Model uncertainty in hierarchical forecasting

Nikolaos Kourentzes

Lancaster University Management School, UK

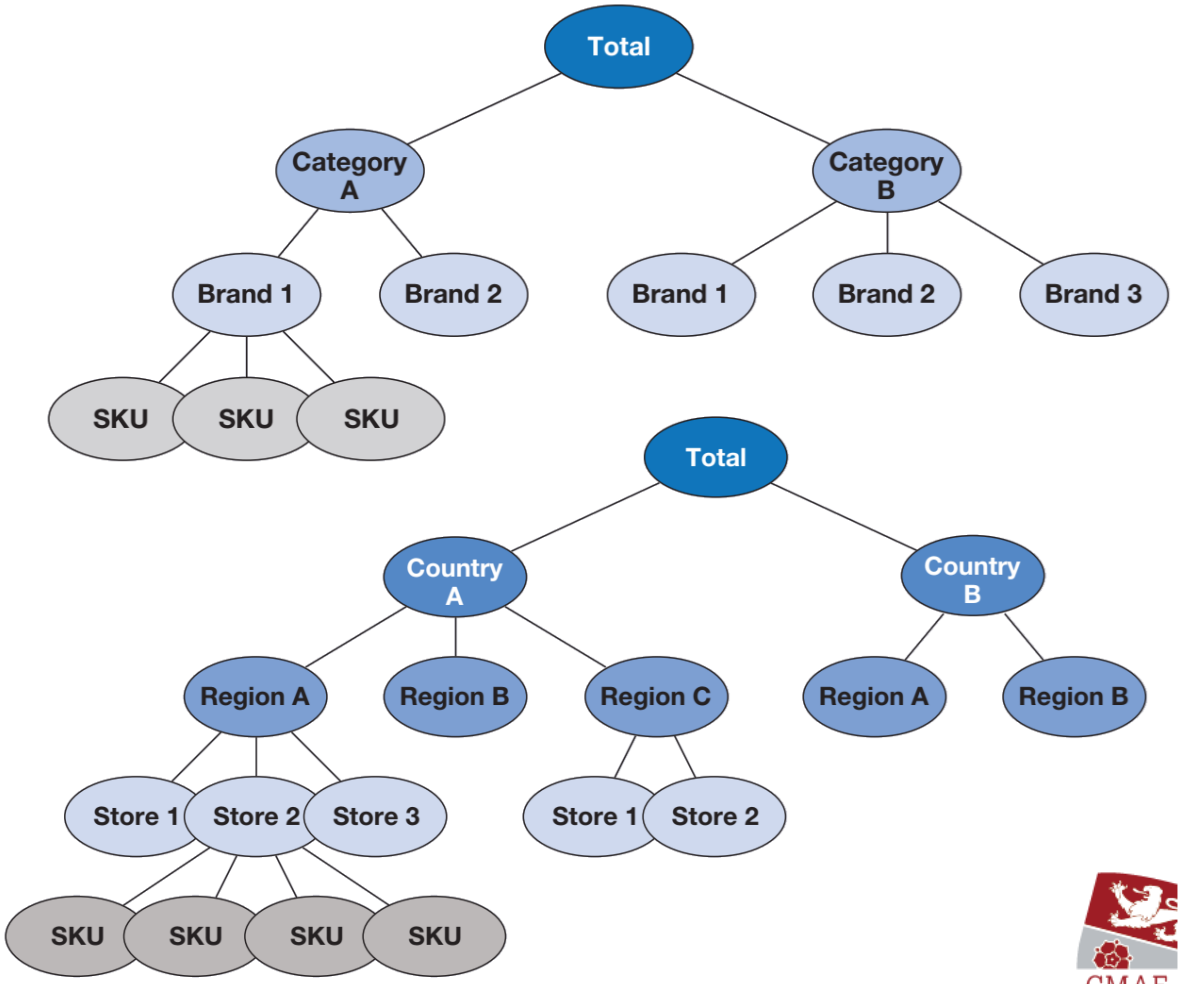
OR60 Lancaster
13/09/2018

Hierarchical forecasting

In practice we often need consistency in our forecasts across:

- product hierarchies;
- market segments;
- planning levels;
- etc.

