

Improving organisational forecasts and decisions with hierarchical forecasting

Nikolaos Kourentzes

Skövde Artificial Intelligence Lab

Skövde University, Sweden

sail.his.se

CMAF Friday Forecasting Talks

12 February 2021*



Agenda

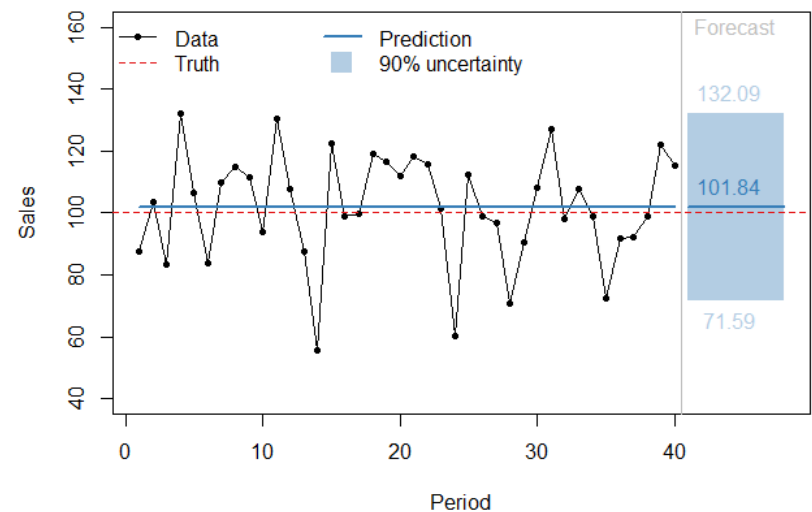
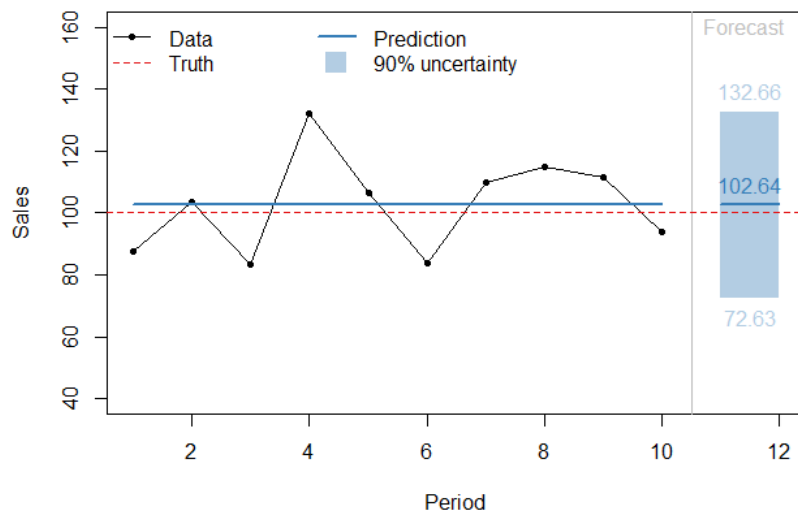
- 1. Context – problems in forecasting**
- 2. Hierarchical forecasting**
 - **Temporal**
 - **Cross-sectional**
 - **Current advances**



Some context: Model Parameters

Let us keep the problem very simple. We forecast ice-cream sales.

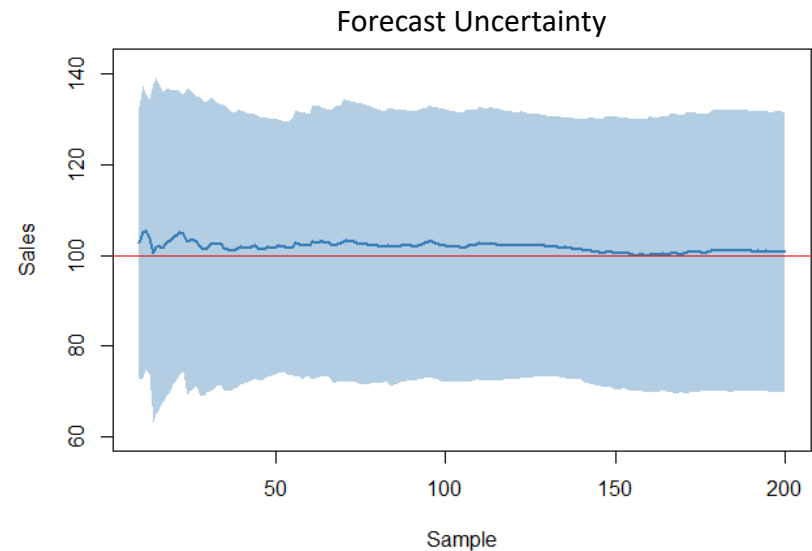
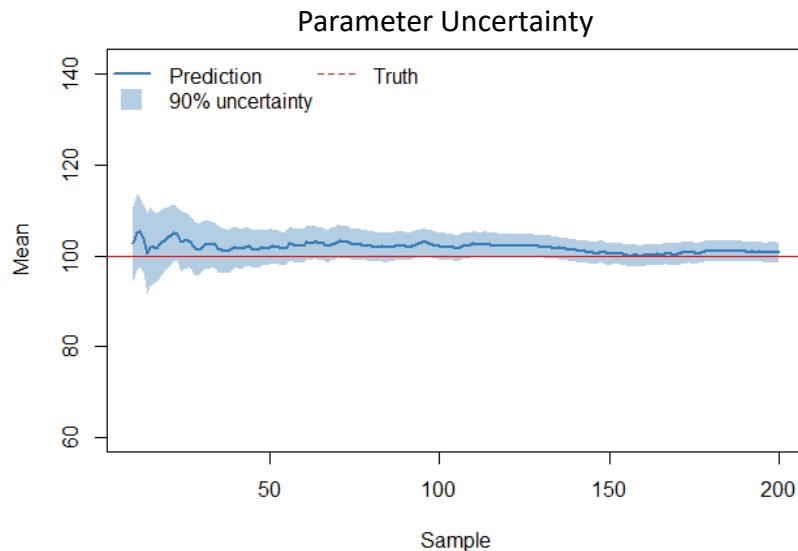
- As it is well known ice-cream sales are independent of temperature/weather/price, so we have a demand with a constant mean of 100 and some randomness.
- We know that the demand is independent of everything, but we do not know the mean
→ We use the correct model, but we do not know the parameters.



- As we have more historical data we get closer to the true value. The uncertainty of the forecast is somewhat more complicated!

Some context: Model Parameters

As we have more data we expect the uncertainty to reduce, however we need to first say which uncertainty!



- Parameter uncertainty reduces
- Forecast uncertainty reduces a bit initially, but then remains rather wide.
 - Forecast uncertainty includes the **noise of the time series**, and the effects from the model form (and its parameters).

