Constructing hierarchies

Abstract

Keywords: Forecasting, Hierarchical time series,

1. Introduction

2. Methodology

2.1. Constructing hierarchies by clustering

2.1.1. Time series representations

2.1.2. Distance measures

2.1.3. Clustering algorithms

2.2. Constructing hierarchies randomly

3. Simulation

In this section, we demonstrate that hierarchies constructed by clustering may not outperform hierarchies constructed randomly even if there are true clusters in the bottom level.

3.1. Simulation design

Time series generation

We assume the bottom-level time series follow an additive time series pattern with a data generating process described as follows:

$$\begin{aligned} y_t &= l_t + s_t + e_t \\ s_t &= s_{t \mod m} \\ l_t &= at + \varepsilon_t, \end{aligned} \tag{1}$$

where l_t represents the trend term which increases or decreases over time at a slope of a. The seasonal component, denoted by s_t, remains constant. Both e_t and ε_t are white noises.

We create 6 clusters at the bottom level by adjusting the directions of the trend and patterns of the seasonal components. The pattern settings for all clusters can be found in Table 1. For an increasing trend we set the slope a to 0.001, and for a decreasing trend, we set it to -0.002. The variances of the corresponding white noise ε_t are set to 2.5×10^{-5} and 4.9×10^{-5}. The terms "Even" and "Odd" indicate the position of the seasonal peak. For "Even" seasonality, the peaks are located at positions $2, 4, \ldots, m$, and vice versa. The values for these seasonal peaks and troughs are uniformly drawn from $[2, 3]$ and $[0, 1]$, respectively. The variance of e_t is set to 0.25. We generate monthly time series data with 20 time series per

Table 1: Parameter setting for all clusters in the simulation experiments.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Trend	Increase	Increase	None	None	Decrease	Decrease
Seasonality	Odd	Even	Odd	Even	Odd	Even

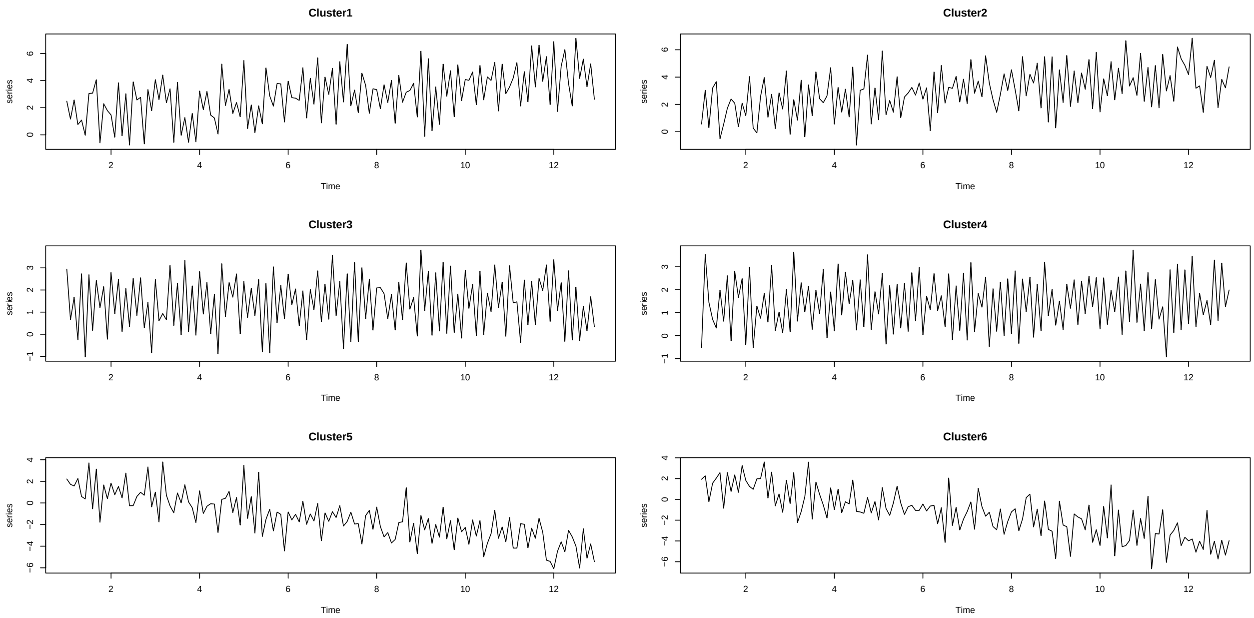

Figure 1: Example time series for each cluster in the simulation experiments.