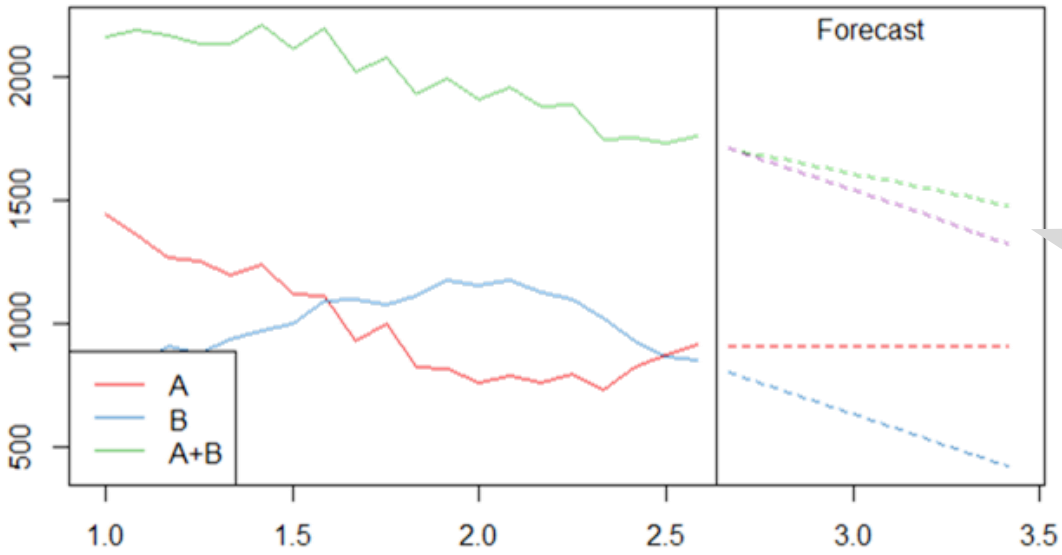
- At each level, the time series carry different (apparent) information and need different modelling. Forecasts will not add up.
- Expert adjustments of forecasts can happen independently, further misaligning forecasts.
- Hierarchical forecasting attempts to impose consistency again.



CMAF

The forecast consistency problem

Suppose we have to forecast two items A and B, which are variants of the same product.



Reconciling this difference imposes the aggregation constraint, and will force changes to the forecasts of A and B.

CMAF

Hierarchical and Grouped series

Note that in the previous examples we assumed that there was one way to get from the lowest level to the highest level, i.e. a single **hierarchy**.

This is not generally true, as there may be many ways to construct the hierarchies, for example:

- SKU → Product group → Total
- SKU → Store → Total
- SKU → Country → Total
- etc.

We can represent all possible pathways from the disaggregate data to the top level aggregate data using the so called **grouped** time series.

CMAF

Hierarchical and Grouped series

An example from a policy problem, managing unemployment is as follows:

- **Sixteen** unemployment time series across the following dimensions:
  - Age {15-24; 25 and above}
  - Country {Denmark; Finland; Norway; Sweden}
  - Gender {Female; Male}

- From these we can construct multiple hierarchies, resulting in 29 unique aggregate series (16 + 29 = 45 series in total).

	Top Level	Level 1	Level 2	Level 3
Hierarchy 1	Total	Country	Gender	Age
Hierarchy 2	Total	Country	Age	Gender
Hierarchy 3	Total	Gender	Country	Age
Hierarchy 4	Total	Gender	Age	Country
Hierarchy 5	Total	Age	Country	Gender
Hierarchy 6	Total	Age	Gender	Country

CMAF

Hierarchical forecasting methods

In the literature there are a number of traditional hierarchical forecasting approaches (Fliedner, 2001; Ord et al., 2017):

- **Top down**: forecast the top level of the hierarchy and disaggregate.
  - Pros: At an aggregate level things become easier to model (typically!)
  - Cons: (i) dissagregation is not trivial; (ii) leads to biased forecasts for lower levels (Hyndman et al., 2011); (iii) trust a single model/method for the whole hierarchy; (iv) typically does not perform well.
- **Bottom up**: forecast the most disaggregate level.
  - Pros: Full view of the details
  - Issues: (i) model at a very difficult level (but also informative level?); (ii) actually it is just that: if we could model the bottom level optimally then there is no hierarchical problem, but this is very difficult to do (e.g. intermittent data, missing or masked expanatory variables, very noisy, etc.).
- **Middle out**: mix of both, compromise between complexity and information available.
- Top down and middle out cannot do grouped hierarchies.

CMAF

Optimal combinations

Hyndman et al. (2011) and Athanasopoulos et al. (2009) introduced a very neat notation for the hierarchical problem that permits more interesting solutions

The forecast consistency problem is the following:

$$\tilde{e}_{ijt} = \hat{y}_{ijt} - \tilde{y}_{ijt}$$

Reconciliation error, the difference between the initial forecasts $\hat{y}_{ijt}$ (for level i, jth series at time t) and the yet unknown reconciled forecasts $\tilde{y}_{ijt}$

