A workshop on

Forecasting with Temporal Hierarchies

(based on my current view on the topic – exercise: forecast the chance I will agree with this next year!)

Nikolaos Kourentzes

International Symposium of Forecasting 2025

What we will talk about

1. Why bother with temporal hierarchies?
2. The theory
3. Applications & observations
4. Newer results

(You also get an example of how to implement temporal hierarchies in R with relatively limited uses of packages so you can follow the logic.)

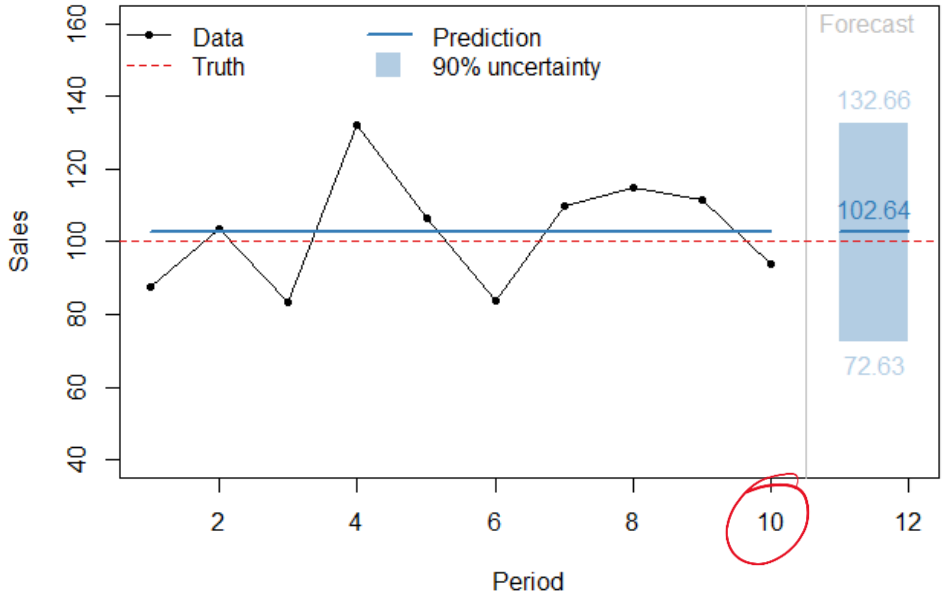
Why bother? The statistical argument

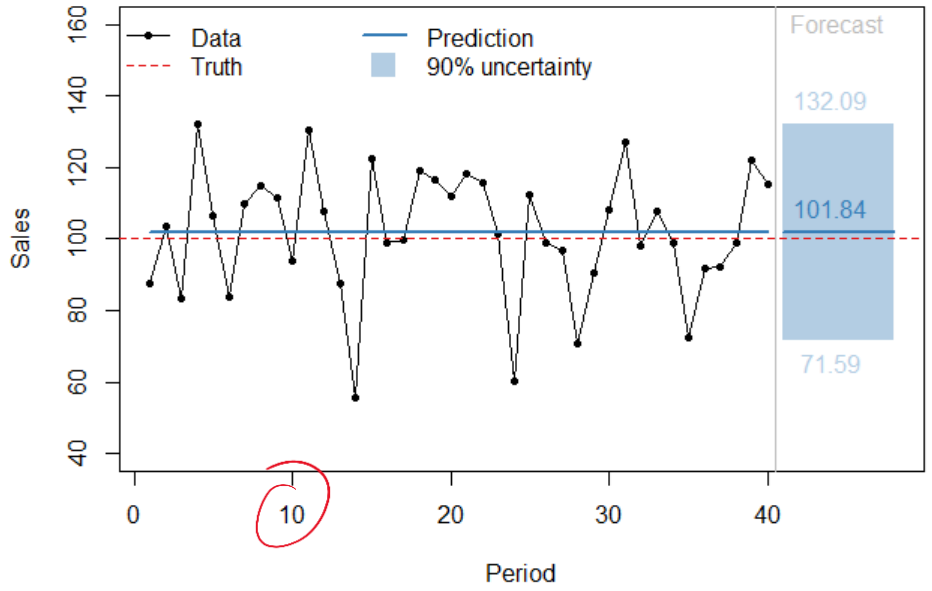
Parameter estimation uncertainty

Let's forecast ice cream sales: as it is well known ice-cream sales are independent of temperature/weather/price, so we have a demand with a constant mean of 100 and some randomness.

This is unknown!

- As they are independent of everything, the mean is the correct model. The only question is to estimate the correct mean (and variance).





Obviously more data, better estimation, But at a period t, how good is our estimation?

Why bother? The statistical argument

Specification uncertainty (+ parameter uncertainty = modelling uncertainty)

In the previous example we knew what was the correct model (constant demand + noise) but realistically this is unknown.

- The task of forecasting involves the selection of a method that approximates well the time series.
- "Well" means: as few terms as possible that still capture most sources of variance (e.g. seasonality, promotions, trends, etc.)

Needs estimation

Forecast = Constant (+ Error)

In this example we knew that all other terms are independent (0 weight)

Forecast = Constant + 0(seasonality) + 0(price) + 0(…) (+ Error)

What if we did not know that?

Forecast = Constant + a(seasonality) + b(price) + c(…) (+ Error)

The error depends on the model and the innovation term

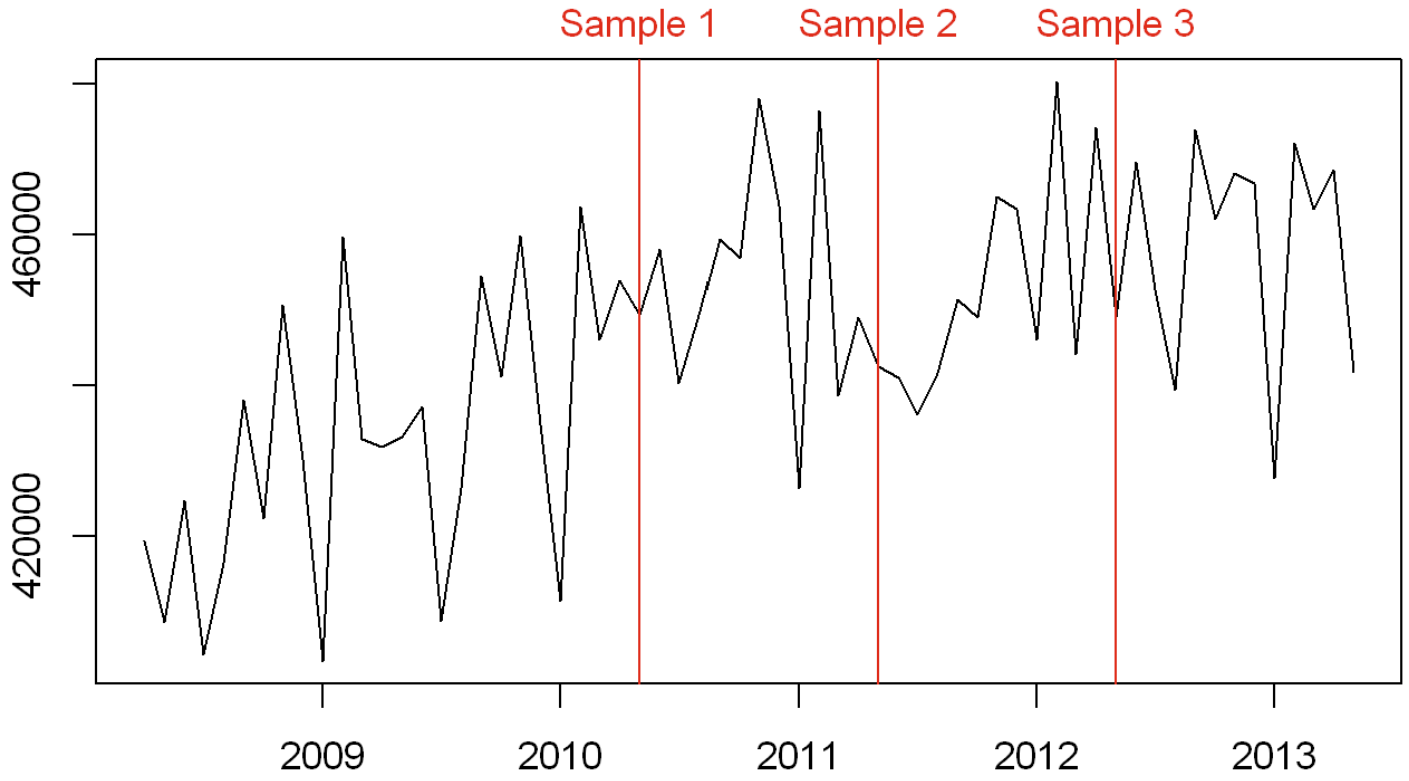
Needs estimation

…and all these add uncertainty

Why bother? The statistical argument

Fit your favourite model family on some data – is the fit consistent over samples?

E.g., choose the best exponential smoothing on Akaike Information Criterion



Sample 1: ETS)A, A, N)
Sample 2: ETS(A, N, A)
Sample 3: ETS(M, A, A)

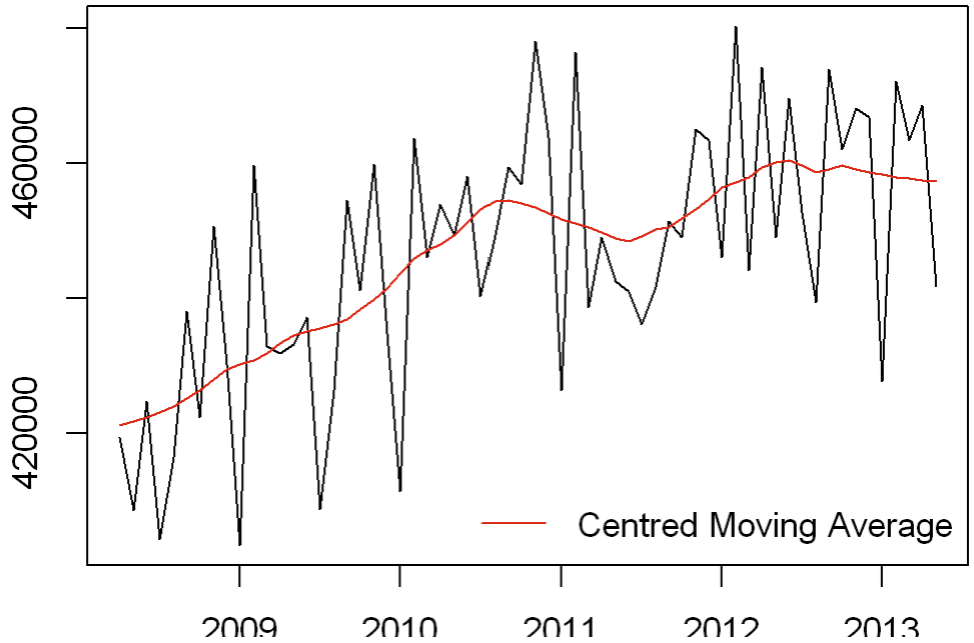
Thank you AIC for the clear answer!

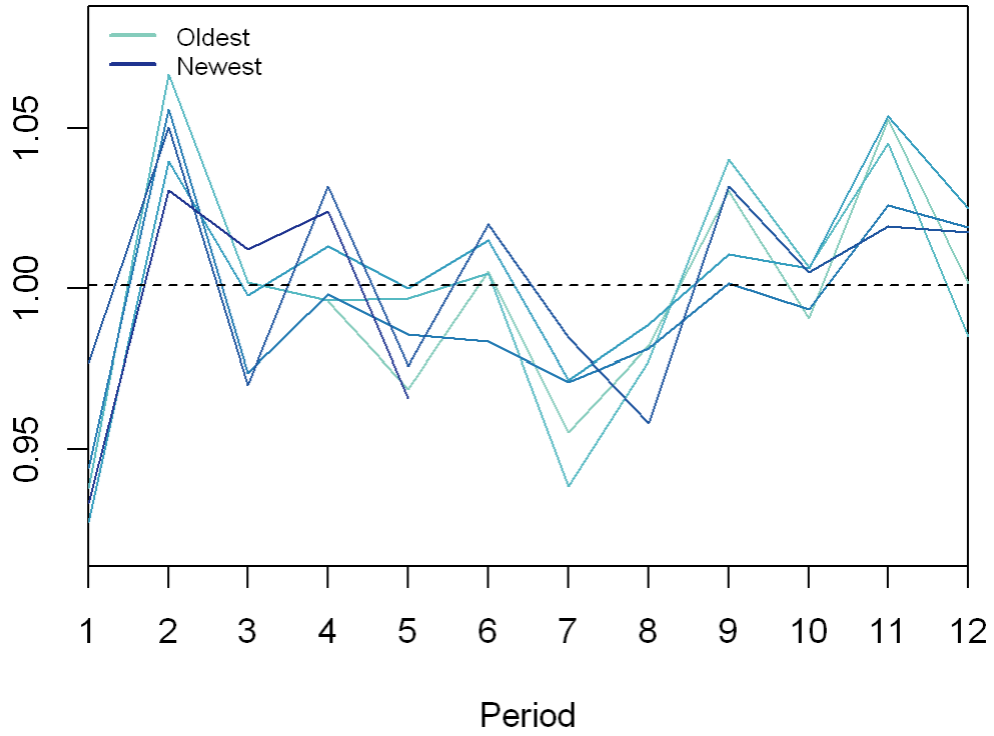
Why did this happen?

Uncertainty in the parameters or the identified components? (It is both, as the parameters are shaping the components)

Why bother? The statistical argument

Let's take a step back and do what we ask our students to do: explore the data!



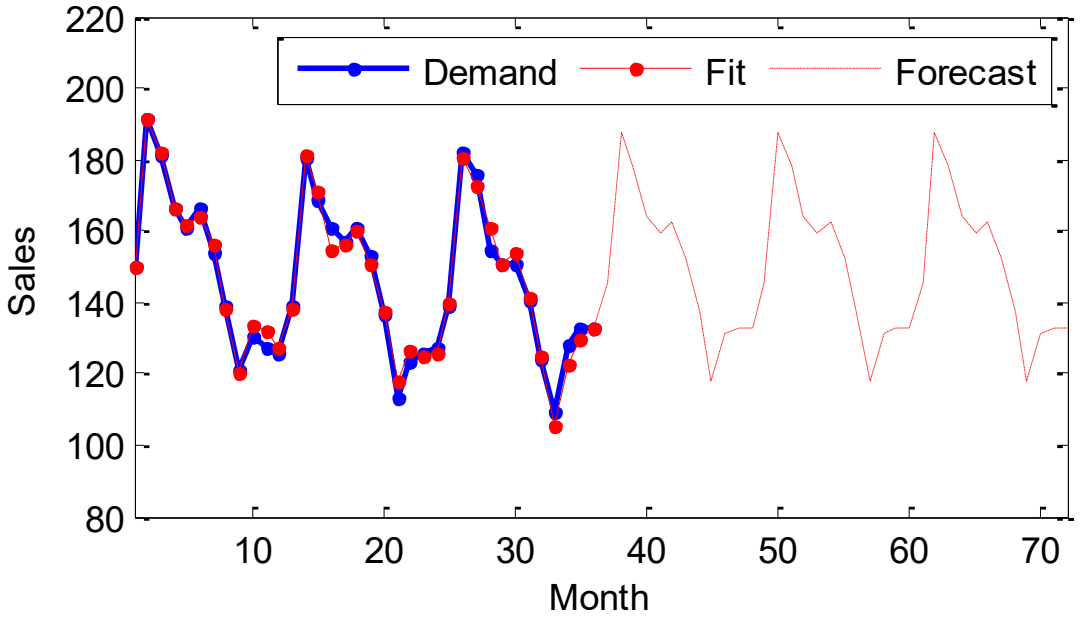


Or if you prefer, it is level and seasonal non-stationary.

I will argue that the time series has stochastic trend and stochastic multiplicative seasonality, so probably ETS(M, A, M)?

Why bother? The statistical argument

I used a “trick” in the previous slide, let me show you another example to spot the trick.



This is a simulated time series, so very easy and clean.

- How does the model fit look like?
- How does the forecast look like?
- Why?

Data Generating Process: ETS(A, **A**, A)

Identified Model: ETS(A, **N**, A)

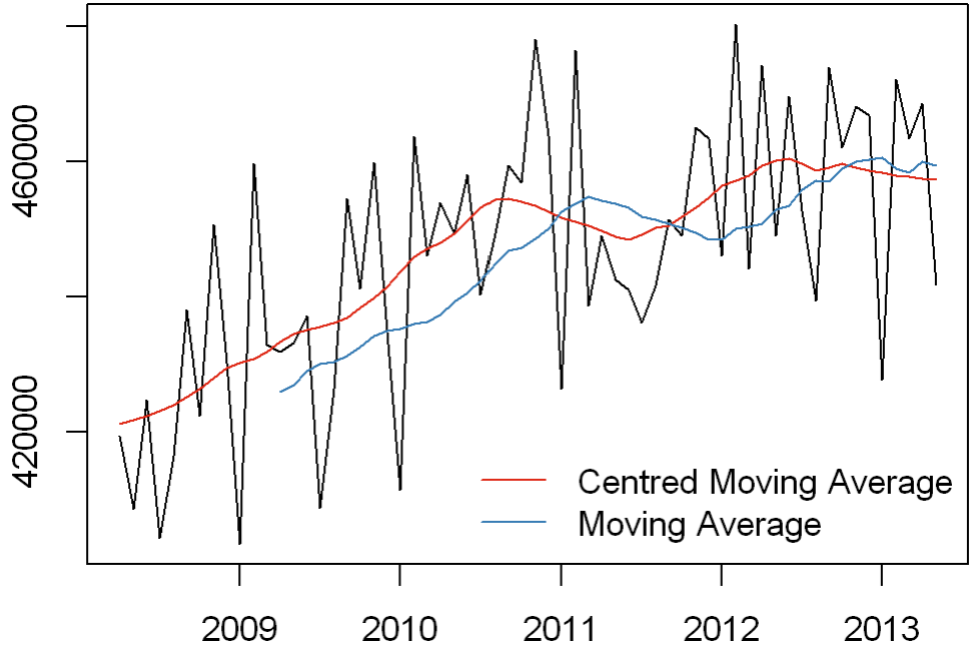
Overwhelmingly, the sample variance is due to the seasonality, a reasonable optimizer or a reasonable statistical methodology will prefer to not “believe” the limited variance due to the trend.

Why bother? The statistical argument

How could we ask the methodology to spot the limited variance component?

If the seasonality was not there the trend becomes apparent. There are a few ways to achieve that, but all of them are effectively filters.

- We could use a centred moving average to remove the seasonality. Issue: a centred moving average uses information from the future.
- We could use a moving average. Issue: it temporally shifts the data by (length of the average)/2.