Some context: Model selection

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 model 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 →

$$\text{Forecast} = \text{Constant} (+ \text{Error})$$

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

$$\text{Forecast} = \text{Constant} + 0(\text{seasonality}) + 0(\text{price}) + 0(\dots) (+ \text{Error})$$

What if we did not know that?

$$\text{Forecast} = \text{Constant} + a(\text{seasonality}) + b(\text{price}) + c(\dots) (+ \text{Error})$$

Needs estimation →

The error depends on the model →

...and all these add uncertainty

Some context: Forecast combination

Finding an appropriate model (I do not use “the” – many will be good enough) remains a major problem in predictive modelling


- But we are both smart and lazy: instead of solving the difficult problem we devise clever tricks to avoid that problem! One such trick is **forecast combination**.
- Suppose we have two forecasts, we can simply take their average:

$$\text{Forecast A} = \text{Constant} \quad (+ \text{Error})$$

$$\text{Forecast B} = \text{Constant} + a(\text{seasonality}) + b(\text{price}) + c(\dots) \quad (+ \text{Error})$$

$$\text{Combined} = \frac{1}{2} \text{Constant} \quad (+ \frac{1}{2} \text{Error})$$

$$\frac{1}{2}\text{Constant} + \frac{1}{2}a(\text{seasonality}) + \frac{1}{2}b(\text{price}) + \frac{1}{2}c(\dots) \quad (+ \frac{1}{2} \text{Error})$$

Terms go in by half 
Shrinkage effect – smaller error

This is messier than it seems 

A helpful over-simplification:
Diverse forecasts combine well!

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.

Some context: The combination paradox

Forecast combinations are overwhelmingly considered to perform better than 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 paradox**, i.e., our elegant combination papers are useless!

Here is the intuition:

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

The diagram shows the equation $F_{combined,h} = b_0 + \sum_{i=1}^k b_i F_{i,h}$ with several handwritten annotations in red and blue ink. A red arrow points from the word 'Estimate' to the term b_0 . A blue arrow points from the word 'Optional to remove bias' to the same term b_0 . A blue arrow points from the phrase 'Combination weights' to the summation term $\sum_{i=1}^k b_i F_{i,h}$. A blue arrow points from the word 'Forecasts' to the term $F_{i,h}$ within the summation. A red arrow points from the word 'Estimate' to the term b_i within the summation.

- 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!**

A new way to combine forecasts

(I am taking a very roundabout way to reach hierarchies – bear with me)

What we have “learned” so far?

- Any estimation is increasing uncertainty – large samples mitigate this.
- The “true” model is unknown, so we try to find an approximate model (and we often fail).
- Forecast combinations can help – especially when they are diverse (they carry different information, so chances are the covariances play to our favour).
- Can we obtain diverse forecasts, without trying all possible models we can think of?
 - Silly forecasts will often harm even combinations (forecast pooling can help).
 - The trick is to look at your data from different viewpoints – just because you were given data in a particular form, nobody told you that it is the best form!

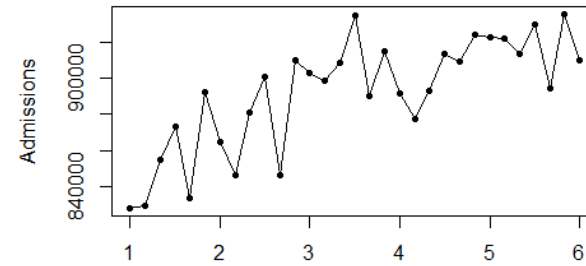
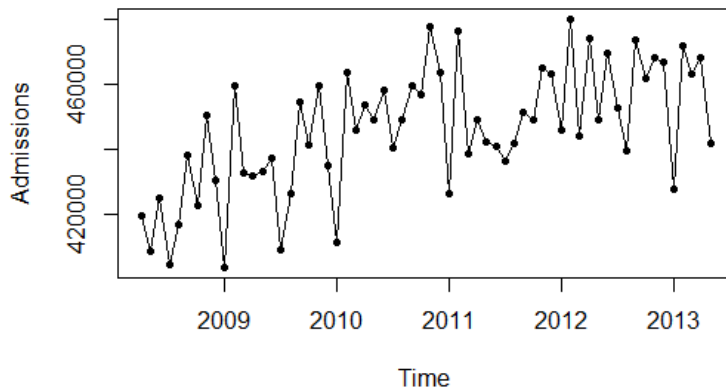


A new way to combine forecasts

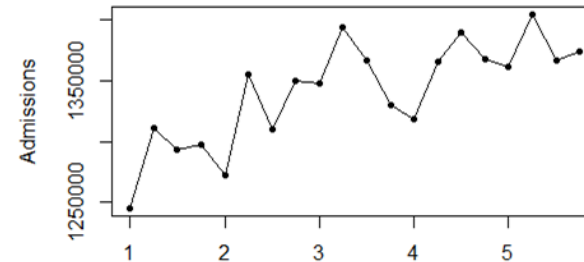
- Temporal aggregation filters high frequency components (e.g. seasonality), strengthening low frequency ones (e.g. trend)
- Reduces sample size, harming estimation.
- In principle we can model all levels:
 - Diverse information
 - More chances at getting a useful model

Monthly

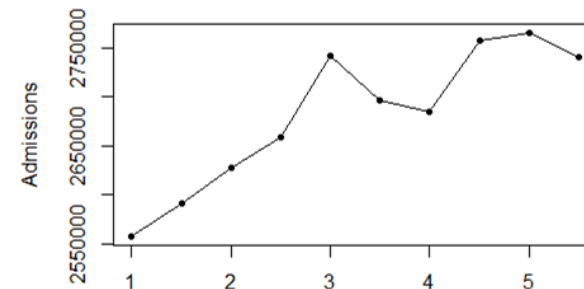
NHS A&E admissions



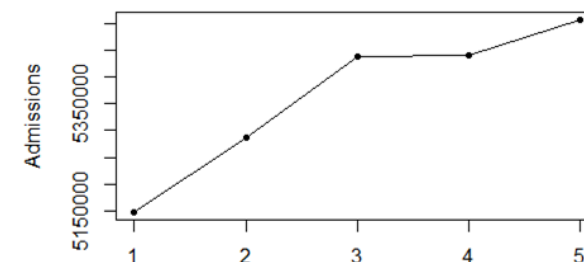
Bi-monthly



Quarterly



Half-annually



Annually

A new way to combine forecasts

What if we combine all these forecasts at different temporal aggregation levels?

- There are some technicalities to overcome, but it is easy to do and:
 - You do not rely on a single model, if you get some wrong it is okay.
 - Accuracy improvements, particularly for long-horizon forecasts → we explicitly include long term information to our forecasts.
 - Combine with a simple average or median → let us not increase uncertainties.
- This became the Multiple Aggregation Prediction Algorithm (MAPA) that is the start of temporal hierarchies (less known about this algorithm is that MAPA forecast in Greek means “rubbish forecast”)
 - We have also shown its good performance on intermittent demand and promotional forecasting.