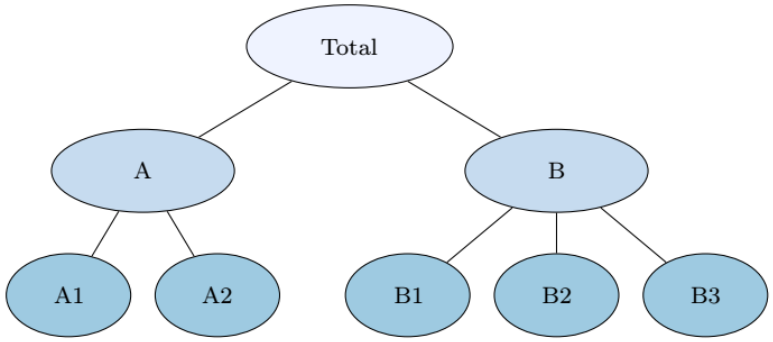
If we minimize $\tilde{e}_{ijt}$ we reach forecasts that are consistent across the whole hierarchy.

CMAF

Optimal combinations

Introduce the summing matrix $\boldsymbol{S}$ to codify the hierarchy (single or grouped)

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ & & I_n & & \end{bmatrix}$$



With $\boldsymbol{S}$ any hierarchy can be encoded, but always the lower part is a diagonal (lowest disaggregate level) and top row is a row of ones (total sum, i.e. top level).

The complete hierarchy can be now compressed in the $\boldsymbol{S}$ matrix and the lowest level time series $\boldsymbol{y}_{Bt} = (y_{B1t}, \dots, y_{Bkt})'$:

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{y}_{Bt}$$

CMAF

Optimal combinations

We can devise a reconciliation model as:

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{c} + \boldsymbol{w}\hat{\boldsymbol{y}}_t + \tilde{\boldsymbol{e}}_t$$
$$\tilde{\boldsymbol{y}}_t = \boldsymbol{c} + \boldsymbol{w}\boldsymbol{S}\hat{\boldsymbol{y}}_{Bt} + \tilde{\boldsymbol{e}}_t$$

This just shows that it is really a bottom level forecasting problem masked as a hierarchical problem

For which we can observe a few things:

- The vector of constants $\boldsymbol{c}$ has to be zeros, otherwise we would get nonzero $\tilde{\boldsymbol{y}}_t$ even when there are no forecasts – alternatively, if you prefer, because we expect $\hat{\boldsymbol{y}}_{Bt}$ to be a collection of unbiased forecasts.
- Matrix $\boldsymbol{w}$ produces a linear combination of the forecasts ($\hat{\boldsymbol{y}}_t$), suggesting that the cross-connections between all series are potential useful, as all forecasts are combined.
- Minimising $\tilde{\boldsymbol{e}}_t$ would give us the minimum change of $\hat{\boldsymbol{y}}_t$ to $\tilde{\boldsymbol{y}}_t$, but this is not an OLS problem, as $\hat{y}_{ijt}$ may have different processes, and surely different mean and variance between aggregation levels.

CMAF

Optimal combinations

We can see this as a GLS problem (Hyndman et al., 2011):

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{\Sigma}^{\circ}\boldsymbol{S}\right)^{-1}\boldsymbol{S}'\boldsymbol{\Sigma}^{\circ}\hat{y}_t$$

where $\boldsymbol{\Sigma}^{\circ}$ is generalised inverse of the variance-covariance matrix of the reconciliation errors.

Wickramasuriya et al. (2018) showed that $\boldsymbol{\Sigma}$ is non-identifiable (the intuitive reason is that it is a chicken and the egg problem) and propose to instead use $\boldsymbol{W_h}$, the variance-covariance of the t+h forecast errors (they show that forecast errors is indeed a reasonable replacement):

$$\boldsymbol{W}_h = E[\hat{\boldsymbol{e}}_{\boldsymbol{t+h}}\,\hat{\boldsymbol{e}}'_{\boldsymbol{t+h}}]$$

- But getting the t+h errors is not trivial, so many approximations have been proposed.

CMAF

Approximations of $\boldsymbol{W}_h$

On the approximations of $\boldsymbol{W}_h$:

- Most approximations attempt to make our life easy in two ways: (i) avoid doing the implied cross-validation exercise; (ii) avoid estimating the full variance-covariance matrix (and inversing it).
- **OLS:** $\boldsymbol{W}_h = \mathbf{I}$, (Athanasopoulos et al., 2009; Hyndman et al., 2011). This is just wrong, as it s has unrealistically restrictive assumptions (everything has equal variances).
- **WLS:** $\boldsymbol{W}_h = \mathrm{diag}(\mathrm{MSE})$**,** (Hyndman et al., 2006, Athanasopoulos et al., 2017). Trivial to calculate, assumes that in-sample t+1 errors are representative and no-cross effects between nodes of the hierarchy. Implicitly considers the quality of model fit.
- **Structural**: $\boldsymbol{W}_h = \mathrm{diag}(\boldsymbol{S1})$, (Athanasopoulos et al., 2017). This considers only the structure of the hierarchy, assuming additivity of variance of the errors. Can be generalised to $\boldsymbol{W}_h = \mathrm{diag}(\boldsymbol{S1})^{\boldsymbol{p}}$ with minimal performance differences.
- **Empirical**: $\boldsymbol{W}_h = \boldsymbol{W}_1$, (Wickramasuriya et al., 2018), which assumes that the empirical covariance of the t+1 forecast errors adequately describes the t+h forecast errors, i.e. everything increases proportionally.
- **MinT:** $\boldsymbol{W}_h = shrink(\boldsymbol{W}_1)$**,** (Wickramasuriya et al., 2018), which recognises that given limited sample and large hierarchies $\boldsymbol{W}_1$ is difficult to estimate.