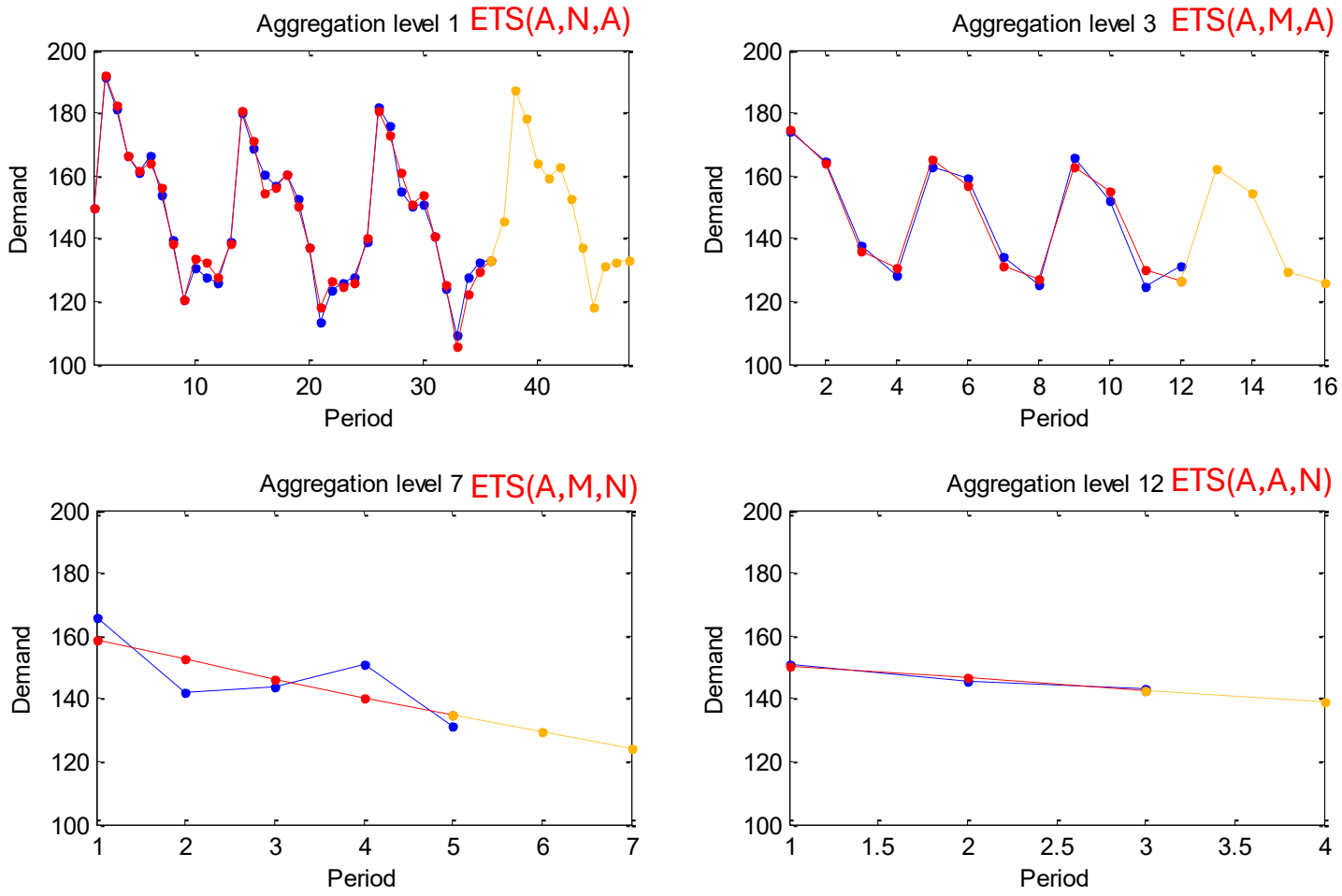


Why bother? The statistical argument

Temporal aggregation is a non-overlapping moving average (so we shift the averaging over k periods, every k periods)

A few things happen:

- High frequency information is filtered (seasonality, outliers, etc.)
- Low frequency information becomes more dominant (trend, cycles, etc.)
- We lose generous amount of sample size – but we have simpler models to estimate.



Why bother? The statistical argument

When you cannot decide between models/methods, you can always combine. Let's do that instead:

- Suppose we have two forecasts. we can simply take their average:

Forecast A = Constant (+ Error)

Forecast B = Constant + a(seasonality) + b(price) + c(…) (+ Error)

Combined = ½ Constant (+ ½ Error)

½Constant + ½a(seasonality) + ½b(price) + ½c(…) (+ ½ Error)

Terms go in by half
Shrinkage effect – smaller error

This is messier than it seems

Errors are distributions, so when we sum them, we need a covariance term: how similarly the two errors move. The total may be smaller or bigger than the two uncertainties on their own.

A helpful over-simplification:
Diverse forecasts combine well!

Why bother? The statistical argument

Forecast combinations are overwhelmingly considered to perform better that choosing a single forecast:

- Intuitively: choosing and weighting base forecasts optimally should lead to a better combined forecast.
- Empirically: a simple average works very well, even when some forecasts are rather silly.

This is the **combination puzzle**, i.e., our elegant combination papers are "potentially useless"!

Here is the intuition:

$$F_{combined,h} = b_0 + \sum_{i=1}^{k} b_i F_{i,h}$$

Estimate

Optional to remove bias

Combination weights

Estimate

Forecasts

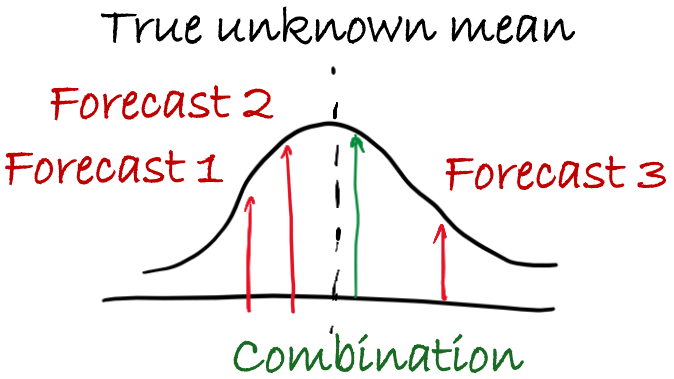
- The weights b_i and b_0 have estimation errors that contaminate the combination. Simple average (fixed weights) does not estimate the weights, so this source of error is not there!
  - **Anything we estimate adds errors!**

Combination sidenote

Forecast combinations is so 00s, do **forecast pooling** in 2025. Also make sure you read up of **stacking**.

If you are interested in understanding combinations quite well keep in mind that there are two equivalent interpretations:

- Variance minimization (from the econometric literature)
- Reducing uncertainty in point estimates (from the ML literature)



A few papers:

- Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754-762.
- Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226-235.
- Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting*, *39*(4), 1518-1547.

Why bother? The statistical argument

Forecasting with temporal hierarchies does a few things at the same time:

1. Uses temporal aggregation to filter high frequency information and noise in the series to help estimating lower frequency components.
2. Does forecast combination
3. Restricts the combination weights
4. Implicitly help us resolve the question of what forecasts to combine.

All these work together to resolve the model identification problem. This is what motivated the original paper in using multiple temporal aggregation levels:

- Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting*, *30*(2), 291-302.

Not the best writing ever, but this paper has a lot of good ideas in there, with many avenues still unexplored, if you are interested in the area (well you are in the workshop..) read it!

Why bother? The business argument

We do not forecast for the sake of forecasting ☹

So, we need to consider what decisions we support with our forecasts.

Example hierarchies of decisions in organisations.

Decision	Frequency	Time series	Output
	Call centre (Koole & Li, 2021)		
Budget planning	Quarterly	Monthly+	Budget
Capacity planning	Monthly	Weekly+	Training and hiring plans
Operational planning	Weekly	Weekly	Outsourced call volume
Scheduling	Weekly	Daily+	Agent schedules per type
Scheduling	Hourly	Intra-daily	Adaptations to schedules
	Tech manufacturer*		
Financial planning	Yearly	Quarterly+	High-level financial goals
Annual operations plan	Yearly	Monthly+	Resource allocation
Production planning	Monthly	Monthly	Aggregate demand planning
Master production plan	Weekly	Weekly	Detailed demand planning
Material planning	Weekly	Weekly	Supply requirements

From Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. *International Journal of Forecasting*, *39*(4), 1502-1511.

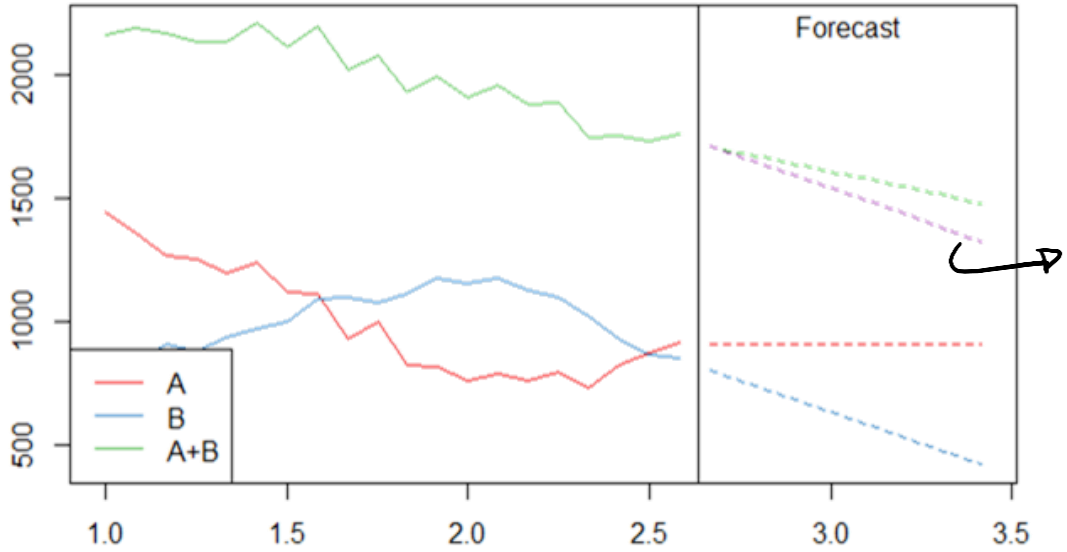
Different planning horizons need the forecasts to agree – to be coherent

Why bother? The business argument

As we aggregate data, some structures become more prominent (e.g., trends, seasonality), while others become less obvious (e.g., promotional activity) and noise is filtered.

Although all series are based on the same information, this does not mean that the same information is useable → different models/parameters/forecasts.

Example: forecasting A and B separately or forecasting their sum does not lead to the same result!



F(A+B) and F(A)+F(B) will typically be different; we need to impose equality (coherency of forecasts).

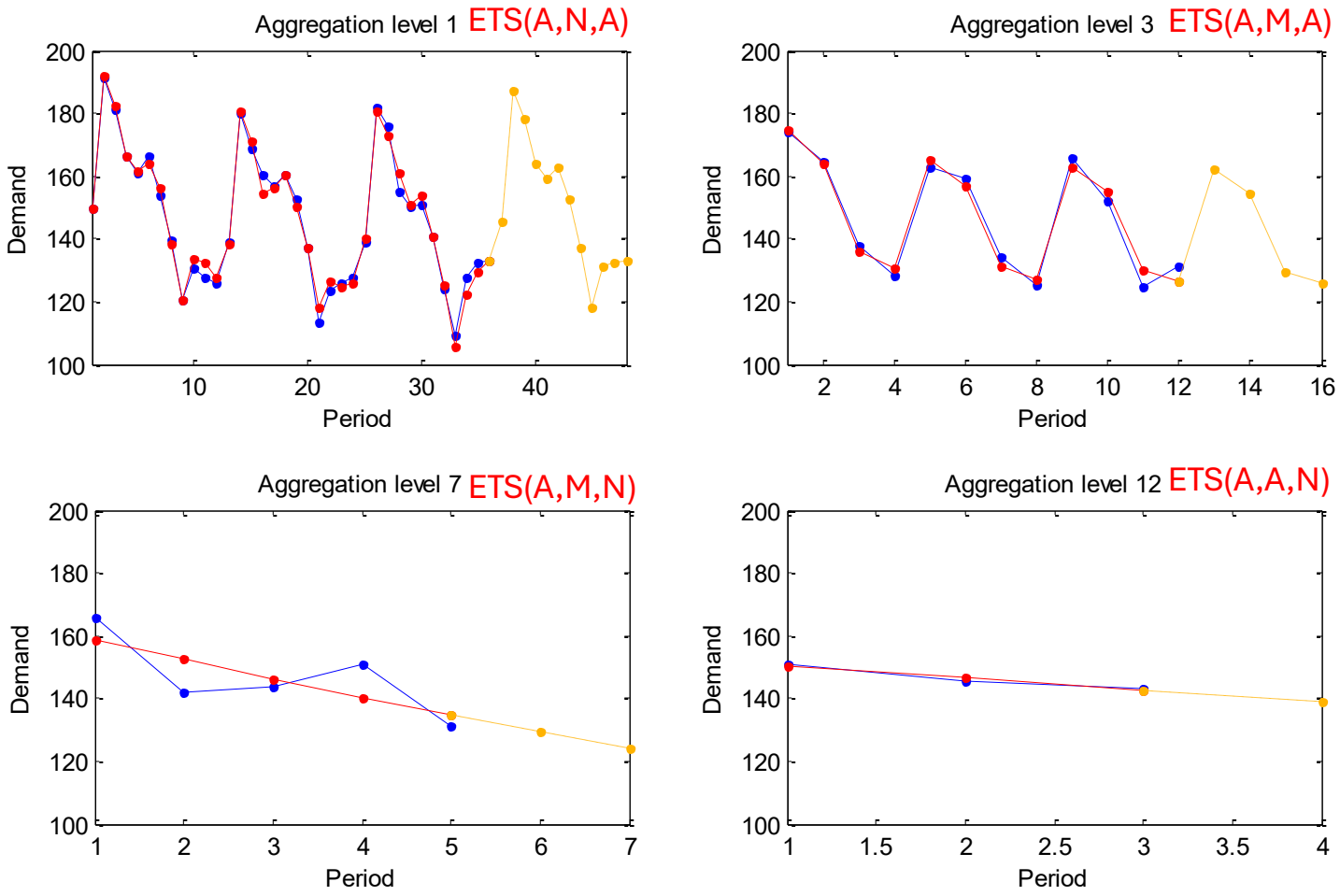
F(A+B) or F(A)+F(B) is correct? Coherency avoids this question

What we will talk about

1. ~~Why bother with temporal hierarchies?~~
2. The theory
3. Applications & observations
4. Newer results

Multiple Aggregation Prediction Algorithm (a.k.a. v1)

Remember this?



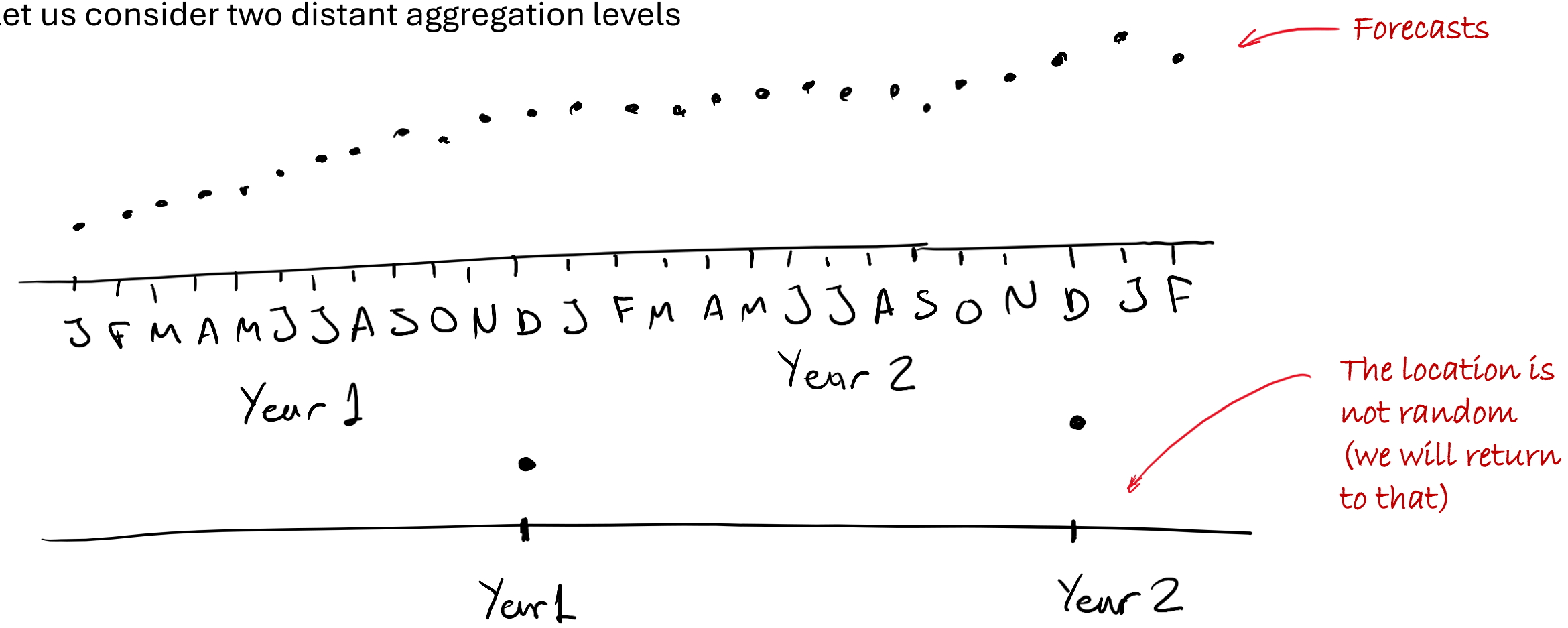
Now combine these!

Two issues:

- Data frequency
- Shrinkage

Multiple Aggregation Prediction Algorithm

Let us consider two distant aggregation levels



Given that we want to be able to forecast at the monthly frequency (as well?), we need to interpolate the yearly forecast.

Multiple Aggregation Prediction Algorithm

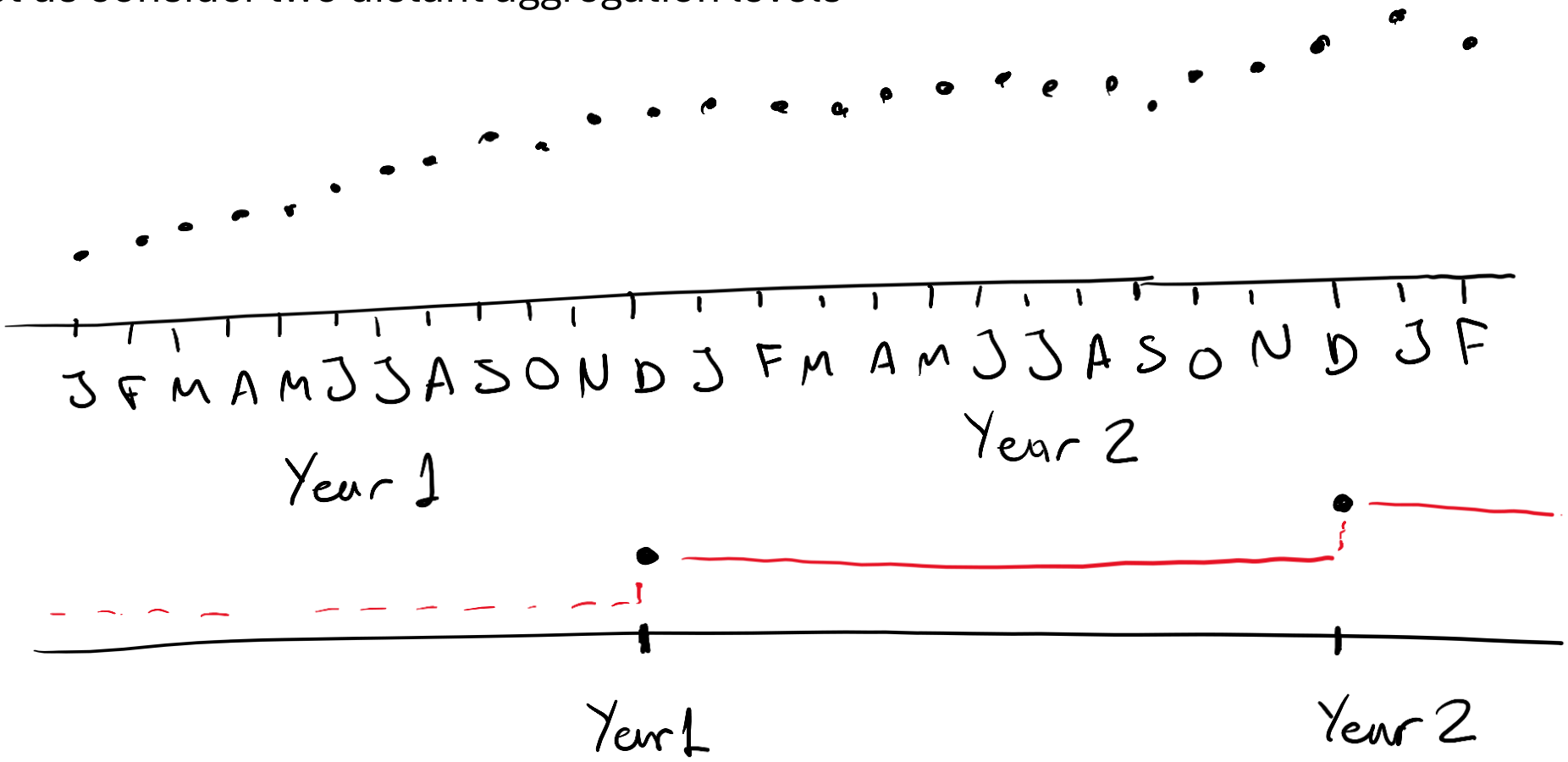
If we want to go fancy, we would consider linear, spline, or other exotic interpolations. However, we are now imposing a model on our data even before we start modelling. That is not a great idea.

