SAS® GLOBAL FORUM 2021

Paper 1047-2021

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

Jason A. Oliver, Senior Compliance Analyst, Canada Revenue Agency (CRA)

ABSTRACT

It may well be a recurring theme of this year's SAS Global Forum that we are faced with more pressure to use flexible thinking - not just critical thinking - and when it comes to time series analysis and forecasting (TSAF) in SAS, it's all about "rethinking the curve".

At the Canada Revenue Agency (CRA) Compliance Programs Branch (CPB), we have grappled with reliable forecasting for macro-level tax variables on a month-to-month basis, even before the COVID-19 pandemic hit. But now we face a particularly difficult challenge. As with many large organizations, it is not easy to foretell what the fallout may be from such a cataclysm.

In setting up SAS to right the trajectory, we must be extra cautious about some of the fallacies in applying TSAF in this context: the lagged effect for tax revenues realized based on audits of the previous tax year, the need to differentiate average tax recovery per case from sum of tax recovery (month-to-month), realizing that industry sectors are not "one size fits all", and accounting for relatively temporary effects of staffing re-orientation in the conversion to a virtual workplace versus the more enduring effects of business disruptions. With SAS Enterprise Miner's abilities to continuously adjust forecasts, sub-categorize datapoints by tax office or industry sector, and apply lagged cross-correlation analysis, we are suitably equipped with the right tools and this can provide abstract learnings for other large organizations.

INTRODUCTION

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. Enterprise Miner™ includes a highly versatile set of functional nodes for configuring and processing time series data. It can decompose time series components such as seasonality and trend, show trend lines and expected forecast within configurable prediction intervals, and demonstrate complex correlation analyses.

While this has been of great benefit to the CRA in gauging the trajectory of macro-variables related to tax revenues and auditor performance, the findings of this research paper could

conceivably be applied in the abstract to large organizations with process-oriented functions, and not just to other foreign tax jurisdictions.

Let us provide a **Glossary of terms** to set the stage:

- **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. It is in NPV (Net Present Value).
- **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 & SETUP

In SAS® Enterprise Miner™, you have six TSAF nodes in the "Time Series" ribbon; but we're only going to use four of them. Below is the Time Series ribbon with the functional nodes in question:

Figure 1. Time Series Functional Nodes

- **TS Data Preparation:** this node allows you to specify basic time series properties including interval, cycle, start/end time, and *accumulation* (i.e. by total, min or max, mean, etc.)
    - Below, the interval is "automatic", so we specify "Month" as the interval.
    - We can leave the seasonal cycle and start/end time as "Default", as SAS® Enterprise Miner™ will auto-determine these parts from the data.
    - In our case, the data was pre-accumulated in SAS® Enterprise Guide™ row-by-row on a per-month basis, so we can leave Accumulation = "Total" (else, we would have to set it "Average").

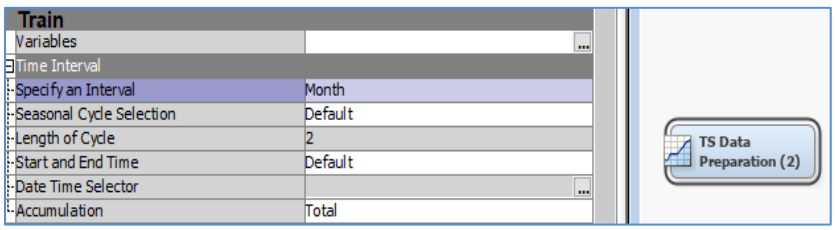


Figure 2. TS Data Preparation node – basic properties

- **TS Decomposition:** this node allows you to specify similar basic settings to that of the TS Data Prep node, but the Number of Periods can be configured, and moreover, you can configure which Export Components you want to display.
    - By default, it will only display "Trend-Cycle" component (=Yes), which is generally regarded as the most salient one.
    - However, in our case, we want to view ALL Components, so we would set that value to "Yes".

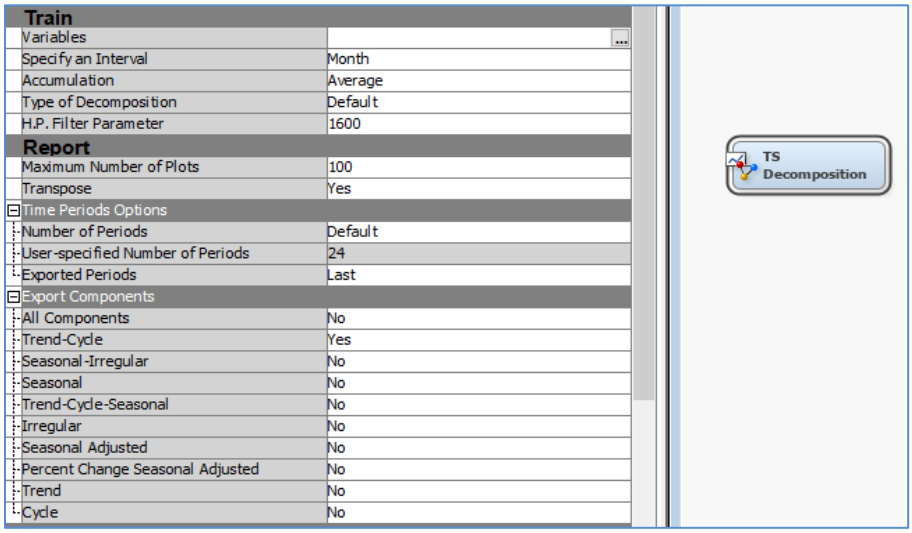


Figure 3. TS Decomposition node –properties

TS Correlation: this node allows you to set up your TSA for autocorrelation analysis, or alternatively for CCA (Cross-correlation analysis). When you select one of those methods, the other one's properties will be greyed out.

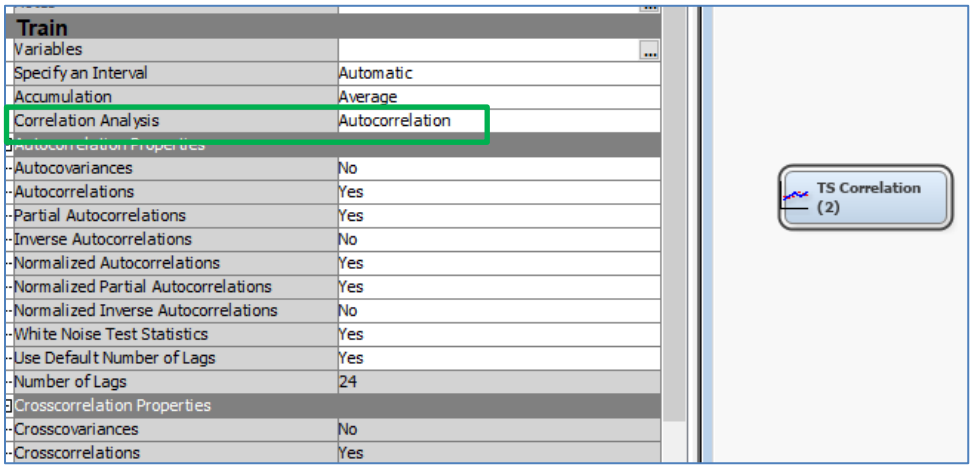


Figure 4. TS Correlation node –properties

Both the TS Correlation and TS Decomposition nodes must be preceded by a *TS Data Preparation* node (which occurs right after the source data node).

TS Exponential Smoothing: this node allows you to conduct forecasting based on your *known data*; as such, you would connect it to a TS Data Preparation node, not directly to your source data node.

- The interval is automatic (which will be *month* in the case of our *pre-accumulated* data), and the accumulation defaults to "Total" (which is OK in our case, for the same reason).
- SAS will pick what it deems to be the best forecasting method.
- The default selection criterion is MSE, or *Mean Squared Error*.
- We will see more on the Forecast lead, back, and significance level parameters during the forecast demonstration in this paper.

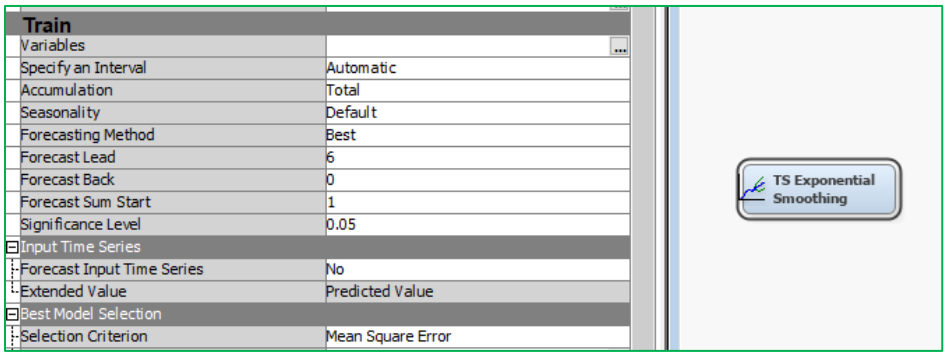


Figure 5. TS Exponential Smoothing node –properties

For our initial workspace setup, we can scrutinize on the C/AR (Case to Action Request) ratio, which as per our glossary is a tentative measure of tax auditor performance. The initial diagram workspace is called "Aggreg_Integras_27mths", which runs from January 2018 to March 2020. This is arranged this way for a reason: because it ends on the month of the COVID shutdown.