cluster. Figure 1 displays example time series from each cluster while Figure 2 visualises these generated time series based on the first two principal components extracted from principal component analysis of the series.

Hierarchies construction

There are a total of 120 bottom series and 1 total series in the hierarchy, referred to as "C0". Additionally, we consider 4 clustering approaches and 2 forecast combination approaches. "C1" creates 6 middle-level series according to the designed correct clusters. We merge clusters with the same trend pattern, resulting in 3 clusters in the middle level, denoted by "C2". Similarly, we construct "C3" and "C4" based on the presence of the trend term and seasonal peak positions, resulting in 2 and 3 middle-level series, respectively. "A1" combines reconciled forecasts obtained from all four clustering approaches ("C1" to "C4") using equal weights. By shuffling the bottom-level series within the hierarchy "C1", we can randomly generate different hierarchies with identical structure. "A2" combines reconciled forecasts from 10 randomly constructed hierarchies using equal weights.

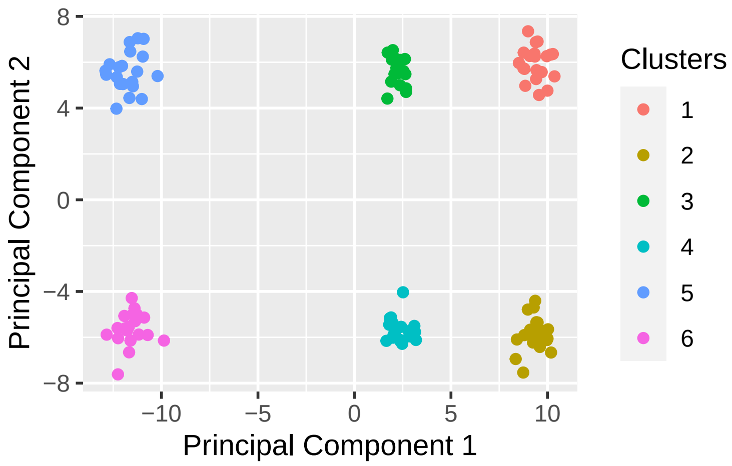

Figure 2: Visualisation of the generated time series in the simulation experiments.

Forecasting

We construct two scenarios by considering different base forecasting models. In the first scenario, exponential smoothing (ETS) is used to generate base forecasts for all time series in hierarchies. This simulates a situation where the time series themselves are inputs to the clustering algorithms. In the second scenario, we use historic mean to generate base forecasts for bottom-level series and ETS for other time series in hierarchies, which simulates the case where in-sample errors are inputs to the clustering algorithms. We refer to these scenarios as "clustering by time series" and "clustering by error", respectively. To reconcile the base forecasts, we employ the minimum trace method with the shrinkage estimator (Wickramasuriya et al., 2019). The shrinkage estimator is effective at capturing the dependence structure within the forecast errors and has demonstrated superior perofrmance in various applications. For each series, we generate 144 observations, and the last 12 observations are reserved for evaluation purpose.

