MANCHESTER 1824
The University of Manchester

The University of Manchester Research

Determining analogies based on the integration of multiple information sources

DOI:
10.1016/j.ijforecast.2018.02.002

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Lu, E., Handl, J., & Xu, D.-L. (2018). Determining analogies based on the integration of multiple information sources. *International Journal of Forecasting*, *34*(2), 507-528. https://doi.org/10.1016/j.ijforecast.2018.02.002

Published in:
International Journal of Forecasting

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:06. Oct. 2024

Determining analogies based on the integration of multiple information sources

Abstract

Forecasting approaches that exploit analogies require the grouping of analogous time series as the first modeling step, but there has been limited research regarding the suitability of different segmentation approaches. We argue that an appropriate analytical segmentation stage should integrate and trade off different available information sources. In particular, it should consider the actual time series patterns in addition to variables that characterize the drivers behind the patterns observed. The simultaneous consideration of both information sources, without prior assumptions regarding their relative importance, leads to a multicriteria formulation of the segmentation stage. Here, we demonstrate the impact of such an adjustment to segmentation on the final forecasting accuracy of the Cross-Sectional Multi-State Kalman Filter. In particular, we study the relative merit of single and multicriteria segmentation stages for a simulated data set with varying noise levels. We find that a multicriteria approach consistently achieves a more reliable recovery of the original clusters, and this feeds forward to improved forecasting accuracy across short forecasting horizons. Using a US data set on income tax liability, we verify that this result generalizes to a real-world setting.

Keywords: Analogy; Bayesian pooling; Kalman Filter; Model selection; Multicriteria clustering

1. Introduction

Forecasting approaches such as the Cross-Sectional Multi-State Kalman Filter algorithm (C-MSKF: Duncan et al., 1993) exploit information from analogies or analogous time series so as to increase the accuracy of point forecasts for a target time series. The identification of suitable analogies is crucial to these approaches, but, despite this, surprisingly little research has been conducted to investigate appropriate analytical modeling approaches for judging similarities between time series (Lee et al., 2007) and supporting the principled selection of analogies (Armstrong, 2001).

The identification of analogous time series typically involves the use of segmentation approaches to partition a set of time series into a set of homogeneous clusters (*e.g.,* Duncan et al., 2001). Segmentation approaches have wide application in areas such as economics, finance, operational research, and public budgeting. Segmentation is typically used to identify meaningful sub-groups (*e.g.,* customers, businesses and countries) and can be useful in terms of identifying, understanding and targeting these groups. The sub-groups identified during segmentation may feed forward into further analysis, including the development of cluster-specific forecasting strategies. Segmentation is often modeled as a single-criterion problem in the traditional marketing literature and in practice, but it is inherently a multicriteria problem as clusters are typically desired to be homogeneous with respect to a set of explanatory as well as response variables (Liu et al., 2010; Myers, 1996; Smith, 1956). Similarly, in the context of forecasting, we may view the segmentation as one involving multiple information sources, as both past realizations of a given time series (response variables) and the associated causal factors (explanatory variables), which describe the underlying causal relationships for the co-movement of the analogous time series (Duncan et al., 2001), need to be considered. For example, a set of products may be considered a group due to the same sphere of influence, similar consumer preferences, promotion levels, or local trends. Ignoring one of these sources of information during the segmentation stage may lead to clusters that are insufficiently differentiated in terms of either time series patterns, or causal factors and thus lead to sub-optimal results in further analysis. To obtain meaningful groups of analogies for forecasting, we need to ensure the identification of clusters that are interpretable at a domain level (represented by similarities in the values of a set of shared causal factors) but simultaneously show similarities in their time-based patterns.

Here, we experiment with a simple prediction process that outlines this idea and contrasts the performance of single-criterion and multicriteria segmentation approaches in the context of forecasting analogous time series, for which both time-based patterns and potential causal factors are known. We illustrate that the segmentation approach using both information sources is preferable in the sense that it can generate, and usually identify, segmentations that boost the performance of pooling in terms of forecasting accuracy.

The remainder of the paper is structured as follows: Section 2 surveys related work, including pooling approaches and popular segmentation approaches in the literature. Section 3 proposes our three-stage prediction process. Section 4 presents experiments that investigate the impact of different segmentation approaches on the performance of pooling approaches. In particular, using simulated data, we investigate the sensitivity of the approaches to changes in the relative reliability of the two information sources. Section 5 summarizes results on a data set describing personal income tax liability data. Finally, Section 6 concludes.

2. Previous research

Analogies have been widely employed in the forecasting field in order to improve the forecasting accuracy (Armstrong, 2006; Green and Armstrong, 2007; Piecyk and McKinnon, 2010). According to Duncan et al. (2001), analogies can be defined as time series that exhibit similarity in time-based patterns due to shared underlying causal factors. They typically co-vary and are thus positively correlated over time.

Most commonly, analogies have been utilized in the context of judgmental approaches (*i.e.,* forecasting by analogy and related work, refer to Nikolopoulos et al., 2015; Savio and Nikolopoulos, 2013). These methods use analogies for the purpose of adjusting statistical forecasts (Webby and O'Connor, 1996) since this may reduce biases due to optimism or wishful thinking (Armstrong, 2001; Petropoulos et al., 2014). There has also been work on the development of statistical methods that can exploit information available from analogies. A well-established model is the Bass model (Bass, 1969; Nikolopoulos et al., 2016), and this has been used to forecast sales of products which have yet to be launched, through the use of information available from similar products (Goodwin et al., 2013). An alternative way of exploiting analogies is to use Bayesian pooling approaches, such as the Cross-Sectional Multi-State Kalman Filter (C-MSKF: Duncan et al., 1993, 2001), which requires a relatively small number of parameters. This method borrows strength from groups of analogous time series to increase the accuracy of point forecasts.

Time series forecasting with respect to the demand of products or services often needs to be robust in situations that are characterized by structural change (i.e. changes to the trend of the time series), *e.g.,* due to external influences such as the action of a competitor. To deal with such situations, methods such as Exponential Smoothing (Brown, 2004) and the Multi-State Kalman Filter (MSKF: Harrison and Stevens, 1971) have been developed, which revise model parameter estimates over time. Such methods must compromise between two different needs, namely the responsiveness to change and the accuracy of forecasts. By utilizing additional information from analogies, the C-MSKF method combines the capability of the MSKF to yield accurate forecasts with a quick responsiveness to change. This approach has proven effective in a number of challenging applications, such as forecasting of churn in telecommunication networks (Greis and Gilstein, 1991), infant mortality rates (Duncan et al., 1995) and tax revenue (Duncan et al., 1993). The C-MSKF can draw strength from the availability of multiple data points for the same time period, across different analogous series, which lends it robustness with respect to outliers. In general, C-MSKF has been said to show competitiveness over conventional time series forecasting methods, such as the Damped Exponential Smoothing (Damped) methods, Exponential Smoothing (ETS), MSKF, the Naïve Drift method (Drift), Random Walk (RW) or the Theta model in situations that satisfy the following three conditions (Duncan et al., 1994, 2001): (i) the number of points that are suitable for extrapolation is small (either due to size or due to a structural change); (ii) analogies are present across several time series; and (iii) at least three observations are available after a structural change due to the impact of an external influence. Finally, a key assumption behind C-MSKF is that time series that are classed as analogous (*i.e.,* that exhibit co-movement during the investigation's estimation period) do not frequently diverge in the forecasting periods. This requirement underlines the importance of accurately determining analogies as the first step of the analysis.

