Constructing hierarchical time series through clustering: Is there an optimal way for forecasting?

Bohan Zhang[1], Anastasios Panagiotelis[2], and Han Li[3]

[1] School of Economics and Management, Beihang University

[2] Discipline of Business Analytics, Business School, University of Sydney

[3] Centre for Actuarial Studies, Department of Economics, University of Melbourne

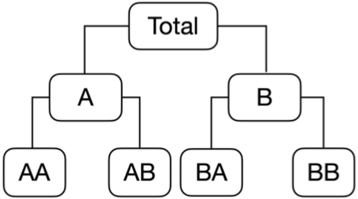
The 44th International Symposium on Forecasting, Dijon, France

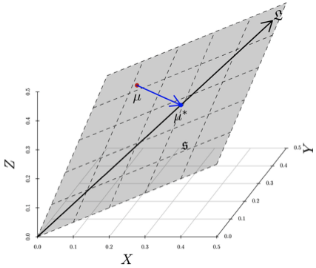
30 June - 3 July, 2024

Outline of the presentation

Outline

Hierarchical forecasting and forecast reconciliation





- (Cross-sectional) Hierarchical time series is multivariate time series whose observations at time t respect **aggregation constraints**.
- Hierarchical forecasting produces *coherent* forecasts for hierarchical time series.
- Forecast reconciliation projects incoherent base forecasts of all time series onto the coherent subspace (Panagiotelis et al. 2021).

Forecast reconciliation: the forecast combination perspective

- Forecast reconciliation can be interpreted from the forecast combination perspective (Hollyman et al. 2021, Di Fonzo & Girolimetto 2024).
- Reconciled forecasts are weighted combination of "direct" and "indirect" base forecasts.

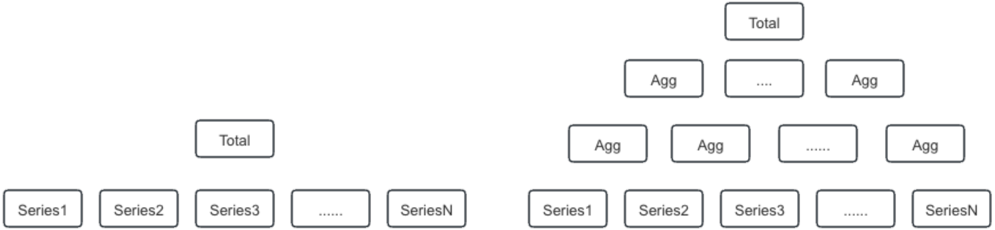
$$\tilde{y}_{AA} = w_1 \hat{y}_{AA} + w_2 (\hat{y}_A - \hat{y}_{AB}) + w_3 (\hat{y}_{Total} - \hat{y}_{AB} - \hat{y}_{BA} - \hat{y}_{BB})$$

- Could we employ more "indirect" forecasts and improve forecast accuracy?
  - Add middle-level series.
- Could we find more accurate "indirect" forecasts and improve forecast accuracy?
  - Cluster similar time series (Li et al. 2019, Pang et al. 2022, Mattera et al. 2023).
  - The rationale behind clustering lies in grouping time series with similar patterns together, thereby creating middle-level series with enhanced signals and consequently, improved forecastability.

Outline

Research questions

RQ1. In terms of forecast performance, can the use of middle level series lead to improvement compared to a two-level hierarchy (consisting of only top and bottom time series)? If so, is it possible to construct hierarchies in a data-driven way that leads to further improvements in forecast accuracy?



- Does creating “Agg”s improve forecast accuracy?
- How to construct “Agg”s in a data-driven approach?

Research questions

The increased accuracy when using data-driven approach may be attributed to two factors.

- Grouping: some correct subsets of series are chosen to form new middle-level series.
- Structure: the number of middle-level series, the depth of the hierarchy, and the distribution of group sizes in the middle layer(s).

RQ2. Should the improved accuracy of clustering-based methods be attributed to grouping together similar time series, or to the structure of the hierarchy?

Research questions

With multiple hierarchies available and inspired by the forecast combination literature (Wang et al. 2023), we consider the last research question:

RQ3. Does an equally-weighted combination of reconciled forecasts derived from multiple hierarchies improve forecast reconciliation performance?