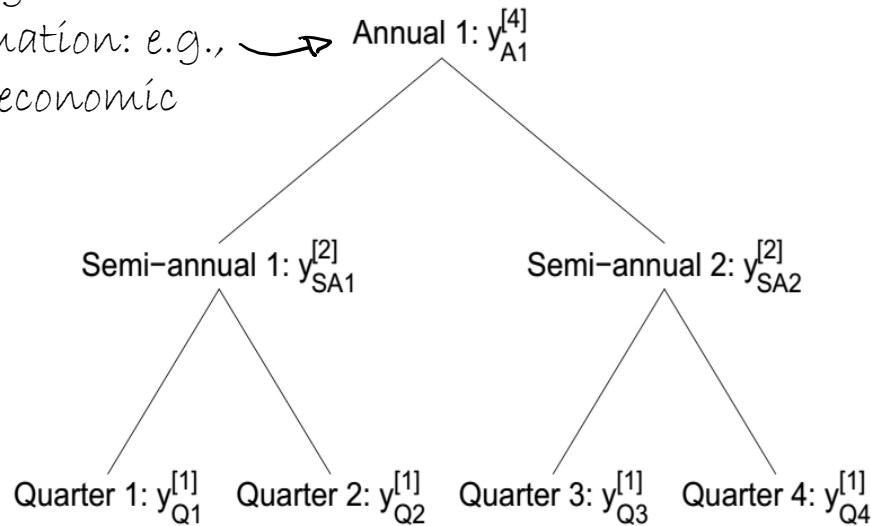


Temporal Hierarchies

If we aggregate across time, we can surely identify a clear structure: a hierarchy of aggregation across time

Temporal hierarchy

Aggregate external
information: e.g.,
macroeconomic



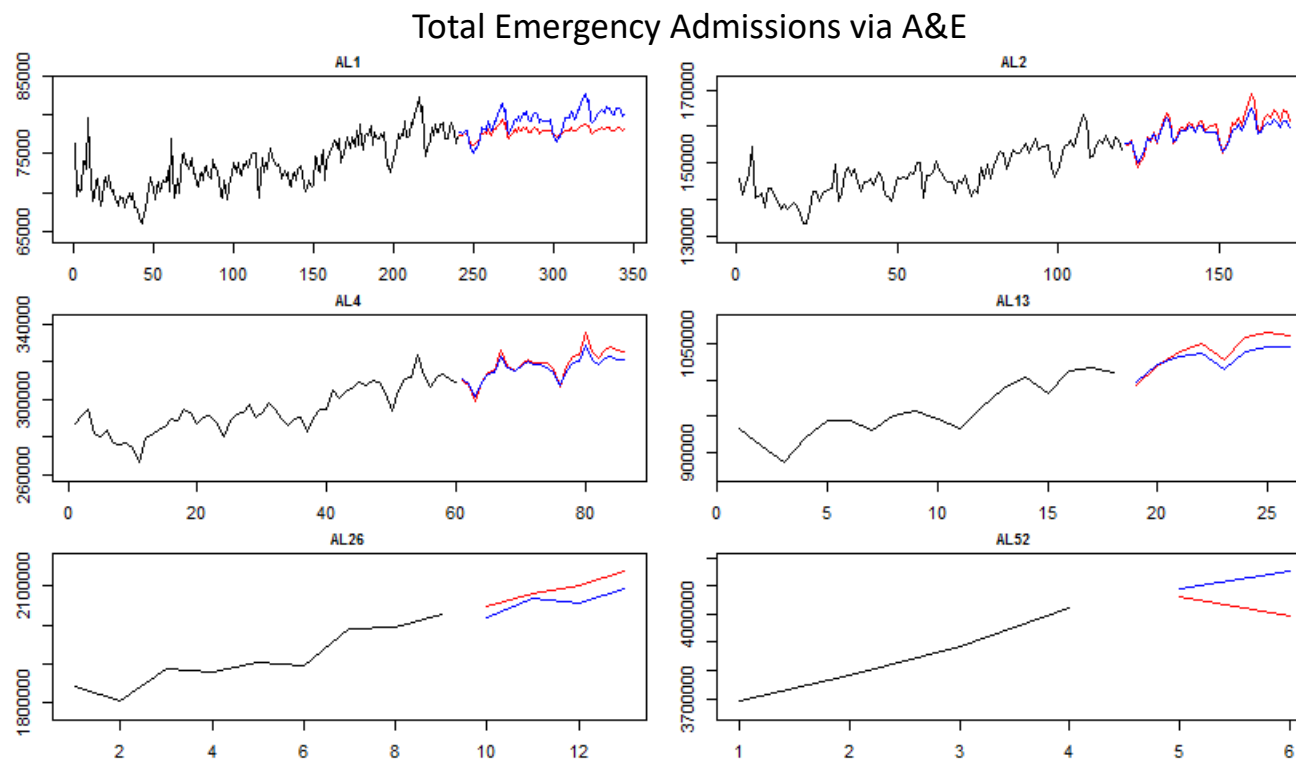
Disaggregate internal
information: e.g.
promotions

The thinking remains the same:

- Model at each level and combine
- Note that different information can be included at each level
- The structure is model independent, obtain the forecasts as you will



Example: Predicting A&E admissions



Red is the prediction of the base model – at each level separately

Blue is the temporal hierarchy forecasts

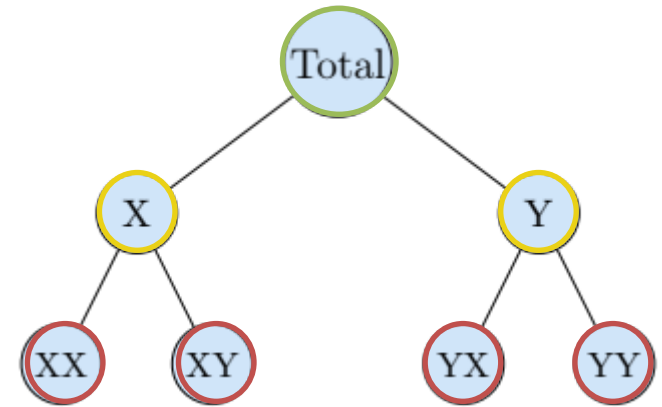
Observe how information is ‘borrowed’ between temporal levels. Base models for instance provide very poor weekly and annual forecasts

Temporal Hierarchy Forecasts (THieF) are typically better than base forecasts.

How does THieF work?

Let us abstract to a general hierarchy, we can name the nodes as we will.

- $\mathbf{b} = (y_{xx}, y_{xy}, y_{yx}, y_{yy})'$ Lower level series
- $\mathbf{y} = (y_{tot}, y_x, y_y, \mathbf{b}')'$ All series
- $\mathbf{y} = \mathbf{S}\mathbf{b}$ Mapping of lower to all



$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ \hline & & \mathbf{I}_m & \end{bmatrix}$$

Top level
Middle level(s)
Bottom level

- $\hat{\mathbf{y}}_h$ h-step ahead forecasts for \mathbf{y} , i.e., all series.

Reconciled coherent forecasts

Then we can write:

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

Summing matrix, i.e. the map of the hierarchy Matrix of base forecasts of all hierarchical nodes Magic!

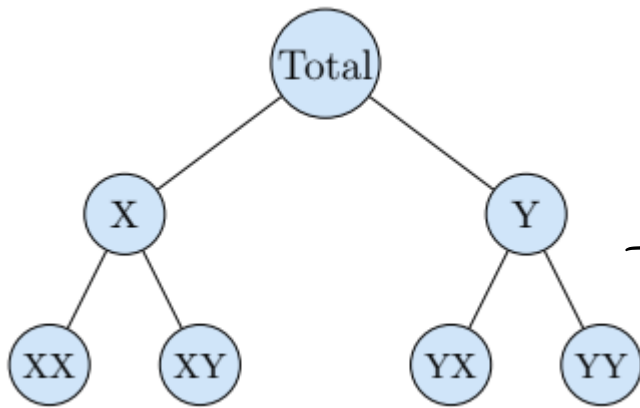
- where \mathbf{G} combines linearly (somehow) the forecasts to the lowest levels, so as to minimise $\tilde{\mathbf{y}}_h - \hat{\mathbf{y}}_h$.

How does THieF work?