The homogeneity of the underlying set of analogous time series is fundamental for the effectiveness of pooling approaches (Stimson, 1985). Previous research (Duncan et al., 2001) has demonstrated that pooling across a homogeneous set of time series gives superior forecasting accuracy to pooling across a heterogeneous set. In this context, three general approaches have typically been considered to identify analogies. These are correlational co-movement, i.e. the grouping of time series based on the correlation between the time series patterns observed; the grouping of time series using model-based approaches (Frühwirth-Schnatter and Kaufmann, 2008); and the grouping of time series based on a set of causal variables associated with each time series (Duncan et al., 2001). These different approaches reflect the fact that time series data often comprise past realizations of the actual time series, as well as additional knowledge regarding the factors that govern the behaviour of these time series and are crucial to a clear understanding of causal relationships (Leitner and Leopold-Wildburger, 2011; Webby and O'Connor, 1996).

Clustering based on time series patterns has been extensively investigated in the field of pattern recognition, but existing approaches differ widely in the way features of the time series are extracted (Liao, 2005). The most straightforward possibility is the use of the raw data points, calculating *e.g.,* correlation. However, previous work such as Granger and Newbold (1974) observed that clustering based on the correlation between time series alone can be problematic for short time series, as temporary correlations between time series may be spurious. More advanced approaches use feature transformations to extract higher level features. For example, model-based clustering approaches, which assume the existence of an underlying physical process, can be powerful in differentiating overlaying time series by modeling time series using Box-Jenkins ARIMA models (see *e.g.,* Kalpakis et al., 2001). However, estimating the parameters of the physical process requires the availability of a sufficient number of historical data points, and model-based approaches are therefore unsuitable for the clustering of very short time series. In general, the performance of different approaches is highly dependent on the setting and purpose of the application considered. When assessing analogies in terms of a set of static (explanatory) variables associated with each time series, the feature representation is usually more straightforward, although suitable distance measures are dependent on the data type. Yet, clustering based on underlying causal factors alone may be affected by the inclusion of irrelevant factors or the omission of relevant ones.

It is evident that characterization of analogies using either of the above approaches will often provide a partial, approximate picture at best. Considering the nature of forecasting problems, we expect that clusters that share similarity in terms of

their patterns are valuable, as they open up opportunities to improve forecasting accuracy by exploiting information from sets of similar time series. On the other hand, clusters that are recognizably similar in terms of the values of hypothesized causal factors are useful, as they may increase the robustness of the analysis and allow for an immediate interpretation of the patterns found. The integration of these two information sources should be valuable, as useful information can potentially be strengthened and noise specific to each individual information source can potentially cancel out.

Furthermore, at an analytical level, there is existing evidence that segmentation approaches that consider multiple aspects of clustering quality may yield more robust discovery of data structure, or uncover more complex structures than single-criterion clustering techniques (Handl and Knowles, 2007). There are some approaches that have specifically investigated the combination of different (complementary) information sources. Vriens et al. (1996) developed a method to consider one criterion at a time in a multi-stage manner. It was capable of producing clusters with a richer interpretation, but they remained sub-optimal as information found in one stage was shared with other stages in a sequential manner (Brusco et al., 2002). For some applications, one option may be the representation of both information sources in a single feature space, but this can be difficult because decisions on relative weighting of information sources need to be made beforehand. Furthermore, this approach is not possible if the distance functions suitable for the two information sources are different, as is the case in our problem. An exact approach to bicriterion data clustering was first proposed in Delattre and Hansen (1980), which was specific to a particular pair of clustering criteria. Ferligoj and Batagelj (1992) described an approach to account for clustering criteria defined with respect to different information sources. More recently, multi-objective evolutionary algorithms were proposed as a more flexible approach that can identify (or at least try to approximate) the full set of Pareto optimal solutions for different choices of objectives (Handl and Knowles, 2007). A simpler way of combining information sources is to combine multiple criteria using a weighted-sum approach (Brusco et al., 2002, 2003), which may be done at the level of the objective or the distance function. Although this methodology is not capable of identifying all Pareto optimal solutions, it has advantages in terms of its simplicity, ease of implementation and time-complexity.

3. Multicriteria clustering for the forecasting of analogous time series

In this section, we detail the elements of our proposed methodological framework, which consists of three components. The first component corresponds to the segmentation stage and is concerned with generating optimal clusters using a multicriteria (weighted-sum) clustering approach. It clusters time series with a concurrent consideration of time series and causal factor data, and generates a set of candidate partitions that trade off the quality of fit to both information sources. The second component employs a forecasting technique – here represented by the C-MSKF algorithm – that is capable of making use of pooled time series data. C-MSKF pools time series data from the identified clusters to inform the forecasting of individual time series. The third component provides suitable model selection. Our segmentation component produces a set of candidate partitions, and further processing is required to identify a single most promising grouping of analogies. We use a combination of internal cluster validation and forecasting accuracy on historical hold-out data, to achieve this. In the following, we describe the relevant methodology in full detail.

3.1. Distance measures for individual information sources

The selection of the most suitable distance measures for clustering generally depends on the data types (*e.g.,* continuous variables, categorical variables, etc) and the particular application considered (Liao, 2005). Our approach permits the separate selection of two distance functions that quantify the difference between time series in terms of (i) the series of data points describing a primary variable of interest; (ii) an additional vector representing levels of (one or multiple) causal factors associated with each time series.

Concerning (i), we use d_{ij}^{TS} to denote the distance between the series of data points making up the time series i and j. Each time series is represented as a vector describing the values of a single variable of interest over time. We adopt a standard correlation-based approach, in which the distance value d_{ij}^{TS} between pairs of time series i and j is calculated based on the correlation between these vectors. Specifically, the Pearson correlation coefficient is defined as:

$$\delta^{TS}(i,j) = 1 - \frac{T(\sum_t x_{it}x_{jt}) - (\sum_t x_{it})(\sum_t x_{jt})}{\sqrt{(T(\sum_t x_{it}^2) - (\sum_t x_{it})^2)(T(\sum_t x_{jt}^2) - (\sum_t x_{jt})^2)}} \quad (1)$$

Here t is the index of time $t = 1, 2, ..., T$; T is the number of time steps used for measuring correlation; and x_{it} and x_{jt} represent the values of time series i and j at time t; The dissimilarity matrix derived from the time series information is defined as $\mathbf{D^{TS}} = (d_{ij}^{TS})$, and each element d_{ij}^{TS} is calculated as $d_{ij}^{TS} = \delta^{TS}(i,j)$.