CMAF

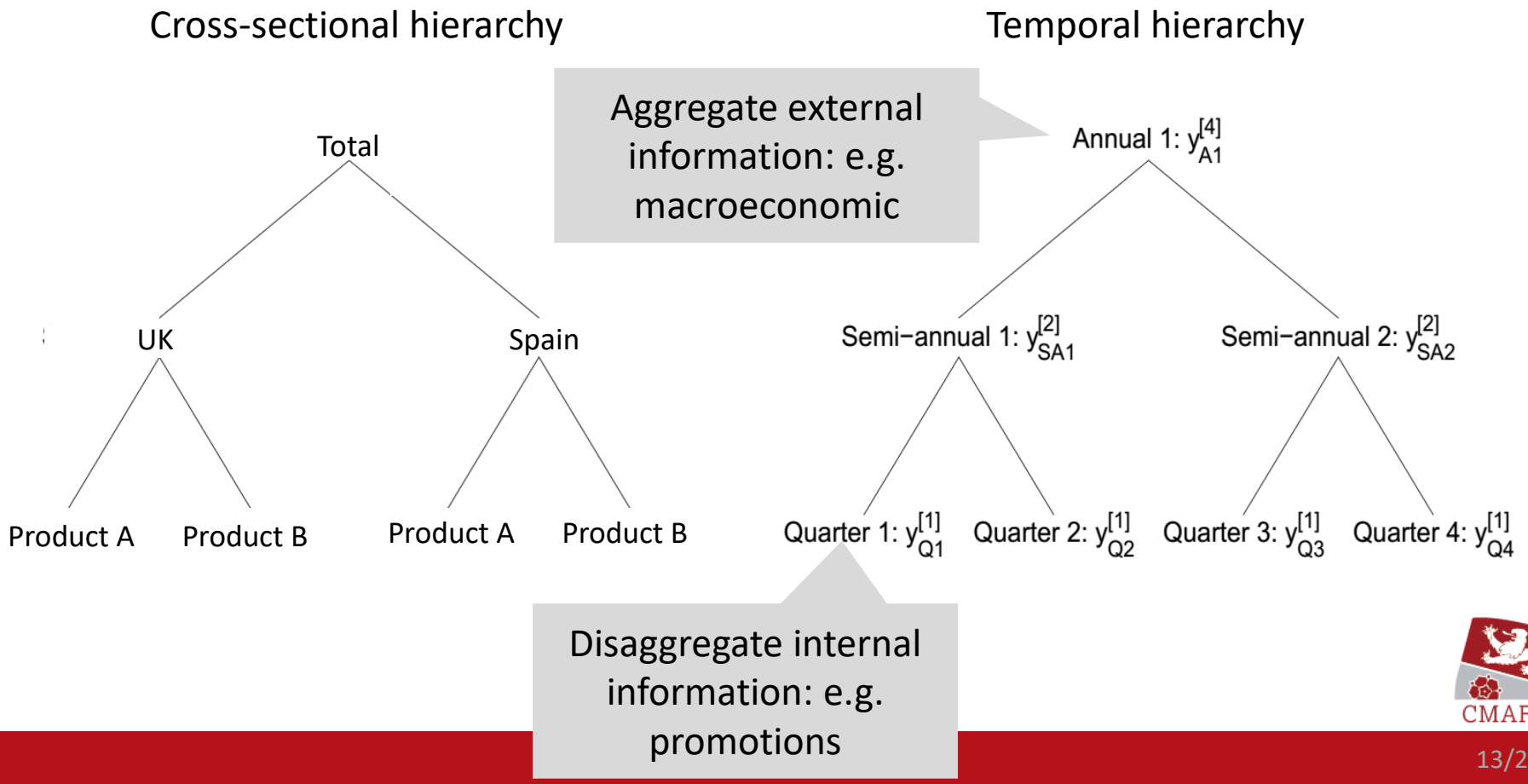
Optimal combinations

- The optimal combinations framework encompasses Top-down, Bottom-up and Middle-out, as these are specific solutions for $\boldsymbol{w}$ (the linear combination weights).
- Using this we can show that:
  - Top-down is always biased and will generally have poor performance.
  - Bottom-up is unbiased and its performance is highly dependent on the variance of the forecast errors of the bottom level → often the bottom level is rather difficult to model.
  - Optimal combinations hedge modelling risk → each node forecast is a linear combination of all the forecasts, therefore mitigating the model misspecification risk.
  - BUT (open research question) if the hierarchy is non-unique, and $\boldsymbol{S}$ is crucial for the calculation of the combination weights, what does this imply for the quality of the hierarchical forecast? *Hypothesis: although hierarchies are not unique, most are dominated specifications of some identifiable hierarchy → ask me next year.*
  - **Empirically, optimal combinations perform very well** (Athanasopoulos et al., 2017; Wickramasuriya et al., 2018; ignore first papers, they used wrong $\boldsymbol{W}_h$).

CMAF

A different flavour of hierarchies: Temporal

- Using the optimal combination hierarchical approach, Athanasopoulos et al. (2017) proposed a generalisation of MAPA (Kourentzes et al., 2014) as a general framework of using **multiple temporal aggregation** to model time series.



CMAF

Temporal Hierarchies

Decisions need to be aligned:

- Operational short-term decisions
- Tactical medium-term decisions
- Strategic long-term decisions

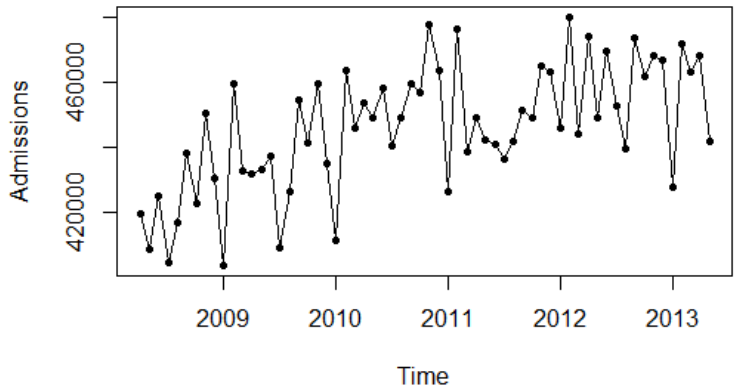
Shorter term plans are **bottom-up** and based mainly on **statistical forecasts** & expert adjustments.
Longer term plans are **top-down** and based mainly on **managerial expertise** factoring in unstructured information and organisational environment.

Given different sources of information (and views) forecasts will differ → plans and decisions not aligned.

Coherent forecasts across planning horizons can lead to less waste & costs, agility to take advantage of opportunities.

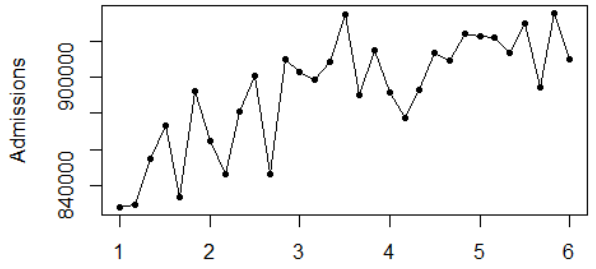
Temporal Aggregation

Consider some historical monthly sales series:

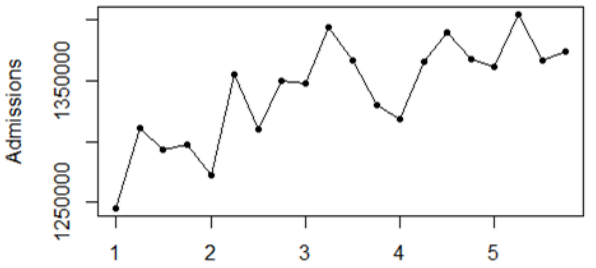


For a long term forecast, we could either produce multi-step ahead forecasts, or aggregate the data and produce single-step ahead forecasts for the long horizon directly:

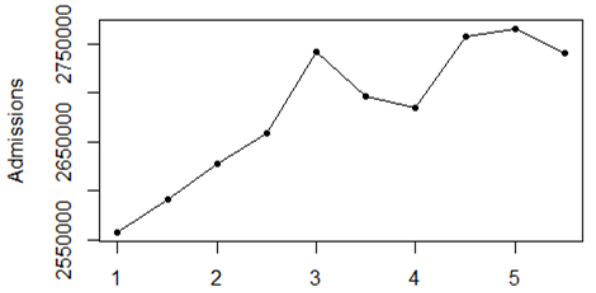
- 12 monthly forecasts vs. 1 yearly!