Instead, we take an information argument: The forecasted points in the more aggregate levels contain all the best available information up to that period. When the information is updated, then we get a new forecast. Therefore, to interpolate, we **only replicate the same value**.

(We will return to this and show that this is actually not a random choice and comes out of the equations naturally.)

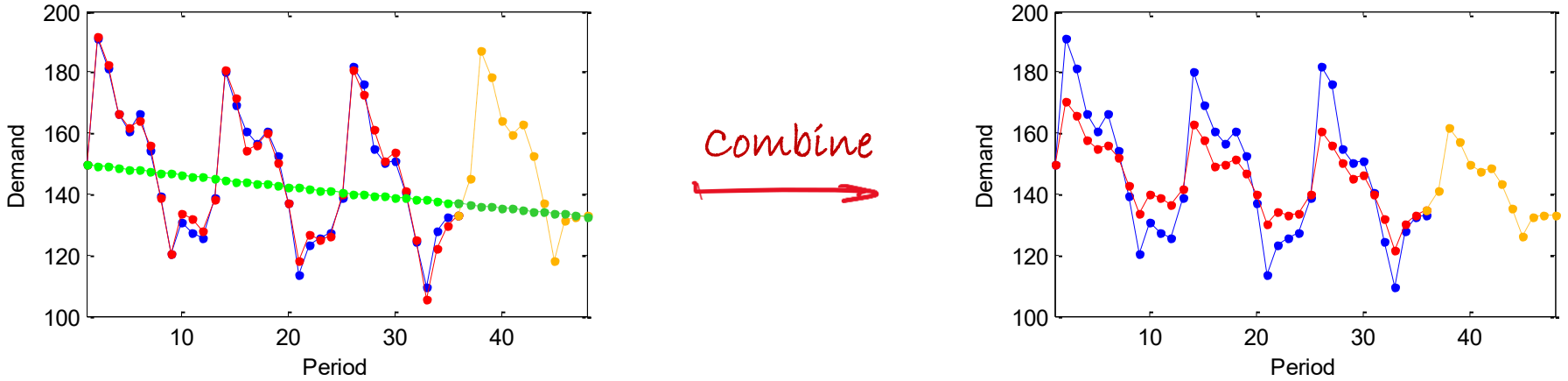
Multiple Aggregation Prediction Algorithm

Let us consider two distant aggregation levels



Multiple Aggregation Prediction Algorithm

Onwards to the shrinkage “issue”



So, straightforward combination will not work, as the lower frequency/more aggregate forecasts contain less of the high frequency information (such as seasonality), so the total seasonal information in the forecast is shrunk towards zero.

We know this behaviour is incorrect as it is an artifact of how we populated our pool of forecasts to be combined.

Multiple Aggregation Prediction Algorithm

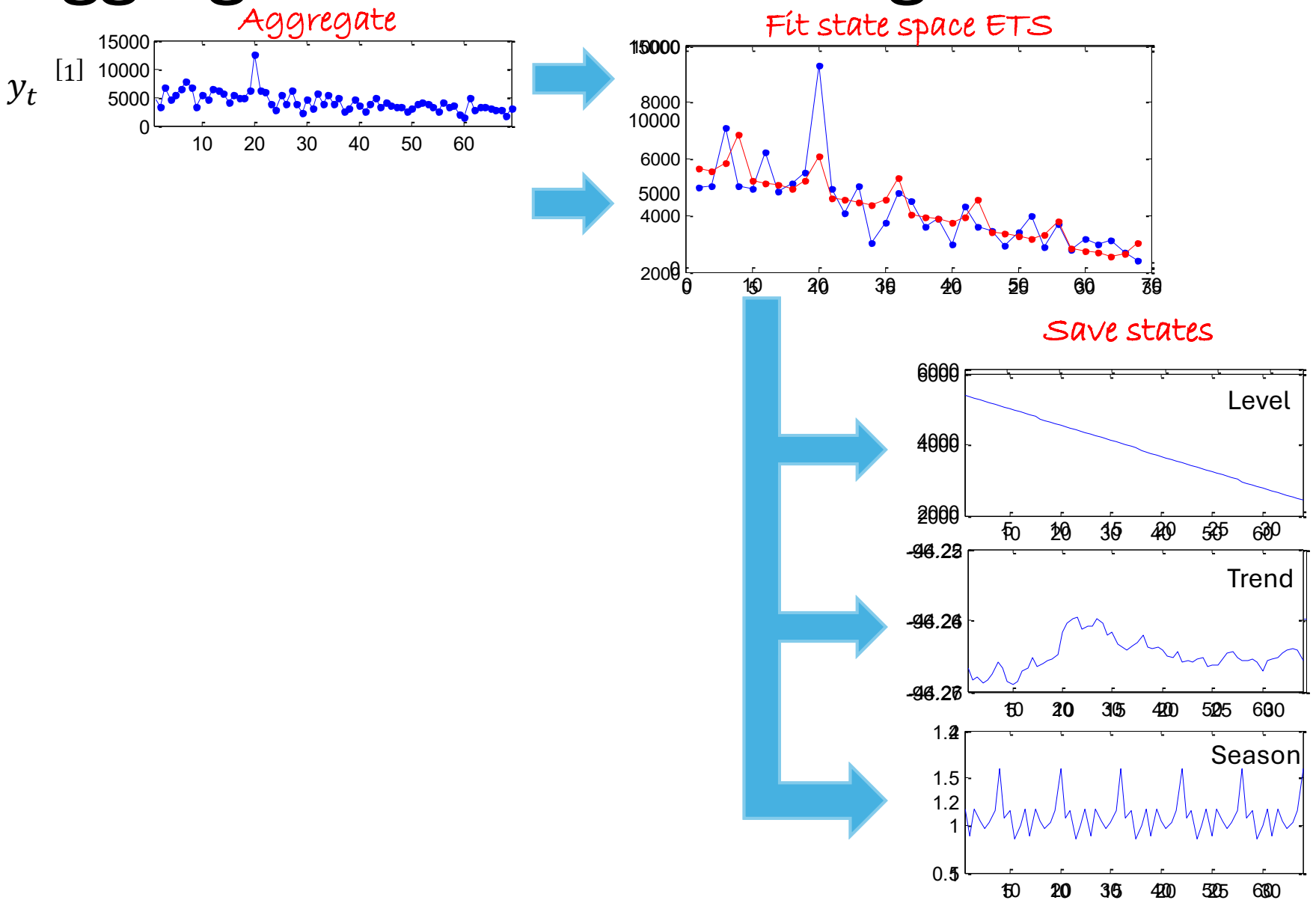
The MAPA algorithm in a nutshell:

Because most information is already filtered

1. Aggregate your original time series in up to the yearly level (and beyond!)
2. Fit a state-space model at each aggregation level independently
3. Extract states and “expand” them to the original sampling frequency by repeating the values as needed. (So, no fancy interpolation.)
4. Combine models by states, not by forecasts! When a state is not present, use zeros. When a state cannot be present, do not combine for that level. (Mititgates the shrinkage issue.)
5. …
6. Profit!

Multiple Aggregation Prediction Algorithm

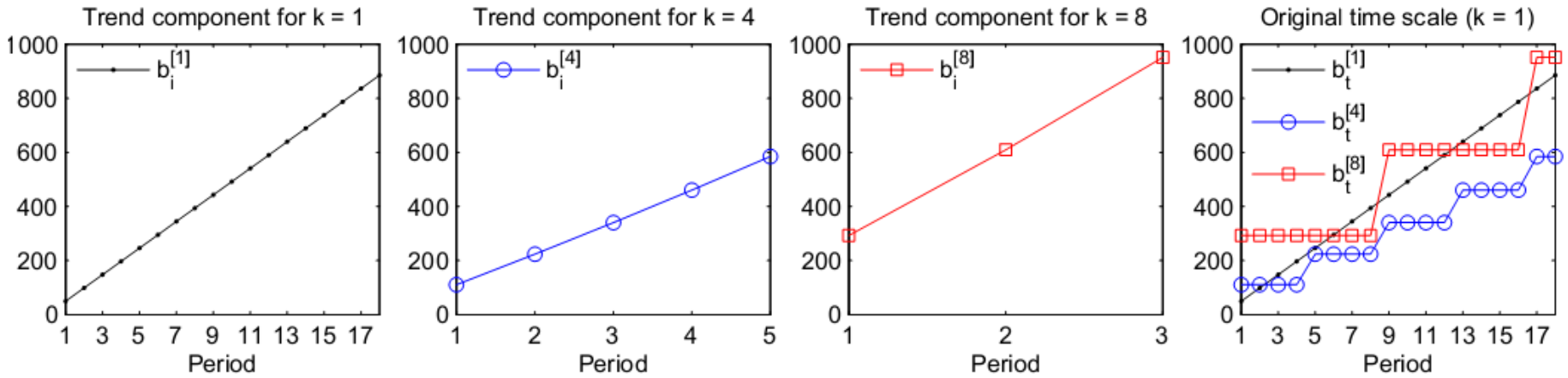
(Part 1)



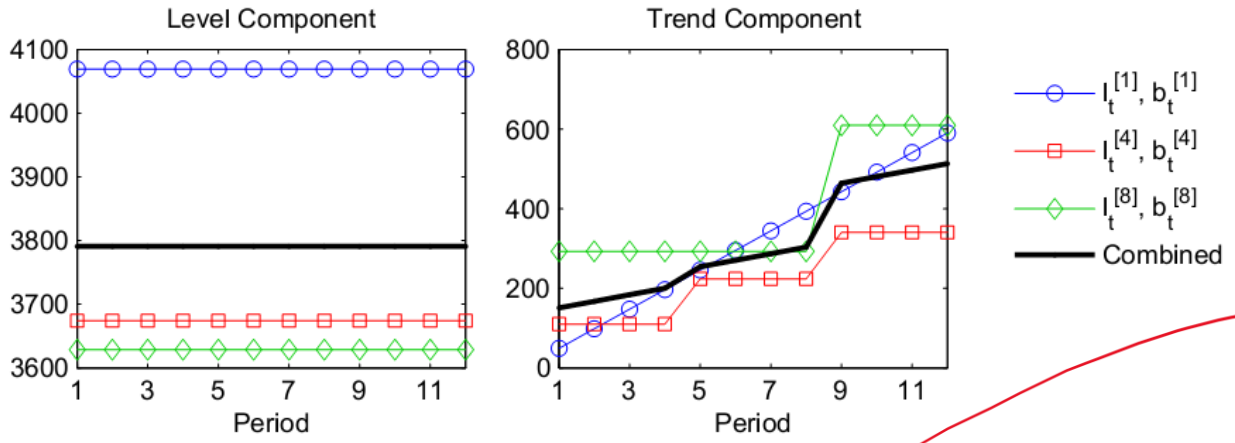
Multiple Aggregation Prediction Algorithm

(Part 2)

Transform states to additive and to original sampling frequency



Combine states (components)

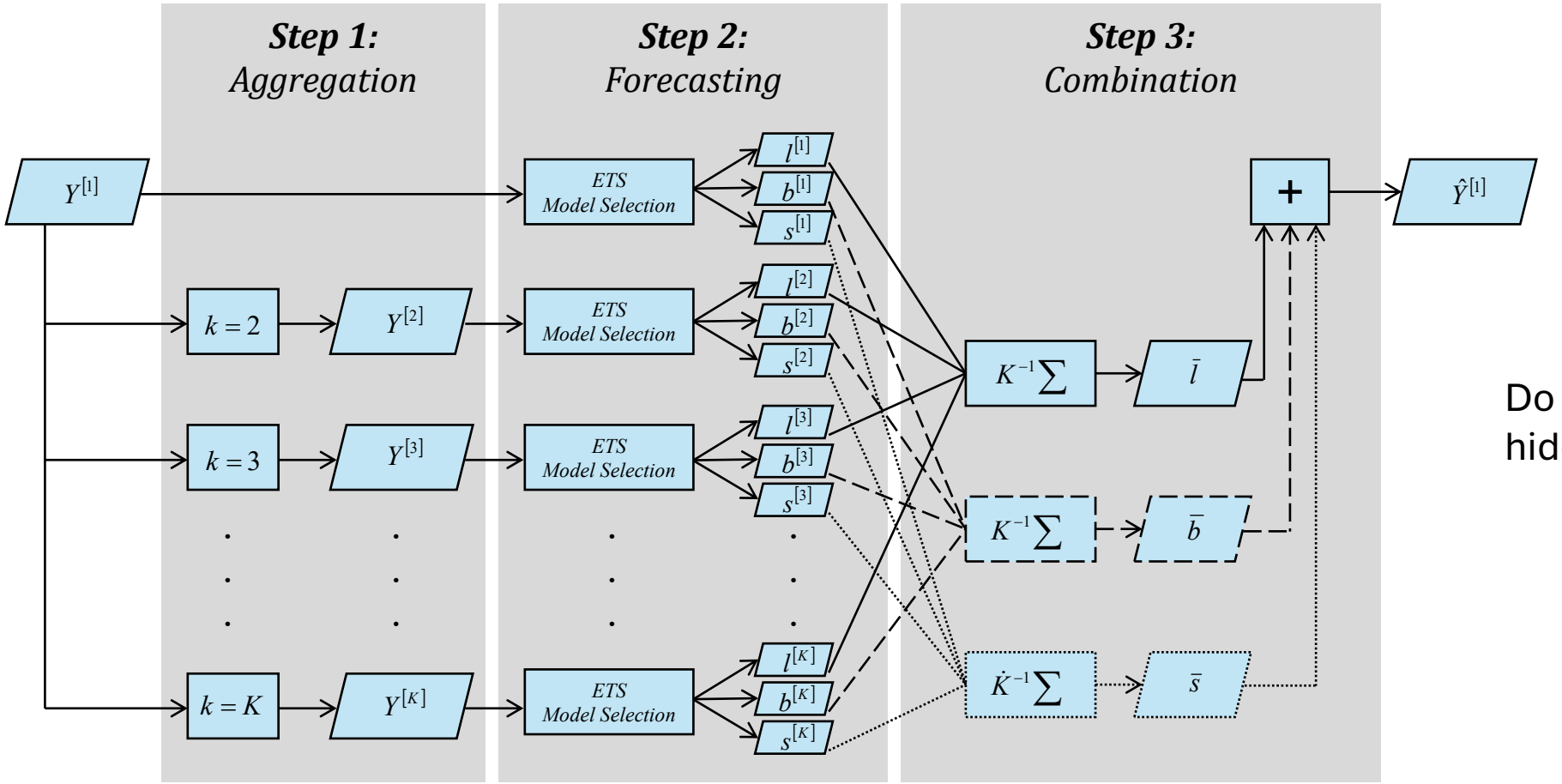


We explored mean and median, but in 2025 use whatever you prefer.

In Greek MAPA is slang for "rubbish", so indeed we are saying here is your rubbish forecast!

Finally, add the combined states together and you get your MAPA forecast.

Multiple Aggregation Prediction Algorithm



Strengthens and attenuates components

Estimation of parameters at multiple levels

Robustness on model selection and parameterisation

Do you see the hierarchy hidden in there?

Multiple Aggregation Prediction Algorithm

The pros:

- Very good forecasting performance – was the first paper after 14 years beating the M3 winning forecasting methods.
- Good long-term forecast accuracy – why?

  We include temporally aggregate forecasts in the combination pool. These model better the low-frequency components and also reduce the number of steps ahead required for long-term forecasts (less accumulation of forecast errors.)

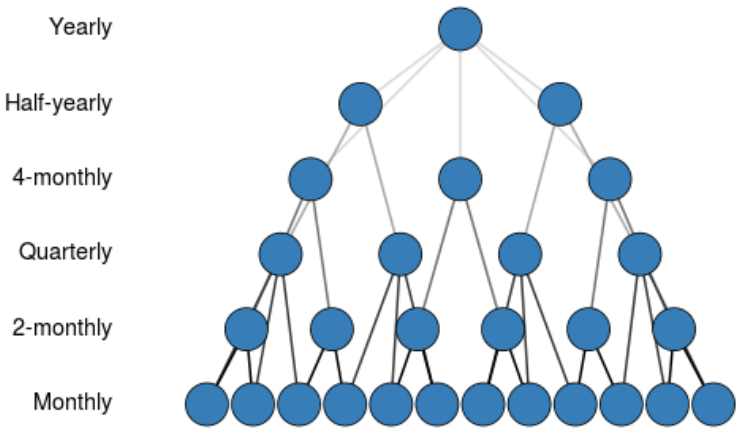
The cons:

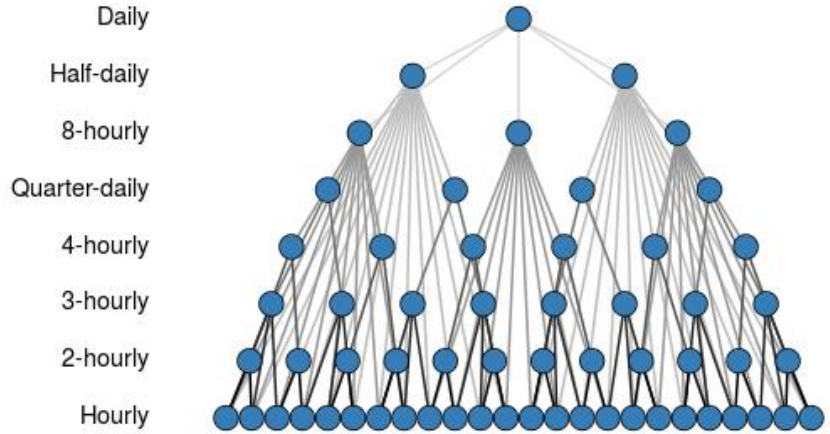
- Very ad-hoc! We did not fully understand what we were doing ☺
- Restricted to state-space models.
- A more subtle issue. Assume monthly data, and aggregate them every 5 months. With MAPA and exponential smoothing we will assume that there is no seasonality there and use only the other states. But there is seasonality, it is just fractional. Have fun modelling that!

The leftover information contaminates the forecasts. An "Easy" fix is to use some ML state space model, where fractional seasonalities are trivial to model.