Our dataset name is **"TSA_AGGREG_SINGLE_LINE_27MTHS".**

So, when I bring this in, I need to set all variables to Role = "Rejected" *except* a) C/AR ratio and b) my MONTH (Time ID) variable.

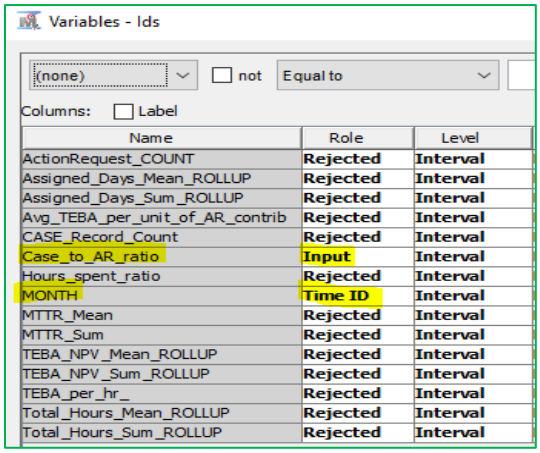


Figure 6. Variable Role selection from data source

You would set your variables once you bring the data source to your diagram (workspace).

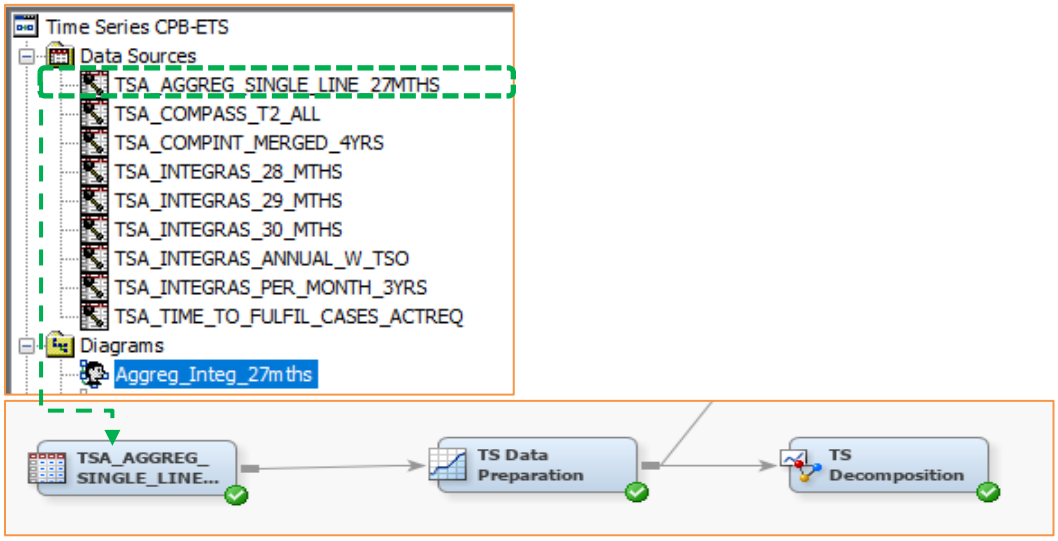


Figure 7. TS Data Source to Diagram flow

NOTE: I do not cover the mechanics behind bringing in a data source, as the principal focus is on conducting TSAF in SAS® Enterprise Miner™. All we need to be concerned with is that as Data Sources become available in the top-left menu, we can drag-and-drop them to our diagram workspace (which are also created by right-clicking 'Diagrams' in the left panel).

In examining the TS Data Preparation node, it is fairly simple: we see the known trajectory of the C/AR variable, simply by right-clicking the node → Run → *Results*.

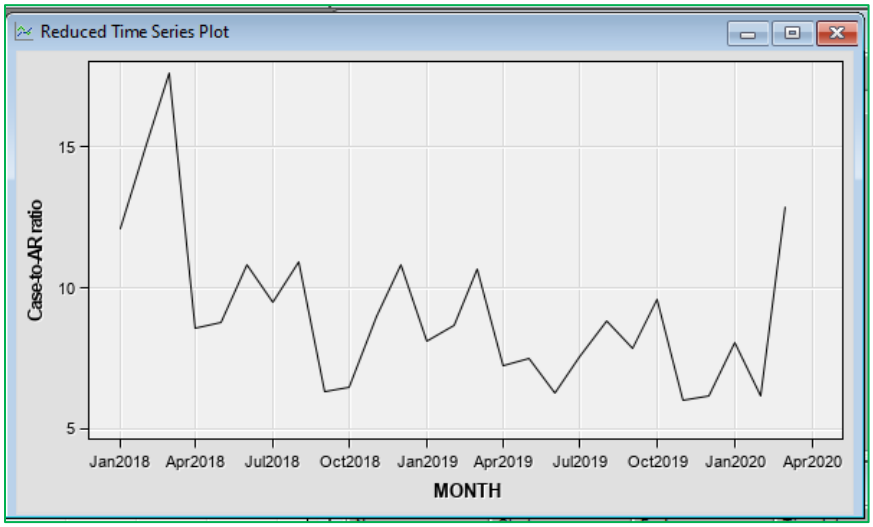


Figure 8. Time Series Plot, for C/AR ratio variable

We can see that the C/AR ratio has fallen off as of mid-2018, and continued on a very gradual downward path. Which means that case auditors are completing disproportionately less cases to the action requests they submit for help, albeit with a seasonal factor and some rebounding of the trend-line in March 2020.

So, we can scrutinize on the more specific components of the time series line by using a TS Decomposition node.

DECOMPOSITION OF TIME SERIES

In running our **TS Decomposition** node, and viewing the results, the first one to examine is the Seasonal Component Plot. When it comes to the C/AR ratio, the seasonal index range is between a high of about 1.3 down to about 0.75.

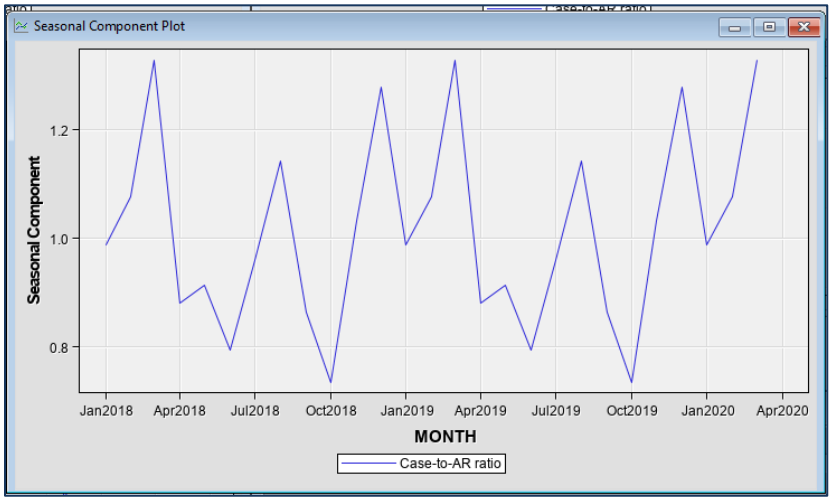


Figure 9. Seasonal Component Plot, for C/AR ratio variable

During the months of March and December, we see fairly high seasonality. This is normal for the time, since the push to complete cases is higher at the end of the CRA fiscal year (March), and ostensibly at the end of the calendar year, also. Auditors are completing proportionally more cases vs. the number of action requests they submit to the service desk. So it is likely that they are fulfilling cases that do not require as many interventions during those months. **Even in March 2020, C/AR still remained high – it was resilient to the initial COVID effects,** due to being a ratio variable and not an absolute sum variable.