3.2. Evaluation

The purpose of constructing middle-level series is to improve the forecast performance of the total-level and bottom-level series by leveraging the strength of these new series. Therefore, we only evaluate the reconciled forecasts at total level and bottom level. To assess the accuracy of single time series, we use root mean squared error (RMSE) as our metric. The simulation is repeated 500 times, resulting in a total of 500×121 RMSEs for each approach. Multiple comparisons with the best (MCB) test is then applied to compute the average ranks of the 7

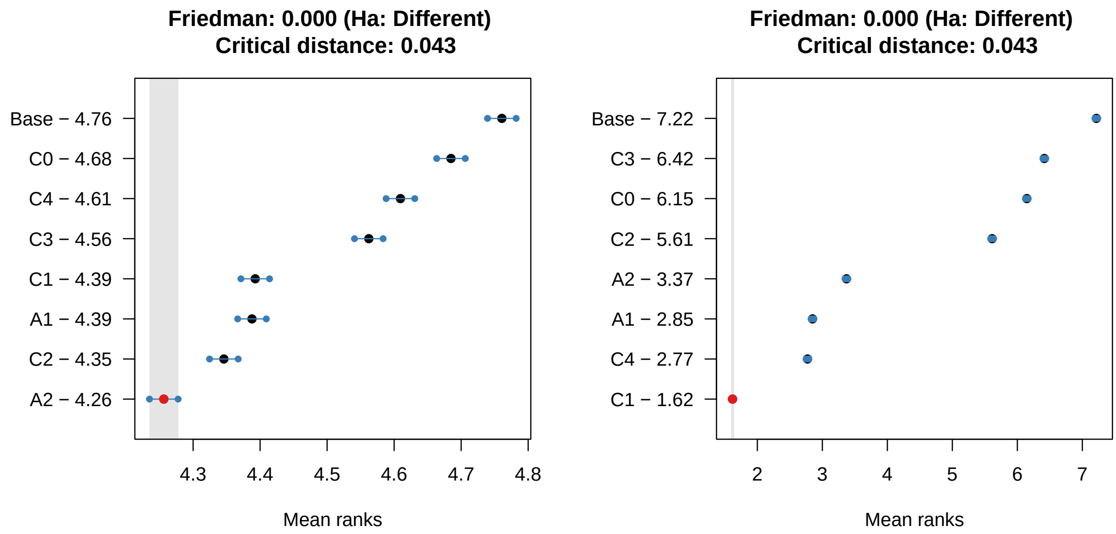

Figure 3: Average ranks and 95% confidence intervals based on the MCB Test for the 8 approaches in the two scenarios of the simulation experiments. Left panel displays test results for the "clustering by time series" scenario and right panel displays test results for the "clustering by error" scenario.

approaches along with base forecasts and to determine whether the performance differences are statistically different (Koning et al., 2005).

3.3. Results

Figure 3 displays the MCB test results for the two scenarios. In both scenarios, most approaches perform better than the base forecasts and the original two-level hierarchy "C0". This suggests that constructing middle-level series generally improves forecast reconciliation performance. In the "clustering by time series" scenario, the correct cluster "C1" ranks 4th, indicates that optimal clusters do not guarantee optimal forecast reconciliation performance. On the other hand, "C1" ranks 1st in the "clustering by error" scenarios. This highlights the importance of carefully selecting time series representations when constructing new middle levels using clustering algorithms. The simple average of 10 random hierarchies ranks 1st in the first scenario and the simple average of 4 clustering hierarchies ranks 3rd in both scenarios, indicating the importance of forecast combination.

4. Empirical studies

4.1. Datasets

We analyse two datasets in our empirical studies. The first dataset is the monthly Australian domestic tourism dataset, which provides visitor nights numbers from 1998 to 2016.

The tourism demand of Australia is geographically disaggregated into 7 states and territories, further divided into 27 zones and 76 regions. Additionally, each geographical series is divided by four travel purposes (Wickramasuriya et al., 2019). This dataset totally contains 555 time series with 304 at the bottom level.

The second dataset focuses on causes of death in the U.S., using the ICD 10 coding system. We obtain monthly cause-specific death counts from the Center for Disease Control and Prevention (CDC) for the period between 1999 and 2019. The coding system forms an unbalanced hierarchy with 137 time series, out of which 113 time series are in the bottom level. To consolidate data with suppressed values, we combine causes that share a parent cause and calculate their death counts by subtracting death counts of sibling causes from death counts of their parent cause. The final dataset includes 120 time series with 98 in the bottom level.