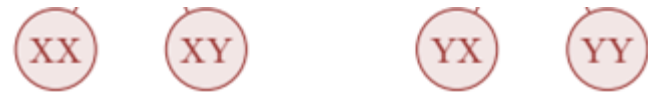
Let us understand this formula more

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

This just says combine all forecasts of the hierarchy using the weights in \mathbf{G} to obtain combined forecasts for the lowest level

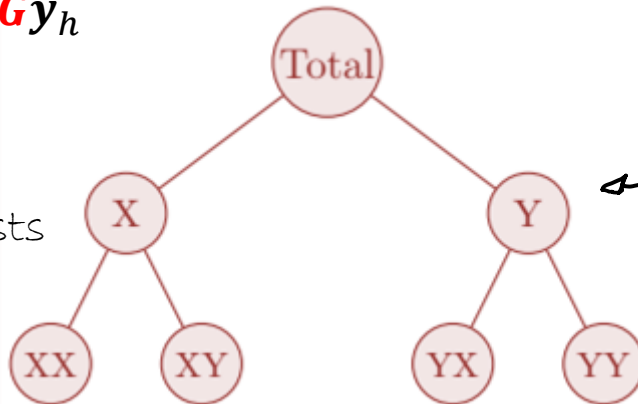


$$\mathbf{G} \hat{\mathbf{y}}_h$$



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

And this re-aggregates the combined forecasts



$$\mathbf{S}$$

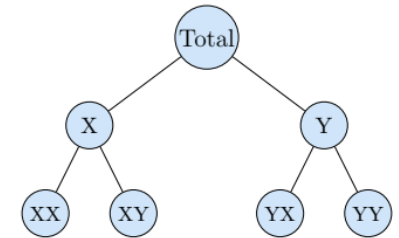
Bonus: we achieved coherency of forecasts. That is, the sum of the lower level agrees with the forecasts of the upper levels

Estimating G

We can show (it is a regression): $G = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$, where W_h^{-1} is the variance-covariance matrix of h-step ahead errors \rightarrow only unknown to achieve coherent forecasts is $W_h \rightarrow$ But getting this is difficult, so we rely on approximations.

Some of the more successful attempts:

- Assume homoscedasticity across everything: $W_{OLS} = I_m$
- Assume proportional increase in variance: Structural scaling.
- Assume no cross-effects: Variance scaling.
- Let the data speak using the full covariance matrix, with shrunk off-diagonals.

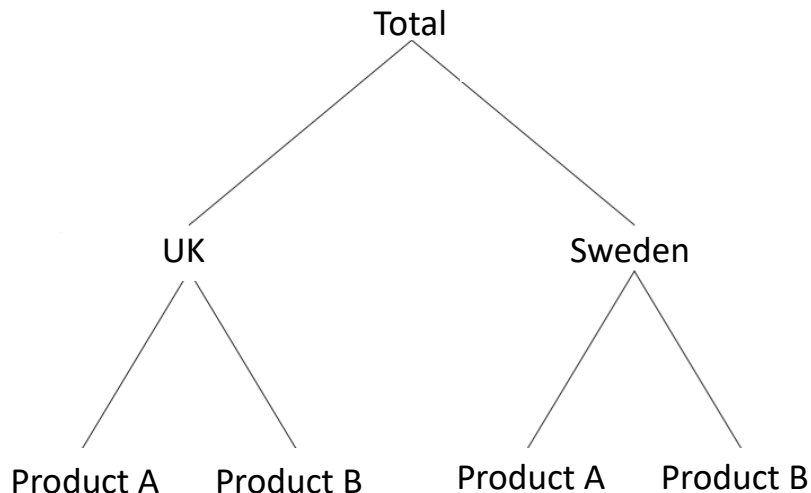


Structural scaling	Variance scaling	MinT shrinkage ($\rho_{i,j} \rightarrow 0$)
$\begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \hat{\sigma}_{Tot}^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\sigma}_X^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}_Y^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\sigma}_{XX}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}_{XY}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{\sigma}_{YX}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{\sigma}_{YY}^2 \end{bmatrix}$	$\begin{bmatrix} \hat{\sigma}_{Tot}^2 & \hat{\rho}_{Tot,X} & \hat{\rho}_{Tot,Y} & \hat{\rho}_{Tot,XX} & \hat{\rho}_{Tot,XY} & \hat{\rho}_{Tot,YX} & \hat{\rho}_{Tot,YY} \\ \hat{\rho}_{X,Tot} & \hat{\sigma}_X^2 & \hat{\rho}_{X,Y} & \hat{\rho}_{X,XX} & \hat{\rho}_{X,XY} & \hat{\rho}_{X,YX} & \hat{\rho}_{X,YY} \\ \hat{\rho}_{Y,Tot} & \hat{\rho}_{Y,X} & \hat{\sigma}_Y^2 & \hat{\rho}_{Y,XX} & \hat{\rho}_{Y,XY} & \hat{\rho}_{Y,YX} & \hat{\rho}_{Y,YY} \\ \hat{\rho}_{XX,Tot} & \hat{\rho}_{XX,X} & \hat{\rho}_{XX,Y} & \hat{\sigma}_{XX}^2 & \hat{\rho}_{XX,XY} & \hat{\rho}_{XX,YX} & \hat{\rho}_{XX,YY} \\ \hat{\rho}_{XY,Tot} & \hat{\rho}_{XY,X} & \hat{\rho}_{XY,Y} & \hat{\rho}_{XY,XX} & \hat{\sigma}_{XY}^2 & \hat{\rho}_{XY,YX} & \hat{\rho}_{XY,YY} \\ \hat{\rho}_{YX,Tot} & \hat{\rho}_{YX,X} & \hat{\rho}_{YX,Y} & \hat{\rho}_{YX,XX} & \hat{\rho}_{YX,XY} & \hat{\sigma}_{YX}^2 & \hat{\rho}_{YX,YY} \\ \hat{\rho}_{YY,Tot} & \hat{\rho}_{YY,X} & \hat{\rho}_{YY,Y} & \hat{\rho}_{YY,XX} & \hat{\rho}_{YY,XY} & \hat{\rho}_{YY,YX} & \hat{\sigma}_{YY}^2 \end{bmatrix}$

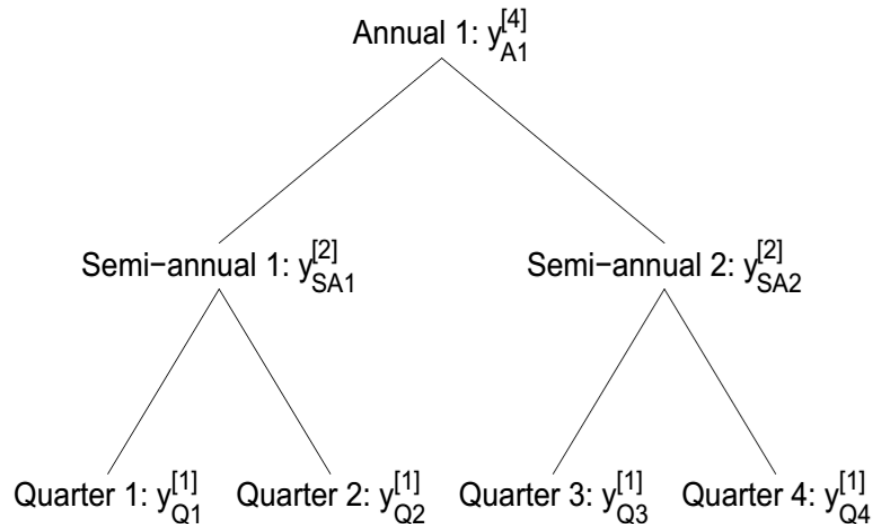
Hierarchical Forecasting

The “temporal” is just a flavour. We can have any sort of hierarchies. We call the other hierarchies cross-sectional.