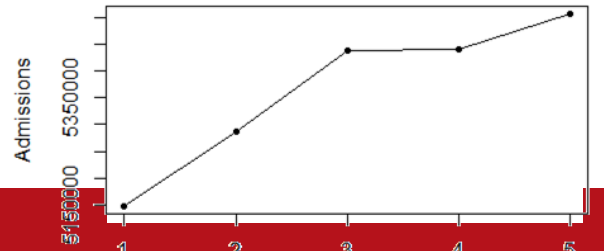
Bi-monthly



Quarterly



Half-annually

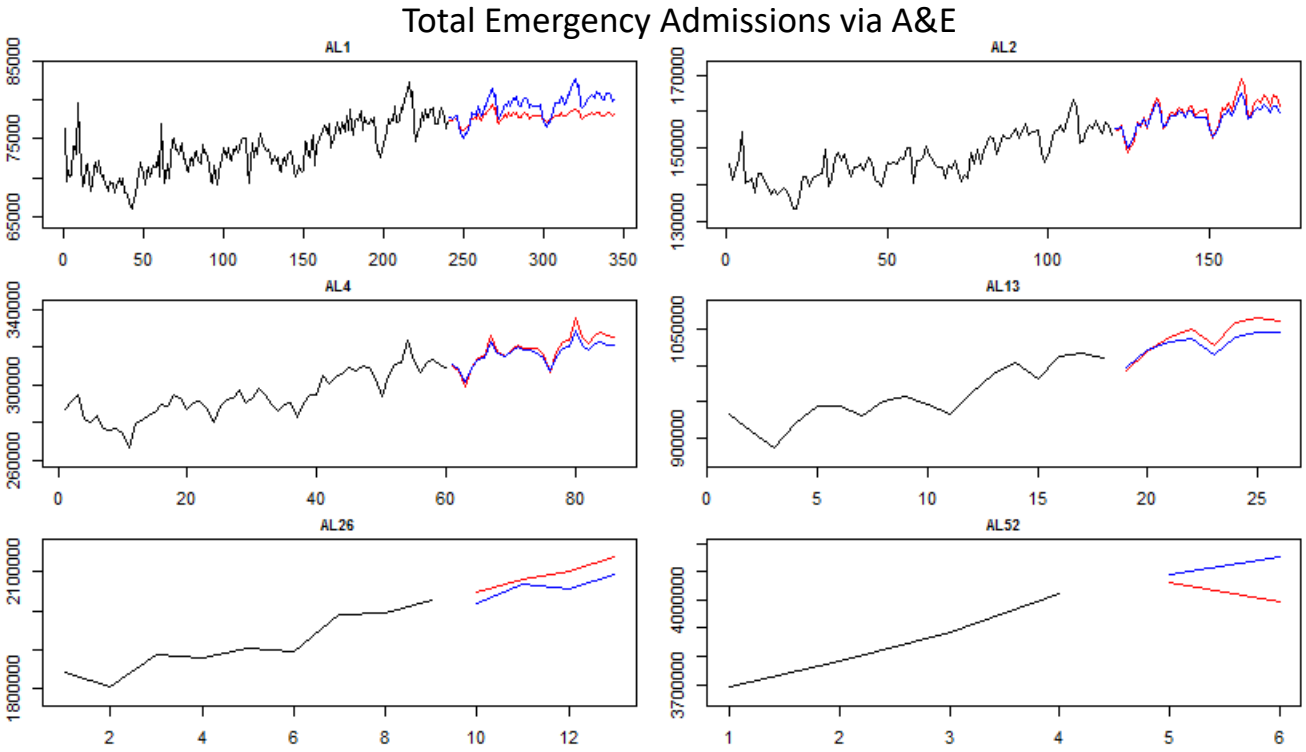


Annually

Temporal Aggregation

- Produce long term forecasts with multi-step predictions is risky: forecast errors accumulate!
  - Temporal aggregation can help to reduce the length of the forecast.
- What does temporal aggregation do to our data?
  - at an aggregate level trend/cycle is easy to distinguish.
  - at a disaggregate level high frequency elements like seasonality and promotions typically dominate.
- Arguably both disaggregate and aggregate views are useful. We can look at both and connect them in a hierarchical way.
- Temporal hierarchies (like MAPA) are very reliable in the face of modelling uncertainty:
  - Good forecasting performance (Kourentzes et al., 2014, Athanasopoulos et al., 2017);
  - Safer option (with relevant accuracy gains) than looking for an optimal aggregation level that econometricians have focused on for 30 years (Kourentzes et al., 2016).