In the decomposed results, we can also examine combinatory components; for instance, the Trend-Cycle Component Plot:

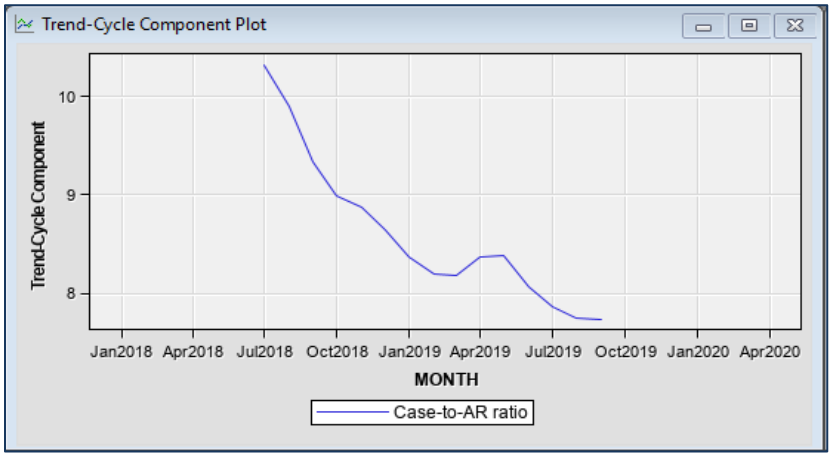


Figure 10. Trend-Cycle Component Plot, for C/AR ratio variable

This tells us what we had surmised from the initial data preparation, that the series has been on a steadily downwards trajectory. Now when it comes to tax-related time series data, there is no real cycle per se; at best, it is an inherited cycle from world economy fluctuations. The proper definition of cycle in a TSA context is not the entity's operational lifecycle; rather, it refers to the boom-and-bust business cycles which are largely unpredictable. Ergo, we are mainly concerned about *trend* here.

Now, if we substitute the **Average TEBA** (tax earned by audit) variable for C/AR [using the Data Source node shown in figure 6 earlier], we can see what emerges in our decomposed time series results.

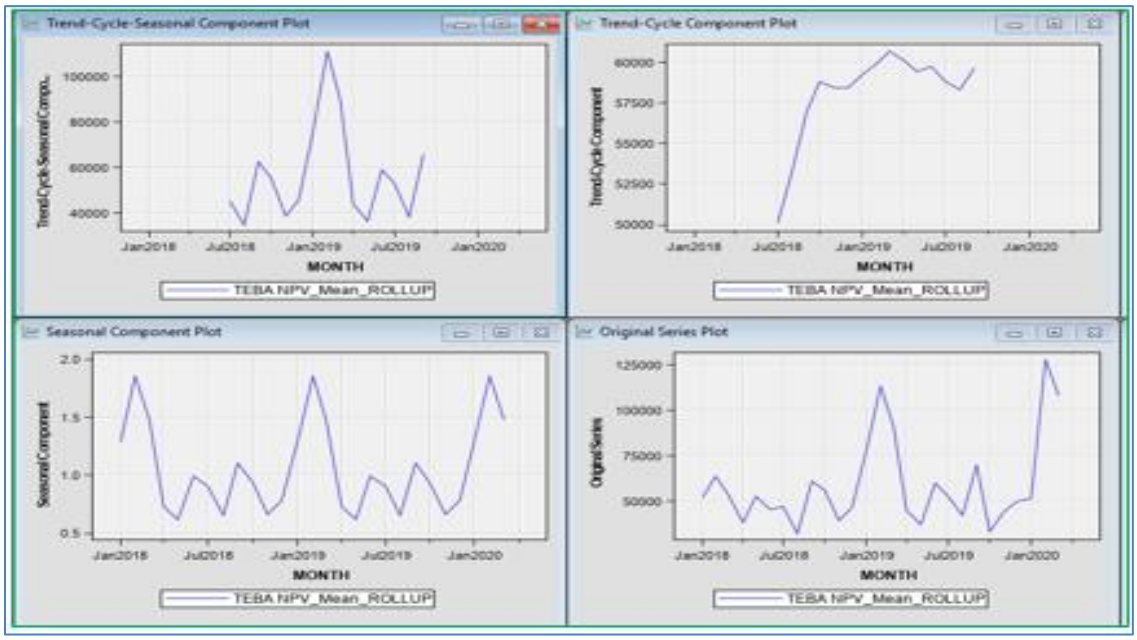


Figure 11. Paneled Component Plots, TS Decomp. for Avg. TEBA

This time, as per the panel graph at bottom-left, we see that our seasonality index is broader than that of C/AR ratio; it goes from a high of about 1.8 to a low of ~0.7. This is largely attributable to the heightened pressures towards fiscal year-end to increase realization of TEBA, which we see in Feb.-March. At the opposite end, we see rather low seasonality for May, August, and November.

For the original series plot, bottom-right, the trend continues gradually upwards with seasonality readily apparent. In the trend-cycle component plot, at top-left, we see that the trend (with cycle, such as it is) is rising steadily upwards but then reaches a virtual plateau.

The key challenge then, has been to resolve and reconcile the expected forecast as of March 2020 with the new COVID-19 realities.

FORECASTING MACRO TAX VARIABLES

AVERAGE TEBA

We can proceed to evaluate the expected trajectory of the AVG. TEBA variable, on a monthly interval. Recall that this variable is pre-accumulated at data source.

When we conduct our forecast, we use the TS Exponential Smoothing node.

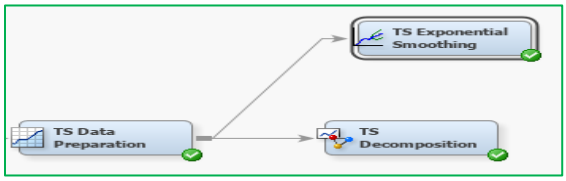


Figure 12. TS Exponential Smoothing node in the TSAF diagram

We let SAS® pick the *best* forecasting method, as well as *selection criterion* (forecast measure). In this case, the latter value is the MSE [Mean Squared Error] as you can see at the bottom of the properties of the node.

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

Figure 13. Properties of the TS Exponential Smoothing node

For our *Significance Level*, we set this to 0.5; it governs the blue bracket around the forecast line, a.k.a. *the prediction interval*. So it is a confidence band of sorts. The way this figure works is the opposite of what some of us might know from frequentist confidence intervals; that is, the lower the "alpha" value, the wider the band (prediction interval) so an "alpha" of 0.01 would produce a very wide band, and an "alpha" value = 0.99 would be virtually limited to just the forecast line itself. So we aim in the middle (which actually is closer to the outline of the trend line, as this figure is more "log-like" in its manifestation).

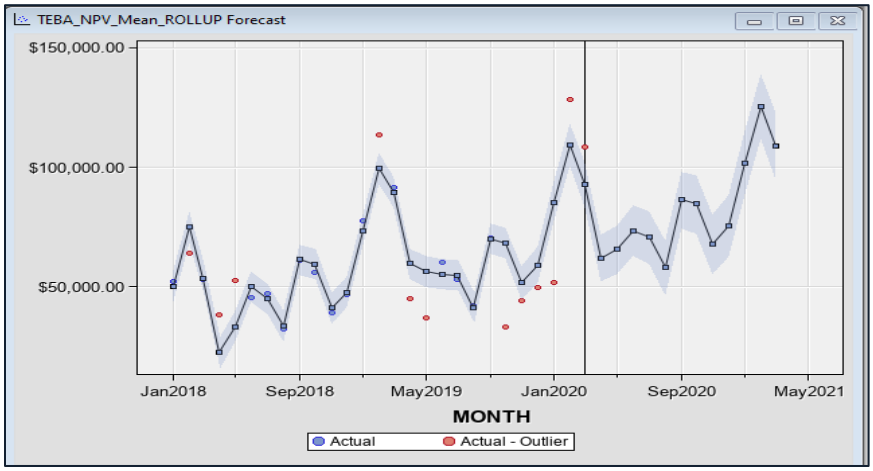


Figure 14. TEBA_NPV_Mean: forecast line from trend

SAS logically expects the trend will continue upwards (while maintaining seasonality, of course) due to "series momentum". Had we began our time series at, say, January 2016 rather than Jan. 2018, that momentum might have been more pronounced. The clichés of

"future behavior is governed by past behavior" and "you can't know where you're going, unless you know where you've been" have never been truer. However, enter COVID-19, and that is a whole new wrench in the gears of the tax-auditing apparatus.

As for the selection of **"Best" *Forecasting Method***: you could try to experiment with different models – there are eight in all, as per fundamental TSAF science – but I can tell from the shape of the forecast line that it's based, appropriately, on the **Additive Winters** method[1]. I ascertained this by running the node with this method selected, and the resulting graph was identical to "best" method. Unlike the *Multiplicative* Winters method, this forecast line is predicated on fairly consistent seasonal "inverted V" shapes in the curve. If those inverted V shapes became noticeable larger (or smaller), then Multiplicative Winters would likely be the "best" method that SAS would auto-select.