Cross-sectional hierarchy

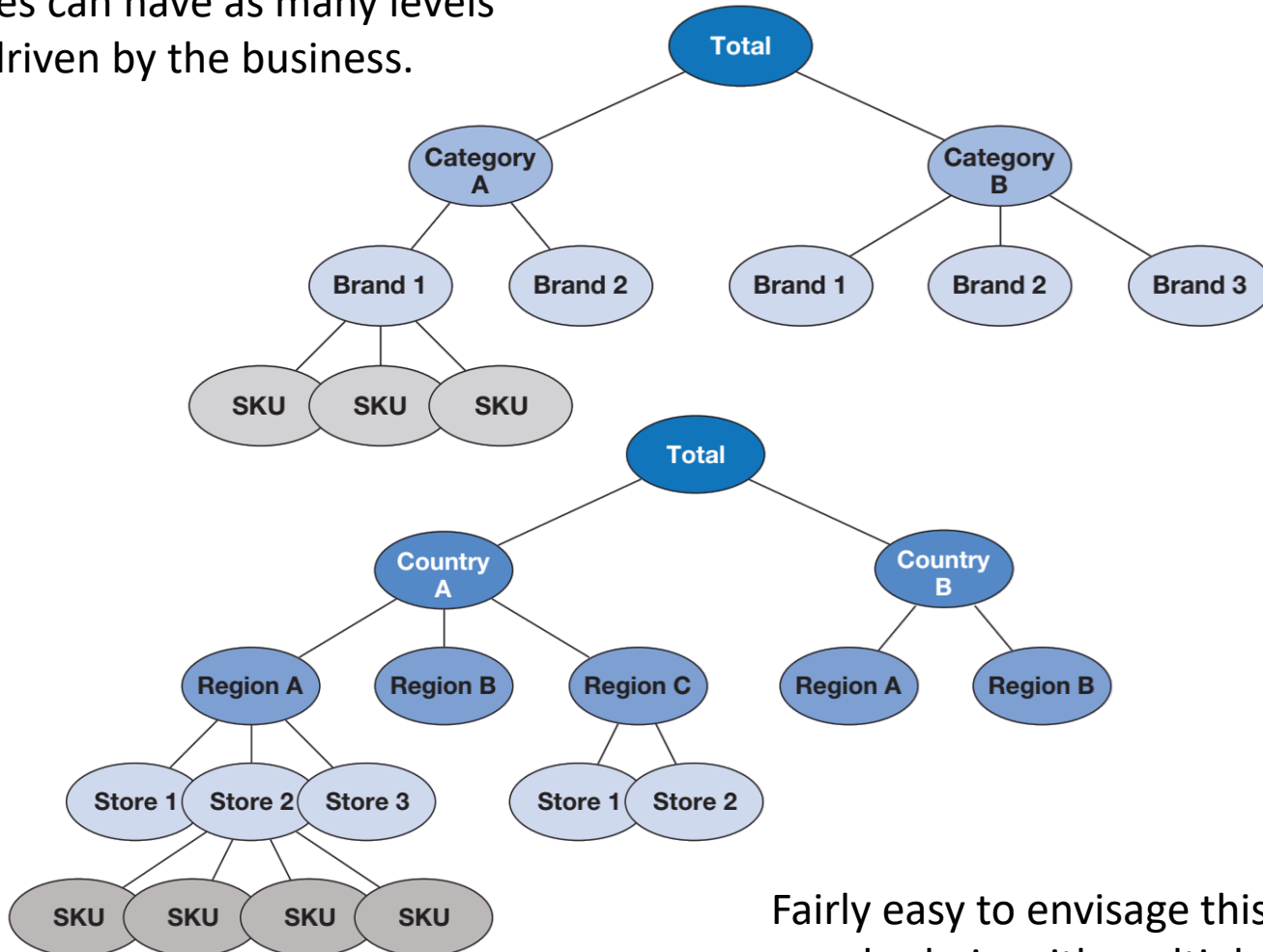


Temporal hierarchy



Hierarchical Forecasting

Our hierarchies can have as many levels as we want, driven by the business.



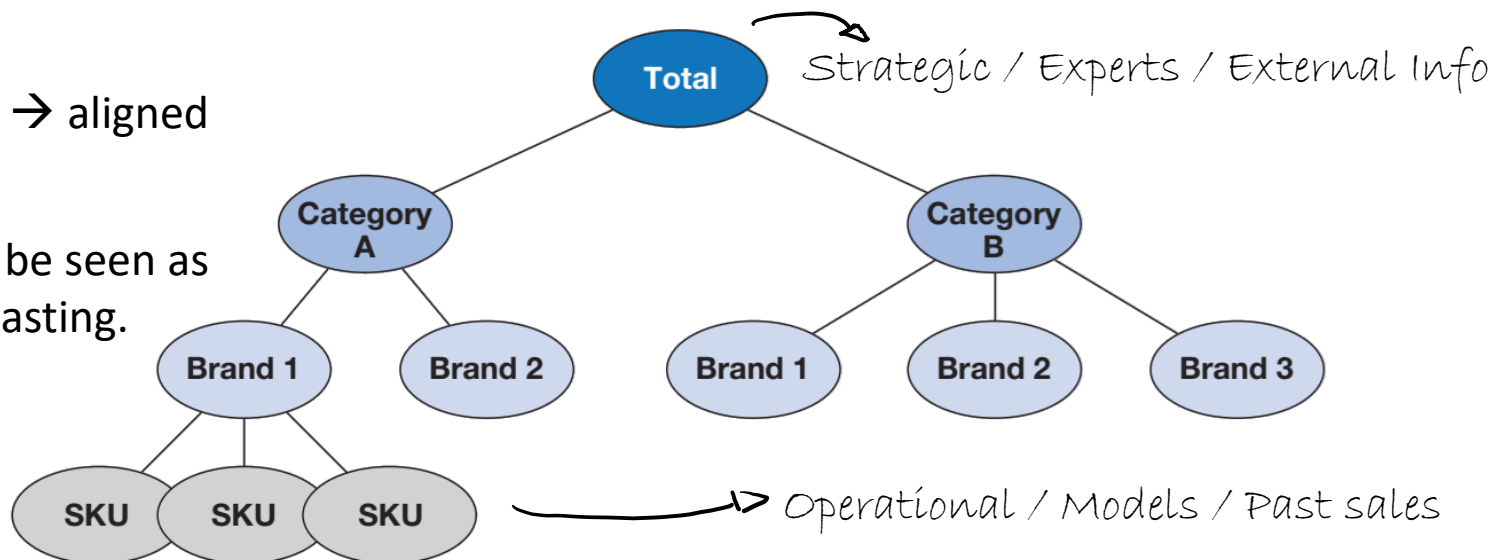
Fairly easy to envisage this across a supply chain with multiple stakeholders

A classic business problem

- Companies rely on forecasts to support decision making at different levels and functions.