4.2. Experiment design

We focus on the performance of total-level series and bottom-level series, disregarding the multiple middle levels in the original hierarchies known as "natural hierarchies". The natural hierarchy is considered one way to construct middle-level series, similar to hierarchies created through clustering. However, we demonstrate that for the purpose of forecast performance, the natural hierarchy may not be the most effective hierarchical structure.

To construct middle levels using clustering, we combine four time series representations with two distance measures and two clustering algorithms. This results in twelve different construction approaches listed in Table 2. In our experiments, dimension reduction of time series representations is crucial due to a small number of time series compared to variable dimensions (i.e., length of time series and number of time series features). Without dimension reduction, undesired clustering outcomes can occur due to the curse of dimensionality when using Euclidean distance. We employ principal component analysis on the time series representations and extract the first k principal components that explain at least 80% variance in the data. These transformed representations are then used as inputs for the clustering algorithms.

Additionally, we consider three forecast combination approaches. In the first approach, we create one middle level with 15 time series by randomly assigning bottom-level series as their children. We ensure that all the middle-level series have approximately an equal number of children. The choice of having 15 middle-level series is arbitrary with the goal of creating

Table 2: Hierarchy construction approaches used in empirical studies.

Approaches	Representation	Dimension reduction	Distance measure	Clustering algorithms
TS-MED	Time series	Yes	Euclidean	k-Medoids
ER-MED	In-sample error	Yes	Euclidean	k-Medoids
TSF-ME	Time series features	Yes	Euclidean	k-Medoids
ERF-ME	In-sample error features	Yes	Euclidean	k-Medoids
TS-HC	Time series	Yes	Euclidean	hierarchical clustering
ER-HC	In-sample error	Yes	Euclidean	hierarchical clustering
TSF-HC	Time series features	Yes	Euclidean	hierarchical clustering
ERF-HC	In-sample error features	Yes	Euclidean	hierarchical clustering
TS-MED-DTW	Time series	No	DTW	k-Medoids
TS-HC-DTW	In-sample error	No	DTW	hierarchical clustering
ER-MED-DTW	Time series	No	DTW	k-Medoids
ER-HC-DTW	In-sample error	No	DTW	hierarchical clustering

a moderate number of groups, each containing a moderate number of series. We repeat this process to create 50 such hierarchies and combine their reconciled forecasts using equal weights. This approach is referred to as "FC-R". In the second approach, we create 10 new hierarchies by randomly shuffling the positions of bottom-level series in the natural hierarchy and combine their reconciled forecasts using equal weights. This approach is denoted by "FC-N". The third approach, labelled by "FC-C", involves equally-weighted combining the reconciled forecasts obtained from the 12 hierarchies shown in Table 2.

The time series features used in this experiment are calculated using the `tsfeatures` package (Hyndman et al., 2022) for R. After filtering out the features that are constant across all series, 56 features are reserved. The complete list of these feature can be found in Appendix. The k-Medoids and hierarchical clustering algorithms are implemented using the `cluster`(Maechler et al., 2022) package for R. The base forecasts are generated using the automatic ETS models, which are then reconciled using the minimum trace method with shrinkage estimator.

We employ the rolling window strategy to evaluate the performance of different approaches for both datasets. We start with 96 observations and fit the base forecasting models using the available data. Then we calculate time series representations and construct new hierarchies by clustering or randomness, and then forecast 12 steps ahead. After that, the training set is increased by one observation and new forecasts are obtained. The procedure is repeated until the last 12 observations are used for evaluation. Finally, we can obtain 121 12-step-ahead forecasts for the tourism dataset and 144 12-step-ahead forecasts for the mortality dataset. We compare the forecast performance of the 15 approaches as well as the natural hierarchy,