Outline

Time series clustering-based forecast reconciliation

To the best of our knowledge, four studies have attempted to improve forecast accuracy in a reconciliation setting by constructing middle levels of the hierarchy using time series clustering.

- Pang et al. (2018) and Pang et al. (2022) propose to employ K-means algorithm to group similar electricity and solar power time series.
- Li et al. (2019) apply agglomerative hierarchical clustering to cause-of-death time series.
- Mattera et al. (2023) utilize Partition Around Medoids algorithms to unveil underlying structures in stock price indexes.

We consider various approaches based on three key components, namely time series representations, distance measures, and clustering algorithms.

Time series representations

The time series representation refers to the object that acts as an input for time series clustering. We consider four representations:

- Raw time series.
- In-sample one-step-ahead forecast error. A key step in MinT reconciliation is to estimate $\boldsymbol{W}_h$ based on in-sample forecast error.
- Time series features of raw time series.
- Time series features of in-sample one-step-ahead forecast error.
  - Features are low dimensional representation of time series, and have been used in various tasks such as clustering and forecasting.
  - 56 features calculated by the `tsfeatures` package in R are included.

Distance measures

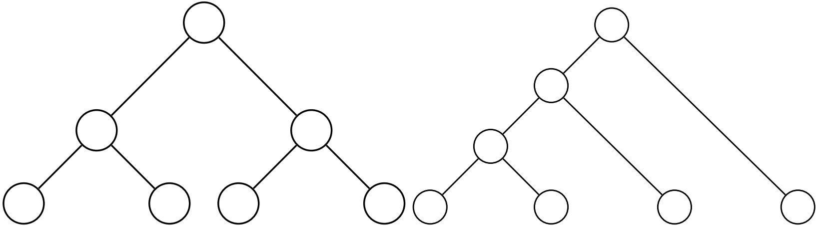
All clustering algorithms we consider require a distance to be defined between the objects that act as inputs to the algorithm.

We consider two widely applied distance measures: Euclidean distance (combined with PCA to reduce dimensionality) and dynamic time warping (DTW).

Clustering algorithms

We focus on two clustering algorithms:

- Partitioning around medoids, which is an algorithm to find a local minimum for the k-medoids problem.
- Agglomerative hierarchical clustering with Ward's linkage.



- PAM (left) constructs a simple hierarchy with a single middle level, while hierarchical clustering (right) generates multiple nested middle levels.
- As the number of bottom-level series increases, these differences become increasingly pronounced, with potential implications for forecast reconciliation.

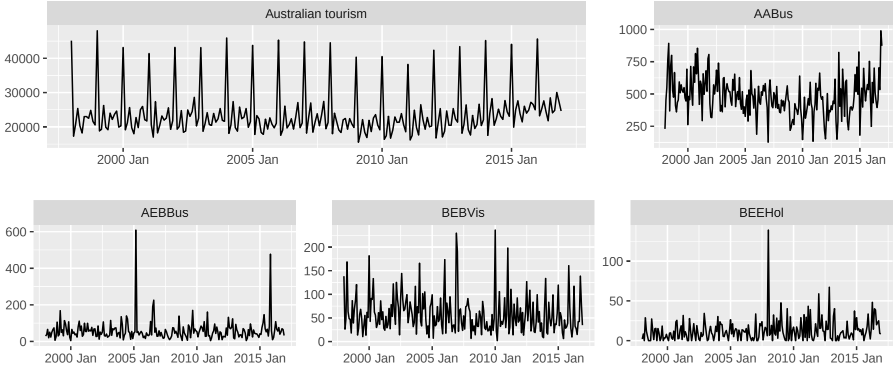
Summary

Table 1: Details of the 12 clustering approaches considered.

Approach	Dimension reduction	Representation	Distance measure	Clustering algorithm
TS-EUC-ME	Yes	Time series	Euclidean	*k*-medoids
ER-EUC-ME	Yes	In-sample error	Euclidean	*k*-medoids
TSF-EUC-ME	Yes	Time series features	Euclidean	*k*-medoids
ERF-EUC-ME	Yes	In-sample error features	Euclidean	*k*-medoids
TS-EUC-HC	Yes	Time series	Euclidean	Hierarchical
ER-EUC-HC	Yes	In-sample error	Euclidean	Hierarchical
TSF-EUC-HC	Yes	Time series features	Euclidean	Hierarchical
ERF-EUC-HC	Yes	In-sample error features	Euclidean	Hierarchical
TS-DTW-ME	No	Time series	DTW	*k*-medoids
TS-DTW-HC	No	In-sample error	DTW	Hierarchical
ER-DTW-ME	No	Time series	DTW	*k*-medoids
ER-DTW-HC	No	In-sample error	DTW	Hierarchical