Towards Forecasting with Temporal Hierarchies

We are using non-overlapping temporal aggregation, so we can easily visualize each aggregation from the original sampling frequency to the most aggregate as a hierarchy



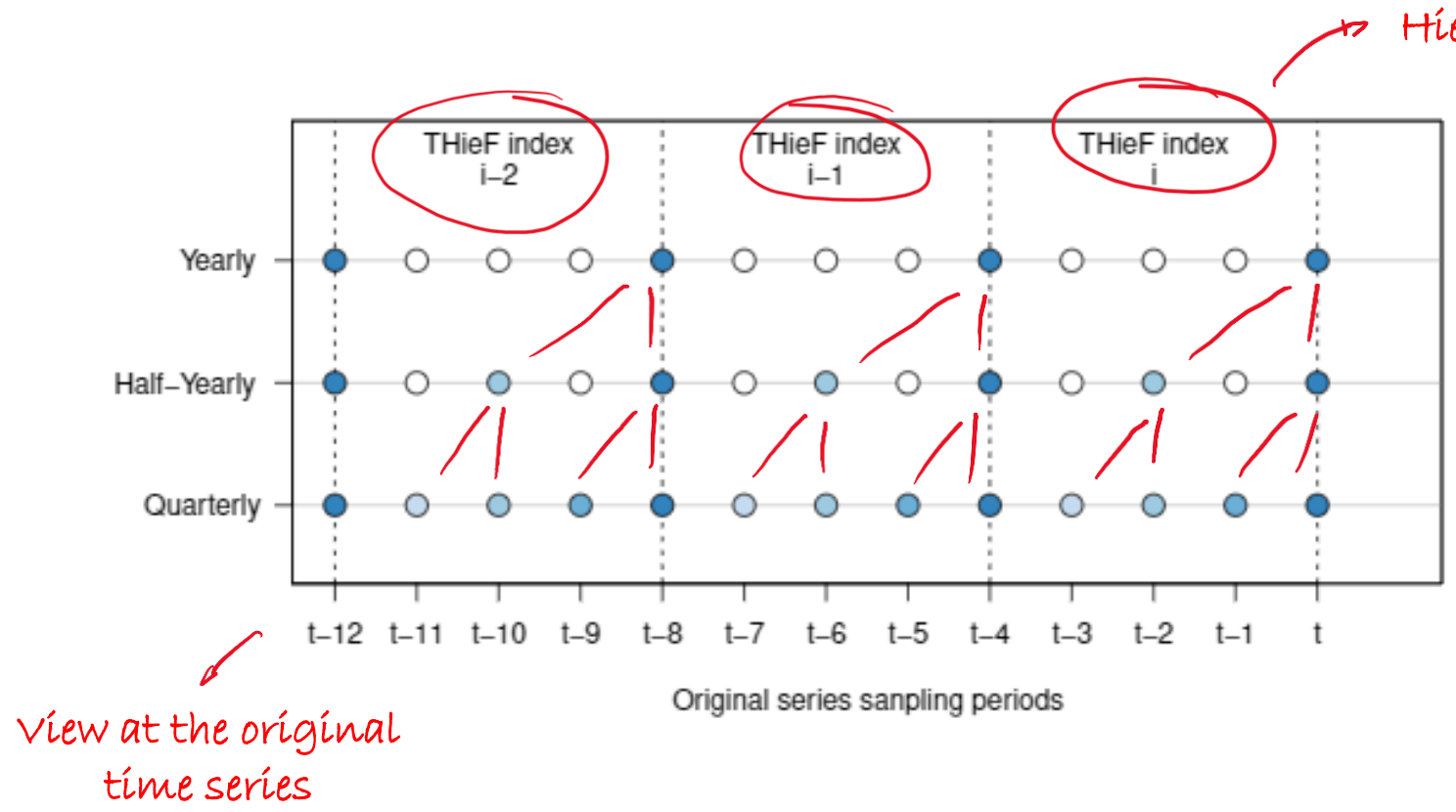


Observe that some possible aggregation levels are missing:

- What is the maximum? Annual makes some sense, because most information is already filtered at that level, but t is not necessary to reach or stop there.
- The in-between levels that would result in fractional seasonality (e.g., aggregate every five months) are avoided. We do not have to, but if we would include these, we would need to have models that can deal with that and also expand the hierarchy so that there is at least one level that everything connects to (not necessarily the top level!)

Towards Forecasting with Temporal Hierarchies

The hierarchy figure is somewhat misleading. Each level runs ar a different time scale, but are they on common time periods (if we would write then, for instance, at the lowest level frequency)?



White dots are empty – no observations exist there.

- They fall in-between the sampling frequency at that level.
- Also note that the hierarchy is not “centred”, we always use the most up-to-date information.

Forecasting with Temporal Hierarchies (THieF)

Let us look at the formulas to get a deeper understanding of the structure of the data

Non-overlapping temporal aggregation

Aggregation level – sum up every k observations

$$y_j^{[k]} = T(y_t, j, k) = \sum_{t=t^*+(j-1)k}^{t^*+jk-1} y_t$$

Observation at period t

Aggregate time series

"Lives" on j periods time index

$$t^* = n - \lfloor n/m \rfloor m + 1$$

Removes some observations from the beginning to ensure we can sum up in complete buckets of k observations

If we were to divide by k, we have a (non-overlapping) moving average. (MAPA does that to simplify the combination.)

Forecasting with Temporal Hierarchies (THieF)

Let us look at the formulas to get a deeper understanding of the structure of the data

We collect all aggregate observations within one period i – the time index at the top level.

$$\boldsymbol{y}_i^{[k]} = \left(y_{m^{[k]}(i-1)+1}^{[k]}, \dots, y_{m^{[k]}i}^{[k]}\right)'$$

We collect all time series together:

$$\boldsymbol{y}_i = \left(y_i^{[m]}, \dots, \boldsymbol{y}_i^{[1]'}\right)'$$

Or we can write down a summing matrix $\boldsymbol{S}$ that maps the original series to all aggregation levels:

$$\boldsymbol{y}_i = \boldsymbol{S}\boldsymbol{y}_i^{[1]}$$

Forecasting with Temporal Hierarchies (THieF)

For example, for quarterly data this looks like:

Which quarter to use

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \} \text{ Yearly} \\ \} \\ \} \text{ Half-yearly} \\ \} \\ \} \\ \} \\ \} \text{ Quarterly} \end{matrix}$$

What aggregation you get

No real reason for the elements of S to be only {0, 1}

Forecasting with Temporal Hierarchies (THieF)

Observationally $\boldsymbol{y}_i = \boldsymbol{S}\boldsymbol{y}_i^{[1]}$ will hold. It just reads as ”aggregate your data from the bottom level”.

There may be cases that we collect the data at the aggregate levels directly, e.g., at a statistics bureau, or they may be forecasts for each level. In that case we cannot assume that the aggregation holds.

$$\boldsymbol{y}_i = \boldsymbol{S}\boldsymbol{y}_i^{[1]} + \delta$$

A slack term for all measurement, estimation, modelling errors across levels.

For more details on this “slack term” see: Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. International Journal of Production Economics, 240, 108221 – more to explore towards that direcrion!

Forecasting with Temporal Hierarchies (THieF)

For forecasts, we can use the formulation of hierarchical forecasting, imposing coherency:

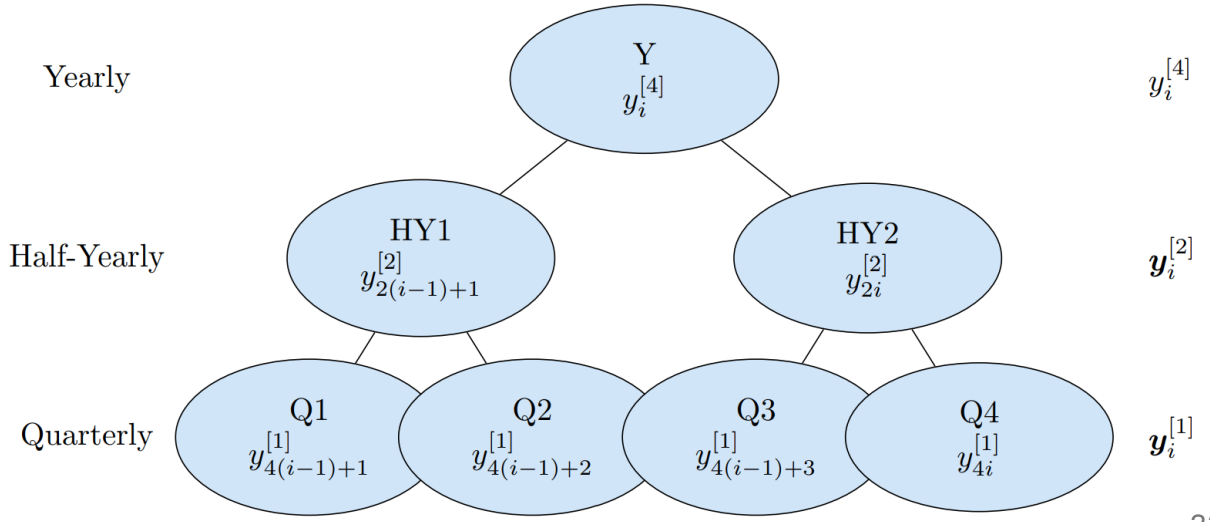
$$\tilde{\mathbf{y}}_i = \mathbf{S}\mathbf{G}\hat{\mathbf{y}}_i$$

Coherent forecasts for all levels

Combination weights

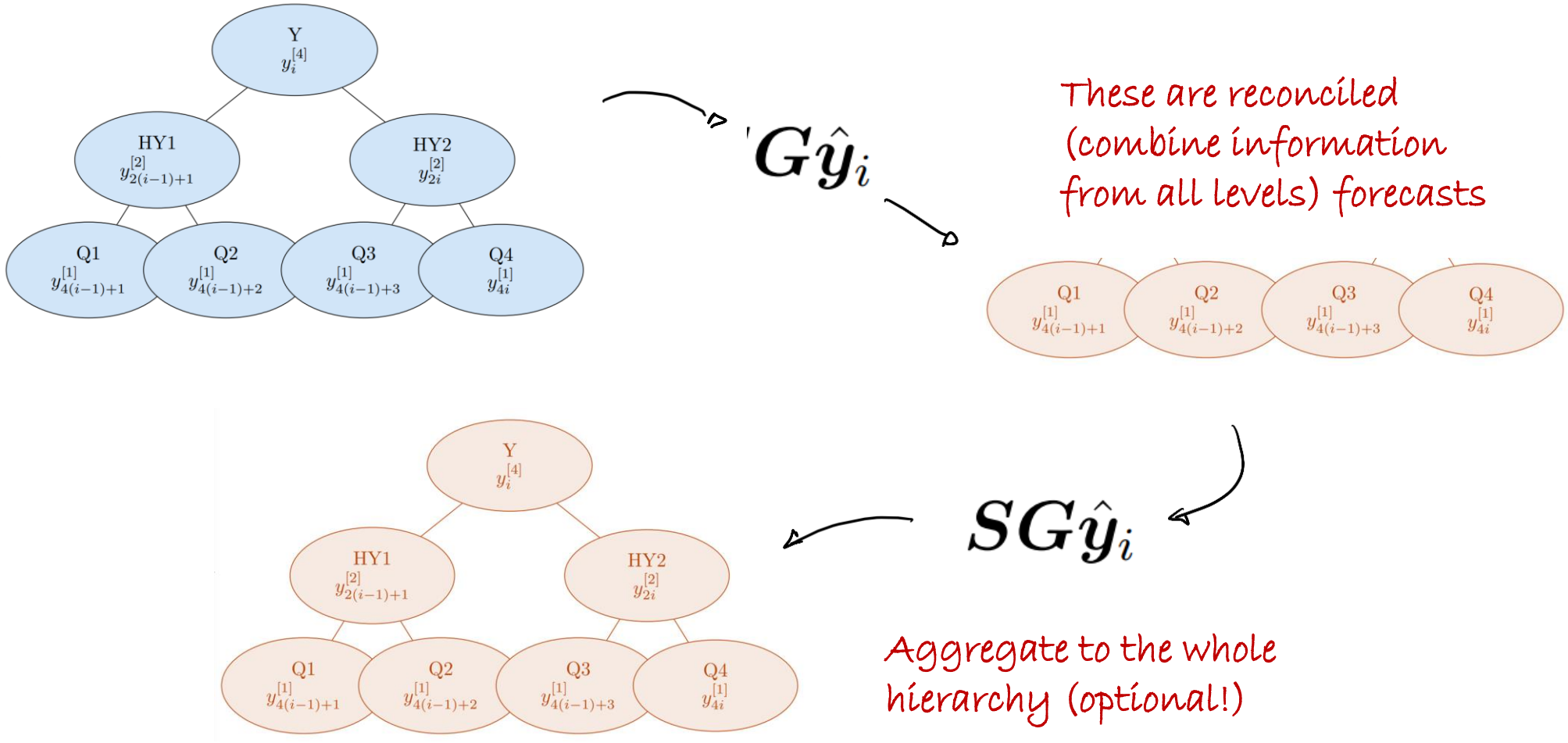
Forecasts across all levels for the i-th temporal hierarchy

$$\hat{\mathbf{y}}_i = \begin{bmatrix} Y \\ HY1 \\ HY2 \\ Q_1 \\ \vdots \\ Q_4 \end{bmatrix}$$



Forecasting with Temporal Hierarchies (THieF)

Visually, we have:



Forecasting with Temporal Hierarchies (THieF)

Let us look at that $\boldsymbol{G}$ matrix

$$\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}^{-1}$$

This is an approximation of (or the exact!) the covariance matrix of the forecast errors of the base forecasts.

- This minimizes the trace $tr[\boldsymbol{SGWG}^T\boldsymbol{S}^T]$ that corresponds to the covariance matrix of the reconciled forecast errors.

- This is subject to $\boldsymbol{SGS} = \boldsymbol{S}$, which in human-speak says, give me unbiased (base) forecasts and I will return to you unbiased (reconciled) forecasts.

Forecasting with Temporal Hierarchies (THieF)

Let us look at some approximations for $\boldsymbol{W}$ (keeping the example of quarterly data).

Structural scaling

$$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Assumes proportional increase in variance and independence across observations,

Variance scaling

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}^2_{X} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}^2_{Y} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}^2_{XX} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}^2_{XY} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YX} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{\sigma}^2_{YY} \end{bmatrix}$$

Assumes ~~proportional increase in variance and~~ independence across observations,

MinT shrinkage ($\rho_{i,j} \to 0$)

$$\begin{bmatrix} \hat{\sigma}^2_{Tot} & \hat{\rho}_{Tot,X} & \hat{\rho}_{Tot,Y} & \hat{\rho}_{Tot,XX} & \hat{\rho}_{Tot,XY} & \hat{\rho}_{Tot,YX} & \hat{\rho}_{Tot,YY} \\ \hat{\rho}_{X,Tot} & \hat{\sigma}^2_{X} & \hat{\rho}_{X,Y} & \hat{\rho}_{X,XX} & \hat{\rho}_{X,XY} & \hat{\rho}_{X,YX} & \hat{\rho}_{X,YY} \\ \hat{\rho}_{Y,Tot} & \hat{\rho}_{Y,X} & \hat{\sigma}^2_{Y} & \hat{\rho}_{Y,XX} & \hat{\rho}_{Y,XY} & \hat{\rho}_{Y,YX} & \hat{\rho}_{Y,YY} \\ \hat{\rho}_{XX,Tot} & \hat{\rho}_{XX,X} & \hat{\rho}_{XX,Y} & \hat{\sigma}^2_{XX} & \hat{\rho}_{XX,XY} & \hat{\rho}_{XX,YX} & \hat{\rho}_{XX,YY} \\ \hat{\rho}_{XY,Tot} & \hat{\rho}_{XY,X} & \hat{\rho}_{XY,Y} & \hat{\rho}_{XY,XX} & \hat{\sigma}^2_{XY} & \hat{\rho}_{XY,YX} & \hat{\rho}_{XY,YY} \\ \hat{\rho}_{YX,Tot} & \hat{\rho}_{YX,X} & \hat{\rho}_{YX,Y} & \hat{\rho}_{YX,XX} & \hat{\rho}_{YX,XY} & \hat{\sigma}^2_{YX} & \hat{\rho}_{YX,YY} \\ \hat{\rho}_{YY,Tot} & \hat{\rho}_{YY,X} & \hat{\rho}_{YY,Y} & \hat{\rho}_{YY,XX} & \hat{\rho}_{YY,XY} & \hat{\rho}_{YY,YX} & \hat{\sigma}^2_{YY} \end{bmatrix}$$

Assumes ~~proportional increase in variance and independence across observations,~~

Forecasting with Temporal Hierarchies (THieF)

Okay, but what do these mean?

For the structural scaling approximation and quarterly data we would get:

Combination weights

Output for

$$\boldsymbol{G}_{\text{Str}} = \begin{bmatrix} 0.083 & 0.208 & -0.042 & 0.708 & -0.292 & -0.042 & -0.042 \\ 0.083 & 0.208 & -0.042 & -0.292 & 0.708 & -0.042 & -0.042 \\ 0.083 & -0.042 & 0.208 & -0.042 & -0.042 & 0.708 & -0.292 \\ 0.083 & -0.042 & 0.208 & -0.042 & -0.042 & -0.292 & 0.708 \end{bmatrix} \begin{matrix} \} \text{Q1} \\ \} \text{Q2} \\ \} \text{Q3} \\ \} \text{Q4} \end{matrix}$$

$$\underbrace{}_{Y} \; \underbrace{}_{HY1} \; \underbrace{}_{HY2} \; \underbrace{}_{Q1} \; \underbrace{}_{Q2} \; \underbrace{}_{Q3} \; \underbrace{}_{Q4}$$

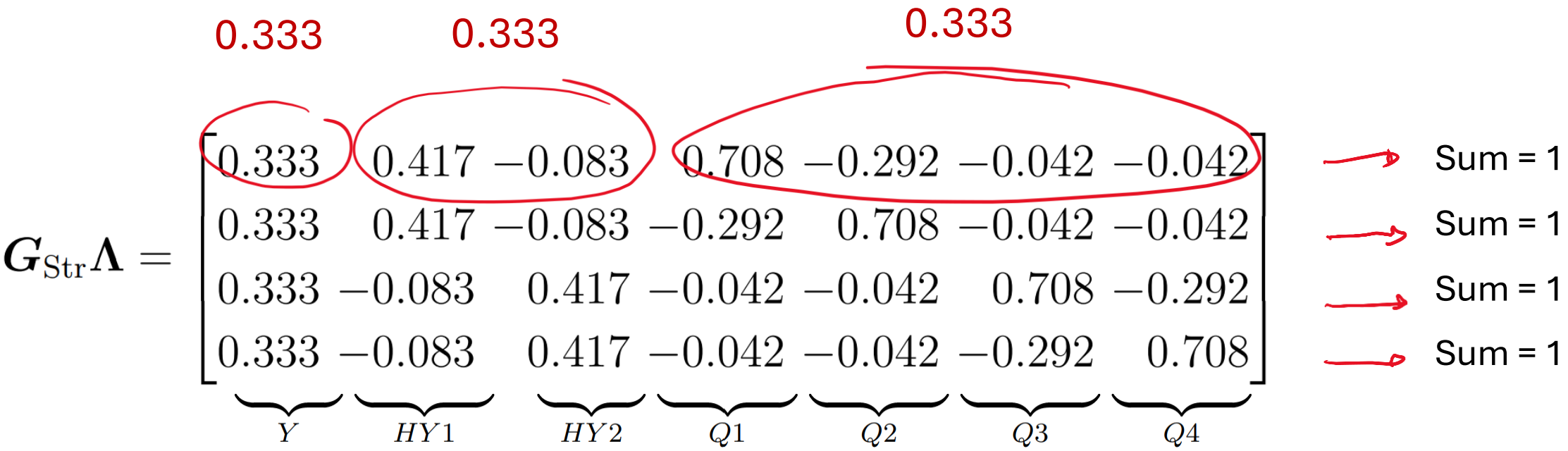
Input from

The reconciled forecast for Q1 is the combination of the base:

0.083 Y + 0.208 HY1 -0.042 HY2 + 0.708 Q1 – 0.282 Q2 – 0.042 Q3 – 0.042 Q4

Forecasting with Temporal Hierarchies (THieF)

Remember: due to the temporal aggregation we have differences in scale across levels of the hierarchy. We can rescale everything.



So, we are equally combining each level of the hierarchy into the reconciled forecast and apply a kind of temporal kernel on top of the target quarter.

(If you squint a lot, you will see the equal combination from MAPA.)