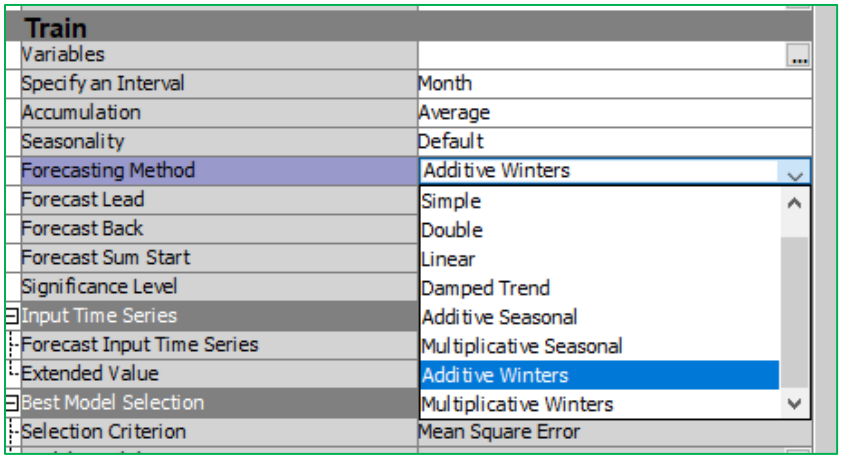


Figure 15. Available Forecasting Methods, properties of TS Exp. Smoothing node

We see that in the resulting forecast, it predicts ahead exactly 12 months. This is the difference between the figures of "Forecast Lead" and "Forecast Back" in the properties. We saw on the previous page that the "Forecast Back" = 6; this acts as our *validation* partition, using the last six months of *known* data (i.e. Oct. 2019 to March 2020). So this gets subtracted from the "Forecast Back" value of 18 to arrive at 12 periods out. Ideally, you want your "back" [validation] period to be between 20-25% of your known data, which it is out of 27 months; even when we increase the known months to 30, it will still be 20% of this.

SUM OF TEBA

When we run a TSAF experiment on the SUM of TEBA – as opposed to its average – we realize a drastic difference in the scale. Because TEBA is a **sum** value, not a ratio (i.e. C/AR, or [Average] TEBA/case), it is simply not as resilient to sudden shocks like COVID-19 – as we will later see when adjusting the forecast based on incremental months (April, May, June) of *known* values.

[1] The essence of the Winters method is to combine discernible trend with seasonality.

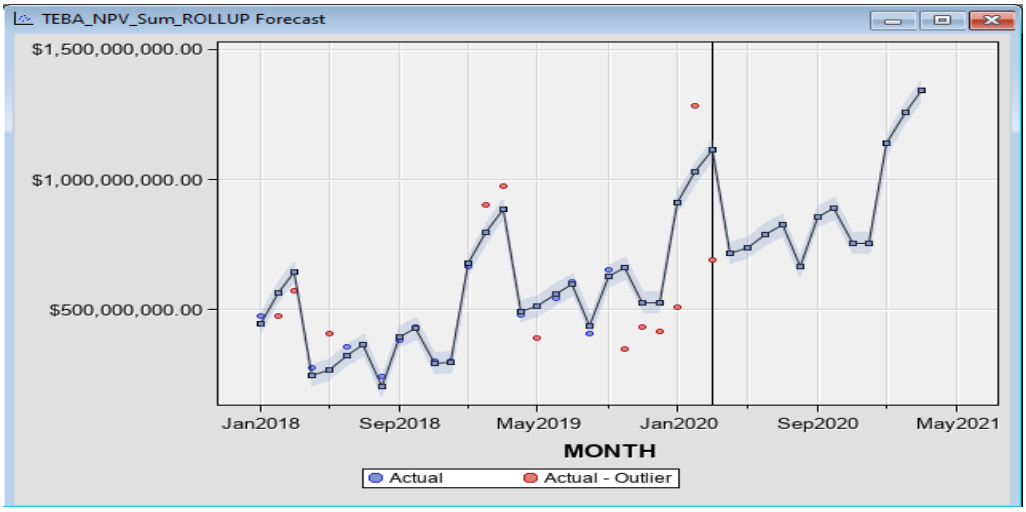


Figure 16. TEBA SUM Forecast (post-March 2020)

Note that the MSE selection criterion (default) graphs a trend line around the known values (which are represented by the red dots here). The SUM TEBA for Feb. 2020 is nearly double what it was for March 2020, as you can see by the relatively large separation of the red dots from the blue dots (on trendline) for those two months. Yet SAS® "thinks" that the trend will continue positively, as it is "COVID-agnostic".

What may also seem shocking to the reader is that the lower limit of the prediction interval for April 2020 (at ~$674.5M) actually ***exceeds* the actual value** for April 2019, which was slightly below $500 million. It is not until the fall until we see that the midpoint of actual 2019 data approximates the LCL (lower confidence limit) of the forecasted band for Sept. 2020. This is ostensibly due to the "positive momentum" of the time series that I alluded to earlier.

C/AR RATIO

Next, we switch out the SUM of TEBA for the C/AR ratio, once again. In forecasting a relatively 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.

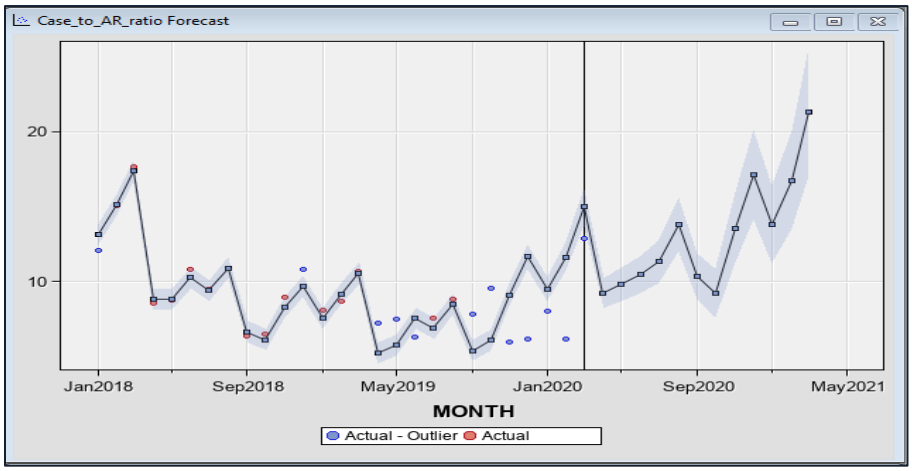


Figure 17. C/AR ratio Forecast

I used the Mean Relative Abs. Error as the forecast metric (selection criterion), which I found to be more appropriate. Regardless, what we see in the actuals for the spring of 2020 is a very low C/AR ratio, telling us that case throughput has suffered as a result of the pandemic AND that Action Requests for help did not decline proportionally; there was still an apparent high need for action requests.

FORECASTING AVG. HOURS PER CASE

For forecasting average hours per [audit] case, I determined that the more ideal Selection Criterion was "Median Relative Abs. Error". No matter what Selection Criterion I used (or Significance Level), the prediction interval still dipped into the negative range. Sometimes, this is unavoidable. But then the prediction interval becomes spurious; you can't have negative hours. So we tend to just focus on the midpoint values in this situation.

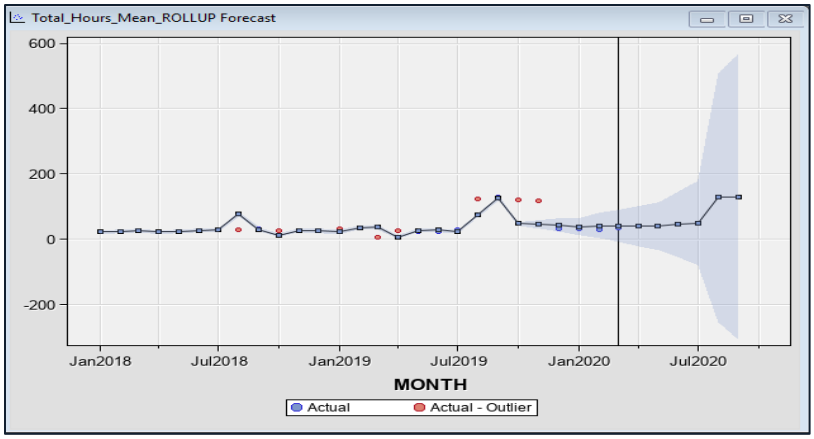


Figure 18. Average hours per case Forecast

We can see that the midpoint goes very subtly upwards for the first few forecasted points (post-March 2020), then sharply up for summer. As it turns out, this is a fairly good approximation of the reality, since the Avg. Hours per case during the middle of 2020 is about 1.5-2.0 times that of the previous year. What is especially pronounced is that the Average Hours of March 2019 were only 6.25, whereas for March 2020, it was 35.44. This was predicated on an Agency policy-induced change; refer to the link and passage below:

https://www.mondaq.com/canada/audit/1030308/cra-moves-forward-with-international-audits-despite-continued-backlog-?email_access=on

In March 2020, the CRA announced that it was suspending the vast majority of audit activity for a minimum of four weeks, other than audits involving the very largest taxpayers. This suspension meant that the CRA ceased requests for information relating to existing audits, finalizing existing audits, and issuing reassessments. Further, deadlines for information or document requests were suspended and no action was required from taxpayers under audit during this time. This suspension remained in effect until June 2020, though audits of small and medium businesses did not resume until late fall.

