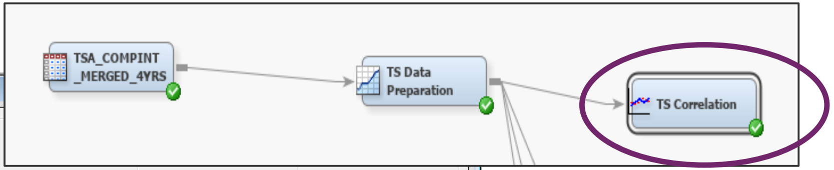
A **step** would work best as an intervention, since the trend line shift is sudden and sustained; it does not happen gradually then return to baseline.



Autocorrelation

(from: 2018-2019)

- When we deal with a significant seasonal and/or trend component, we usually find a greater degree of **autocorrelation** (abbrev. “ACF”).
- As the name suggests, this is the tendency of a variable to *self-influence*. It could also be regarded as momentum, or “muscle memory”.
- This uses the **TS Correlation** node.



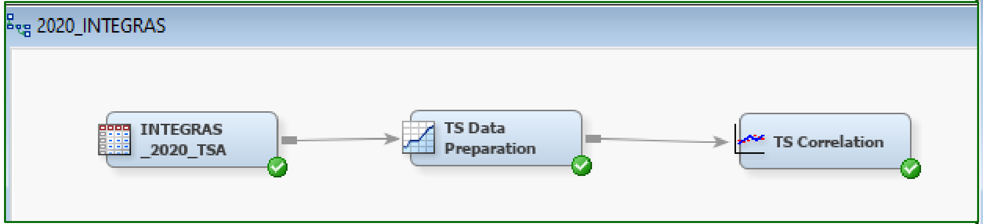
From these three variables, Est. TAR-AI has low ACF, TEBA has moderately high ACF, and Case Hours has very high ACF.

At lag t=5, TEBA reaches the zero line; but Total Hours is still at ACF=0.45.

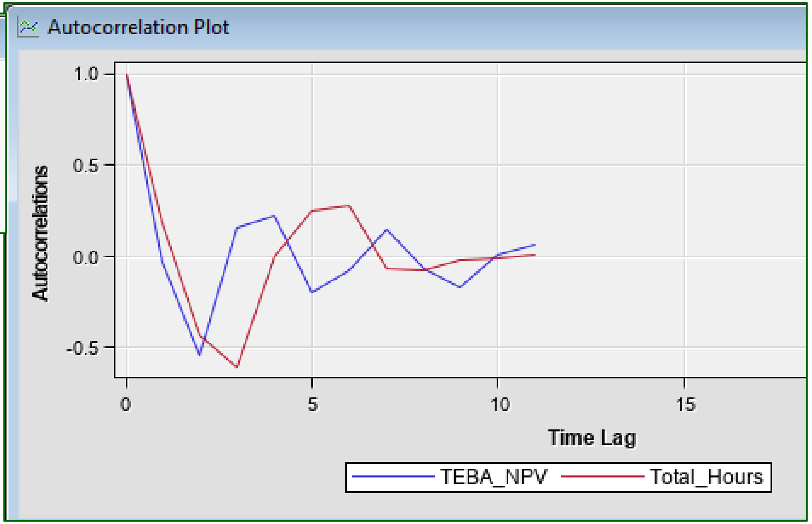
Autocorrelation

(in 2020)

- By contrast, the ACF for both Avg. TEBA and Total Hours in 2020 is very weak overall. In fact, both drop precipi-tously at the very outset of 2020, just before COVID-19.



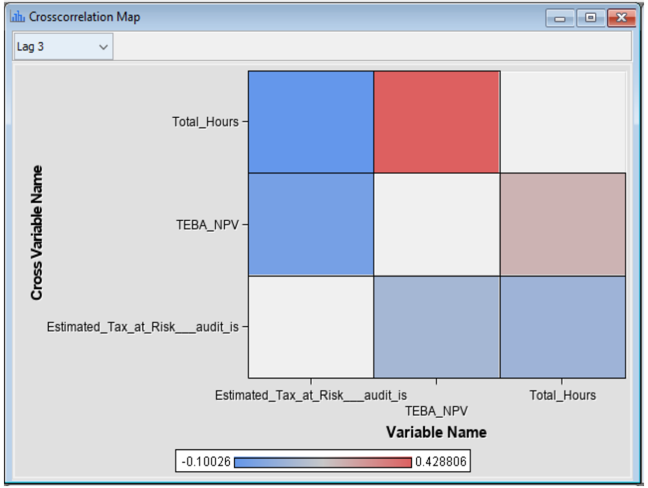
Train	
Variables	
Specify an Interval	Month
Accumulation	Average
Correlation Analysis	Autocorrelation