Example: Predicting A&E admissions

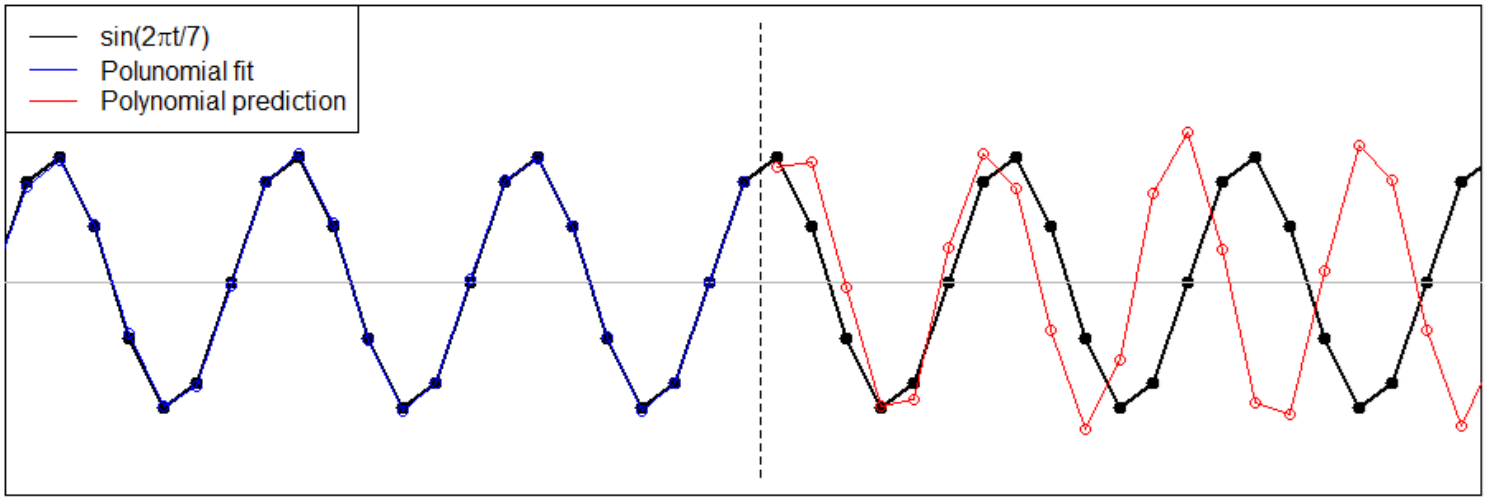


Red is the prediction of the base model – at each level separately
Blue is the temporal hierarchy forecasts

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

Back to approximations of W_h

- All approximations (with the exception of the structural) are based on in-sample quadratic errors. This is dangerous → overfitting.
- Low in-sample MSE means nothing about our out-of-sample modelling risk and forecast errors.



- Strong evidence of low correlation of in- and out-of-sample forecast errors. Barrow and Kourentzes (2016) showed a 10%+ underestimation of uncertainty when in-sample errors are used.

CMAF

My issue with uncertainty calculations

- Models are great, as they give you prediction intervals conditional on data and model parameters → but assume the model to be true!
  - We account for the inherent uncertainty in the data;
  - We account for estimation uncertainty (lately);
  - **We do not account for the uncertainty of picking the wrong model though!**
- t+1 in-sample errors would be representative of t+1 out-of-sample errors if the model was true. If not (so, in reality!) then minimising t+1 errors only results in an approximation of the behavior at t+1 and not at t+h, when h >> 1.
  - Minimising t+h likewise approximates the t+h behavior and not the t+1 (and is a new family of univariate shrinkage estimators!). Similarly in-sample do not correlate well with out-of-sample.
- Back to hierarchies: approximations of $\boldsymbol{W}_h$ suffer from the same underestimation of uncertainty.
  - Unless we cross-validate the out-of-sample errors, but this seriously limits available sample.

CMAF

Revising the approximation of $\boldsymbol{W}_h$

- What we are missing?
  - We want the weights of the linear combination $\boldsymbol{w}$ to account for:
    (i) Difficulty to forecast a specific node of the hierarchy;
    (ii) The level of the node in the hierarchy (mean and variance);
    (iii) The model uncertainty (over-fitting/under-fitting).
  - In-sample errors do not penalise model complexity (or overfitting).
    - Some practical issues: how do you reconcile a LASSO forecast with a MSE-optimal forecast? The MSE forecast will be wrongly overweighted! Sagaert et al. (2018) demonstrated the inventory gains for hierarchies with top level LASSO forecast with leading indicators and bottom level univariate exponential smoothing forecasts.

CMAF

Revising the approximation of W_h

- Balance model fitness and over-fitting, if only there was something that could do that…

  … but there is! Information criteria, such as AIC.

  Why we like AIC?