Forecasting with Temporal Hierarchies (THieF)

There are many options for estimating $\boldsymbol{W}$, giving us different combination weights

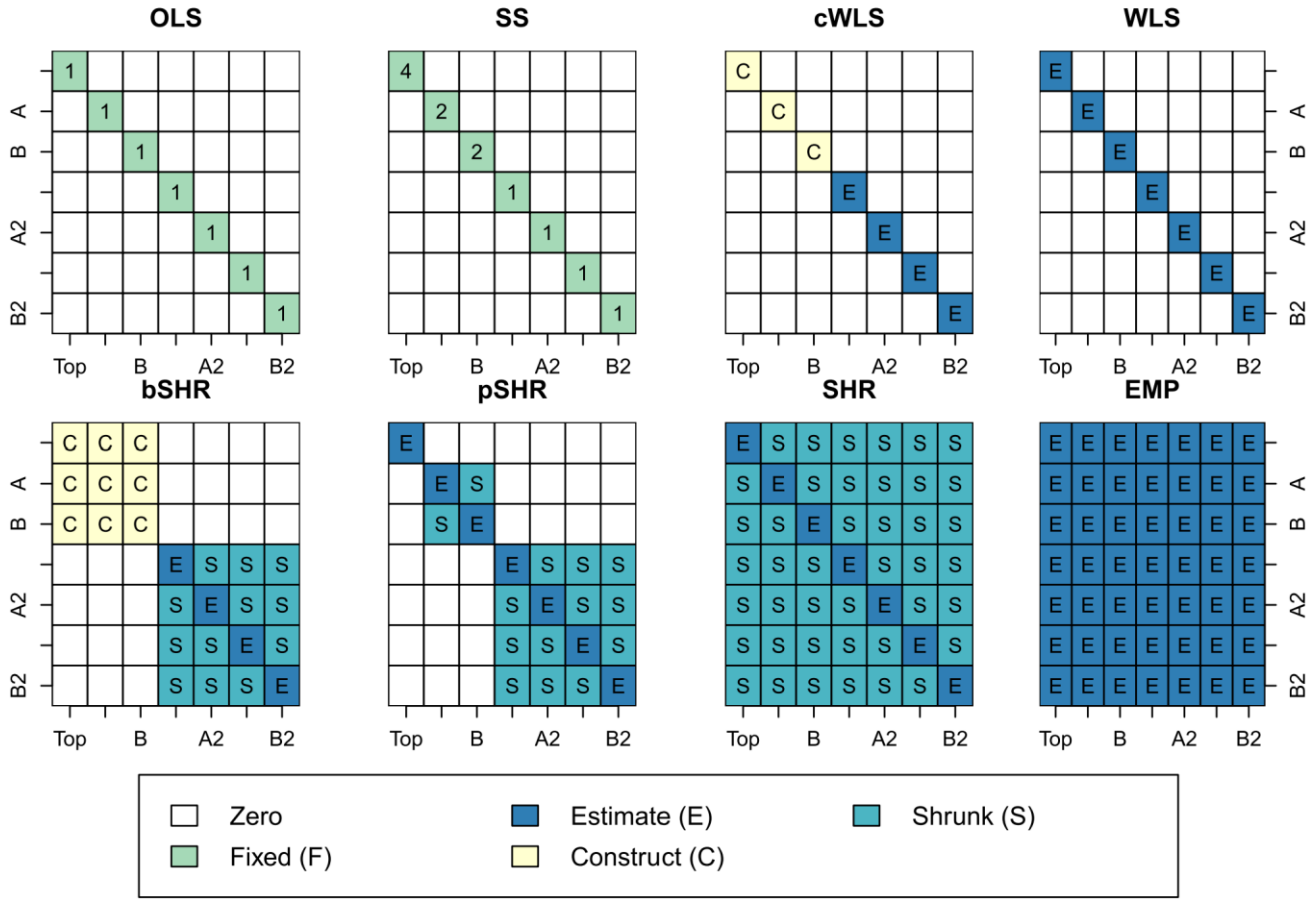
- Structural may look as if it has too many assumptions, but it "is" the equal weight combination. Nothing to estimate → remember the combination puzzle.
- Variance scaling corrects for scale as estimated by the variances, so it weights less forecasts with disproportionally large errors.

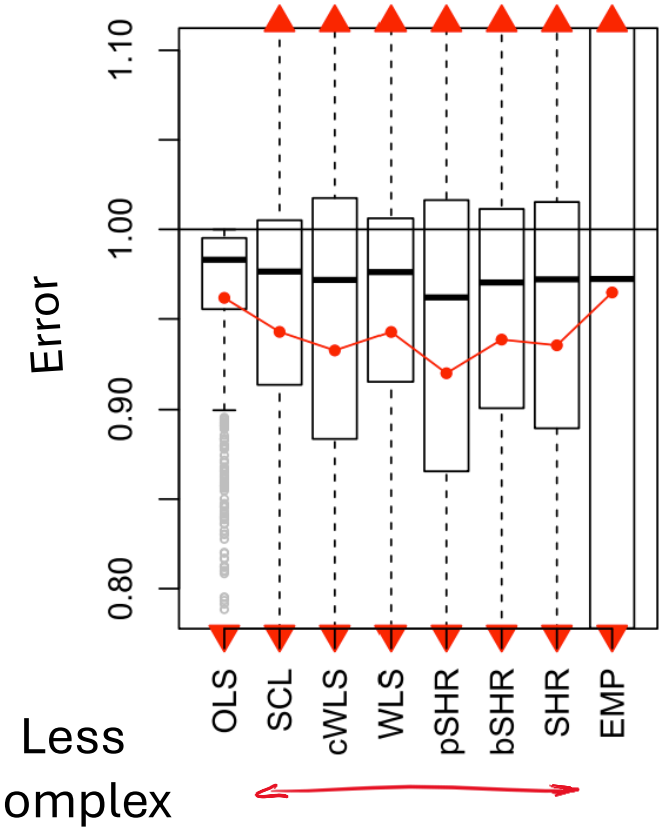
There is an insidious detail: the estimation of the variance is not done on samples in t, but on samples in i (complete hierarchies). This means that if for instance you are working with 4 years of quarterly data, you have only 4 data points to estimate a 7x7 covariance matrix.

If you have plenty of sample, then by all means, go for complex approximations of $\boldsymbol{W}$, otherwise you want to keep it fairly simple.

Forecasting with Temporal Hierarchies (THieF)

Plenty of approximations in the literature.



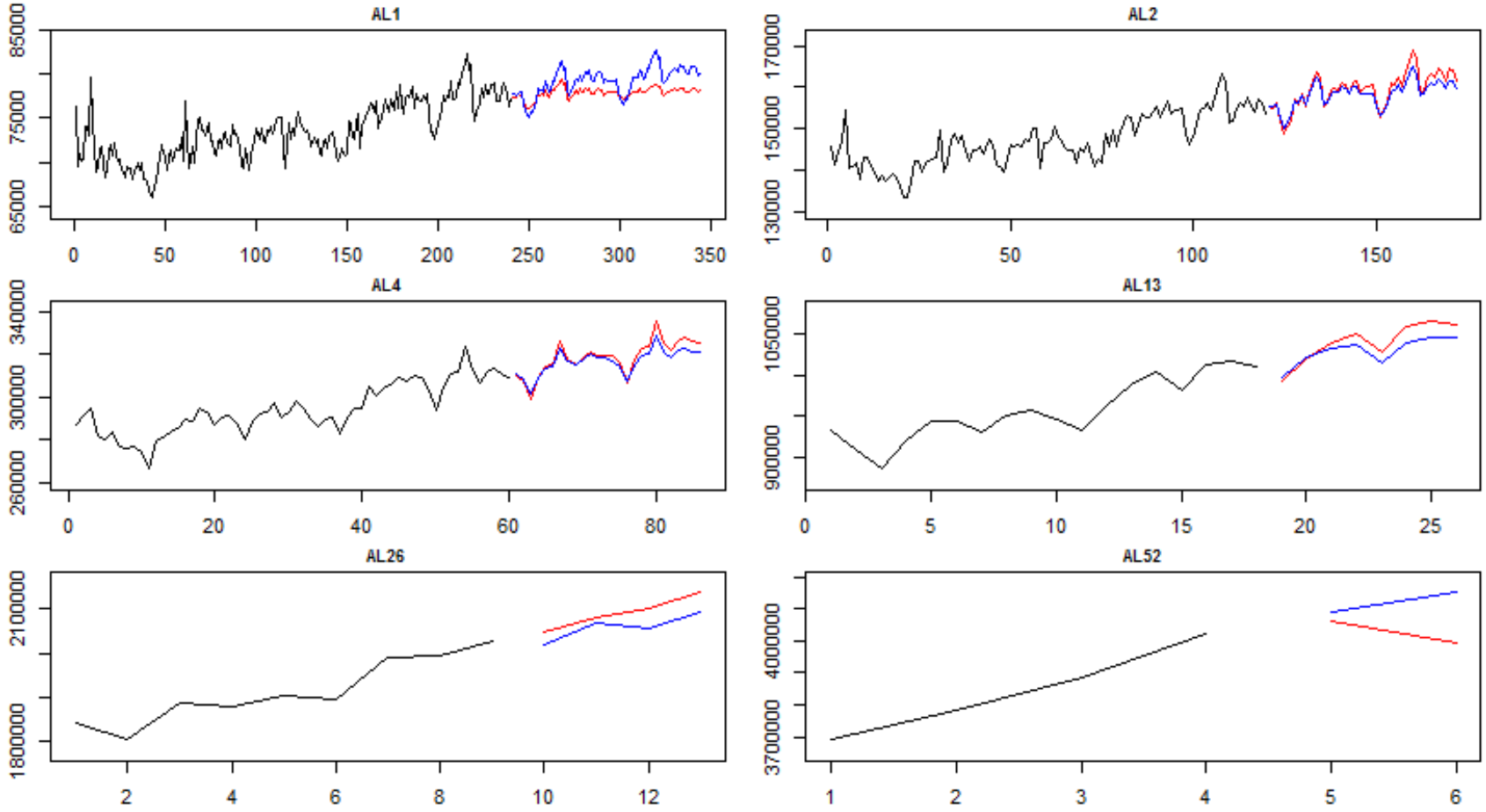


More complex may reduce error, but can increase variability of performance

Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. International Journal of Production Economics, 240, 108221

Forecasting with Temporal Hierarchies (THieF)

An example of what we have achieved

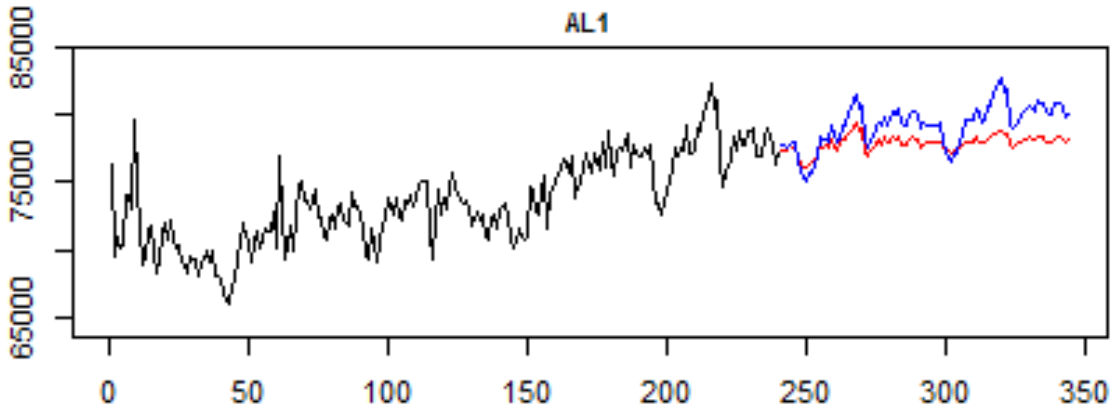


Red is the prediction of the base model (ARIMA) – at each level indepedently

Blue is the temporal hierarchy forecasts

Forecasting with Temporal Hierarchies (THieF)

There is no seasonal shrinkage, without any of the tricks from MAPA – how?



Two factors at play:

- Eliminated fractional levels (how many times can we fit 5-month buckets in a year? There is always some leftover information)
- The negative weights

Forecasting with Temporal Hierarchies (THieF)

Some published results*

- Kourentzes et al. (2014): 4-12% gains
- Kourentzes & Petropoulos (2016): 5-17% gains
- Athanasopoulos et al. (2017): 5-42% gains
- Kourentzes & Athanasopoulos (2019): 4-9% gains

Some unpublished results

- Schaer et al.: 15-23% gains
- Kourentzes & Athanasopoulos: 15-20% gains

Gains over competitive benchmarks (second best, or literature benchmark)

* see list of references at the end

What we will talk about

1. ~~Why bother with temporal hierarchies?~~
2. ~~The theory~~
3. Applications & observations
4. Newer results

Modelling uncertainty and THieF

Some interesting implications for the Data Generating Process (DGP)

- Can THieF model the data generating process?
  - Depends on how we want to understand the combination. If we take the shrinkage interpretation, sure, it is just a peculiar shrinkage estimator of the underlying process.
  - But it really asks us: do you believe that the data generating process happens at the sampling rate of your data? What are the chances of that? Because if it is not happening, THieF also tells you that you cannot revover the full signal only from the temporally aggregate levels.

- And if I look at it philosophically, there is no reason to believe that all components of the DGP operate on the same time scales. That does not mean we cannot write the DGP at the most disaggregate level, but it will probably not be written in its most efficient form!

- A more practical question is: how does THieF behave under modelling uncertainty? (Out initial motivation for all this work!)

Modelling uncertainty and THieF

Simulate with full knowledge of the data generating process (and algebraically aggregate it)

Four scenarios:

- No estimation or model specification errors.
- Estimation uncertainty (but the model equation is known)
- Only the model family is known, so if we are lucky we can get the correct one
- We apply the wrong modelling family (the realistic case).

Look at different estimation samples and compare against the base forecasts.

Details in: Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. European Journal of Operational Research, 262(1), 60-74.

Modelling uncertainty and THieF

Sample size: specified at the annual aggregation level
(Forecast horizon: specified at the annual aggregation level)

(negative values are percentage lower than the base forecast)

	4 (1)	12 (3)	20 (5)	40 (10)	4 (1)	12 (3)	20 (5)	40 (10)	4 (1)	12 (3)	20 (5)	40 (10)	4 (1)	12 (3)	20 (5)	40 (10)
	Scenario 1				Scenario 2				Scenario 3				Scenario 4			
	WLS combination forecasts using variance scaling															
Annual	−0.3	0.0	0.0	0.0	−4.3	−7.9	−6.1	−3.3	−66.2	−5.1	−2.6	−0.4	−24.7	1.6	0.5	−1.8
Semi-annual	−0.1	−0.1	0.0	0.0	−5.2	−3.5	−1.6	−0.2	−50.6	−4.9	−2.6	−1.2	−42.6	−5.5	−2.7	−1.1
Four-monthly	−0.1	0.0	0.0	0.0	−3.8	−1.5	−0.4	−0.1	−10.1	−6.2	−2.0	−1.2	−9.4	−6.7	−2.7	−4.3
Quarterly	−0.1	0.0	0.0	0.0	−3.9	−0.6	−0.2	−0.1	−16.4	−4.1	−1.9	−0.8	−1.2	−8.3	−5.5	−6.0
Bi-monthly	0.0	0.0	0.0	0.0	−1.1	0.0	0.1	0.0	−7.5	−3.3	−0.7	−0.9	−1.0	−8.3	−9.3	−8.6
Monthly	0.0	0.0	0.0	0.0	1.0	0.5	0.1	0.0	−0.9	−0.5	−0.8	−1.9	−1.4	−7.3	−11.3	−17.0
	Bottom-up															
Annual	−0.7	−0.1	0.2	0.1	−5.3	−9.5	−7.1	−3.4	−64.2	−1.2	5.9	27.9	−20.9	69.1	101.6	150.4
Semi-annual	−0.5	−0.1	0.1	0.0	−7.6	−4.8	−2.4	−0.2	−48.5	−2.8	2.3	13.8	−40.0	35.5	63.8	105.3
Four-monthly	−0.2	−0.1	0.1	−0.1	−5.5	−2.7	−1.0	−0.2	−7.1	−5.1	1.4	8.7	−5.8	23.4	47.8	73.1
Quarterly	−0.2	0.0	0.0	0.0	−6.1	−1.8	−0.7	−0.2	−14.0	−3.0	0.4	6.5	2.3	15.5	33.4	54.9
Bi-monthly	−0.1	−0.1	0.0	0.0	−2.8	−0.9	−0.2	−0.1	−5.8	−2.4	1.2	3.8	1.9	8.2	16.1	32.7
Monthly	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Zeros!

Only modelling original data

No uncertainty

Estimation uncertainty

Model uncertainty

Incorrect model

It is no worse than full knowledge, and shines when there is modelling uncertainty!

Sidenote: Modelling uncertainty and MAPA

We compared MAPA with choosing a single optimal aggregation level when there is an analytical formula for that and looked at the relative accuracy.

- MAPA performed better even when the theory matched the simulated case fully – why?
- First, MAPA is by no means optimal in those cases, and unlike THieF, there is no reason to anticipate that it can achieve zero differences from the base forecasts under full information.
- What it does, just like THieF, is mitigate estimation uncertainty via combination across aggregation levels.
- Unless we plug in the true process parameters, chances are that MAPA will outperform model parameters from standard estimators.

Details in: Kourentzes, N., Rostami-Tabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: using optimal or multiple aggregation levels? Journal of Business Research, 78:1-9.

Using Multiple Temporal Aggregation levels is beneficial in mitigating modelling uncertainty no matter if you go with the first or the second version of using Multiple Temporal Aggregation (MTA) level approaches!

Sidenote II: Spotting modelling uncertainty

Do you see it?

	Sample size: specified at the annual aggregation level *(Forecast horizon: specified at the annual aggregation level)*															
	4 *(1)*	12 *(3)*	20 *(5)*	40 *(10)*	4 *(1)*	12 *(3)*	20 *(5)*	40 *(10)*	4 *(1)*	12 *(3)*	20 *(5)*	40 *(10)*	4 *(1)*	12 *(3)*	20 *(5)*	40 *(10)*
	Scenario 1				Scenario 2				Scenario 3				Scenario 4			
	WLS combination forecasts using variance scaling															
Annual	−0.3	0.0	0.0	0.0	−4.3	−7.9	−6.1	−3.3	−66.2	−5.1	−2.6	−0.4	−24.7	1.6	0.5	−1.8
Semi-annual	−0.1	−0.1	0.0	0.0	−5.2	−3.5	−1.6	−0.2	−50.6	−4.9	−2.6	−1.2	−42.6	−5.5	−2.7	−1.1
Four-monthly	−0.1	0.0	0.0	0.0	−3.8	−1.5	−0.4	−0.1	−10.1	−6.2	−2.0	−1.2	−9.4	−6.7	−2.7	−4.3
Quarterly	−0.1	0.0	0.0	0.0	−3.9	−0.6	−0.2	−0.1	−16.4	−4.1	−1.9	−0.8	−1.2	−8.3	−5.5	−6.0
Bi-monthly	0.0	0.0	0.0	0.0	−1.1	0.0	0.1	0.0	−7.5	−3.3	−0.7	−0.9	−1.0	−8.3	−9.3	−8.6
Monthly	0.0	0.0	0.0	0.0	1.0	0.5	0.1	0.0	−0.9	−0.5	−0.8	−1.9	−1.4	−7.3	−11.3	−17.0
	Bottom-up															
Annual	−0.7	−0.1	0.2	0.1	−5.3	−9.5	−7.1	−3.4	−64.2	−1.2	5.9	27.9	−20.9	69.1	101.6	150.4
Semi-annual	−0.5	−0.1	0.1	0.0	−7.6	−4.8	−2.4	−0.2	−48.5	−2.8	2.3	13.8	−40.0	35.5	63.8	105.3
Four-monthly	−0.2	−0.1	0.1	−0.1	−5.5	−2.7	−1.0	−0.2	−7.1	−5.1	1.4	8.7	−5.8	23.4	47.8	73.1
Quarterly	−0.2	0.0	0.0	0.0	−6.1	−1.8	−0.7	−0.2	−14.0	−3.0	0.4	6.5	2.3	15.5	33.4	54.9
Bi-monthly	−0.1	−0.1	0.0	0.0	−2.8	−0.9	−0.2	−0.1	−5.8	−2.4	1.2	3.8	1.9	8.2	16.1	32.7
Monthly	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

If THieF improves a lot over the base forecasts, chances are the base forecasts are very misspecified

- Maybe try different base forecasts!

Forecasting with Temporal Hierarchies (THieF)

THieF – 5 stars! Would forecast with it again!

Pros:

- More flexible than MAPA as it does not require the state space formulation.
- Model independent – plug in your stats, ML/AI, or judgmental forecasts.
- Cheaper than MAPA – less aggregation levels to model.
- Good long-term accuracy (so the same reasons as MAPA).

Cons:

- Turns out that MAPA remains competitive in terms of accuracy
- My view is that the hierarchical treatment helped to develop its formulation, but the combination view helps more in understanding what THieF does (and does not).
- Just like with combinations or hierarchical forecasts, we are missing quick ways to obtain probabilistic forecasts (we will return to that).

Tricks with THieF – Intermittent demand

An issue with many intermittent demand methods/models is that they are “unnatural”: they do not let standard time series patterns emerge from them.

Let us assume we deal with some seasonal time series. If we sample it fast enough, it will become intermittent (e.g., tick data of supermarket sales).