This is also arguably responsible for the "pulse" effect we see in *actual* Avg. TEBA for July 2020, as per the monthly incremental analysis that comes next.

INCREMENTAL ALIGNMENT

APRIL 2020, KNOWN VALUES

Now when we add the month of April 2020 to our data (making it 28 mths total), we would expect the **AVG. TEBA *actuals*** for subsequent months to become closer to / within forecast range. As an example in the graph cross-section that follows, the forecast for September, October, and December 2020 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. This was likely due to the resumption of standard large business audit as of June 2020 (see previous page article/passage).

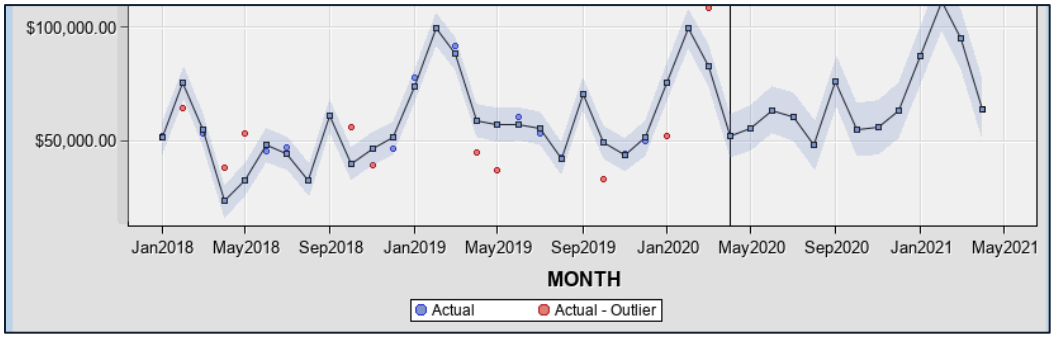


Figure 19. Revised AVG. TEBA forecast, incremental inclusion of APRIL 2020

Again, we typically use the measure of MSE [Mean Squared Error] in gauging efficacy or proximity of a forecast to actual [values]. See the *Appendix* tables at the end of this paper for a breakdown of this analysis, where I illustrate monthly incremental effect on accuracy of the last six months of the calendar year (i.e. from July to Dec. 2020).

MAY 2020, KNOWN VALUES

Clearly, the addition of April wasn't enough to right the trajectory of the expanding "COVID window". So in continuing our analysis of monthly incremental effect, I added May 2020's known data 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 Avg. TEBA is an irregular value (~$122K), since July 2018 had Avg. TEBA =~$45K, and July 2019's Avg. TEBA was ~$57K. *It is clear that this is a COVID-adjustment spike.*

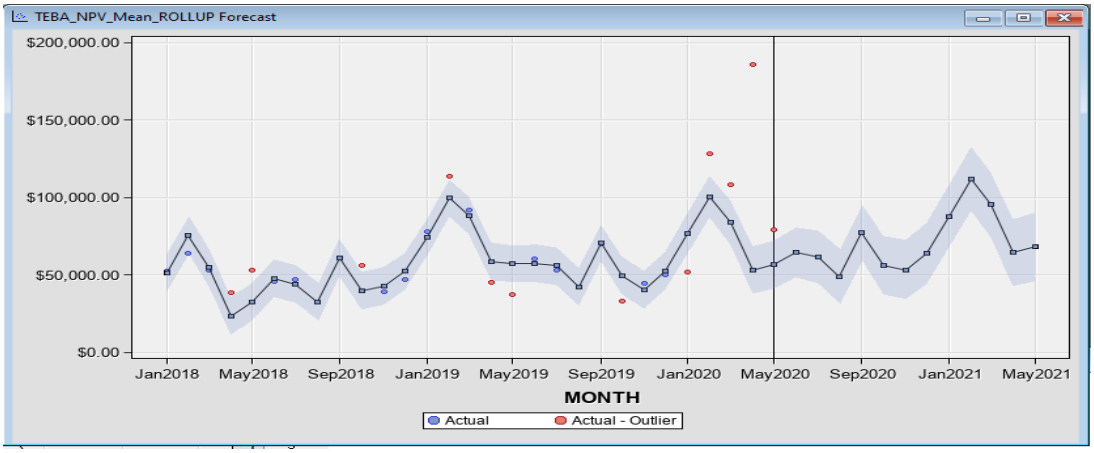


Figure 20. Revised AVG. TEBA forecast, incremental inclusion of MAY 2020

We can therefore define July 2020 as a ***pulse***, or a one-time brief event, that caused a spike in the accumulated time series value for that month. This emphasis on larger business for audit while suspending SMB audits at the time is further substantiated by the fact that in July 2020, there was an average of 50.75 hrs per case completed, which is extremely high. For April, which had a very high Average TEBA of $185.5K, the figure was 52.16 average hours per case.

JUNE 2020, KNOWN VALUES

Predictably, for the addition of June 2020, it didn't improve the forecast band to include the actual Avg. TEBA for 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 (being a ratio value, in essence). To wit: observe below that the April-May-June line for the original forecast (left) and actual data points (right) is just above the $50K line, and follows the same trajectory.

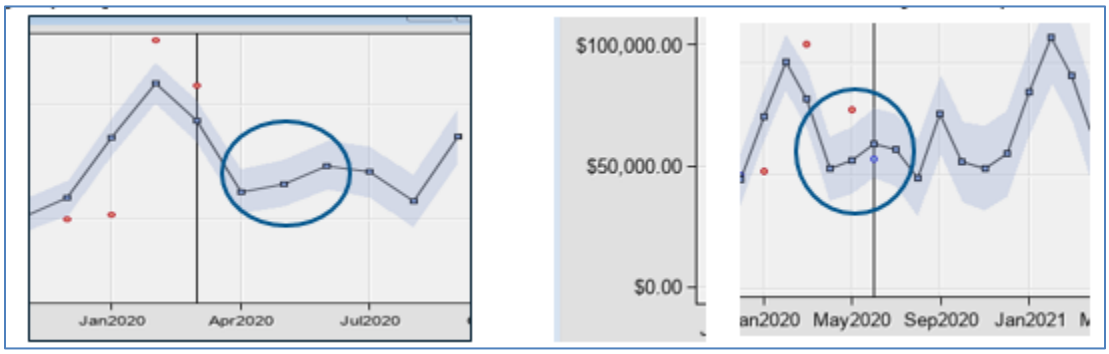


Figure 21. Comparing Q1 of FY2020-21 forecast vs. actual data points

In taking MSE and RMSE (R is "root") measurements for both the as-of-March and as-of-June forecasts, we only note a slight improvement (reduction) in that value. Which also goes to show the resilience of this variable, and the "pulse" nature of July's spike.

MEASURE / as of MONTH	MARCH 2020	JUNE 2020
AVG. TEBA (MSE)	$ 954,467,257.64	$ 888,454,004.34
RMSE	$ 30,894.45	$ 29,806.95

Table 1. Point-in-time [R]MSE for AVG. TEBA forecast-to-actual: July to Dec. 2020

Refer to the *Appendix* at the end of this paper for a more detailed month-by-month breakdown of these calculations.

FALLACY: COMPARING SUM OF TEBA SHIFT TO AVG. TEBA CHANGES

TSAF works best when you accumulate data records **by *average***, not by sum total. If we tried this exercise using SUM TEBA per month, it would not turn out very well, because sum totals are immediately impacted by any severe transition, i.e. auditor work re-arrangements and temporary audit case policy due to COVID-19 fallout as of March 2020.

Evaluating the March 2019-2020 comparison in the following table, the **TEBA_SUM** and **Case Count** have dropped significantly in March 2020, yet the **C/AR ratio** has augmented.

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

Table 2. Year-over-Year March comparison, key macro-variables in TSA

However, as the staffing situation has attempted to stabilize in the intervening months (April to June 2020), the C/AR ratio has dropped dramatically. (Not shown in above table.) The same is true for the TEBA/AR pattern.

SUM OF TEBA: DRASTIC CHANGE

We now compare the SUM TEBA forecast as of March 2020 (left image) and that of June 2020 known data points (right image).

Figure 22. Comparison of SUM of TEBA forecast as of March vs. as of June (2020)

For the first image, none of the actuals of the last six months of 2020 fall in the forecast band. Whereas, for the second image, two of the actuals of the last six months (Oct., Nov.) fall in the forecast band.