Regarding (ii), we use the notation $\delta^{CF}(i,j)$ to refer to the distance function between the vectors of causal factor levels associated with time series i and j. In a situation where the levels of all causal factors can be described on a ratio scale, the squared Euclidean distance can be used to measure distance between the vector of values associated with each time series. In this case, $\delta^{CF}(i,j)$ is defined as:

$$\delta^{CF}(i,j) = \sum_m (a_{im} - a_{jm})^2 \quad (2)$$

Here a_{im} and a_{jm} represent the values of causal variable m associated with time series i and j, respectively, for $m = 1, 2, ..., M$, and M represents the number of causal factors. To eliminate scale differences, all variables are standardized using z-scores.

The dissimilarity matrix derived from causal variables is defined as $\mathbf{D^{CF}} = (d_{ij}^{CF})$, and each element d_{ij}^{CF} is calculated as $d_{ij}^{CF} = \delta^{CF}(i, j)$.

Alternatively, where all causal factors are of a categorical nature, the Euclidean distance may be replaced by the Hamming distance. The Hamming distance calculates the number of places in which the values of two vectors differ, leading to the following definition of $\delta^{CF}(i, j)$:

$$\delta^{CF}(i, j) = \#\{m : a_{im} \neq a_{jm}, m = 1, ..., M\} \tag{3}$$

3.2. Combination of distance measures

To combine the two information sources, we deploy a weighted-sum method on the standardized dissimilarity matrices. To achieve standardization (0-1 transformation), we update each element of the dissimilarity matrices as follows:

$$d_{ij}^{CF} \leftarrow \frac{d_{ij}^{CF} - \min(\mathbf{D^{CF}})}{\max(\mathbf{D^{CF}}) - \min(\mathbf{D^{CF}})} \tag{4}$$

$$d_{ij}^{TS} \leftarrow \frac{d_{ij}^{TS} - \min(\mathbf{D^{TS}})}{\max(\mathbf{D^{TS}}) - \min(\mathbf{D^{TS}})} \tag{5}$$

Subsequently, a new dissimilarity matrix can be defined as a weighted combination of these standardized dissimiliarity matrices. Specifically, for a given choice of the weight ω, each element in $\mathbf{D}_{\omega}^{\mathbf{MC}}$ is obtained as follows:

$$d_{ij\omega}^{MC} = (1 - \omega) \times d_{ij}^{CF} + \omega \times d_{ij}^{TS} \tag{6}$$

Separate dissimilarity matrices are obtained for values of ω=0 to 1 in steps of 0.10.

While this weighted-sum approach is limited in terms of its ability to reach all optimal trade-off solutions, it creates flexibility in terms of the choice of clustering methodology, as any clustering approach that works on a dissimilarity matrix can be employed.[1] Here, we proceed by applying a standard clustering technique, namely PAM clustering (Kaufman and Rousseeuw, 2009). An advantage of this approach is its availability in all standard software packages. Furthermore, this technique has a tendency to produce partitions consisting of equally-sized clusters, which we consider advantageous in our application context. As this method can converge to local optima, we repeat the clustering step 30 times and return the clustering solution which minimizes the sum of within-cluster dissimilarities.

3.3. Model selection

3.3.1. Selection of the number of clusters

We typically have no prior knowledge regarding the number of analogous sets present in a given time series data set. Our approach therefore includes a model selection component that uses an automatic approach to the determination of the number of clusters, based on the Silhouette Width.

[1] Clustering methods that are not applicable here are those that operate directly in the feature space, *e.g.,* by using a centroid-based representation.

The Silhouette Width is an established internal method of cluster validation that assesses the quality of a partitioning based on its structure alone. In particular, it takes into account elements of cluster cohesion and cluster separation.

More specifically, given a candidate clustering solution, the Silhouette value (Rousseeuw, 1987) for an individual data item i is defined as:

$$\mathrm{Sil}(i) = \frac{b_i - c_i}{\max(c_i, b_i)} \tag{7}$$

where c_i denotes the average distance between i and all data items in the same cluster, and b_i denotes the average distance between i and all data items in the closest other cluster, which is defined as the one generating the minimum b_i. The Silhouette Width (Rousseeuw, 1987) of the entire partition is then calculated as the mean Silhouette value of all data elements. The resulting index can take values in the range [-1,1], with a higher value reflecting a better partitioning.

In the context of our experiments, we apply the Silhouette Width as follows: Assume a data set contains N items and, it can be partitioned into $k \in [3, 9]$ clusters by employing a clustering algorithm. The Silhouette values will be calculated for the partitions resulting from all choices of k. The clustering solution with the largest mean Silhouette value, and the associated optimal cluster number k^*, will be fed forward to the forecasting stage.

3.3.2. Weight selection

The use of a multicriteria clustering approach introduces an additional challenge for model selection, as several different partitions may be obtained for the same number of clusters. Specifically, in our analysis, we allow the weight ω to take 11 different values. Given the choice of the number of clusters k^* (determined using the Silhouette Width), we may still face a choice of up to 11 different partitions that reflect different trade-offs between the quality of fit with respect to the different information sources.

As discussed in Guyon et al. (2009), the success of clustering is best assessed in the context of the overall success of a particular application. In our scenario, the optimal ω^* for the distance function d_{ij}^{MC} should produce partitions that yield the best forecasting accuracy of a given forecasting algorithm for relevant lead time periods. We propose a simple methodology that aligns model selection with this overarching aim: specifically, we apply C-MSKF to each set of analogies, and assess its forecasting accuracy for the last in-sample time step. The partition producing the best average forecasting accuracy for this time step is selected for the prediction of future data points.

In this context, the measure employed to determine forecasting accuracy is the Mean Square Error (MSE), which is given as:

$$\mathrm{MSE} = \mathrm{mean}(e_t^2) \tag{8}$$

Here t indicates the forecasting time period, $e_t = X_t - F_t$, X_t is the observation of the time series X at time t, and F_t is the respective forecast.

3.4. Forecasting

In the forecasting stage, we employ the C-MSKF algorithm as our prediction method. In brief, C-MSKF is a Bayesian pooling approach, which combines parameter estimates from a univariate time series forecasting method (Dynamic Linear Model) with the parameter estimates derived from pooled data. The C-MSKF algorithm is an extension of the MSKF with the Conditionally Independent Hierarchical Model (CIHM: Kass and Steffey, 1989) using the DGS shrinkage formula (DGS's shrinkage: Duncan et al., 1993).

A full description of the C-MSKF algorithm is available in the literature (Duncan et al., 1993) and a summary is included in the Appendix. The aim of this paper is to demonstrate the advantage obtained by considering multiple sources of information during the clustering stage. Specifically, we aim to demonstrate that the resulting, more accurate, partitions lead to improvements in a pooling approach. Here, C-MSKF was chosen as a representative example, but experiments with other types of pooling approaches would be useful, and the general principles of our approach are expected to generalize to other forecasting methods that exploit analogies.