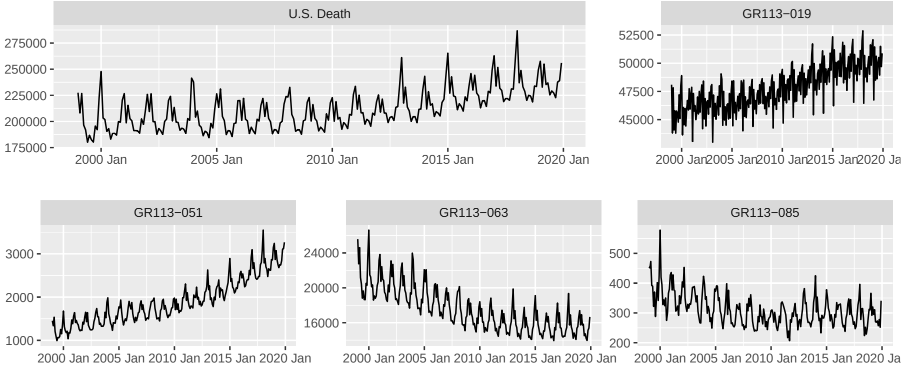
Outline

Australian tourism dataset



- Monthly Australian domestic tourism dataset, covering the period from January 1998 to December 2016 (Wickramasuriya et al. 2019).
- Consists of 555 time series with 304 of those at the bottom level. The middle-level series are constructed based on state, region, city and travel purpose.

U.S. cause-of-death count dataset



- Monthly cause-specific death count data of U.S. for the period between January 1999 and December 2019.
- Consists of 120 time series, with 98 of those being bottom-level series. The middle-level series are constructed based on major cause-of-death groups.

Outline

Evaluation

- We employ the expanding window strategy, resulting in 121 12-step-ahead forecasts for tourism dataset, and 145 12-step-ahead forecasts for mortality dataset.
- We only focus on the total series and bottom-level series.
- Forecast accuracy is evaluated based on average RMSSE of all time series (total and bottom).
- Three benchmarks: Base forecast, Two-level hierarchy (total and bottom), Natural hierarchy (constructed based on attributes of the bottom-level series)

Cluster hierarchies vs benchmarks

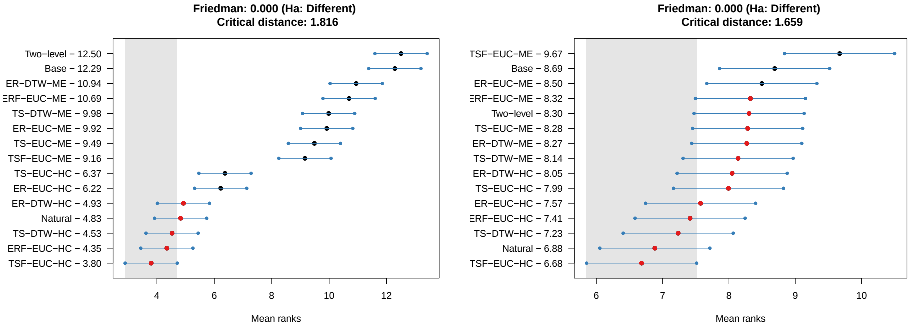


Figure 1: Average ranks and 95% confidence intervals for twelve cluster hierarchies and three benchmarks on tourism dataset (left) and mortality dataset (right) based on MCB test.

Cluster hierarchies vs benchmarks

- The natural hierarchies provide better results than the base forecasts and the two-level hierarchies, and comparable results with cluster hierarchies.
- For tourism dataset, ten out of twelve cluster hierarchies significantly outperform two-level hierarchy. However for mortality dataset, none of the cluster hierarchies significantly outperform two-level hierarchy.
  - The performance of clustering-based forecast reconciliation depend on the characteristics of bottom-level series.
  - The tourism dataset predominately contains volatile and noisy bottom-level time series with weak trend and seasonality.
  - The bottom-level series in mortality dataset exhibit stronger trend and seasonality patterns, meaning that the addition of middle-level series is less beneficial.
- The hierarchies constructed via hierarchical clustering algorithms outperform the hierarchies based on k-medoids when using the same representation and distance metric.