Level	Horizon	Scope	Forecasts	Methods	Information
Operational	Short	Local	Way too many	Statistical	Univariate/Hard
Tactical	Medium	Regional	↕	↕	↕
Strategic	Long	Global	Few expensive	Experts	Multivariate/Soft

- The challenge: Forecasts must be aligned.
- Aligned forecasts → aligned decisions.
- The problem can be seen as hierarchical forecasting.

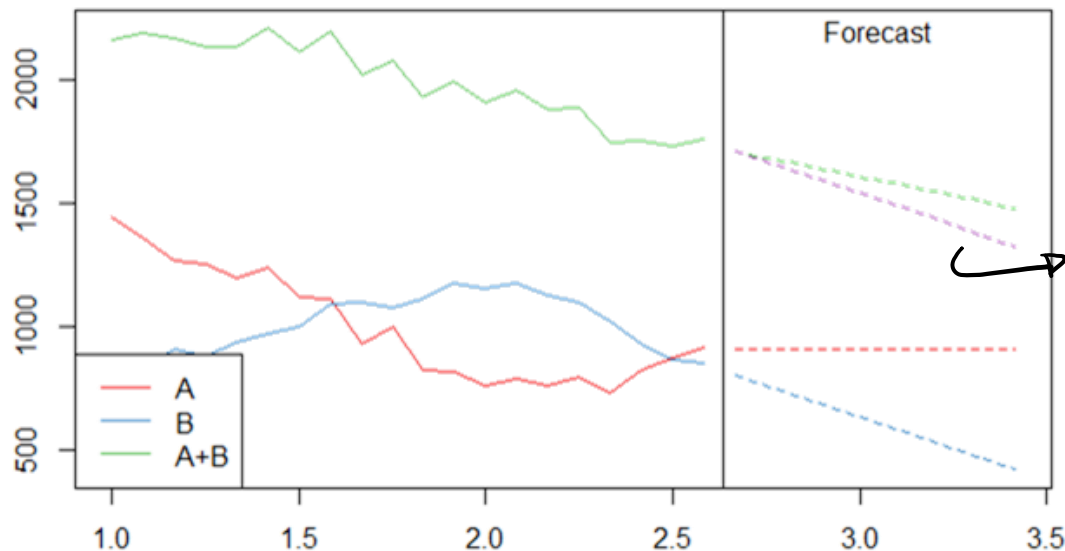


Coherent Forecasts

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!

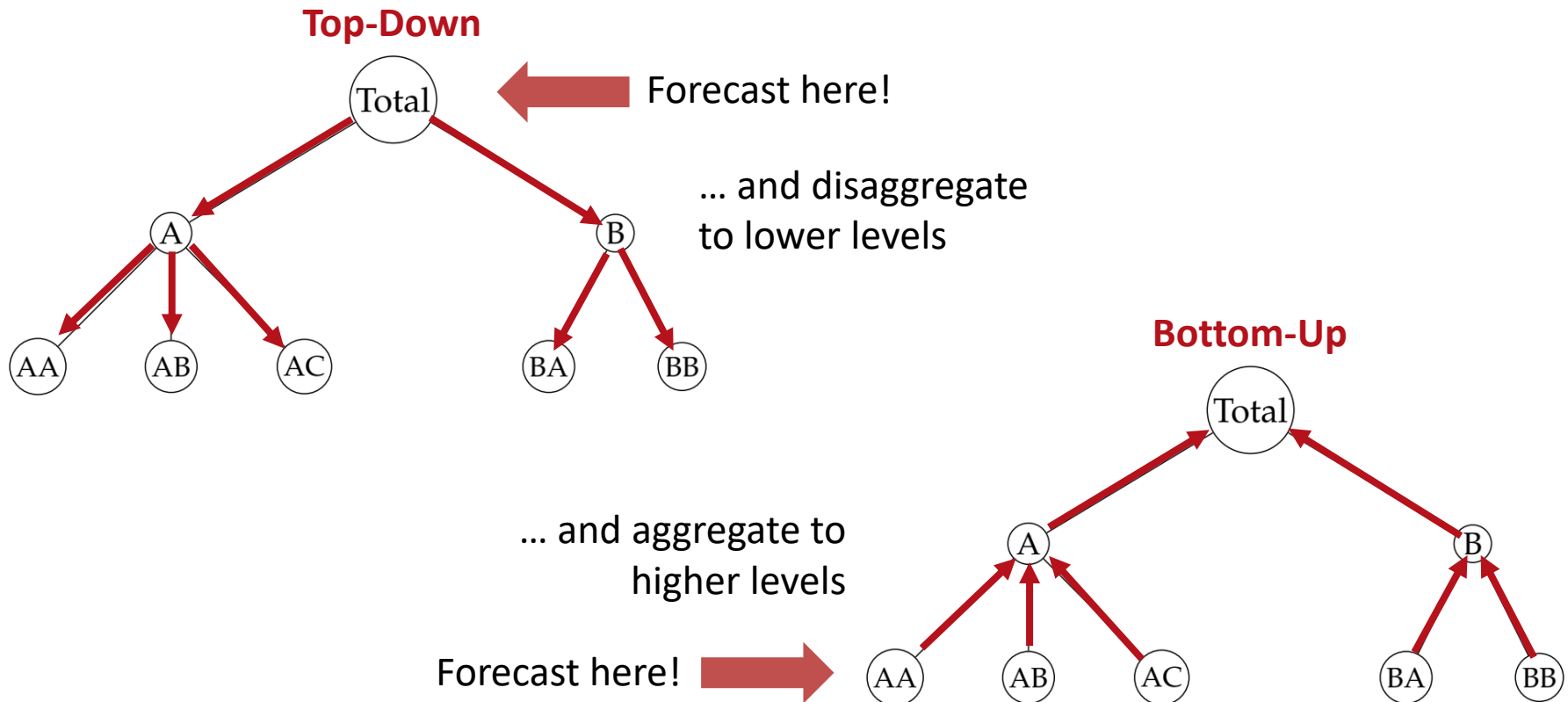


$F(A+B)$ and $F(A) + F(B)$ will typically be different, we need to impose equality (coherency of forecasts).

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

Hierarchical Forecasting Archeology

The two principal approaches to achieve hierarchical forecasts have been the **Top-Down** and **Bottom-Up**.



Key limitations:

- No forecast combination
- Rely on a few models (risky)

Do we need coherent forecasts?

The argument of alignment historically pushed towards coherent forecasts

- Top-Down & Bottom-Up only achieve coherency
- Reconciliation based hierarchical forecasts imply a forecast combination and therefore can offer accuracy benefits as well – even if coherency is a gimmick.

- A few questions arise:

Q For temporal the combination argument is straightforward, how does it play on cross-sectional hierarchies?

A Two elements are at work here. Combination is shrinkage, so you eliminate extremes in your forecasts. The other part comes from the **G** matrix, that introduces cross-series contemporaneous relations, i.e. cross-effects between products!

Q That **G** matrix is pretty big, you said before estimation introduces uncertainty, how does it work?

A That is a very interesting part of hierarchical forecasting, most research is here.