In a forecasting context, the forecasting origin T denotes the most recent data point used during model construction, while the forecasting horizon denotes the number of time steps into the future that predictions are made. In our experiments, C-MSKF is used to make forecasts for a range of prediction horizons. Specifically, for a given forecasting origin T, the h-step ahead forecast (for $h \geq 2$) is obtained by iteratively updating C-MSKF using the forecasts obtained for the $(T+1), \ldots, (T+h-1)th$ time steps, and predicting the succeeding time point.

3.5. Implementation

Our methods were implemented using a combination of R and Java. A full implementation is available through our repository at https://github.com/EmiaoLu/Analgoies

4. Empirical evaluation

4.1. Simulated data

For the initial testing of our methodology, simulated data sets are used. The advantages of simulated data lie in the full control over the properties of the data; in our case, it allows investigation into the algorithms' sensitivity to time series length and noise. A relevant real-world application, and results for this application setting, are presented later in this manuscript, in Section 5. For the simulated data, we generate data representing two information sources, *i.e.,* time series data as well as information about static variables (playing the role of causal factors) associated with each time series. We use a fairly simple setup at this point.

For the time series data, we aim to generate a set of time series that are correlated across an initial time interval but later display differing trend changes, due to an external influence that is shared across sub-sets of analogous series. In particular, we use a linear, logarithmic and piece-wise linear function, respectively, to describe these trend changes as a function of time t. Conceptually, the linear model can be interpreted as a time series that exhibits a stable increasing trend, while the logarithmic model reflects a decreasing rate of growth. Finally, the piece-wise linear function reflects a pattern change from a positive slope to a negative slope. The specific models used for these three generating functions $f_g(t)$, $g = 1, 2, 3$, are defined as follows:

$$f_1(t) = 0.8t + 2.8, \quad \text{if } 1 \leq t \leq q \tag{9}$$

$$f_2(t) = 4ln(t) + 2, \quad \text{if } 1 \leq t \leq q \tag{10}$$

$$f_3(t) = \begin{cases} 0.7t + 2.8, & \text{if } 1 \leq t \leq p \\ -0.9t + 25, & \text{if } p+1 \leq t \leq q \end{cases} \tag{11}$$

where parameter q defines the number of time points, and p defines the time of the trend change for the piece-wise linear function.

To obtain a set of analogous time series from a given generating function, we added normally-distributed noise to the trend at each time point.[2] Specifically, the noisy time series pattern X_{it} for time series i at time t, associated with generating function g, is obtained as follows:

$$X_{it} = \begin{cases} f_g(0) + N(f_g(t+1) - f_g(t), \sigma^2_{TS}), & \text{if } t = 1 \\ X_{i(t-1)} + N(f_g(t+1) - f_g(t), \sigma^2_{TS}), & \text{if } 1 < t \leq q-1 \end{cases} \tag{12}$$

where g represents the choice of generating function. The notation $N(\mu_{TS}, \sigma^2_{TS})$ describes a random variate drawn from a normal distribution with mean μ_{TS} and variance σ^2_{TS}; here σ^2_{TS} is static, but μ_{TS} changes over time and, for each time step t, is defined by the slope of the generating function $f_g(t+1) - f_g(t)$.

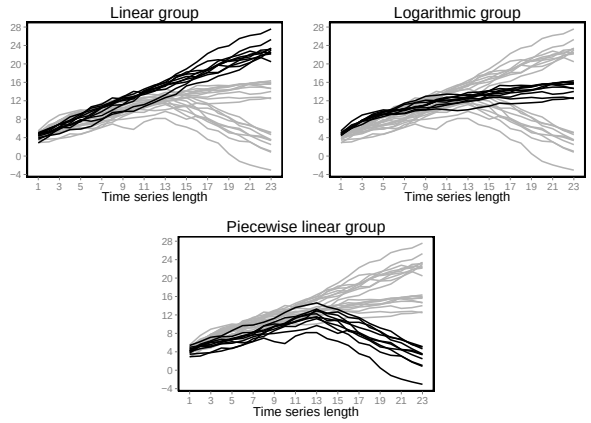
Using Equation (12), each generating function is used to obtain a set of I analogous time series of length $q-1$, exhibiting additive noise. An example of the resulting time series data is shown in Figure 1, and it is evident that differentiation between these series is challenging for earlier time intervals. Following Duncan et al. (1993), all time series are standardized individually using the z-score to improve the CIHM cross-sectional adjustment and remove any scale differences between clusters.

To obtain the second information source, we assume the presence of a single causal factor that governs the differences in behaviour between the time series.[3] In our simulated data, the ground truth (*i.e.,* the nature of the generating model for each time series) is known; this information could therefore be used

[2] This approach ensures the validity of a key assumption behind the C-MSKF algorithm which, due to its base in Kalman Filters, assumes normally-distributed noise.

[3] While a single factor is used in our experiments, the methodology generalizes to a feature space of arbitrary dimension (which may be categorical), as long as a suitable distance measure can be defined. The core property modelled here is simply the availability of two different, incommensurable and noisy feature spaces.

Figure 1: Illustration of raw time series data generated from a linear, logarithmic, and piecewise linear function.



to derive suitable (informative but noisy) data for the causal factor. Specifically, the value of the causal factor for time series i is drawn from the normal distribution $N(\mu_{CF}, \sigma^2_{CF})$, where $\mu_{CF} \in \{1, 2, 3\}$ corresponds to the index g of the generating function $f_g(t)$, associated with time series i (*i.e.,* it takes value in $1, \ldots, 3$).

It is evident that the use of two information sources is superfluous in the absence of noise in the individual information sources, and can only become beneficial in the presence of uncorrelated noise. To assess the impact of varying reliability of the different information sources, we adjust the levels of σ_{CF} and σ_{TS} relative to each other (see Table 1). Specifically, σ_{CF} is fixed at 0.35 while σ_{TS} is increased from 0.35 to 1.15 in steps of 0.2.

Table 1: Standard deviation used to generate simulated time series and causal factor data.

Scenarios	σ_{CF}	σ_{TS}
1	0.35	0.35
2	0.35	0.55
3	0.35	0.75
4	0.35	0.95
5	0.35	1.15

All other parameters are kept constant in the experiments, and are summarized in Table 2. The forecasting origin T is fixed at 17 throughout our analysis. This choice allows for the observation of more than 3 data points after the trend change of the time series, thus meeting one of the key assumptions behind the C-MSKF algorithm (see Section 1). The parameter l (Length selection) reflects the fact that we systematically drop the earliest historical points one at a time, while keeping the forecasting origin fixed, to consider the effect of shorter time series.

Overall, the above setup is used to obtain a set of 30 replicates (*i.e.,* 30 sets of 30 time series each), to support statistically sound analysis of the results.