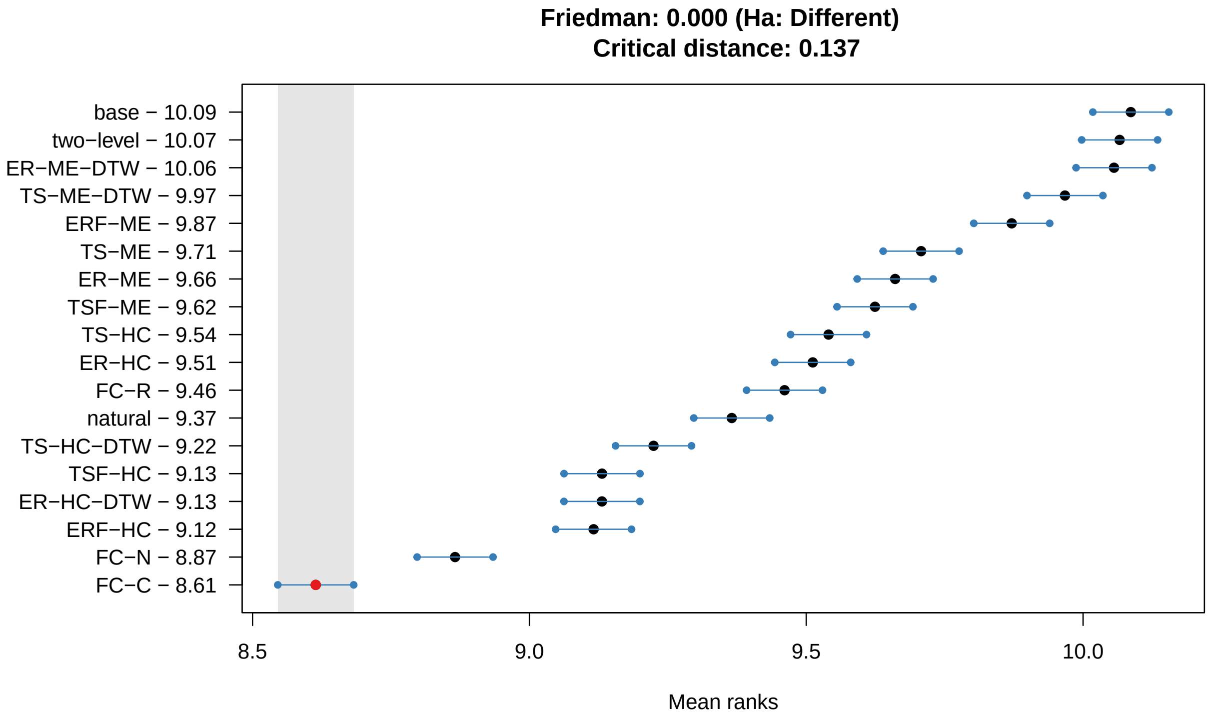

Figure 4: Average ranks and 95% confidence intervals based on the MCB Test for the 18 approaches on the tourism dataset.

two-level hierarchy and base forecasts for both total-level and bottom-level series.

4.3. Results

Same as the evaluation measure described in Section 3.2, we combine the forecasts of total-level and bottom-level series across all evaluation windows and conduct the MCB test. The test results for the tourism dataset and mortality dataset are shown in Figure 4 and Figure 5, respectively.

Most hierarchy construction approaches outperform both base forecasts and the two-level hierarchy, except for three approaches on the mortality dataset. For both datasets, the simple average of randomised natural hierarchies ("FC-N") performs better than both natural hierarchy and all clustering-based hierarchies. Additionally, the simple average of 12 clustering-based hierarchies ("FC-C") significantly outperforms all other approaches. These results indicate that using a clustering-based hierarchy can improve forecast performance. However, forecast combination is even more beneficial, as evidenced by the high ranking of "FC-N" and "FC-R".