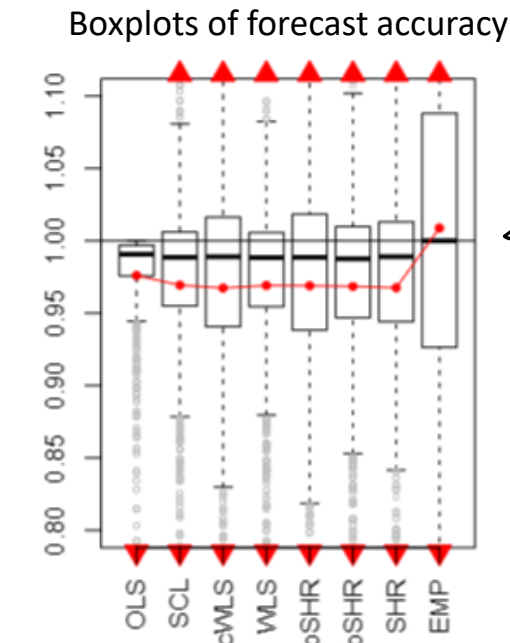
Implications of G for performance

I presented one methodology to obtain G . For its derivation we need coherency.

- Coherency introduces a set of linear constraints. This keeps the solution “under control”.

$G = (S'W_h^{-1}S)^{-1}S'W_h^{-1}$ this is what all these S do in there, and that term is fixed.

- But the approximation of the covariance W can still cause issues.



Anything below 1 better than base forecasts

As the complexity increases:

- Up to a point accuracy on average increases.
- Variability of forecast errors increases. Too much estimation remains a problem.
- Simpler approximations simply introduce restrictions to mitigate this.

Simpler W \longleftrightarrow More complex W

Implications of G for performance

Any alternatives?

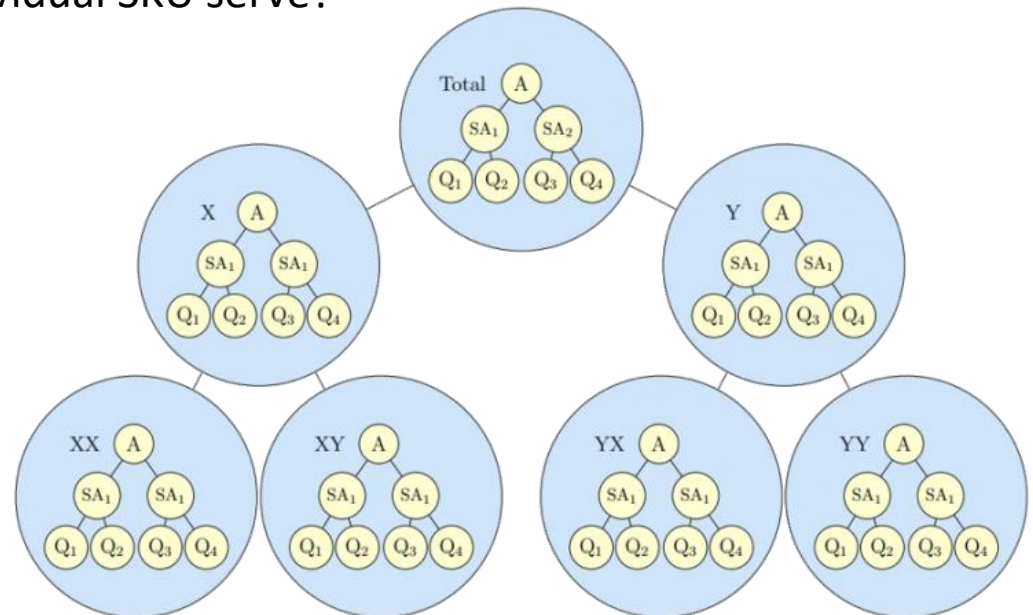
- MAPA just uses a mean/median and works quite well! There is nothing to restrict us from doing the same with Hierarchical Forecasting.
- Throw coherency requirements away (or not) and estimate somehow combination weights
 - Fixed weights.
 - Estimated weights, with restrictions:
 - using standard linear combinations
 - allowing for flexible forms \rightarrow rely on ML/AI to do the combination
 - Use the implied geometry of the hierarchical problem to obtain more efficient combinations (more efficient = estimate better with same amount of data).
- Or use coherency but with some slack? Allow for small violations of coherency, but do not throw it completely away.

Cross-temporal Hierarchies

The two sides of hierarchical forecasting have their limitations as well:

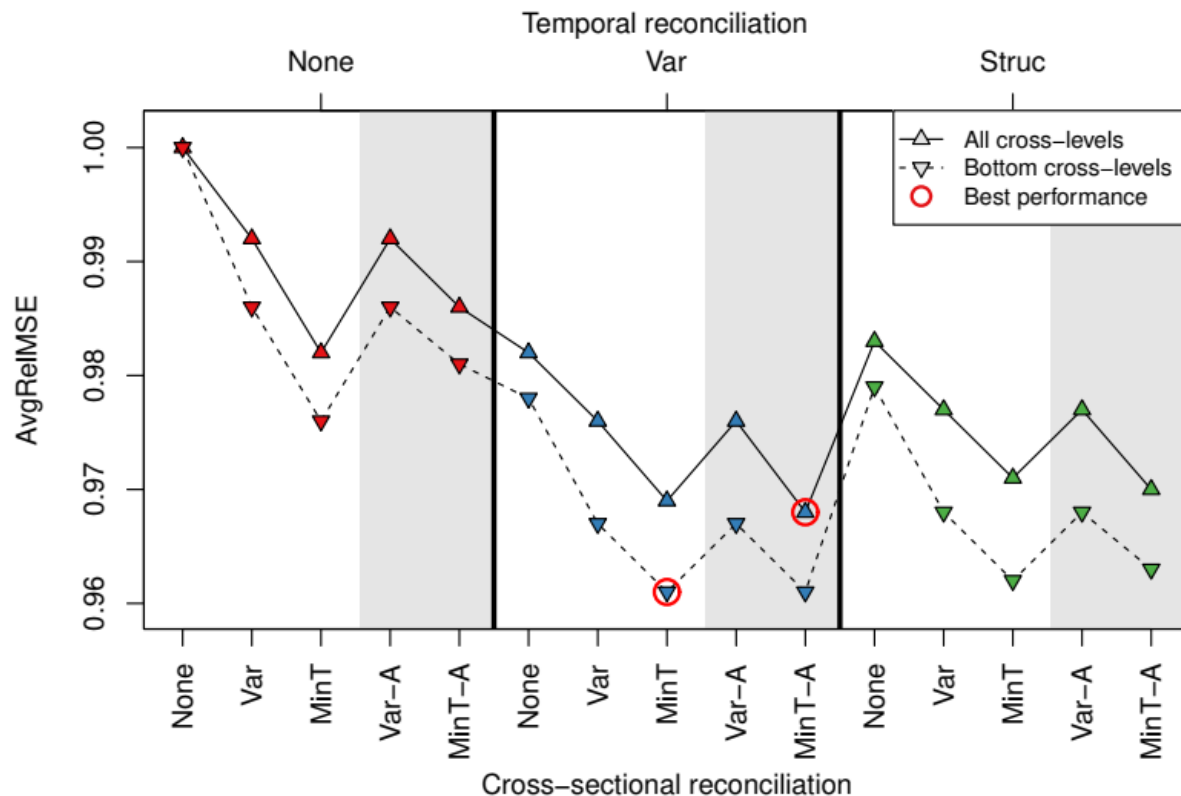
- Cross-sectional: we forecast a SKU at individual (lowest) and total sales at top level, i.e. do we need total sales at short term forecasting, which is fine for SKU level?
- Temporal: we forecast a SKU at short-term and long-term, i.e. sales of X at size Y for short-term (operational) and long-term (strategic) horizons. What decision would long-term forecasting of sales of individual SKU serve?