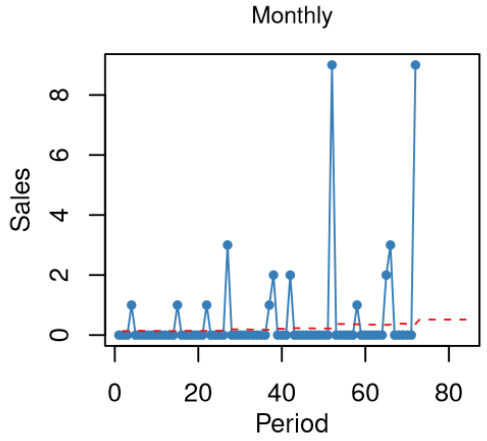
- If we were to model that data with standard intermittent demand approaches, then we would not be able to recover the seasonality when we aggregate the forecasts.

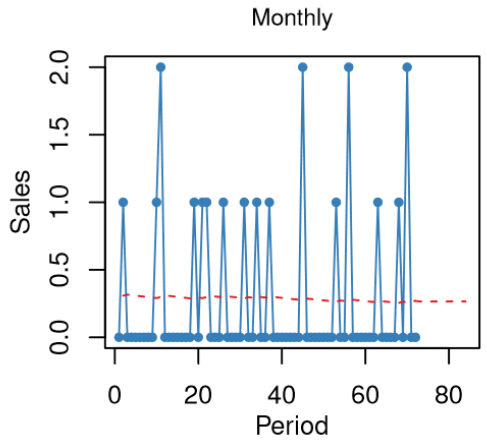
Let us impose coherency between temporal aggregation levels: at some level we only see intermittency, at some other level we see a clear seasonal pattern.

- Pass the independent forecasts through THieF, seasonality will not pop up on your intermittent level (frequency is too high!), but if you were to aggregate it, now it will become seasonal.
- Consider that in many of these cases we are interested in the data over the lead time, which is a summation of the demand over periods – that s a temporal aggregation.

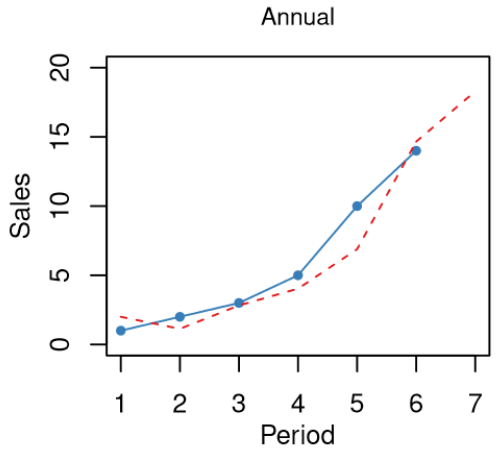
Tricks with THieF – Intermittent demand

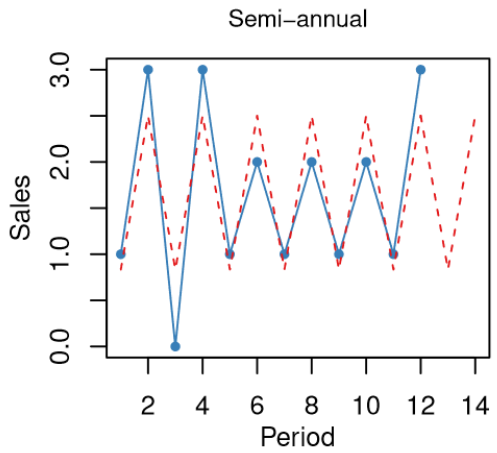
Original data





Aggregate data





Patterns emerge even in well studied intermittent datasets that are supposed to have no classical structure. (These are monthly RAF spare parts series,)

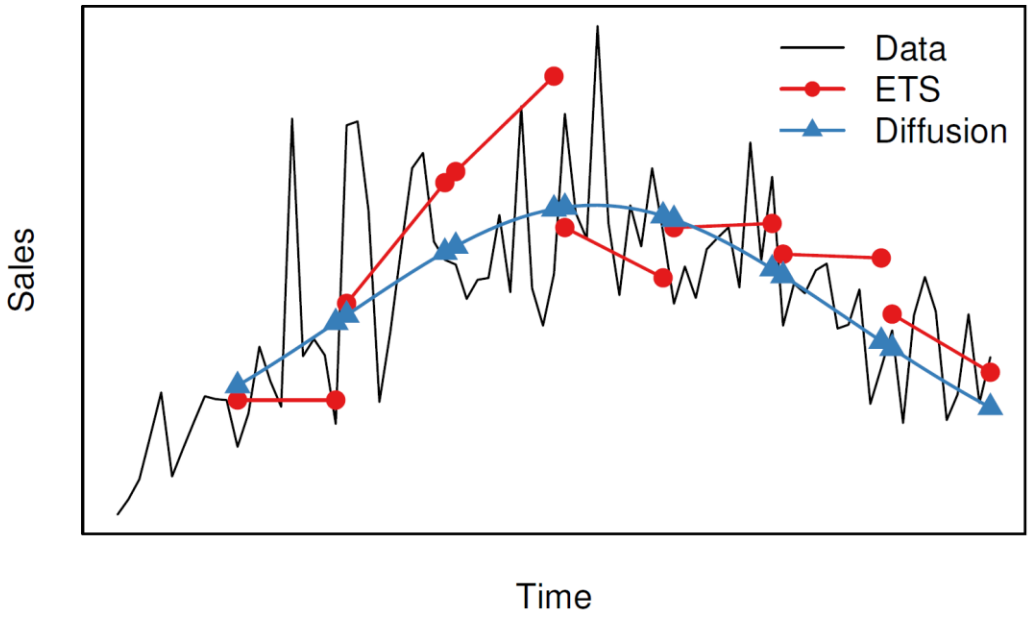
Details: Kourentzes, N., & Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. European Journal of Operational Research, 288(1), 141-152.

Tricks with THieF – Lifecycle modelling

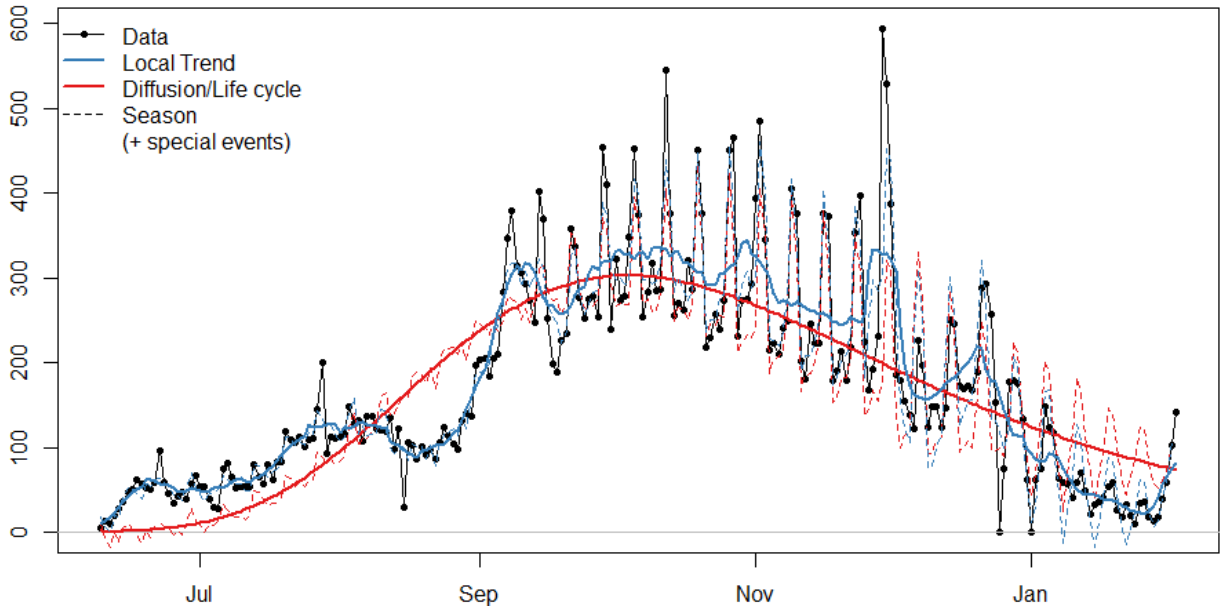
Many products (e.g., tech, fashion) have short lifecycles that makes their modelling complicated:

- The trend is dominated by the lifecycle dynamics. The adoption process is easy to model with diffusion curves, by difficult with time series models.
- Lifecycle forecasting methods (diffusion curves) typically do not handle "local" effects, such as promotions, stochastic trends, seasonality, etc.

Tech product



Fashion product

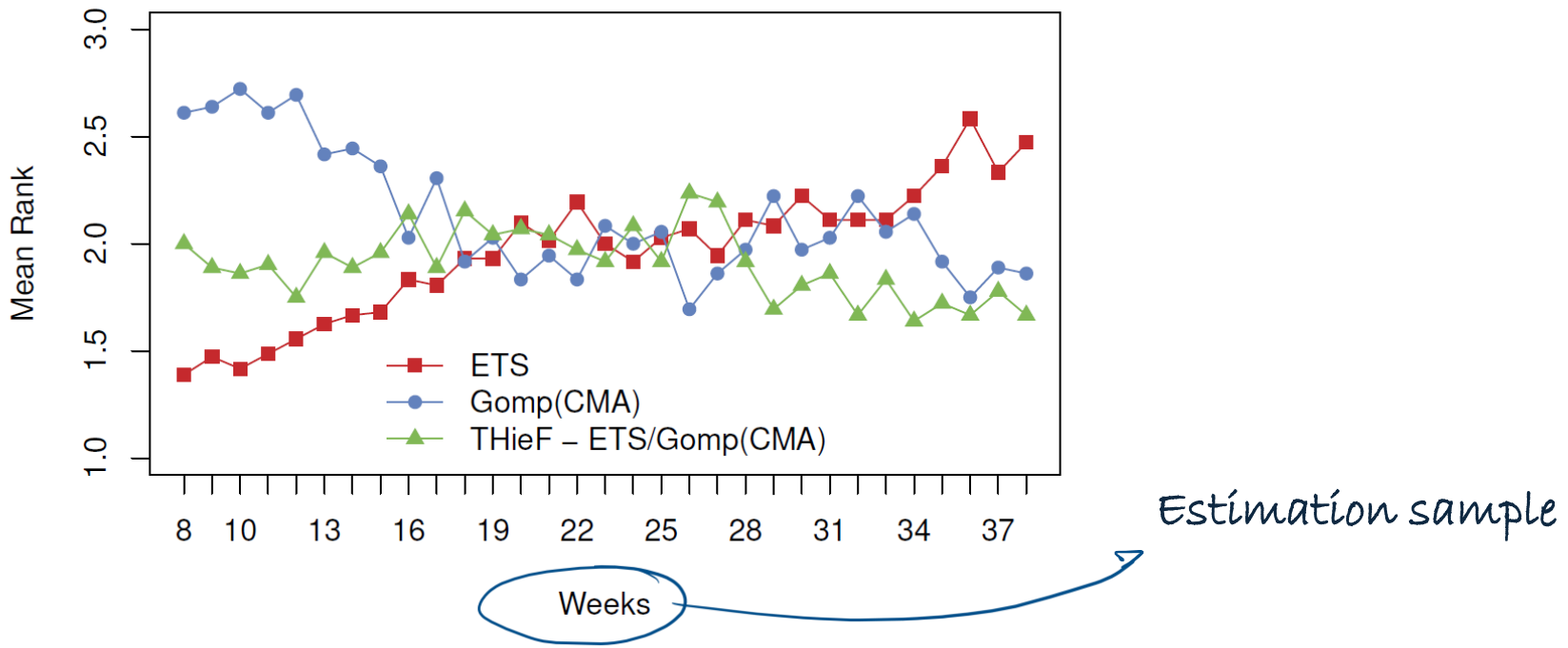


Tricks with THieF – Lifecycle modelling

Model higher sampling frequency data with time series models

Model aggregate view of the data with diffusion curves (lifecycle models)

Use THieF to combine the information between the modelling approaches:



The performance of THieF remains consistent across sample sizes.

Tricks with THieF – Lifecycle modelling

Writing the equation of the combined model, we can show that we capture composite trends that do not exist in either diffusion or time series methods.

We take a very simple example of a semi-annual to annual data, with the annual data modelled as a diffusion (Gompertz) curve, and the semi-annual as linesr trend exponential smoothing.

$$\mathbf{G} = \begin{bmatrix} 0.25 & 0.75 & -0.25 \\ 0.25 & -0.25 & 0.75 \end{bmatrix}$$

$$\begin{aligned}
\tilde{y}_{SA_1 i} &= 0.25\,[\text{Gompertz}(A_i)] + 0.75\,[\text{ETS}(SA_{1i})] - 0.25\,[\text{ETS}(SA_{2i})] \\
&= 0.25(g_\alpha \exp(-g_\beta \exp(-g_\gamma i))) + 0.75(l_{2(i-1)} + b_{2(i-1)}) - 0.25(l_{2(i-1)} + 2b_{2(i-1)}) \\
&= 0.25(g_\alpha \exp(-g_\beta \exp(-g_\gamma i))) + 0.5 l_{2(i-1)} + 0.25 b_{2(i-1)}).
\end{aligned}$$

We get one part Gompertz, one part ETS trend, and two parts ETS level

Tricks with THieF and Hierarchies: cross-temporal

Remember this?

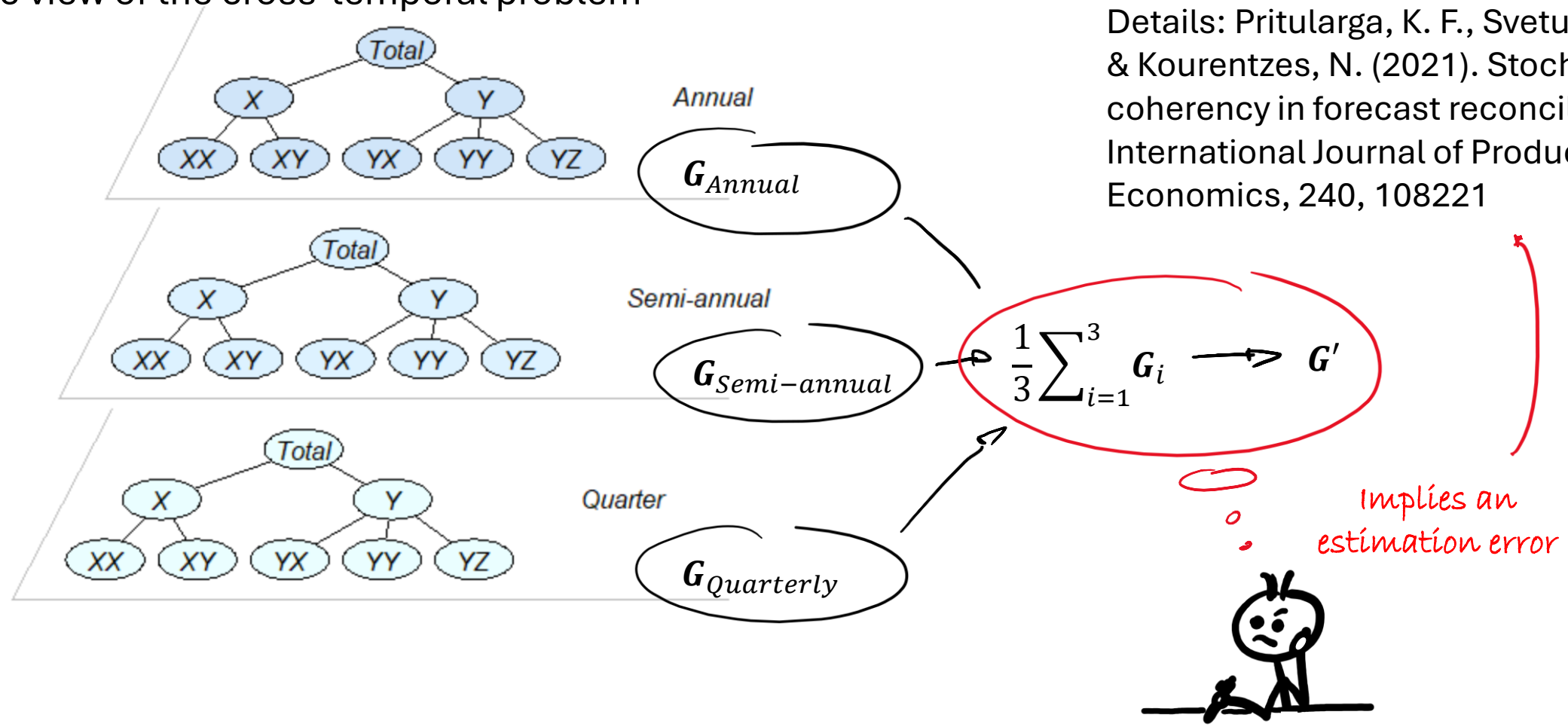
Example hierarchies of decisions in organisations.

Decision	Frequency	Time series	Output
	Call centre (Koole & Li, 2021)		
Budget planning	Quarterly	Monthly+	Budget
Capacity planning	Monthly	Weekly+	Training and hiring plans
Operational planning	Weekly	Weekly	Outsourced call volume
Scheduling	Weekly	Daily+	Agent schedules per type
Scheduling	Hourly	Intra-daily	Adaptations to schedules
	Tech manufacturer*		
Financial planning	Yearly	Quarterly+	High-level financial goals
Annual operations plan	Yearly	Monthly+	Resource allocation
Production planning	Monthly	Monthly	Aggregate demand planning
Master production plan	Weekly	Weekly	Detailed demand planning
Material planning	Weekly	Weekly	Supply requirements

Observe that the scale across the example hierarchies changes both in time and unit of analysis. Coherency should be achieved both across the temporal and the cross-section hierarchies.

Tricks with THieF and Hierarchies: cross-temporal

A simplistic view of the cross-temporal problem



Details: Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. International Journal of Production Economics, 240, 108221

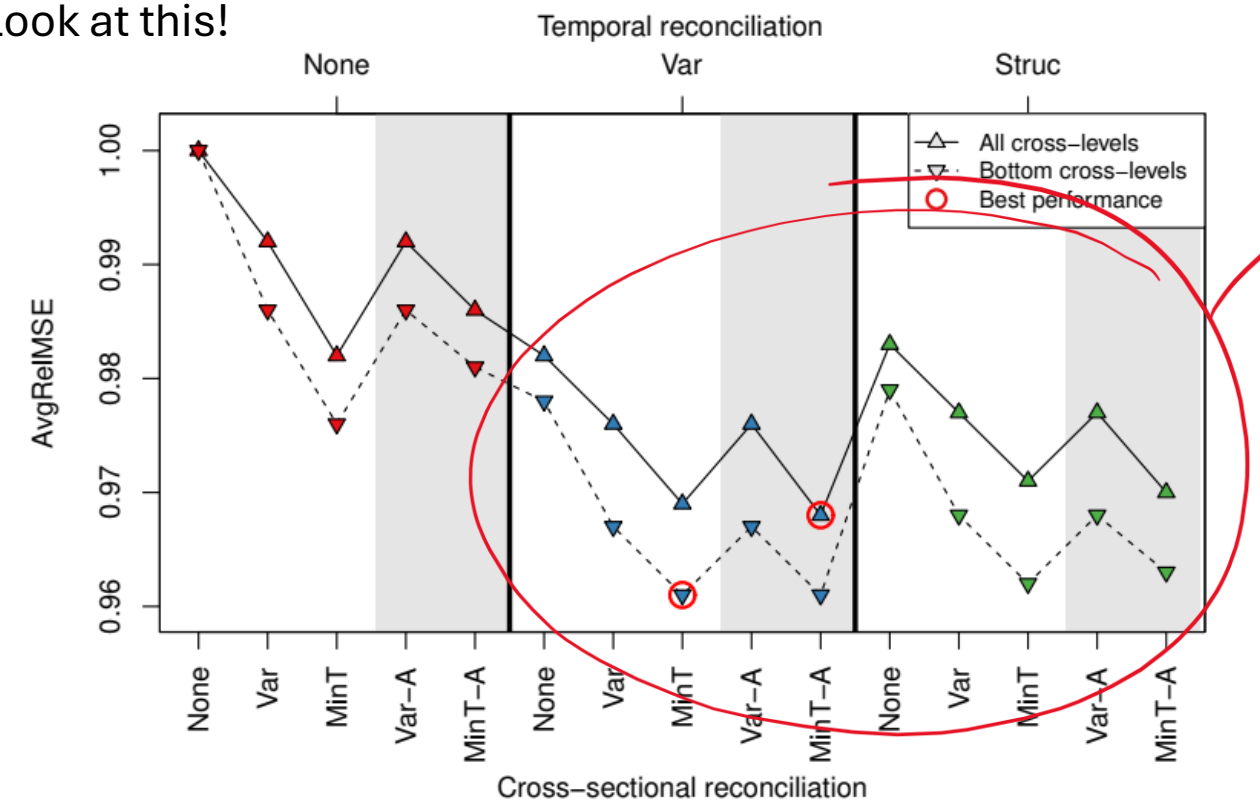
Details: Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research, 75, 393-409.