The approach based on hierarchical clustering outperforms the approach based on k-Medoids on the tourism dataset when using the same representation and distance metric,

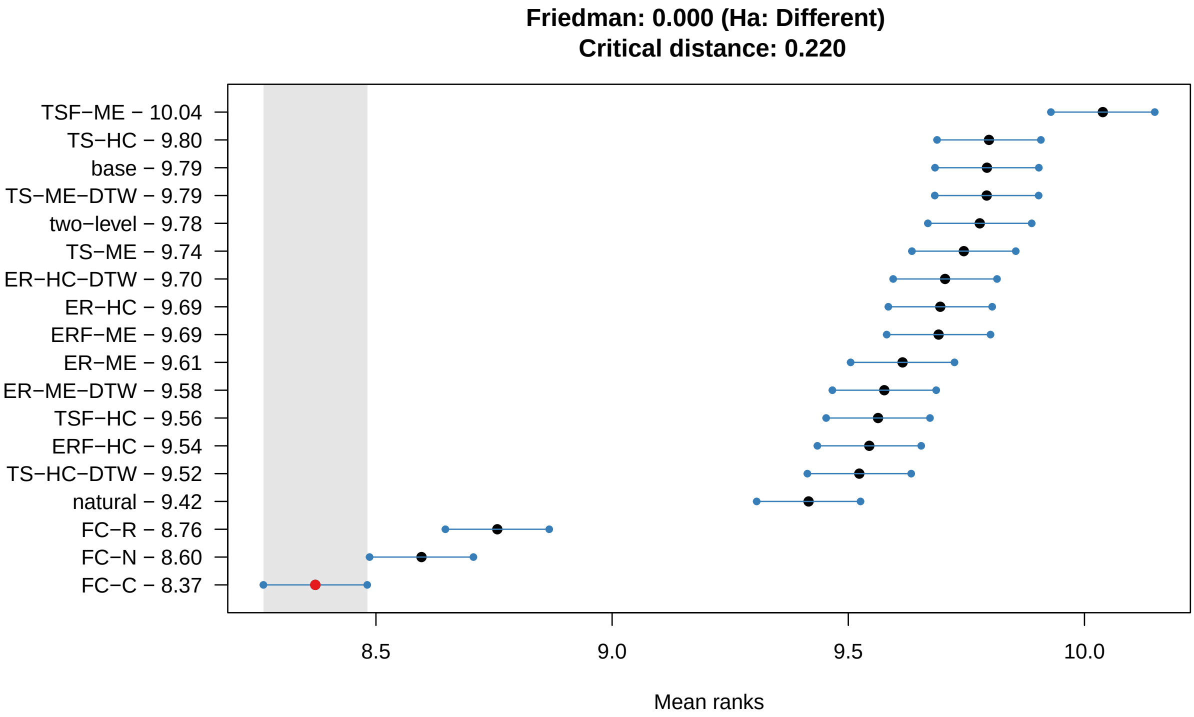

Figure 5: Average ranks and 95% confidence intervals based on the MCB Test for the 18 approaches on the mortality dataset.

e.g., "TSF-HC" significantly ourperforms "TSF-ME". However, this is not consistently observed for every combination of representation and distance metric on the mortality dataset. We believe this discrepancy is due to different time series patterns at the bottom level of these two datasets. In the case of the mortality dataset, we combine rare causes of death that have suppressed values and record death counts at a national level. As a result, most series at the bottom level exhibit strong seasonality and trend. On the other hand, in the tourism dataset, the tourism demand is disaggregated for both travel purposes and geographical regions. This leads to numerous bottom-level series with high volatility and many zeros, making them more challenging to forecast. Therefore, having a hierarchy with more middle-level series can be more advantageous for reconciliation in the tourism dataset compared to the mortality dataset. Additionally, we suspect that the inferior performance of "FC-R" on the tourism dataset may be partly attributed to an insufficient number of middle-level series.

In most cases, the in-sample error representation outperforms the time series representation when using the same distance metric and clustering algorithms. However, this performance difference is not as significant as what was observed in Section 3.3. Intuitively, hierarchies constructed based on clustering in-sample error should be similar to random hierarchies when all base models are correctly specified. However, due to ubiquitous model misspecification in practice, vague patterns emerge in the in-sample error, which leads to contradictory results.