  (i) It respects that we do not have infinite data to waste on cross-validation

  (ii) It does not need hyperparameters (okay, go with BIC if you must, but remember that AIC is equivalent to t+1 out-of-sample cross-validated errors; Stone, 1977)

  Why we do not like AIC?

  (i) it does not tackle scale across time series well. So let us deal with that.

From the definition of AIC we can show that:

$e^{\frac{AIC}{n}} = L\frac{1}{e^{-2k/n}}$, so what we need for W_h is $e^{AIC/n}$. (Yes, 30 minutes for 1 line!)

Empirical evaluation

- Test the idea in two settings: (i) cross-sectional hierarchies; (ii) temporal hierarchies
- Benchmarks:
  - OLS approximation
  - MSE approximation – often the best
  - MinT (shrunk covariance; not possible for temporal) – most elegant
  - Structural reconciliation (i.e. weight by the inverse of the number time series in that level) – this makes a lot of sense in temporal, but not cross-sectional
- New:
  - Scaled AIC approximation
  - Cross-validated MSE approximation (Jeon et al., 2018)
  - Shrunk cross-validated covariance approximation

CMAF

Hierarchical forecasting

Cross-sectional hierarchies – AvgRelMAE results

Dataset	Base	OLS	MSE	CV	AIC	Structural	MinT	MinT-CV
ETS								
Unempl.	1.000	1.012	1.051	0.980	0.980	**0.979**	1.137	0.994
Tourism	1.000	1.015	1.078	0.985	0.986	0.983	2.491	**0.977**
A&E	1.000	1.112	1.062	0.993	**0.991**	0.998	18.016	1.002
Infant	1.000	1.015	1.116	**0.955**	0.961	0.961	1.612	0.939
Overall	1.000	1.038	1.076	0.978	**0.979**	0.980	3.012	0.978
ARIMA								
Unempl.	1.000	1.011	0.964	0.964	**0.956**	0.985	0.973	0.971
Tourism	1.000	1.035	0.979	0.984	0.974	0.978	**0.972**	0.981
A&E	1.000	1.074	0.992	0.995	**0.991**	0.995	1.001	**0.991**
Infant	1.000	0.989	0.987	0.987	**0.985**	0.981	1.022	1.013
Overall	1.000	1.027	0.980	0.982	**0.976**	0.985	0.992	0.989

- CV and AIC results are very close! But AIC is trivial to calculate.

Datasets (picked to give me a hard time – no M3 here):

- Unemployment: 45 time series, 312 observation, 120 test, horizon t+12;
- Tourism: 89 time series, 36 observations, 12 test, t+6;
- A&E: 413 time series, 27 observations, 10 test, t+4;
- Infant mortality: 27 time series, 71 observations, 30 test, t+5

Hierarchical forecasting

Temporal hierarchies – AvgRelMAE results

Dataset	Base	OLS	MSE	Structural	AIC
ETS					
Unemployment	1.000	1.050	1.025	0.991	**0.982**
Tourism	1.000	0.996	0.998	**0.989**	0.995
A&E	1.000	0.983	1.006	**0.972**	0.989
Overall	1.000	1.009	1.010	**0.984**	0.989
ARIMA					
Unemployment	1.000	0.990	0.939	0.934	**0.926**
Tourism	1.000	0.983	0.984	**0.981**	1.002
A&E	1.000	0.997	0.962	**0.949**	0.995
Overall	1.000	0.990	0.961	**0.954**	0.974

- Structural is very hard to beat: it takes into account the special hierarchical structural, even though it is tremendously simplistic. It works very well in cross-sectional data as well.
- Otherwise, AIC is the way, as it is very sceptical to complex models.

CMAF

Concluding remarks

- Hierarchies are very common – even with a single time series (temporal hierarchies).
- However, **hierarchical forecasting requires the covariance of the reconciliation errors**, which we can replace with the t+h forecast errors, which we can approximate with…
- Typical **approximations ignore modelling uncertainty**:
  - We do a lot of that, each node in the hierarchy implies a model selection step!
  - The bottom level is quite crucial, and it has large and unaccounted model uncertainty.
- **Cross-validated errors** can overcome this, but **not convenient to estimate and at times not possible** (especially for temporal hierarchies).
- We can manipulate **information criteria** to do that for us:
  - **Very simple** to calculate;
  - **Very fast** to calculate: real cases can have massive hierarchies;
  - Cross-temporal hierarchies (the fabled “one number forecast”) can make hierarchies stupendously large.
- $\boldsymbol{W_{AIC}}$ **performs very well → improves further performance of hierarchical**
  → Temporal hierarchies hard to beat plus decision making gains.

CMAF

Thank you for your attention!
Questions?

Nikolaos Kourentzes
email: nikolaos@kourentzes.com
twitter @nkourentz
Blog: http://nikolaos.kourentzes.com