Tricks with THieF and Hierarchies: cross-temporal

But this is not a workshop about cross-temporal. If we wanted that we would read:

- Di Fonzo, T., & Girolimetto, D. (2023). Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. International Journal of Forecasting, 39(1), 39-57.

Sure, but look at this!

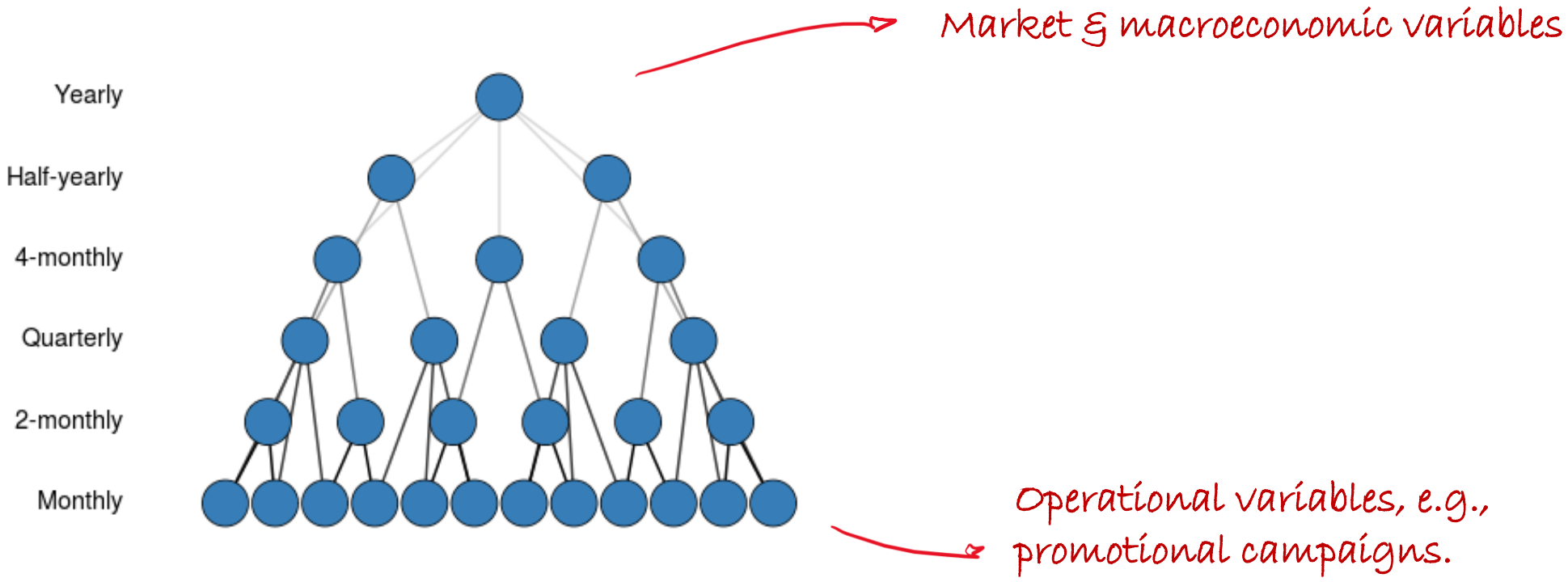


The temporal side brings larger benefits in accuracy, why?

- Because the forecasts in the pool to be combined are actually meaningful! This is not a given for cross-sectional hierarchies.
- The point here is that temporal is a beast of its own, because of what goes in the pool of forecasts – it performs typically much better than hierarchical forecasting.

Tricks with THieF and exogenous variables

Forget statistics for a second. Consider the business context



THieF provides a very easy way to incorporate variables available at different sampling rates.

What does THieF optimise?

Let us first understand what happens at each temporal aggregation level. For sake of simplicity, let us optimize each forecasting method on Mean Squared Errors (MSE).

- The forecasting method at the original data is MSE optimal.
- The forecasting method at the k=2 aggregation level is MSE optimal on two-period buckets → On forecasting the sum of two periods. We have theory for that!

In: Svetunkov,I., Kourentzes,N., & Killick,R. (2024). Multi-step estimators and shrinkage effect in time series models. Computational Statistics 39 (3), 1203-1239, we show that optimizing on MSE with horizons > 1 imposes a type of univariate shrinkage on model parameters:

- Your autoregressive and moving average related coefficients will regularize, but not your exogenous variable related coefficients.
- The longer the horizon, the stronger the regularization effect. `

So THieF imposes a double shrinkage effect: (1) from the combination across temporal aggregation levels; and (2) from the implicit loss function at each temporally aggregate level.

Careful though: we have no control on the second shrinkage, as its strength is connected to k, the aggregation level.

Sequential or joint estimation?

Disclaimer: more work is needed here!

If THieF is a forecast combination, should we estimate the combination weights (the $\boldsymbol{G}$ matrix, or the $\boldsymbol{W}$ matrix) together after the generation of the forecasts, or jointly with the forecasting method parameters?

- In machine-learning-speak: are end-to-end approaches meaningful for THieF-like implantations?

Since we are framing THieF (or more generally hierarchical forecasting) as a forecast combination, then the question is the same as: should we do forecast combination of already calculated forecasts, or do ensemble learning?

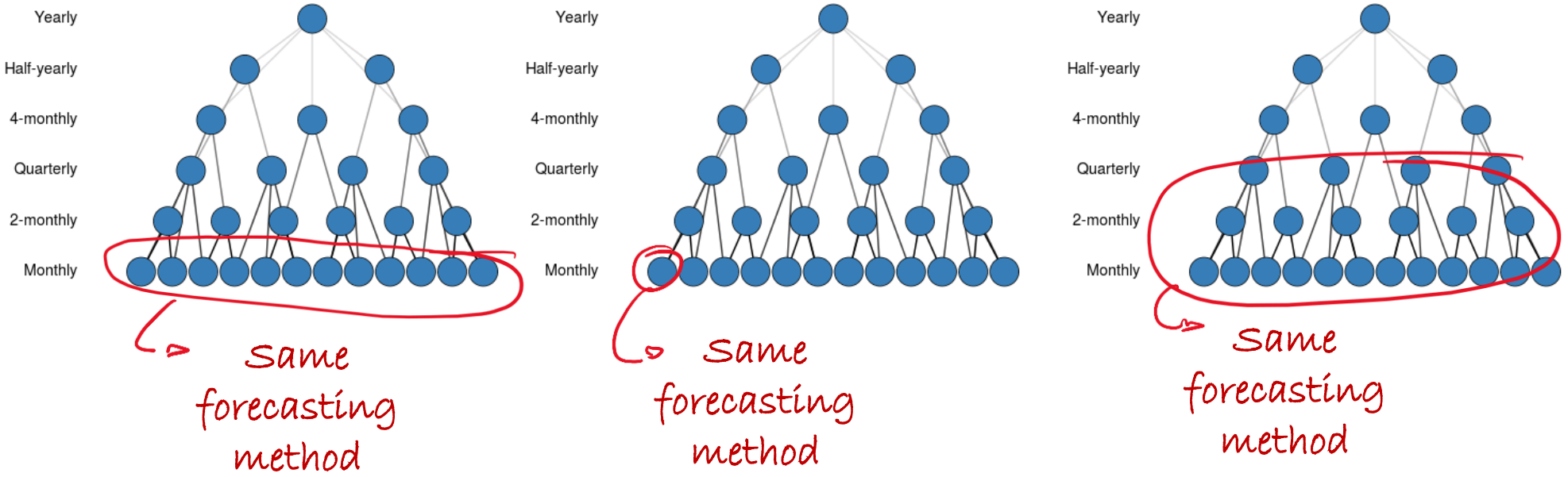
- Both are useful with substantial evidence that they work – but they are not the same!
- The THieF discussed so far is clearly a forecast combination.
- THieF-style ensemble learners are meaningful as they structure the information processing in a useful way to recover the information (recall the motivation of MAPA), but they (may?) also have different properties – I will show some evidence today.
- **More research is needed!**

ML/AI & THieF

THieF is model independent: there is nothing stopping you from plugging ML/AI forecasts as base forecasts.

Given how we think of ML/AI forecasts, we can ask some more fun questions:

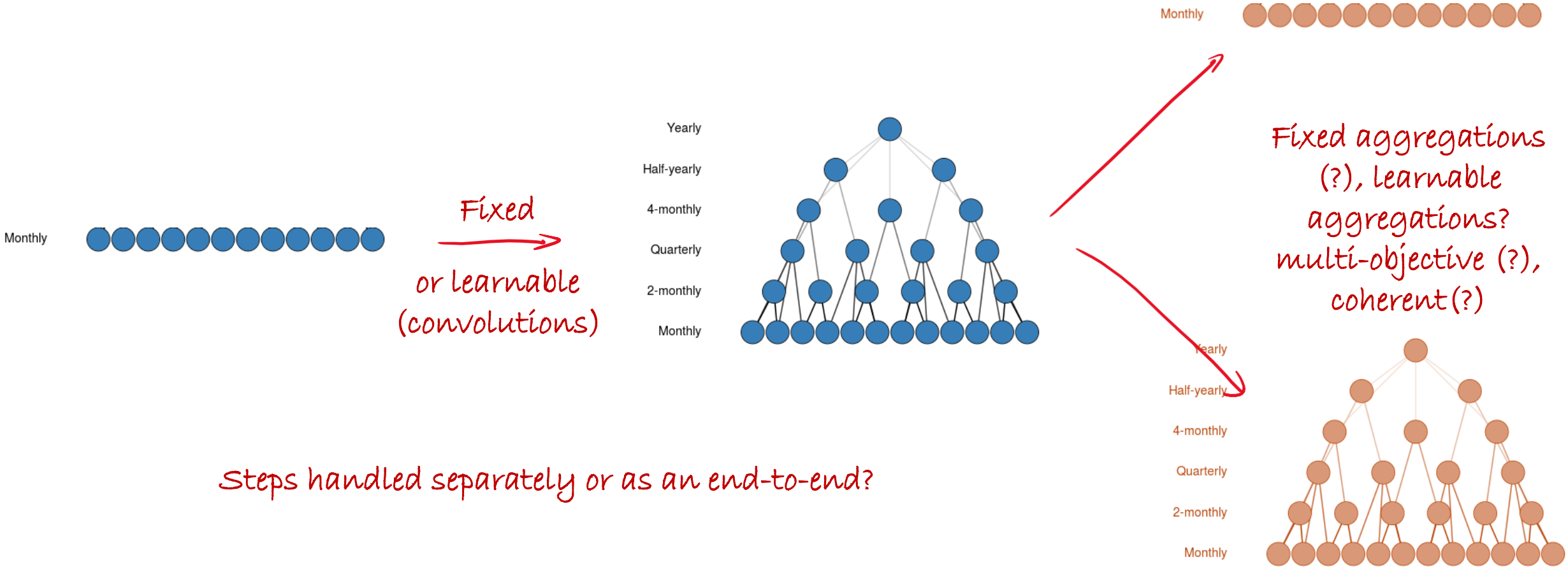
- Do forecasts need to be per level? Per node? On larger groupings?



This is not really an AI/ML question. The vector with the base forecasts is simply a multivariate object. Forecast it as you will! Global or local, uni- or multivariate, with whatever covariance matrix you can estimate.

ML/AI & THieF

Here is the logic we see in some ML/AI implementations:



Some limited discussion in: Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Panagiotelis, A. (2024). Innovations in hierarchical forecasting. International Journal of Forecasting, 40(2), 427-429.

ML/AI & THieF – some of my thoughts

Interesting research direction, BUT

- Let's understand where the benefits are coming from (if any): is it the:
  1. End-to-end? Are we doing combinations or ensemble learning?
  2. The learnable aspects?
  3. The multivariate/multi-target? A lot to be said about loss functions.
  4. The ML/AI time series part?
  5. The ML/AI reconciliation part?
  6. The use of multiple temporal aggregation levels?
  7. Something else in the architecture?

I would like to see more work towards these directions, as they would give us the toolset for better modelling propositions. Currently, it is very difficult to disentangle where the benefits (if any) come from.

THieF and predictive distributions

Multiple ways to obtain predictive distributions:

- Lean on the hierarchical literature (coherent simulated paths, copulas, empirical)
- Lean on the forecast combination literature
- Just do empirical: obtain THieF forecasts, calculate residuals, sample from that distribution.

Example of empirical (some useful details): Kourentzes, N., & Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. European Journal of Operational Research, 288(1), 141-152.

I do not see predictive distributions and THieF as a separate research question from hierarchical forecasting or forecast combinations and predictive distributions, but THieF provides a very structured way to explore this further – there is a very clear way how the information in the different aggregation levels is connected!

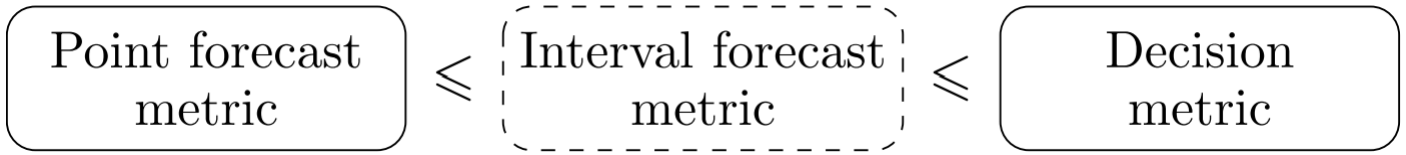
- It may be easier that either of the more general questions → open question

THieF and forecast evaluation

Why are we using THieF?

- Better accuracy for the target series in the original sampling frequency?
- Coherent forecasts across aggregation levels/horizons?

Match the evaluation to your objective – always!

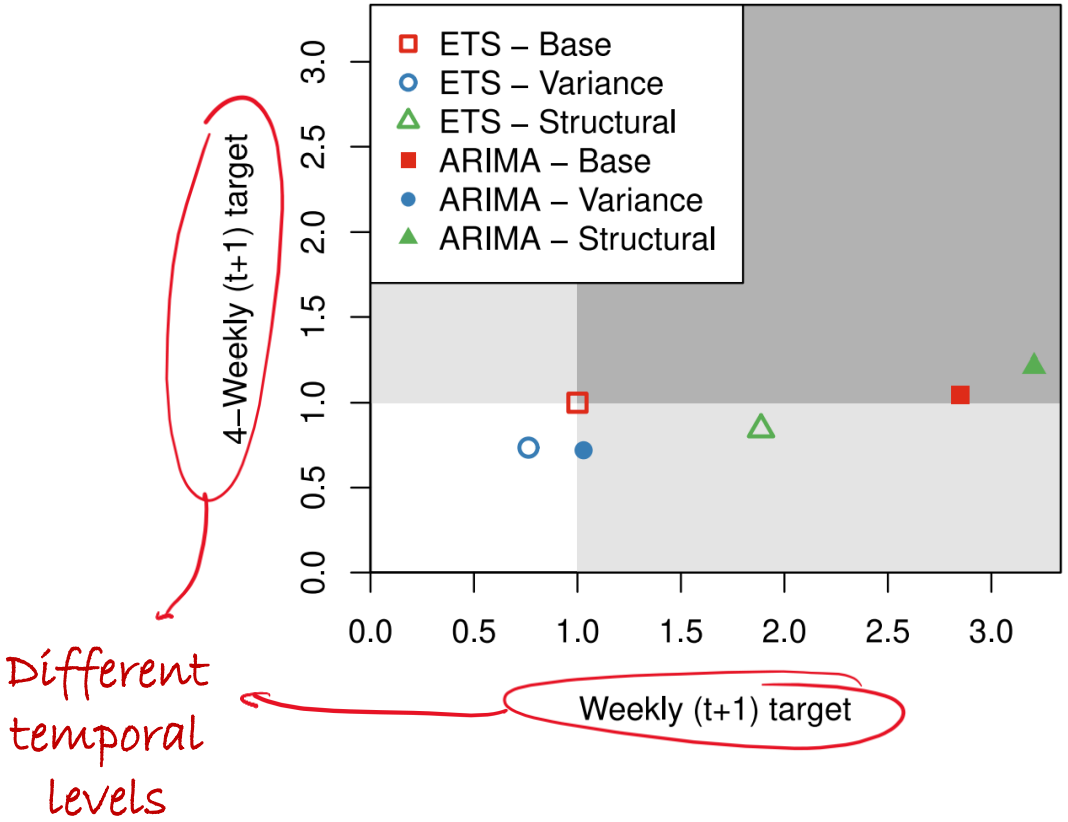


Average performance over the whole temporal hierarchy is most probably meaningless – many levels are statistical devices to help us get better forecasts → disconnected from decisions.

THieF and forecast evaluation

If you are evaluating on different levels: do you need different metrics across levels? Is an average meaningful

- Personally, I think the problem as a multi-objective problem.



- Is there a globally dominant solution or forecasts are only partially dominant?
- Different levels may have different loss functions – weighting can become rather messy – that's okay.
- What do we want – Coherency? Accuracy? Both? Some other "meta-metric"? (meta-metrics: metrics that get value from the application context, e.g., reliability)

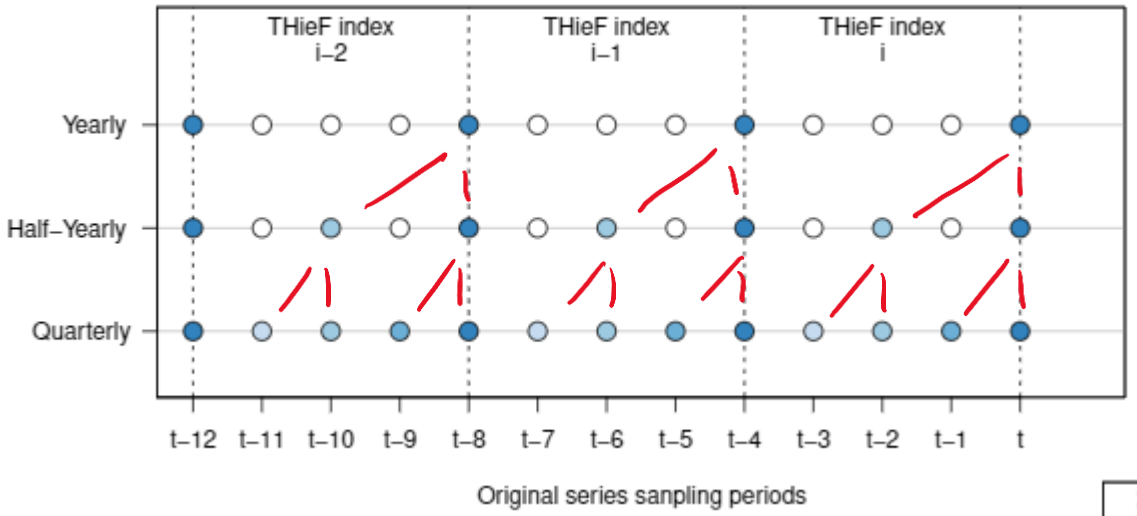
A good discussion in (no bias here!): Athanasopoulos, G., & Kourentzes, N. (2023). On the evaluation of hierarchical forecasts. International Journal of Forecasting, 39(4), 1502-1511.