Table 2: Constant parameters for the generation of simulated data

Parameter name	Value
Forecasting horizon	h=1, 2,. . . ,6
Forecasting origin	T=17
Length selection	l=12, 13,...,17
No. of time series in a group	I=10
Total No. of time points	q=24
Turning point	p=14

4.2. Contestant techniques

Our primary aim here is to analyze and compare the forecasting accuracy of prediction processes that employ analogies. We therefore define approaches based on the single-criterion clustering of causal factors (CF clustering), the single-criterion clustering of time series data (TS clustering) and the multicriteria clustering of both information sources (MC clustering). The multicriteria approach is described in detail in Section 3. The single-criterion approaches follow the same methodology, but differ in the choice of dissimilarity matrix (defined in Equation (4) and (5), rather than Equation (6)). Furthermore, they do not require the additional weight selection step outlined in Section 3.3.2, as a single partition is obtained for each choice of K.

In addition, we also benchmark our method against the basic MSKF algorithm (which makes no use of analogies), as well as a number of standard univariate forecasting approaches. Specifically, we employ Damped Exponential Smoothing (Damped), Drift, Exponential Smoothing (ETS), Random Walk (RW), and the Theta model. Brief details of these contestant techniques are provided in the Appendix. For the ETS method, we employed the automated implementation in the *forecast* R package.

4.3. Performance evaluation

In analyzing our results, we consider both the accuracy of the segmentation stage and the forecasting stage.

Forecasting error is evaluated using the Mean Squared Error, previously defined in Equation (8). Additionally, we also employ the Symmetric Mean Absolute Percentage Error, sMAPE (Bergmeir et al., 2016). This is slightly different from the version described in Makridakis and Hibon (2000), which makes no use of absolute values in the denominator. This modified version can correctly account for situations in which observations and forecasts have equal magnitude but opposite signs, and is given as:

$$\text{sMAPE} = \text{mean}(200\frac{|e_t|}{|X_t| + |F_t|}) \tag{13}$$

where all relevant variables have been defined previously (see Equation (8)). We assess forecasting error by calculating the average MSE and sMAPE across different prediction horizons, replicates, time series, and time series lengths. In order to provide further insight, some of our results are broken up by key aspects that are found to influence forecasting accuracy, specifically the noise scenario, the number of clusters, and the prediction horizon.

The accuracy with which analogies are identified is expected to have an impact on final forecasting accuracy. To evaluate the correctness of clustering solutions, we use the Adjusted Rand Index (ARI: Hubert and Arabie, 1985), an established cluster validation index that evaluates the agreement between two different groupings. Specifically, the ARI is employed to measure the consistency between each clustering solution and the ground truth, as defined by the generating models for the time series.

Using a representation based on the $L \times K$ contingency table defined by two partitions (of the same data) with L and K clusters, respectively, the Adjusted Rand Index between the two partitions is given as

$$\text{ARI} = \frac{\sum_{l,m}\binom{N_{lm}}{2} - [\sum_{l}\binom{N_{l\cdot}}{2} \cdot \sum_{k}\binom{N_{\cdot m}}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_{l}\binom{N_{l\cdot}}{2} + \sum_{m}\binom{N_{\cdot m}}{2}] - [\sum_{l}\binom{N_{l\cdot}}{2} \cdot \sum_{m}\binom{N_{\cdot m}}{2}]/\binom{N}{2}} \tag{14}$$

where N is the size of the data set, N_{lm} denotes the entry in row l and column m of the contingency table (*i.e.,* the number of data items that have been assigned to both cluster l and cluster m), and $N_{l\cdot}$ and $N_{\cdot m}$ represent row and column totals for row l and column m of the table, respectively.

The ARI has been constructed so that the expected value of two random partitions is 0, with the generalized hypergeometric distribution as the model of randomness. The ARI takes a maximum value of 1 and an expected minimum value of 0, with higher values indicating a closer match between the partitions considered. Values reported in our analysis are averages across different replicates.

4.4. Results

4.4.1. Preliminary experiments

Our initial focus is to understand whether better segmentation leads to improved forecasting. For this purpose, we eliminate the complicating aspect of automatic model selection (see Section 3.3), as this selection stage is likely to introduce additional errors.

Specifically, we analyze performance of the model associated with the best final MSE for a given number of clusters. We consider a range of different choices of K (in steps of 2)[4]. For each number of clusters, we report averages across a range of setups, namely variations in time series length and forecasting horizon.

With respect to the clustering performance, measured by average ARI, our findings (see Figure 2) show that, as may be expected, clustering performance decreases for all three approaches, as the number of clusters increases significantly beyond the ground truth. Yet, for the range of cluster sizes considered here, the MC clustering shows a superior clustering performance to the single-criterion clustering approaches (CF and TS clustering approaches) for the range from 3 to 8 clusters. This indicates that this method continues to benefit from the use of two complementary information sources, even in a scenario where the correct number of clusters is overestimated.

Comparing the forecasting results for C-MSKF based on the CF, MC and TS partitions (see Figure 3), we observe that MC's improved segmentation does translate into improved forecasting accuracy, for both evaluation measures.

These results are promising, as they highlight that our approach has the ability to generate better quality partitions and forecasts, in principle. Furthermore, the consistent performance advantage across a range of cluster numbers demonstrates that performance is not overly reliant on prior knowledge (or exact estimation) of the number of clusters.

4.4.2. Performance comparison across different noise levels

Generally, the selection of best forecasting results, as done in the previous experiment, is not feasible. In a practical scenario, use of the two model selection steps outlined in Section 3.3 will typically be fundamental, both in order to reduce computational cost and to identify a single forecasting model in the absence of access to future forecasting accuracy.

Evidently, both model selection steps in our approach can be expected to cause a drop in final forecasting accuracy, as additional room for error is introduced. However, the previous experiment indicates that performance is fairly robust with respect to the number of clusters, hence automated weight selection is likely to present a more problematic issue.

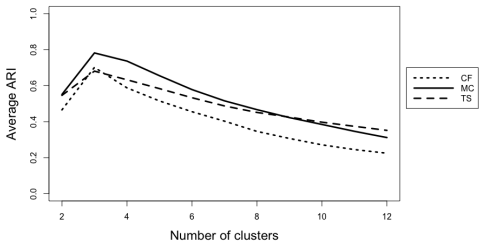
To explore the impact of automated weight selection in more detail, this section contrast the results obtained after the first model selection step (MC, which continues to select the weight for a given K by considering the best possible forecasting accuracy), with a fully automatic approach, $\text{MC}_{SilHist}$, that implements both of the model selection steps outlined in Section 3. To provide context to these results, we compare to the performance of CF and TS, MSKF and a range of established forecasting approaches. Key results are presented and discussed in the following, but additional analysis (mean and standard error of the difference for each pair of forecasting methods) is included in the Appendix.

Table 3 demonstrates that MC clustering generally continues to produce the best results (as established by MSE and sMAPE), after accounting for automatic K-selection alone. The performance of the fully automated approach $\text{MC}_{SilHist}$ is more mixed: for four out of five noise scenarios (specifically those scenarios where noise levels are not excessive), this method outperforms the single-criterion approaches (CF and TS). On the other hand, for the higher noise levels (S_3, S_4 and S_5), $\text{MC}_{SilHist}$ is alternatively outperformed by Damped, Drift or MSKF, pointing to limitations of our current weight selection step in dealing robustly with the increasingly noisy nature of the time series data.