Effect of number of random hierarchies

In our previous experiments, we set the number of random hierarchies in "FC-R" and "FC-N" to 10 for fair comparisons with "FC-C". However, we believe that increasing the number of random hierarchies could further enhance performance. In this subsection, we explore the effects of using 20 and 50 random hierarchies, referred to as "FC-N-20", "FC-N-50", "FC-R-20", and "FC-R-50". We compare these variations with "FC-R", "FC-C", and "FC-N" using the same evaluation procedure described earlier. The results from the MCB test are presented in Figure 6, where the left panel shows results on the tourism dataset and the right panel displays results on the mortality dataset.

Generally, we observe that forecast performance improves as the number of random hierarchies increases, except for "FC-N-20" and "FC-N" on the mortality dataset. Surprisingly, on the tourism dataset, "FC-N-50" performs slightly better than "FC-C". It is important to note that employing more random hierarchies comes at a cost of increased computational resources. Therefore, one must carefully consider trade-offs between efficiency and performance when making decisions.

5. Conclusion

Acknowledgements

References

Hyndman, R. J., Kang, Y., Montero-Manso, P., Talagala, T., Wang, E., Yang, Y. and O'Hara-Wild, M. (2022), *tsfeatures: Time Series Feature Extraction*. R package version 1.1.0.9000. **URL:** *https://pkg.robjhyndman.com/tsfeatures/*

Koning, A. J., Franses, P. H., Hibon, M. and Stekler, H. O. (2005), 'The M3 competition: Statistical tests of the results', *International Journal of Forecasting* **21**(3), 397–409.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. (2022), *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.4. **URL:** *https://cran.r-project.org/web/packages/cluster/index.html*

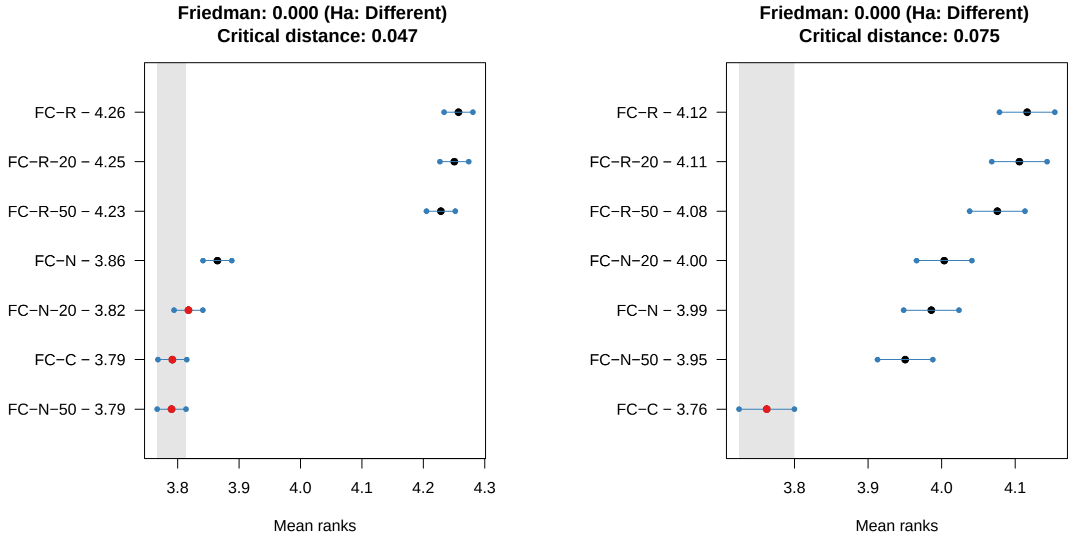

Figure 6: Average ranks and 95% confidence intervals based on the MCB Test of different number of random hierarchies on two datasets. Left and right panels display test results on tourism dataset and mortality dataset, respectively.

Wickramasuriya, S. L., Athanasopoulos, G. and Hyndman, R. J. (2019), 'Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization', *Journal of the American Statistical Association* **114**(526), 804–819.