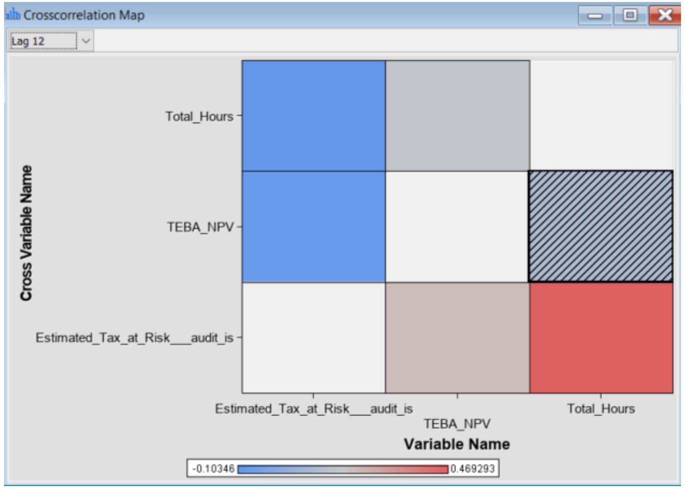
CCA – Cross Correlation Analysis

- For CCA (2016-2019), we can explore lagged effects between estimated TAR (tax-at-risk) and TEBA, as well as those considering Total Hours (on audit cases).

Time LAG 3



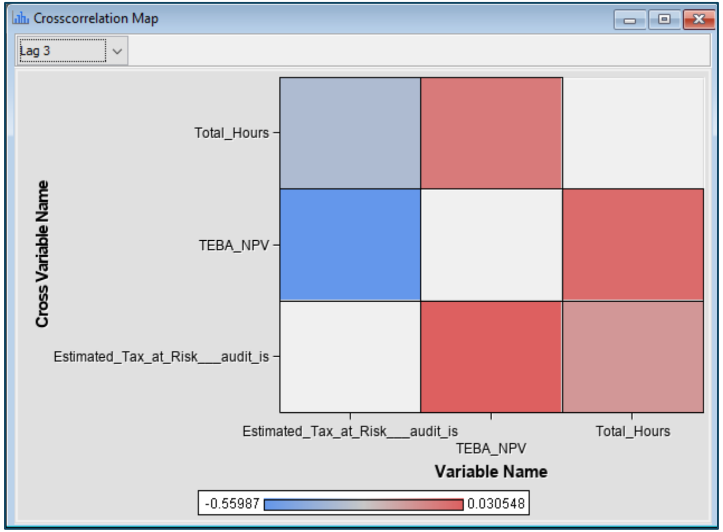
Time LAG 12

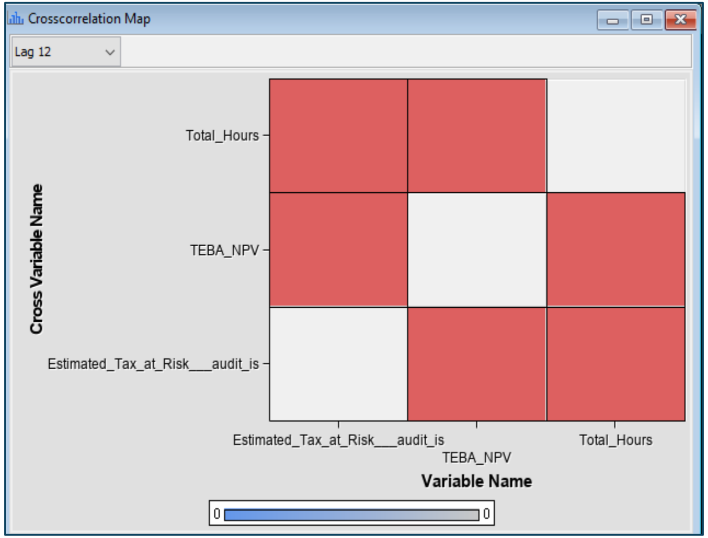


CCA, continued

(During COVID)

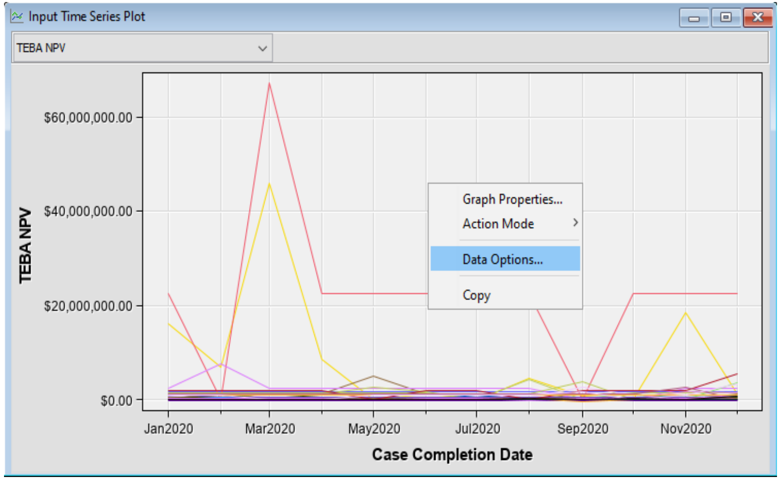
- When we run CCA for lagged effects of TAR (during 2018-2019) on TEBA for 2020, we find a very different pattern at time lag=3 and 12.
- For time lag=3, at left, the best we can get is 3% influence.
- For t=12, at right, it's absolutely nothing.