Breaking up the results by prediction horizon (see Table 4), we can confirm the consistent advantage of C-MSKF when employing partitions that have been generated based on multicriteria clustering (MC and $\text{MC}_{SilHist}$), as compared to TS or CF clustering. Only for the highest noise level is $\text{MC}_{SilHist}$ method

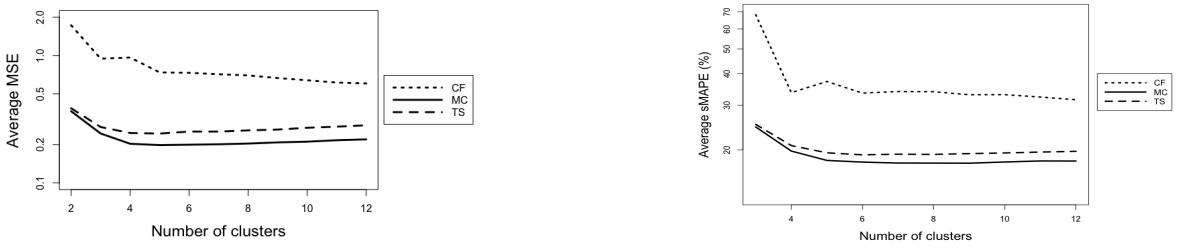
[4]Given the small scale of the data sets considered here, a maximum cluster size of 12 is employed, as further increases would encourage the identification of singleton clusters. For such clusters, C-MSKF will operate equivalently to MSKF, as no analogous series are available.

Figure 2: Comparison of clustering accuracy between CF, TS, MC methods (without weight selection) across different numbers of clusters. Data are generated using $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.35$. The expected results are reported here by taking the mean over 30 sets of simulated data, and 6 time series lengths for each set.



a)

Figure 3: Comparison of forecasting accuracy after the implementation of CF, TS, MC clustering methods (without weight selection) as the number of clusters sincreases from 2 to 12 in steps of 2. The data are generated using $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.35$. The expected results are obtained by taking the mean over 30 sets of simulated data, 6 forecasting horizons, 30 series and 6 time series lengths (to facilitate comparison, the y-axis is presented on a log-scale).



a) b)

Table 3: Summary of forecasting results for different noise levels (scenarios) of the time series patterns in the simulated data. For each noise scenario, average forecasting results are calculated by taking the mean across 30 replicates, 6 different time series lengths and forecasting horizon ranging from 1 to 6. For the $MC_{SilHist}$ method, the optimal weight is selected based on optimal (historical) forecasting accuracy, specifically the best MSE achieved for the forecasting origin $t = 17$. The best performance obtained for each setting is highlighted in bold face, with the second best performance highlighted in italic bold face.

	Scenarios	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta	TS
Average MSE	S_1	0.51	0.2	0.88	0.44	**0.16**	**0.16**	***0.17***	0.8	0.83	***0.17***
	S_2	0.83	0.64	1.07	0.96	**0.47**	***0.59***	0.72	1.04	1.06	0.68
	S_3	1.18	***0.98***	1.24	1.16	**0.80**	1.00	1.14	1.20	1.26	1.25
	S_4	1.57	1.39	***1.37***	1.52	**1.16**	1.47	1.82	1.39	1.47	1.80
	S_5	1.67	1.90	***1.56***	1.74	**1.37**	1.78	2.46	1.57	1.65	2.24
Average sMAPE (%)	S_1	34.38	21.75	56.77	29.28	***20.80***	21.22	**19.60**	57.73	56.79	21.89
	S_2	46.61	38.09	59.73	51.03	**34.23**	***37.6***7	39.26	61.15	60.53	39.09
	S_3	62.79	52.43	64.34	61.58	**49.77**	54.97	***50.62***	65.22	65.2	58.75
	S_4	68.76	58.1	63.69	65.64	**55.09**	60.90	***58.08***	65.48	65.66	62.57
	S_5	71.40	***65.14***	66.82	70.70	**61.19**	67.59	66.42	69.24	69.52	72.75

outperformed by the single-criterion CF approach, as the segments used in that approach remain unaffected by the noise on the TS data.

In summary, our results on simulated data confirm the hypothesis that the integration of two information sources, at the segmentation stage, can improve the forecasting accuracy of approaches that exploit analogies. This result holds even after the integration of automatic model selection. Importantly, this result relies on two key assumptions, including reasonable noise levels for both information sources and the absence of correlation of the noise across sources. If noise is either absent or damagingly high for one of the information sources, MC can only be expected to reach the performance achieved for the better of the single-criterion techniques.

5. Forecasting real data: personal income liability tax

Revenue forecasting for local governments is an important topic in the field of public budgeting research. It is regularly performed each fiscal year for the purpose of budget preparation and future planning of expenditure. In this section, we describe experiments conducted on annual personal income tax liability, covering the time period 1994 to 2007. The data was collected from the US Department of Taxation for multiple states. This

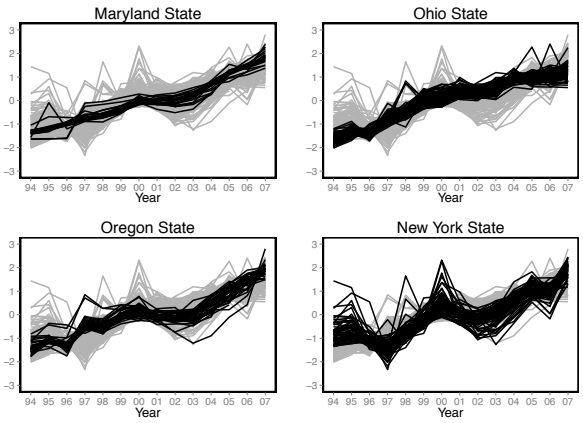
Table 4: In-depth comparison of the impact of different segmentation methods on C-MSKF's forecasting accuracy on the simulated data, broken up by noise level (scenario) and forecasting horizon ranging from 1 to 6. Shown are averages across 30 replicates and 6 different time series lengths. The best performance obtained for each setting is highlighted in bold face, with the second best performance highlighted in italic bold face.