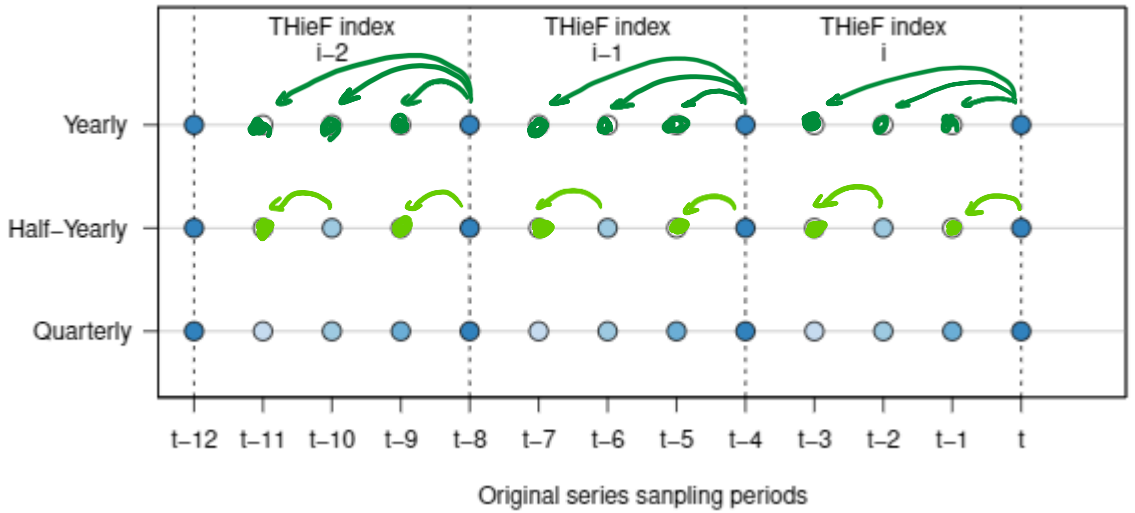
What we will talk about

1. ~~Why bother with temporal hierarchies?~~
2. ~~The theory~~
3. ~~Applications & observations~~
4. Newer results

Some newer (unpublished*) ideas/results

Let us revisit the interpolation/disaggregation of MAPA. We recast THieF as a forecast combination.



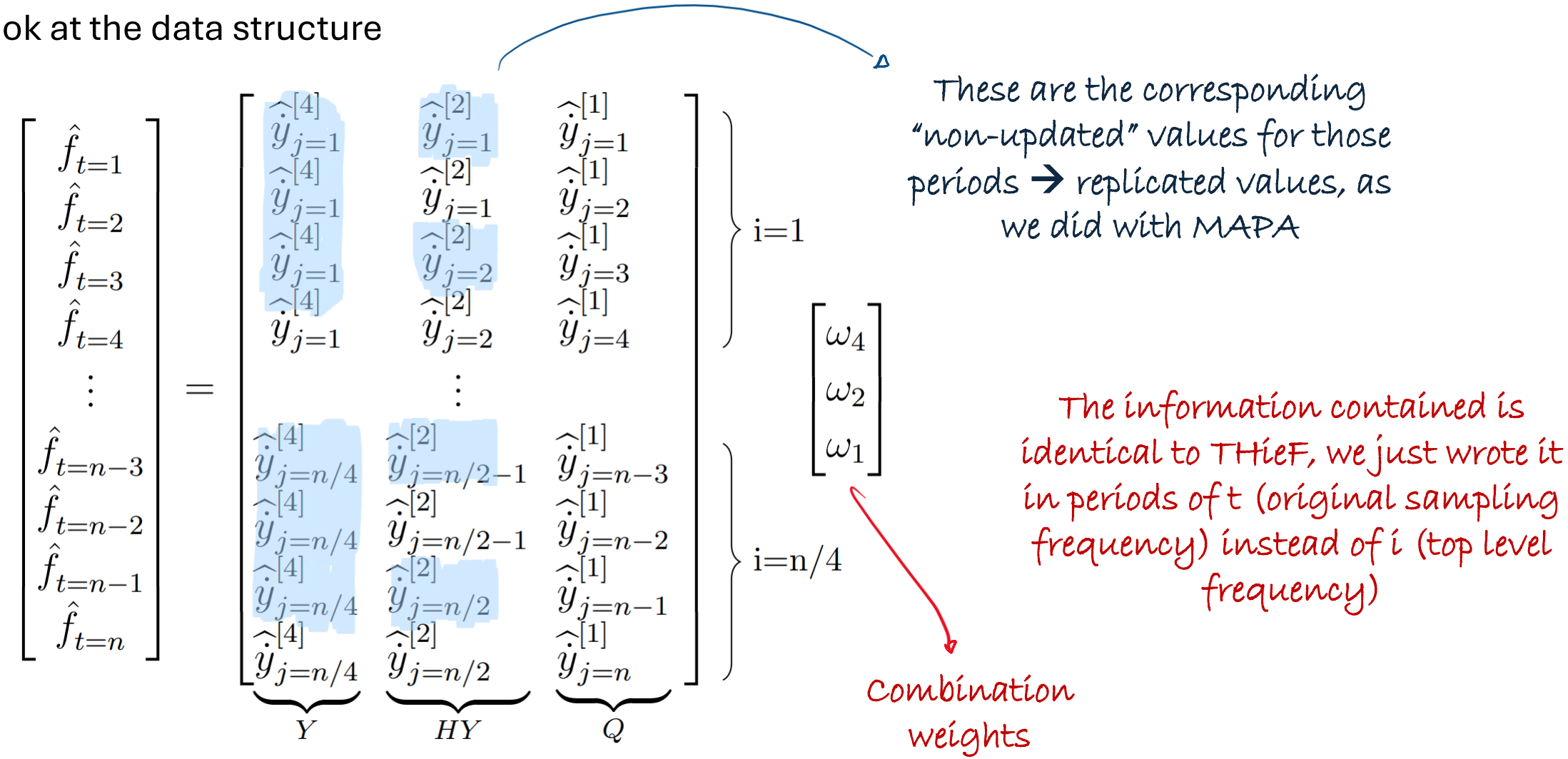


MAPA-style imputation

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Let us look at the data structure



*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

The combination weights are different are different

Because SGS=S

THieF: Structural

	Annual	SA1	SA2	Q1	Q2	Q3	Q4	Sum
Q1	0.333	0.417	-0.083	0.708	-0.292	-0.042	-0.042	1
Q2	0.333	0.417	-0.083	-0.292	0.708	-0.042	-0.042	1
Q3	0.333	-0.083	0.417	-0.042	-0.042	0.708	-0.292	1
Q4	0.333	-0.083	0.417	-0.042	-0.042	-0.292	0.708	1

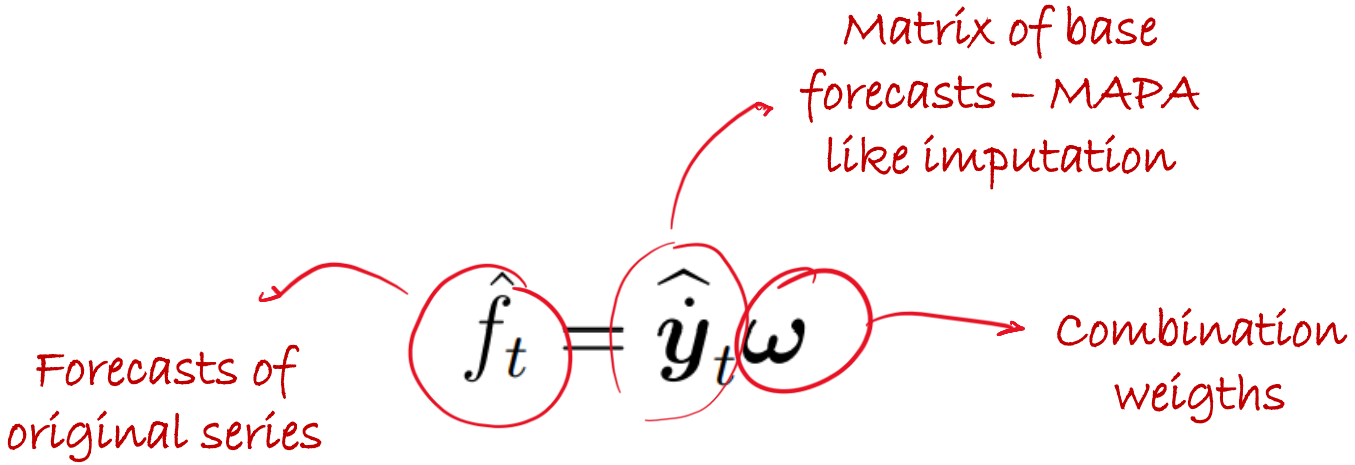
MAPA-like

	Annual	SA1	SA2	Q1	Q2	Q3	Q4	Sum
Q1	0.333	0.333	0	0.333	0	0	0	1
Q2	0.333	0.333	0	0	0.333	0	0	1
Q3	0.333	0	0.333	0	0	0.333	0	1
Q4	0.333	0	0.333	0	0	0	0.333	1

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

We can now write a v3 of Multiple Temporal Aggregation approaches that progresses on both MAPA and THieF:



With THieF we need the covariance matrix, which at minimum requires $\sum(\boldsymbol{k})$ elements. The new formulation requires only as many weights as temporal levels. E.g., for quarterly data $\sum(\boldsymbol{k}) = 7$, while the size of the $\boldsymbol{w}$ vector is 3 → increased estimation efficiency.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

With the reformulation we can probe some properties of MTA approaches.

The variance of combined forecasts is

$$\mathrm{Var}\left(\hat{f}_t\right) = \boldsymbol{\omega}' \boldsymbol{\Sigma}_{\hat{\boldsymbol{y}}} \boldsymbol{\omega}$$

Covariance

We split this into two parts

$$\boldsymbol{P}_{\hat{\boldsymbol{y}}} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{y}}} - \boldsymbol{\Sigma}_{\dot{\boldsymbol{y}}}$$

Leftover due to model misspecification

Covariance of aatA generating process across temporal levels

$$\mathrm{Var}\left(\hat{f}_t\right) = \boldsymbol{\omega}' \boldsymbol{\Sigma}_{\dot{\boldsymbol{y}}} \boldsymbol{\omega} + \boldsymbol{\omega}' \boldsymbol{P}_{\hat{\boldsymbol{y}}} \boldsymbol{\omega}.$$

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Since temporal aggregation is a moving average:

$$\text{Var}\left(\dot{y}_j^{[k+1]}\right) \leq \text{Var}\left(\dot{y}_j^{[k]}\right)$$

"The variance drops as we aggregate (we divide by k the aggregate series)"

and by Cauchy-Schwartz inequality

$$\left|\text{Cov}\left(\dot{y}_j^{[k+1]}, \dot{y}_j^{[k]}\right)\right| \leq \sqrt{\text{Var}\left(\dot{y}_j^{[k+1]}\right)\text{Var}\left(\dot{y}_j^{[k]}\right)}$$

we get that all elements will be at maximum 1 in

$$\boldsymbol{\Sigma}_{\dot{\boldsymbol{y}}}/\text{Var}\left(y^{[1]}\right)$$

"The data covariance divided by the variance of the original series"

If the combination weights follow usual restrictions ($w\_i \leq 1, \sum(\boldsymbol{w}) = 1$):

$$\boldsymbol{\omega}'\boldsymbol{\Sigma}_{\dot{\boldsymbol{y}}}\boldsymbol{\omega} \leq \text{Var}\left(y^{[1]}\right)$$

"The temporal hierarchy variance is small than the original series"

Duh... you aggregated, right?

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Let us focus on the remainder part. There are two cases:

1. If these are proportional $\boldsymbol{P}_{\hat{\boldsymbol{y}}} \propto \boldsymbol{\Sigma}_{\boldsymbol{y}}$ (i.e., are forecasts in the different levels are pretty good)

$$\boldsymbol{\omega}'\boldsymbol{\Sigma}_{\hat{\boldsymbol{y}}}\boldsymbol{\omega} \leq \operatorname{Var}\left(\hat{y}^{[1]}\right)$$

"THieF forecasts more accurate than base forecasts"

2. If they are not, i.e., the base forecasts at some levels fail, then from forecast pooling we know that we should set some combination weights to zero → motivation for sparse temporal hierarchies.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Can we make a statement that THieF will be more accurate than not?

Kind of. We start from the easy case

Base forecast for original series

Base forecast for aggregated for k=2

$$\hat{f}_t = \omega_1 \hat{y}_t^{[1]} + \omega_2 \hat{y}_t^{[2]}$$

Not THieF errors

$$e^2 = \left(y - \hat{y}^{[1]}\right)^2$$

THieF errors

$$u^2 = \left(y - \hat{f}\right)^2$$

So, all we need to show is that there is a w_2 (since $w_1 + w_2 = 1$) where $e^2 > u^2$

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Let's show it then!

$$
\begin{aligned}
e^2 - u^2 &= \left(y - \hat{y}^{[1]}\right)^2 - \left(y - \hat{f}\right)^2 \\
&= \left(y - \hat{y}^{[1]}\right)^2 - \left(y - \left((1-\omega_2)\hat{y}^{[1]} + \omega_2\widehat{\dot{y}}^{[2]}\right)\right)^2 \\
&= -\omega_2^2\left(\hat{y}^{[1]}\right)^2 + 2\omega_2^2\hat{y}^{[1]}\widehat{\dot{y}}^{[2]} - \omega_2^2\left(\widehat{\dot{y}}^{[2]}\right)^2 \\
&\qquad + 2\omega_2\left(\hat{y}^{[1]}\right)^2 - 2\omega_2 y\hat{y}^{[1]} - 2\omega_2\hat{y}^{[1]}\widehat{\dot{y}}^{[2]} + 2\omega_2 y\widehat{\dot{y}}^{[2]} \\
&= -\omega_2^2\left(\left(\hat{y}^{[1]}\right)^2 - 2\hat{y}^{[1]}\widehat{\dot{y}}^{[2]} + \left(\widehat{\dot{y}}^{[2]}\right)^2\right) \\
&\qquad - 2\omega_2\left(-\left(\hat{y}^{[1]}\right)^2 + y\hat{y}^{[1]} + \hat{y}^{[1]}\widehat{\dot{y}}^{[2]} - y\widehat{\dot{y}}^{[2]}\right) \\
&= -\omega_2^2\left(\widehat{\dot{y}}^{[2]} - \hat{y}^{[1]}\right)^2 - 2\omega_2\left(-y + \hat{y}^{[1]}\right)\left(\widehat{\dot{y}}^{[2]} - \hat{y}^{[1]}\right) \\
&= -\omega_2^2\left(\widehat{\dot{y}}^{[2]} - \hat{y}^{[1]}\right)^2 + 2\omega_2 e\left(\widehat{\dot{y}}^{[2]} - \hat{y}^{[1]}\right)
\end{aligned}
$$

Blah blah blah... it's unpublished, so you get the maths to follow the argument if you want, otherwise skip to the next slide!

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

This is the condition that we get:

$$0 < \omega_2 < \frac{2e}{\hat{y}^{[2]} - \hat{y}^{[1]}}$$

What does it tell us?

- There is always an w_2 that can give $u^2 < e^2$, as long as, the contribution of the additional forecast can offset the **bias of the base forecast**.
- This is irrespective of the quality of the base forecasts.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

- Remember that we know that the variance of the aggregate levels lowers (temporal aggregation is a moving average).
- Therefore, chances are that the aggregate forecasts remain closer to a “centre”.
- Let’s do a small simulation, between two forecasts.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

Chance of THieF being more accurate

r	Misspecification due to probabilistic bias (q)										
	None	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0.1	97%	97%	97%	99%	100%	100%	100%	100%	100%	100%	100%
0.2	94%	93%	94%	97%	99%	100%	100%	100%	100%	100%	100%
0.3	91%	90%	91%	94%	97%	99%	100%	100%	100%	100%	100%
0.4	88%	87%	86%	90%	95%	98%	99%	100%	100%	100%	100%
0.5	85%	83%	81%	85%	90%	94%	97%	99%	99%	100%	100%
0.6	83%	80%	76%	78%	83%	89%	93%	95%	97%	98%	99%
0.7	81%	76%	71%	72%	76%	81%	85%	89%	92%	94%	96%
0.8	78%	74%	66%	65%	67%	71%	75%	78%	81%	84%	87%
0.9	77%	71%	61%	58%	59%	61%	62%	64%	67%	69%	70%
1.0	75%	68%	57%	52%	50%	50%	50%	50%	50%	50%	50%

Ratio of variances aggregate/ original

Add a random jitter to the forecasts. The value is the size of that jitter compared to the mean.

25% chance that the condition is violated with no variance reduction and bias

50% chance with no variance reduction.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

In human-speak:

- THieF is expected to improve your base forecast accuracy with a fairly high chance – as the variance reduction is expected.
- The chance of gains is higher as the misspecification increases (i.e., smaller reduction in variance is needed to achive almost 100% probability of accuracy gains).

This is the best argument I can offer you so far that using MAPA, THieF, or similar MTA approaches, should be the default in time series modelling.

*the reviewers have not taken them apart yet!

Some newer (unpublished*) ideas/results

We can obtain sparse temporal hierarchies with modified shrinkage approaches (the median in MAPA kind of did that as well)

% gains over base forecast

Forecast	RMSsE		MAsE		MsE		$SGS = S$
	Mean	Median	Mean	Median	Mean	Median	
Base	0	0	0	0	0	0	-
Structural	3.78	0.05	4.27	0.26	2.23	7.38	✓
Variance	2.84	-0.17	3.46	0.14	3.91	6.23	✓
Mean	13.81	15.35	11.06	10.96	2.23	7.38	✗
Median	14.73	16.3	12.71	12.48	9.18	17	-
OLSstep	2.64	0.47	2.87	0.58	-3.8	1.08	✗
OptQP	3.29	0.8	3.44	0.88	-1.7	1.55	✗
OptC	13.81	15.35	11.06	10.96	2.23	7.38	✗
OptH	-3.33	-0.95	-3.87	-1.64	-7.2	-5.22	✓
OptCH	13.81	15.35	11.06	10.96	2.23	7.37	✗
RidgeEgl	10.06	8.77	8.62	6.74	-1.13	3.17	✗
LassoEgl	8.93	7.54	7.57	5.48	-4.79	-5.38	✗
RidgeCn	15.28	16.74	13.11	13.89	2.74	15.66	✗
LassoCn	12.59	11.25	12.36	10.92	8.68	19.08	✗

The hierarchical unbiasedness restriction

Sparse hierarchies

In brief…

- Using Multiple Temporal Aggregation levels in your modelling is a way to mitigate (and take advantage of) model misspecification.
- MAPA is v1 on how to do it, fairly restrictive, but still very accurate.
- THieF is v2: the flexible way to do it.
- Do not be confused by the name “Temporal Hierarchies”, ir builds on hierarchical forecasting, but it is really a constrained forecast combination, with tricks on how to obtain the forecast pool.
- Best see it as a device to combine information that “lives” on different temporal aggregation levels. That may be:
  - univariate components that are otherwise difficult to identify;
  - exogenous variables that are not sampled in higher frequencies.

- But should we use it? Chances are it will be more accurate than not using it!
- Is it hard to use? There are packages (MAPA, thief, foreco), but really it is only a few lines of code if you already have the base forecasts.

Some references

- Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. International Journal of Forecasting, 30(2), 291-302.
- Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. International Journal of Production Economics, 181, 145-153.
- Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. European Journal of Operational Research, 262(1), 60-74.
- Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for Australian tourism. Annals of Tourism Research, 75, 393-409.
- Kourentzes, N., & Athanasopoulos, G. (2021). Elucidate structure in intermittent demand series. European Journal of Operational Research, 288(1), 141-152.
- Pritularga, K. F., Svetunkov, I., & Kourentzes, N. (2021). Stochastic coherency in forecast reconciliation. International Journal of Production Economics, 240, 108221.
- Athanasopoulos, G., & Kourentzes, N. (2022). On the evaluation of hierarchical forecasts. International Journal of Forecasting
- Kourentzes, N. (2022). Toward a one-number forecast: cross-temporal hierarchies. Foresight: The International Journal of Applied Forecasting, 67, 32-38.

First paper

Here we made the theory cool

Introduce the concept

A good read!

Practitioner perspective

Thank you for your time! Questions?

nikolaos@kourentzes.com