Disentangling grouping and structure

To assess whether “grouping” or “structure” has relatively more importance, we propose to construct “twin” hierarchies constructed by randomly permuting the bottom-level series.

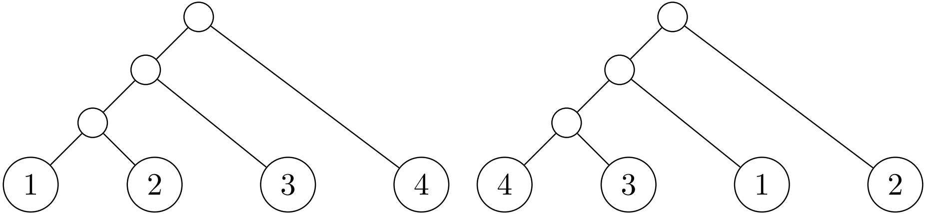
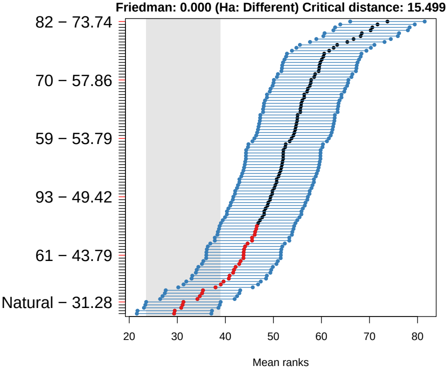


Figure 2: Examples of a given hierarchy and its “twin”.

Natural hierarchy vs its twins

We construct 100 twin hierarchies for each of the two natural hierarchies and compare their performance.



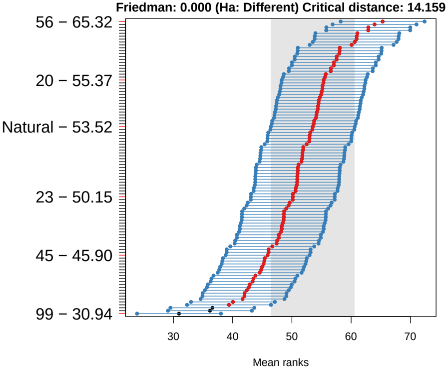


Figure 3: Average ranks and 95% confidence intervals for natural hierarchy and its 100 twins, tourism dataset (left) and mortality dataset(right).

Best cluster hierarchy vs its twins

We construct 100 twin hierarchies for each of the two best cluster hierarchies and compare their performance.

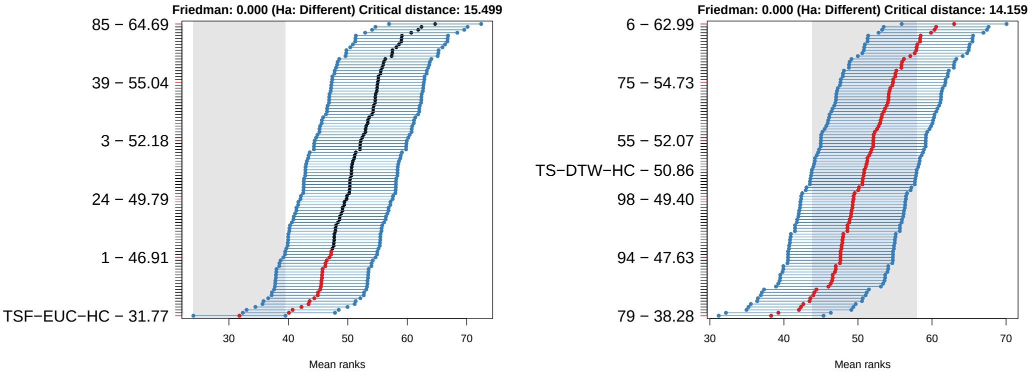


Figure 4: Average ranks and 95% confidence intervals for the best cluster hierarchy and its 100 twins, tourism dataset (left) and mortality dataset(right).