	Scenarios	Methods	1-period ahead	2-period ahead	3-period ahead	4-period ahead	5-period ahead	6-period ahead
Average MSE	S_1	CF	0.27	0.36	0.45	0.55	0.66	0.78
		MC	**0.08**	**0.11**	**0.14**	**0.17**	**0.20**	**0.24**
		$MC_{SilHist}$	***0.09***	**0.11**	**0.14**	**0.17**	***0.21***	***0.25***
		TS	***0.09***	***0.12***	***0.15***	***0.19***	0.22	0.27
	S_2	CF	0.45	0.57	0.71	0.89	1.08	1.27
		MC	**0.23**	**0.31**	**0.40**	**0.52**	**0.63**	**0.75**
		$MC_{SilHist}$	***0.25***	***0.35***	***0.47***	***0.64***	***0.81***	***1.00***
		TS	0.29	0.40	0.54	0.74	0.95	1.18
	S_3	CF	0.61	0.80	1.02	1.26	1.54	1.85
		MC	**0.37**	**0.52**	**0.70**	**0.87**	**1.06**	**1.27**
		$MC_{SilHist}$	***0.40***	***0.58***	***0.82***	***1.08***	***1.38***	***1.72***
		TS	0.51	0.73	1.02	1.36	1.73	2.15
	S_4	CF	0.83	1.06	1.35	1.68	2.05	2.44
		MC	**0.51**	**0.73**	**1.00**	**1.28**	**1.57**	**1.88**
		$MC_{SilHist}$	***0.59***	***0.85***	***1.20***	***1.61***	***2.05***	***2.53***
		TS	0.71	1.04	1.45	1.95	2.50	3.11
	S_5	CF	0.94	1.18	***1.44***	***1.77***	***2.15***	***2.56***
		MC	**0.66**	**0.91**	**1.17**	**1.48**	**1.82**	**2.17**
		$MC_{SilHist}$	***0.75***	***1.08***	1.46	1.91	2.44	3.01
		TS	0.92	1.38	1.86	2.43	3.08	3.81
Average sMAPE (%)	S_1	CF	36.67	35.19	34.18	33.64	33.37	33.24
		MC	**23.23**	**21.68**	**20.65**	**20.06**	**19.68**	**19.49**
		$MC_{SilHist}$	***23.58***	***22.03***	***21.02***	***20.47***	***20.17***	***20.06***
		TS	24.19	22.66	21.69	21.16	20.87	20.77
	S_2	CF	48.32	47.20	46.35	46.07	45.90	45.82
		MC	**35.22**	**34.55**	**34.12**	**33.96**	**33.83**	**33.67**
		$MC_{SilHist}$	***38.35***	***37.62***	***37.30***	***37.39***	***37.59***	***37.75***
		TS	39.56	38.92	38.72	38.87	39.15	39.34
	S_3	CF	62.68	62.82	62.84	62.65	62.76	63.01
		MC	**49.46**	**50.32**	**50.15**	**49.66**	**49.57**	**49.44**
		$MC_{SilHist}$	***54.53***	***55.03***	***55.02***	***54.94***	***55.07***	***55.24***
		TS	57.24	58.43	58.88	59.04	59.33	59.60
	S_4	CF	69.10	68.68	69.02	68.70	68.59	68.43
		MC	**52.62**	**54.17**	**55.45**	**55.86**	**56.14**	**56.32**
		$MC_{SilHist}$	***58.38***	***59.56***	***60.95***	***61.60***	***62.23***	***62.70***
		TS	59.55	60.99	62.37	63.30	64.22	64.97
	S_5	CF	73.40	72.11	71.03	70.50	70.57	70.75
		MC	**60.10**	**61.17**	**61.23**	**61.38**	**61.59**	**61.68**
		$MC_{SilHist}$	***65.58***	***66.91***	***67.54***	***68.06***	***68.57***	***68.88***
		TS	68.80	71.37	72.74	73.64	74.55	75.39

Table 5: Summary of forecasting results for the personal income tax liability data, broken up by forecasting horizon ranging from 1 to 3. For the $MC_{SilHist}$ method, the optimal weight is selected based on optimal (historical) forecasting accuracy, specifically the best MSE achieved for the time step $t = 11$ (Year 2007). The best performance obtained for each setting is highlighted in bold face, with the second best performance highlighted in italic bold face.

Methods	Average MSE			Average sMAPE (%)		
	1-year ahead	2-year ahead	3-year ahead	1-year ahead	2-year ahead	3-year ahead
CF	0.45	***0.82***	***0.89***	27.13	***30.78***	***30.59***
Damped	0.69	1.08	1.58	36.45	37.75	41.77
Drift	0.48	0.82	1.22	30.23	32.41	36.30
ETS	0.74	1.29	2.04	40.55	44.10	50.98
MC	***0.41***	**0.76**	**0.87**	***25.46***	**29.81**	**30.16**
$MC_{SilHist}$	***0.41***	**0.76**	**0.87**	***25.46***	**29.81**	**30.16**
MSKF	**0.38**	0.89	1.15	**24.80**	31.09	32.47
RW	0.63	1.13	1.87	34.10	37.88	45.51
Theta	0.74	1.18	1.77	36.69	38.18	42.85
TS	0.51	0.88	1.05	29.66	33.41	34.04

Figure 4: Standardized time series of personal income tax in 208 counties in Maryland, New York, Ohio and Oregon State from 1994 to 2007.



type of forecasting task meets the conditions for the applicability of the C-MSKF algorithm, as summarized in Section 2.

In total, tax liability data for four states (namely Maryland, New York, Ohio and Oregon) is used, comprising a total of 208 counties. Note that two time series corresponding to Baltimore city and Somerset County (Maryland State) are excluded from the analysis as they show uncharacteristic income tax patterns, compared to all other time series. The set of time series (after standardization) is presented in Figure 4 and shows that counties pertaining to different states exhibit different sensitivity to the recession of the early 2000s (2001-2003) in the US. We can observe a small pattern change (a general slight slope change) for counties in Maryland and Ohio, while Oregon and New York show much bigger slope changes around this point in time.

5.1. Problem formulation

For the purpose of our analysis, the whole time period (1994-2007) is divided into two parts. The first 11 time points (1994-2004) of the time series are regarded as historical observations, while the hold-out forecasting period is defined to span 2005 to 2007. This choice is made to allow for more than 3 observations after the trend change caused by the economic recession. Thus, as the main conditions for use of C-MSKF are met, it is expected that C-MSKF may outperform conventional univariate time series forecasting methods in this scenario.

In the US, income tax is positively correlated with GDP and local economy, but also influenced by state-level policy. The particular patterns of income tax liability are therefore expected to differ in terms of different federal states, *i.e.,* state membership can be thought to represent a key driver behind differences in tax liability patterns. As the state of origin can be expected to be a noisy predictor of trend alone, we expect time series forecasting to benefit from the integration of all available data. In other words, the fiscal variable (federal states) and the historical time series points are considered as two separate information sources, which we aim to integrate using our multicriteria clustering approach.

To define the set of causal factors, the state name is recorded as a categorical variable associated with the time series of income tax liability, for each county. All other aspects of the methodology follow the description previously provided in Section 3 and Section 4.

5.2. Results

Table 5 shows forecasting accuracy of different methods across the three relevant prediction horizons. Additional analysis (mean and standard error of the difference for each pair of forecasting methods) is provided in Table 11 in the Appendix.