Also observe how some of the accumulated data points in the forecast are more "depressed" in the latter graph; while there is a discernible peak, it doesn't quite have the same buoyancy or upwards momentum as the former graph. (We must keep in mind, though, that this is still using *the MSE method,* i.e. taking a line of best fit, where the red dots are the actual values.)

So, there is little point in using the MSE to gauge efficacy of the monthly adjustment, simply because the values would be so huge (as opposed to those in the *Avg.* TEBA MSE).

ADVERSE IMPACTS AND DELAYED EFFECTS

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*. The graph below covers known Avg. TEBA trend data points right up to December 2020, the lowest point.

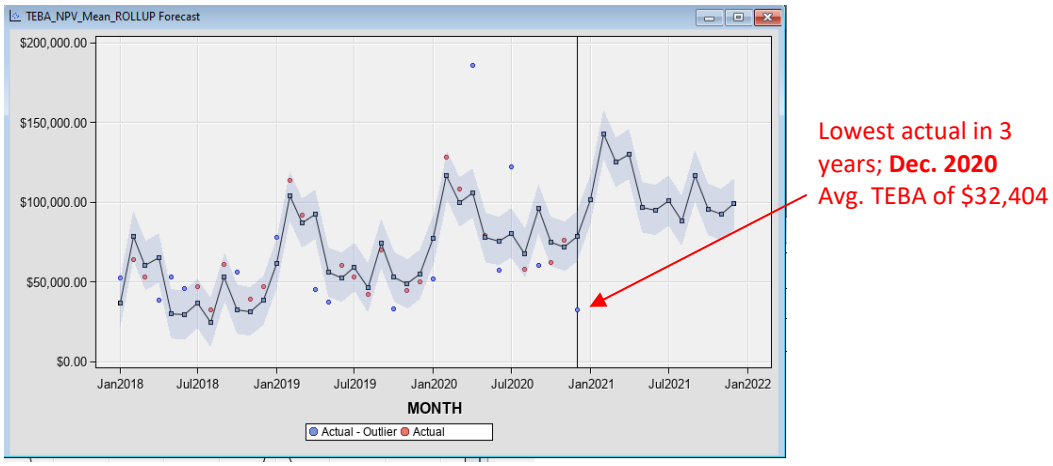


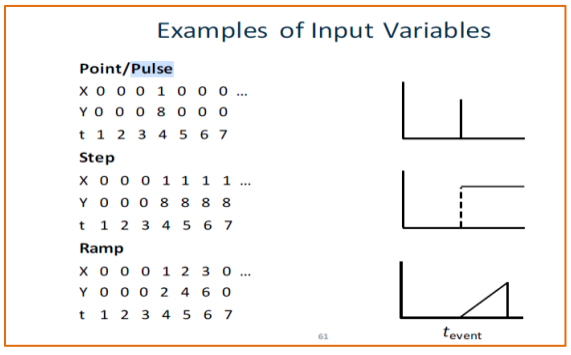
Figure 23. Calendar-year-end (2020) Avg. TEBA; lowest point

This extremely low Average TEBA of ~$32,000 per case *could be* a harbinger of further average TEBA decline, but we'd have to observe the last quarter of the fiscal year – January to March 2020, once available – and validate that theory. (*Then* we might apply an intervention to the time series line.)

Incidentally, when it comes to SUM of TEBA with actuals up to Dec. 2020, the forecast trend line for 2021 is far more credible, showing all datapoints as being well under $1 billion, and mostly under $500 million.

INTERVENTIONS

As alluded to before, a TSAF exercise may use **interventions**, if the extreme or irregular event is known in advance (or shortly thereafter). This is an adjustment to the "regular" time series, using a "dummy" variable for the period of observation. In this case study, we'd recommend an intervention for the SUM of TEBA as of March 2020, and possibly for AVG TEBA as of Dec. 2020. Plus, we might use a "pulse effect" for July 2020. However, programming an intervention requires SAS® Studio™, which is out of scope for this paper.



A **step** would work best as an intervention (for March 2020 and Dec. 2020), since the trend line shift is sudden and sustained; it does not happen gradually then return to baseline.

Figure 24. Basic denotation of input variables (interventions) by type

TS CORRELATION NODE

AUTOCORRELATION

When we deal with a significant seasonal and/or trend component, we usually find a greater degree of **autocorrelation** factor (abbreviated "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".

In a similar vein, when frontline auditing teams are performing well, some of that momentum carries over from one period to the next, as they build "muscle memory" and are better-equipped to deal with more trying scenarios that have [abstract] aspects in common with recent cases worked on. This presents opportunities for "boilerplate" copying and pasting of common findings from one case to another, adjusting for specifics, and accelerating average time to complete as well as garnering more average TEBA per case.

Clearly, during the current COVID-19 climate at this writing, and the embargo of SMB case audit during the spring 2020 period, we can expect some of that momentum to be adversely impacted – since auditors were working on more complex large business cases overall. But first, let us examine a baseline from the years 2018-2019, below:

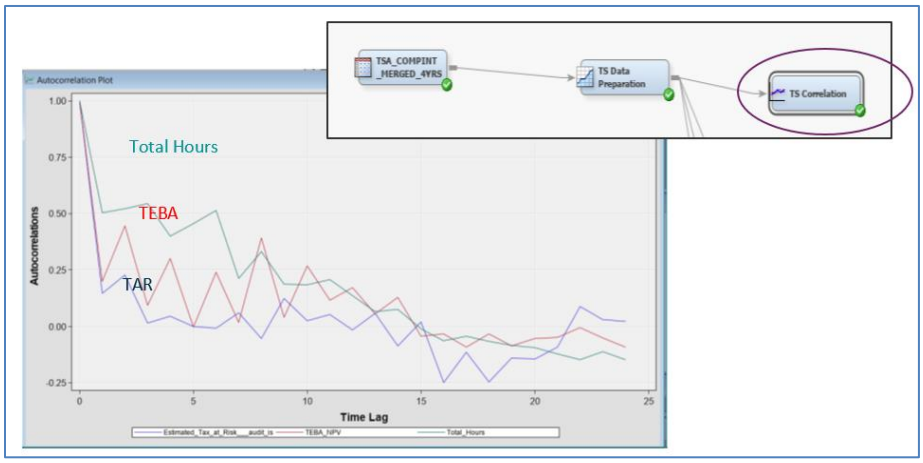


Figure 25. ACF Plot, three key tax-related macro-variables (2018-2019)

From the three variables plotted above, Est. TAR-AI (tax-at-risk – audit issue) has low ACF, TEBA has moderately high ACF, and Total [Avg. Case] Hours has very high ACF. To wit: at lag t=5, TEBA reaches the zero line; but Total Hours is still at ACF=0.45.

By stark contrast, in 2020 (below), the ACF for both Avg. TEBA *and* Case Hours is very weak overall. In fact, both drop precipitously at the very outset of 2020, just prior to COVID-19.

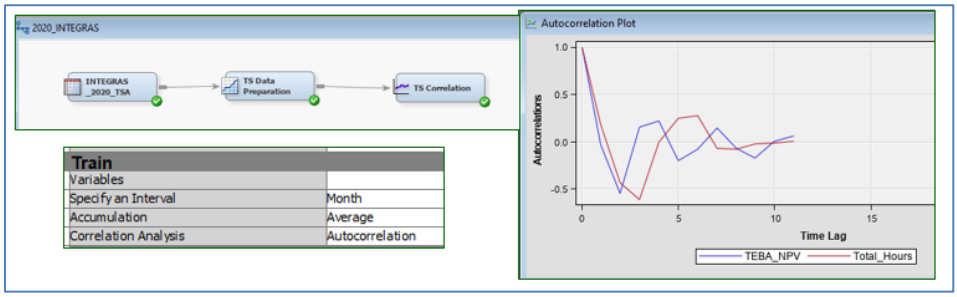


Figure 26. ACF Plot, same macro-variables, for 2020

CCA – CROSS-CORRELATION ANALYSIS

When we explore lagged effects between risk-related variables – in this case, TAR (tax-at-risk) and TEBA (tax earned by audit) – we would use a CCA plot. We are also considering Total Hours (on audit cases) here. The plots below are at t=3 months and t=12 months out, with the influencing variables on the vertical axis, and the influenced variables on the X-axis. The color shading is somewhat counterintuitive, whereby red means more positively cross-correlated, and blue means less so. Again, we set a baseline of expectations using tax data from 2016 to 2019 (48 months) here.