Disentangling grouping and structure

- For mortality dataset, the performances of the natural hierarchy and the best cluster hierarchy are statistically indistinguishable from most of their twins. "Structure" is the primary contributor to the improvement in forecast accuracy over the two-level hierarchy.
- For tourism dataset, both natural and best cluster hierarchies are statistically indistinguishable from around 30 of their twins, but significantly better than the remaining 70.
- For tourism dataset, a data-driven method for grouping time series plays a more prominent role in forecast improvement. This may be attributed to the noisier nature of bottom level tourism data, suggesting that similar weak signals are strengthened when aggregated.
- However, there roughly a 30% chance that a random twin performs similarly, once again highlighting that structure is the main contributor to improved forecast performance.

Forecast combination

- Although it is possible to significantly improve forecast performance through clustering, the selection of the best performing combination of time series representation, distance measure, and clustering algorithm remains an open question.
- We consider averaging forecasts across different hierarchies, as an alternative to hierarchy selection.

Forecast combination

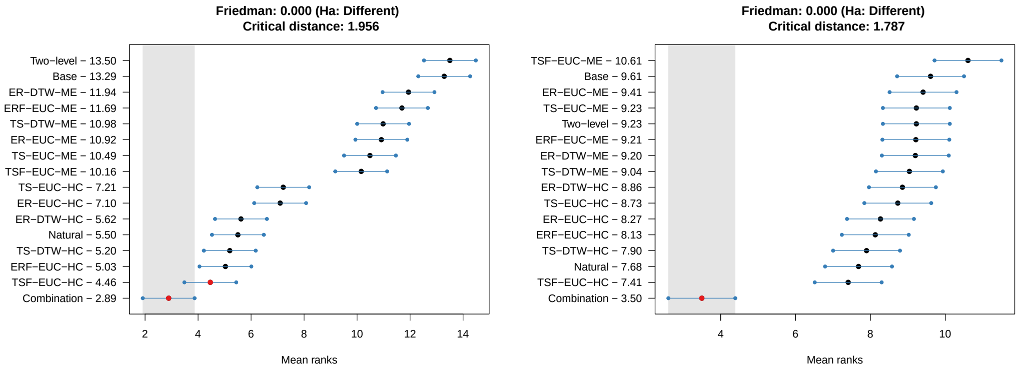


Figure 5: Average ranks and 95% confidence intervals for all approaches on tourism dataset(left) and mortality dataset(right) based on MCB test

- On both datasets, forecast combination improves forecast performance compared to any single hierarchy. The improvement on the mortality dataset is more pronounced.

Outline

Conclusions

- We consider a general framework, incorporating three distinct approaches: cluster hierarchies, permutation hierarchies, and combination hierarchies.
- Although adding middle-level time series via clustering or natural hierarchy improve forecast accuracy over two-level hierarchy, the primary contributing factor varies across datasets. For both datasets, “structure” plays an important role.
- Our main practical recommendation is to use multiple clustering methods and combine forecasts across these methods using equal weights combination. This mitigates the uncertainty of selecting the best clustering approach and is shown to significantly outperform all benchmarks across both datasets that we consider.

Questions and discussions

Contact email: zhangbohan@buaa.edu.cn

References

Di Fonzo, T. & Girolimetto, D. (2024), 'Forecast combination-based forecast reconciliation: Insights and extensions', *International Journal of Forecasting* **40**(2), 490–514.

Hollyman, R., Petropoulos, F. & Tipping, M. E. (2021), 'Understanding forecast reconciliation', *European Journal of Operational Research* **294**(1), 149–160.

Li, H., Li, H., Lu, Y. & Panagiotelis, A. (2019), 'A forecast reconciliation approach to cause-of-death mortality modeling', *Insurance: Mathematics and Economics* **86**, 122–133.

Mattera, R., Athanasopoulos, G. & Hyndman, R. J. (2023), 'Improving out-of-sample forecasts of stock price indexes with forecast reconciliation and clustering'.
URL: *https://robjhyndman.com/publications/dow_hts.html*

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P. & Hyndman, R. J. (2021), 'Forecast reconciliation: A geometric view with new insights on bias correction', *International Journal of Forecasting* **37**(1), 343–359.

Pang, Y., Yao, B., Zhou, X., Zhang, Y., Xu, Y. & Tan, Z. (2018), Hierarchical Electricity Time Series Forecasting for Integrating Consumption Patterns Analysis and Aggregation Consistency, *in* 'Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence', pp. 3506–3512.