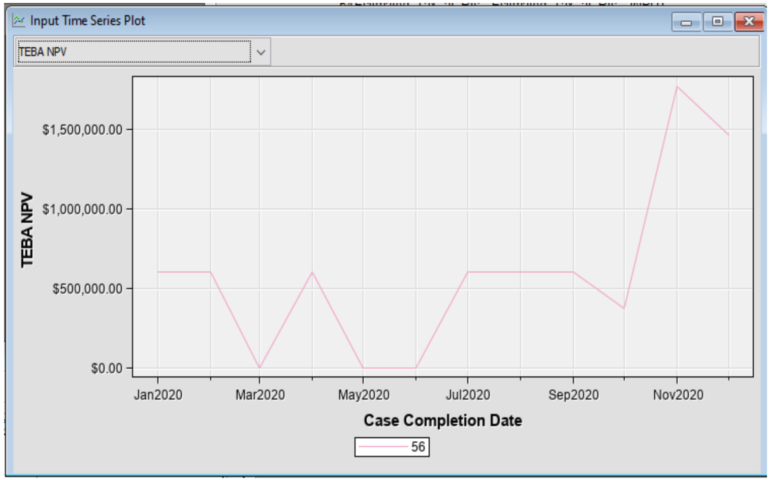




Industry Profiling Analysis

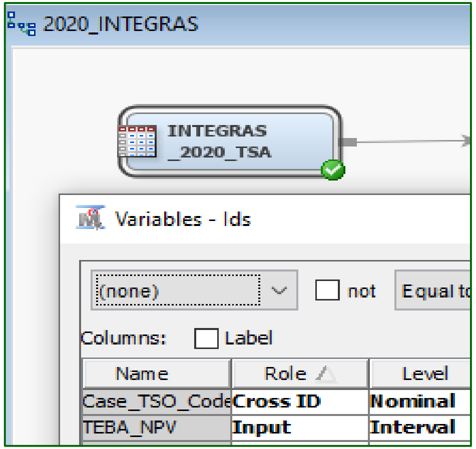
- Using the same data for CCA, we can subdivide our dataset by industry sector, or **NAICS** code. I can set this input to "Cross ID" in the data source's variables list, then re-run the flow.
- From the **TS Data Prep** node's *Results*, right-click in the Time Series Plot and select Data Options. We'll pick a NAICS code at random. And you can see that it took a tumble at the outset of COVID, and struggled to regain its footing – yet exceeding it at calendar year-end.

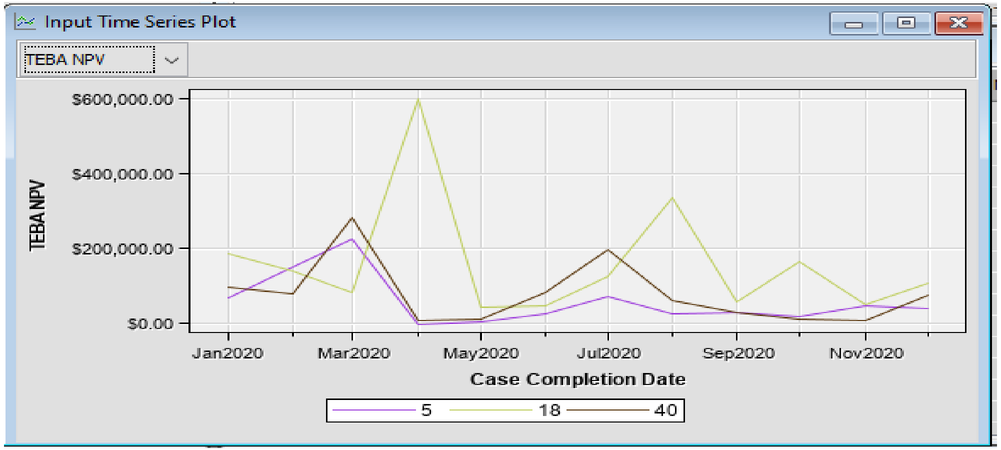




Subsetting by *Tax Service Office*
(the TSO)

- If I want to subset my analysis by a TSO in Canada, I can easily do so by setting the Case_TSO_ID input to **"Cross ID"** at the data source node. (Then re-run the flow.)
- However, by default this displays *all* TSOs in the Input Time Series Plot; so I need to right-click this plot area and select "Data Options" to specify WHERE conditions (where the TSO = 5, 18, **or** 40).





Thank you!

Contact Information
jason.oliver@cra-arc.gc.ca