What we need is to combine both using **cross-temporal hierarchies**. The same formulation/logic applies.



Example: Tourism Forecasting

- Total to regional monthly tourism flows for Australia. 111 series, spanning 10 years.
- Test set 6 years, with rolling origin evaluation. Relative RMSE (<1 better) to base forecast.
- Forecast using exponential smoothing.



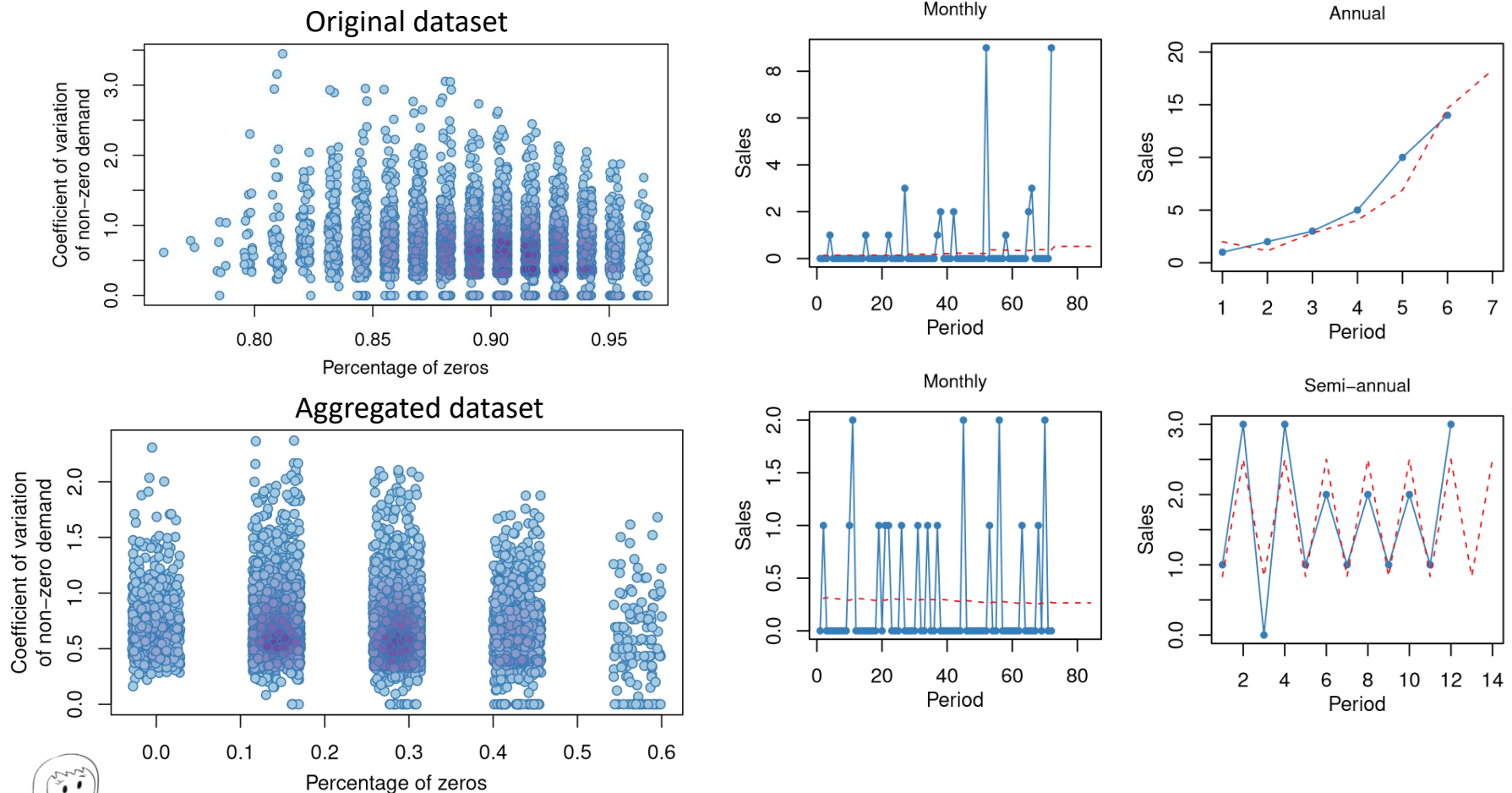
Figures in grey are cross-temporally coherent

Observe: temporal provides larger benefit



Temporal hierarchies & intermittent demand

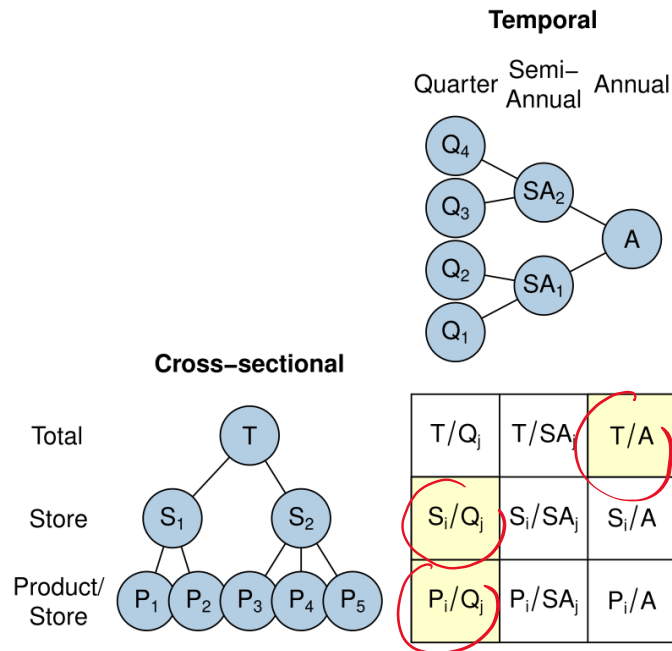
- Nobody told us to solve the problem as it was handed to us!



Hierarchies: Back to the business problem

Match forecasting to the supported decision!

- What level of aggregation (SKU? Brand? Market?)
- What data frequency? (Daily? Monthly? Quarterly? Annual?)



Decision relevant levels

The rest are statistical devices (forecast combination)

- Are the decisions equally important?
- What information is available at each level?
- What models at each level?
- What evaluation metrics?

Concluding remarks

- Hierarchical approaches can help meld information supporting different decisions and planning horizons to achieve an aligned view of the future throughout organisations.
 - Hierarchical forecasting is at its core a forecast combination with tricks
 - Cross-sectional, temporal and cross-temporal flavours. They can also be used as statistical devices to increase accuracy.
 - Organisational implications of hierarchical forecasting: "one-number" forecast, different functions/plans are based on a common view of the future.
- Hierarchical forecasting is model independent
 - Use with conventional models
 - Machine learning/AI
 - Multivariate models (AI or VAR models for complete levels of the hierarchy)
 - Expert adjustments
- For almost everything I mentioned there is an open-source R package ☺
 - MAPA, thief, hts (examples at my blog; <http://nikolaos.kourentzes.com>)

Thank you for your attention!
Questions?

Nikolaos Kourentzes

email: nikolaos@kourentzes.com

twitter [@nkourentz](https://twitter.com/nkourentz)

Blog: <http://nikolaos.kourentzes.com>