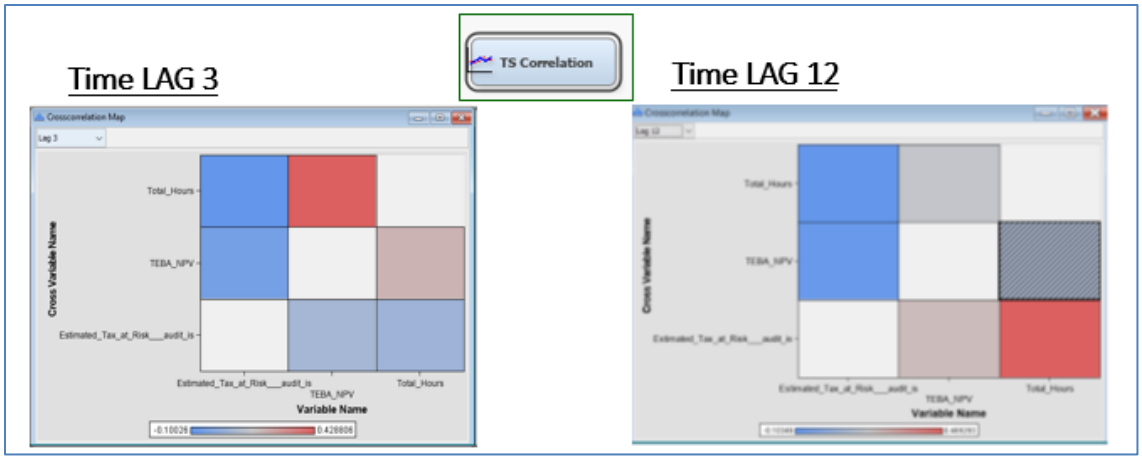


Figure 27. CCA Map, at time lags 3 and 12, key macro-variables

Note the pronounced difference in CCA factor: for time lag 3, the Estimated TAR has virtually no effect on TEBA or Total Hours per case (because it's too close time-wise), but 12 months out (at right) it has a very pronounced effect on total case hours, and a moderate effect on TEBA (~22%). Also, in the first graph for time lag 3, TEBA highly influences Total Hours and to a noticeable degree vice-versa too. **But** when we get to 12 months out, Total Hours has virtually no lagged effect on TEBA, and vice-versa.

If we repeat the experiment from 2018 data up to 2020 (COVID window) data, evaluating lagged effects of TAR on TEBA for 2020, we find a very different pattern at t=3 and t=12. For time lag=3, the best we get is ~3% influence; for t=12, it's absolutely nothing.

Figure 28. CCA Map, at time lags 3 and 12, inclusive of COVID-19 period

SUBSETTED ANALYSIS

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 fell at the outset of COVID, and struggled to regain its footing – yet exceeding it by calendar year-end.

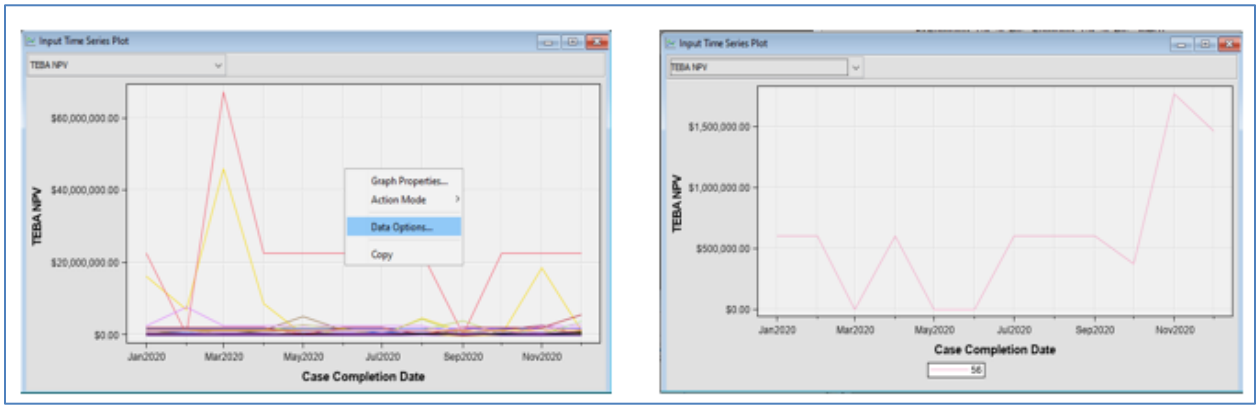


Figure 29. Industry Profile (NAICS) subsetting of Avg. TEBA in TS Plot (in 2020)

Note that when you have over 100 categorical values – as in the case of NAICS industry codes here – it will only allow you to select from the first 100. In my opinion and experience, I prefer SAS VIYA when it comes to subsetting TSA by key categories.

BY TSO (TAX SERVICES OFFICE)

So let us examine a subsetting TSA for an under-100 categorical set. I use the **TSO**, or *Tax Service Office* parameter, so again I set the Case_TSO_ID input to "Cross ID" at the data source node. Then I re-run the flow and access the *Results*.

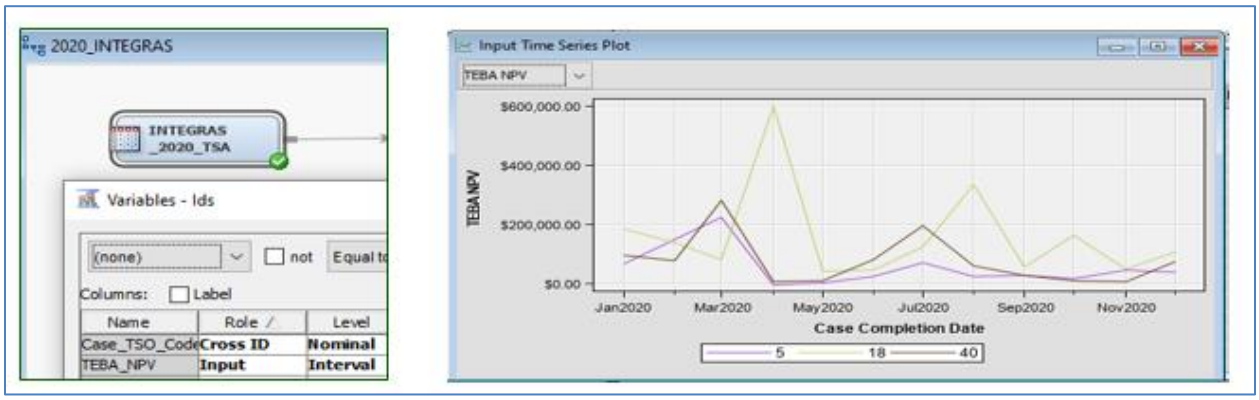


Figure 30. Tax Services Office (TSO) subsetting of Avg. TEBA in TS Plot (in 2020)

By default, this will display all TSO IDs in the Input TS Plot; so I have to right-click the plot area and select **"Data Options"** to specify filters (WHERE TSO = 5, 18, **or** 40). Note that while all of these TSOs converge at various points, in the month of April we find a very strange anomaly: TSO 18 has AVG. TEBA =~ $600K, but the other two TSOs have TEBA just under $10,000. Yet all three of them re-converge later in 2020.

CONCLUSION

We have seen the power and versatility of SAS® Enterprise Miner™ for conducting TSAF exercises. It is clear that not all macro-variables in the Canada Revenue Agency exhibit the same behaviors or resilience at various points in the turbulent COVID-19 period, but a good deal of this can be attributed to whether they were pure sum variables, or derived ratio-like variables. Some disruptions – prompting the insertion of intervention effects – were ostensibly due to policies in place to "take the edge off" more vulnerable business.

Many of us can also take away abstract learnings from this paper, even if such individuals are not employed in the tax sector – because in the end, it is all about maintaining a certain buoyancy of the macro-variables that matter most, to the extent possible – these are not easy times to navigate and we wish those adversely impacted the most clement journey to a regained prosperity.

REFERENCES

Sarma, Kattamuri S., PhD. Copyright © 2017. Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition. Cary, NC, USA: SAS Institute, Inc.

ACKNOWLEDGMENTS

I am grateful to my family for their encouragement on this endeavor. I am also grateful to the numerous staff of the CRA who were the audience in my internal presentation of this TSAF subject matter. I also acknowledge and admit defeat to the spell checker in insisting on the spelling of "endeavor" as it is, not like it ought to be as it is on the space shuttle. Which, unlike CRA time series, must be expected to follow a known trajectory.

RECOMMENDED READING

- Milhøj, Anders. *Practical Time Series Analysis Using SAS®.* Copyright © 2013, SAS Institute Inc., Cary, NC, USA.
- Shumway, Robert H. and Stoffer, David S. *Time Series Analysis and its Applications.* 4th ed. © Springer International Publishing AG, 2017, Univ. of California at Davis. Davis, CA, USA.
- Brocklebank, John C., Dickey, David A, and Choi, Bong S. *SAS® for Forecasting Time Series.* 3rd ed. Copyright © 2018, SAS Institute Inc., Cary, NC, USA.
- Svolba, Gerhard. *Applying Data Science: Business Case Studies Using SAS®.* Copyright © 2017, SAS Institute Inc., Cary, NC, USA.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jason A. Oliver, Senior Compliance Analyst & Data Scientist
Canada Revenue Agency
Jason.oliver@cra-arc.gc.ca

APPENDIX: TABLES OF ACTUAL-TO-FORECAST ANALYSIS

This contains detailed breakdowns of the incremental monthly additions of accumulated data to the COVID-19 observation window.

AVERAGE TEBA

This begins with **Average TEBA,** being subject to both MSE and RMSE (Mean Squared Error, and Root Mean Squared Error).

	as of: March 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
July	$ 59,647.53	$70,662.16	$ 81,676.79	$ 121,878.42	$ 51,216.26	$ 2,623,105,288.39	
Aug	$ 49,593.40	$58,008.86	$ 69,424.33	$ 57,904.40	-$ 104.46	$ 10,911.89	
Sept	$ 74,489.56	$86,293.00	$ 98,096.45	$ 60,307.46	-$ 25,985.54	$ 675,248,289.09	
Oct	$ 72,423.54	$84,604.20	$ 96,784.87	$ 61,904.43	-$ 22,699.77	$ 515,279,558.05	
Nov	$ 55,312.29	$67,858.72	$ 80,405.14	$ 76,306.67	$ 8,447.95	$ 71,367,859.20	
Dec	$ 62,417.96	$75,320.48	$ 88,223.00	$ 32,404.38	-$ 42,916.10	$ 1,841,791,639.21	
						$ 5,726,803,545.84	TOTAL
						$ 954,467,257.64	MSE
						30,894.45	RMSE

	as of: April 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
July	$ 49,924.61	$60,574.97	$ 71,225.33	$ 121,878.42	$ 61,303.45	$ 3,758,112,981.90	
Aug	$ 36,886.63	$47,890.41	$ 58,894.18	$ 57,904.40	$ 10,013.99	$ 100,279,995.72	
Sept	$ 64,798.25	$76,145.10	$ 87,491.95	$ 60,307.46	-$ 15,837.64	$ 250,830,840.77	
Oct	$ 43,404.28	$55,084.79	$ 66,765.30	$ 61,904.43	$ 6,819.64	$ 46,507,489.73	
Nov	$ 43,684.45	$55,691.02	$ 67,697.60	$ 76,306.67	$ 20,615.65	$ 425,005,024.92	
Dec	$ 50,796.25	$63,119.87	$ 75,443.48	$ 32,404.38	-$ 30,715.49	$ 943,441,325.94	
						$ 5,524,177,658.98	TOTAL
						$ 920,696,276.50	MSE
						30,342.98	RMSE

At this juncture, between April and May 2020 known data, the MSE / RMSE actually regresses slightly, telling us that we might as well have gone straight to June 2020's data.

	as of: May 2020 known				Reduced to 0.25 P.I.		
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
July	$ 44,725.59	$61,594.64	$ 78,463.70	$ 121,878.42	$ 60,283.78	$ 3,634,134,131.09	
Aug	$ 31,464.59	$48,912.12	$ 66,377.65	$ 57,904.40	$ 8,992.28	$ 80,861,099.60	
Sept	$ 59,159.81	$77,185.80	$ 95,211.78	$ 60,307.46	-$ 16,878.34	$ 284,878,361.16	
Oct	$ 37,554.69	$56,133.76	$ 74,712.83	$ 61,904.43	$ 5,770.67	$ 33,300,632.25	
Nov	$ 34,100.64	$53,217.84	$ 72,335.05	$ 76,306.67	$ 23,088.83	$ 533,094,070.77	
Dec	$ 44,237.81	$63,881.18	$ 83,524.55	$ 32,404.38	-$ 31,476.80	$ 990,788,938.24	
						$ 5,557,057,233.10	TOTAL
						$ 926,176,205.52	MSE
						30,433.14	RMSE

In the end, this substantiates our earlier findings, that because Average TEBA is in essence a ratio variable and more resilient to initial COVID window – especially since it is predicated

on audits of *past year's* tax filings – there was no real near-future benefit to forecast alignment based on incremental monthly additions for spring.

	as of: June 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
July	$ 45,199.85	$61,424.49	$ 77,649.14	$ 121,878.42	$ 60,453.93	$ 3,654,677,652.44	
Aug	$ 31,883.79	$48,757.98	$ 65,632.17	$ 57,904.40	$ 9,146.42	$ 83,656,998.82	
Sept	$ 59,528.38	$77,029.25	$ 94,530.11	$ 60,307.46	-$ 16,721.79	$ 279,618,260.80	
Oct	$ 37,876.10	$55,983.14	$ 74,090.18	$ 61,904.43	$ 5,921.29	$ 35,061,675.26	
Nov	$ 34,377.42	$53,072.14	$ 71,766.86	$ 76,306.67	$ 23,234.53	$ 539,843,384.32	
Dec	$ 40,302.48	$59,568.07	$ 78,833.66	$ 32,404.38	-$ 27,163.69	$ 737,866,054.42	
						$ 5,330,724,026.07	TOTAL
						$ 888,454,004.34	MSE
						29,806.95	RMSE

C/AR RATIO

This, once again, is the *Cases [Completed] to Action Requests [Submitted]* ratio. Here I break down the monthly forecast measure, using MSE (no RMSE), of the last six months of calendar year 2020 and incrementing known months from March up to June. For March to May, I include the spring months not yet arrived at in each incremental forecast.

	as of: March 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
April	7.53	10.91	9.22	5.24	5.67	32.1489	
May	2.95	11.73	9.84	2.57	9.16	83.9056	
June	8.36	10.48	12.59	3.87	6.61	43.6921	
July	8.91	11.33	13.74	4.41	6.92	47.8864	
Aug	10.74	13.80	16.85	4.49	9.31	86.6761	
Sept	7.77	10.37	12.96	5.43	4.94	24.4036	
Oct	6.39	9.21	12.03	7.38	1.83	3.3489	
Nov	9.63	13.56	17.49	5.07	8.49	72.0801	
Dec	12.07	17.16	22.65	5.62	11.54	133.1716	
						367.57	TOTAL
						40.84	MSE

	as of: April 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
May	10.02	11.98	13.94	2.57	9.41	88.5481	
June	11.03	13.19	15.35	3.87	9.32	86.8624	
July	11.86	14.26	16.66	4.41	9.85	97.0225	
Aug	14.69	17.64	20.58	4.49	13.15	172.7910	
Sept	10.49	13.05	15.60	5.43	7.62	57.9882	
Oct	11.82	14.76	17.70	7.38	7.38	54.4644	
Nov	14.66	18.69	22.72	5.07	13.62	185.5044	
Dec	18.51	23.61	28.70	5.62	17.99	323.4602	
						891.23	TOTAL
						111.40	MSE

From adding April known data, the forecast actually worsens; this is arguably due to having been accustomed to high C/AR values for so long. It is not until we add MAY that it becomes more realistic.

	as of: May 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
June	3.35	5.67	7.98	3.87	1.80	3.2400	
July	3.24	5.64	8.03	4.41	1.23	1.5129	
Aug	4.51	6.99	9.46	4.49	2.50	6.2500	
Sept	1.66	4.26	6.76	5.43	-1.17	1.3689	
Oct	2.52	5.15	7.77	7.38	-2.23	4.9729	
Nov	1.90	4.60	7.30	5.07	-0.47	0.2209	
Dec	3.95	6.73	9.50	5.62	1.11	1.2321	
						15.56	TOTAL
						2.22	MSE

Given this extremely low MSE value, brought on by the actual 2.57 C/AR value of May, we have reached the optimum point – as evidenced by adding June to known values:

	as of: June 2020 known						
	LCL	Midpoint	UCL	ACTUAL	DIFF.	SQRD.	
July	2.80	4.84	6.88	4.41	0.43	0.1849	
Aug	4.11	6.19	8.27	4.49	1.70	2.8900	
Sept	1.30	3.42	5.53	5.43	-2.01	4.0401	
Oct	2.20	4.35	6.49	7.38	-3.03	9.1809	
Nov	1.62	3.80	5.98	5.07	-1.27	1.6129	
Dec	2.60	4.81	7.03	5.62	-0.81	0.6529	
						18.56	TOTAL
						3.09	MSE

CASE HOURS

Lastly, in speaking to *Hours per [audit] case* forecast, I provide a condensed analysis using a simplified MAE [Mean Absolute Error] criterion.

- As of March 2020; forecast of April to Dec. 2020: MAE = 78.52
- As of April 2020; forecast of May to Dec. 2020: MAE = 95.83
- As of May 2020; forecast of June to Dec. 2020: MAE = 107.99
- As of June 2020; forecast of July to Dec. 2020: MAE = 71.51

So, all in all, this proved a very difficult variable to effectively forecast.