In line with previous work (Duncan et al., 1993), the MSKF method performs better than C-MSKF methods for the shortest forecasting horizon (1-step ahead), but its performance decreases as the prediction horizon increases. Considering all 1-step forecasts, MSKF achieves the best performance among all of the candidates, as measured by both average MSE and sMAPE. For the 2-step and 3-step ahead forecasts, our MC-based C-MSKF method outperforms all other approaches, both with and without automated model selection. In particular, the C-MSKF method using multicriteria clustering partitions outperforms the forecasting results obtained for the CF and TS partitions across all forecasting horizons considered, suggesting that the segments obtained are beneficial for forecasting.

6. Conclusions

This paper considers the selection of analogies, using clustering, in the context of time series forecasting. Specifically, we illustrate the sensitivity of a specific pooling approach, C-MSKF, to the segmentation stage and outline a methodology that enables the simultaneous consideration of multiple complementary information sources. Our experiments illustrate that this approach has the potential to feed through to distinct improvements in forecasting accuracy. The specific contributions of this manuscript are as follows: (i) We propose the concept of multicriteria segmentation in the context of forecasting analogous time series; (ii) We describe an automated approach to model selection in this setting; (iii) We illustrate the potential of our approach in improving forecasting accuracy for short time series; (iv) We provide new insights into the relationship between the accuracy of the segmentation stage and the performance of a forecasting algorithm that makes use of analogies. The use of pooling approaches has been previously shown to be appropriate in applications involving short time series or significant trend changes, and this is where we see the main applicability of our approach.

Our experiments using simulated data consider variations in relative noise levels of the available information sources, and the resulting impact on the performance of forecasting. As expected, both single-criterion forecasting approaches show an increased sensitivity to such variation, as compared to our multicriterion approach, which is flexible in catering for changes in the reliability of the sources. In the concrete real-world application considered here, causal factor information (*i.e.,* federal states) happens to carry a more reliable signal than time series information, as evident from the performance of the CF and TS methods. In general, the relative importance of the two sources is expected to vary by application domain, time series length and the amount of domain knowledge applied in defining appropriate causal factors. Exploring the impact of these factors in the context of other application areas presents an exciting area for future research.

In considering and varying the noise of different information sources, we have attempted to highlight one of the key factors likely to affect the viability of our approach. However, further benchmarking of our approach on other (simulated or real) data will be useful to further understand its strengths and limitations. In this context, it may be interesting to introduce varying levels of correlation into the noise models, to investigate the sensitivity of the approach to this aspect.

Our experiments do highlight a remaining sensitivity of our model selection approach to increasing noise levels in the time series data. This is likely to be caused by the fact that weight selection is currently achieved through the consideration of historical time series data and is thus directly affected by noise in this particular information source. In future work, we will be investigating alternative approaches to automating model selection.

References

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*, volume 30. Springer Science & Business Media.

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*, 22(3):583–598.

Assimakopoulos, V. and Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International journal of forecasting*, 16(4):521–530.

Bass, F. M. (1969). A new product growth for model consumer durables. *Management science*, 15(5):215–227.

Bergmeir, C., Hyndman, R. J., and Benítez, J. M. (2016). Bagging exponential smoothing methods using stl decomposition and box–cox transformation. *International Journal of Forecasting*, 32(2):303–312.

Brown, R. G. (2004). *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

Brusco, M. J., Cradit, J. D., and Stahl, S. (2002). A simulated annealing heuristic for a bicriterion partitioning problem in market segmentation. *Journal of Marketing Research*, 39(1):99–109.

Brusco, M. J., Cradit, J. D., and Tashchian, A. (2003). Multicriterion clusterwise regression for joint segmentation settings: An application to customer value. *Journal of Marketing Research*, 40(2):225–234.

Delattre, M. and Hansen, P. (1980). Bicriterion cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):277–291.

Duncan, G., Gorr, W., and Szczypula, J. (1993). Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting. *Management Science*, 39(3):275–293.

Duncan, G., Gorr, W., and Szczypula, J. (1994). Comparative study of cross sectional methods for time series with structural changes.

Duncan, G. T., Gorr, W., and Szczypula, J. (1995). Bayesian hierarchical forecasts for dynamic systems: Case study on backcasting school district income tax revenues. In *New Directions in Spatial Econometrics*, pages 322–358. Springer.

Duncan, G. T., Gorr, W. L., and Szczypula, J. (2001). Forecasting analogous time series. In *Principles of forecasting*, pages 195–213. Springer.

Ferligoj, A. and Batagelj, V. (1992). Direct multicriteria clustering algorithms. *Journal of Classification*, 9(1):43–61.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89.

Gardner, J., Everette, S., and McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, 31(10):1237–1246.

Goodwin, P., Dyussekeneva, K., and Meeran, S. (2013). The use of analogies in forecasting the annual sales of new electronics products. *IMA Journal of Management Mathematics*, 24(4):407–422.

Granger, C. W. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120.

Green, K. C. and Armstrong, J. S. (2007). Structured analogies for forecasting. *International Journal of Forecasting*, 23(3):365–376.

Greis, N. P. and Gilstein, C. Z. (1991). Empirical bayes methods for telecommunications forecasting. *International Journal of Forecasting*, 7(2):183–197.

Guyon, I., Von Luxburg, U., and Williamson, R. C. (2009). Clustering: Science or art. In *NIPS 2009 workshop on clustering theory*, pages 1–11.

Handl, J. and Knowles, J. (2007). An evolutionary approach to multiobjective clustering. *IEEE transactions on Evolutionary Computation*, 11(1):56–76.

Harrison, J. and West, M. (1999). *Bayesian forecasting & dynamic models*, volume 1030. Springer New York City.

Harrison, P. and Stevens, C. F. (1971). A bayesian approach to short-term forecasting. *Operational Research Quarterly*, pages 341–362.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 273–280. IEEE.

Kass, R. E. and Steffey, D. (1989). Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.
Kool, C. J. (1983). Forecasts with multi-state kalman filters. *Eduard Bomhoff, Monetary Uncertainty, app*, 1.
Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., and Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting*, 23(3):377–390.
Leitner, J. and Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information–a review of the literature. *European Journal of Operational Research*, 213(3):459–469.
Liao, T. W. (2005). Clustering of time series data-a survey. *Pattern recognition*, 38(11):1857–1874.
Liu, Y., Ram, S., Lusch, R. F., and Brusco, M. (2010). Multicriterion market segmentation: a new model, implementation, and evaluation. *Marketing Science*, 29(5):880–894.
Makridakis, S. and Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476.
Myers, J. H. (1996). Segmentation and positioning for strategic marketing decisions. American Marketing Association.
Nikolopoulos, K., Buxton, S., Khammash, M., and Stern, P. (2016). Forecasting branded and generic pharmaceuticals. *International Journal of Forecasting*, 32(2):344–357.
Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukos, V., and Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of Business Research*, 68(8):1785–1791.
Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). horses for courses in demand forecasting. *European Journal of Operational Research*, 237(1):152–163.
Piecyk, M. I. and McKinnon, A. C. (2010). Forecasting the carbon footprint of road freight transport in 2020. *International Journal of Production Economics*, 128(1):31–42.
Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
Savio, N. D. and Nikolopoulos, K. (2013). A strategic forecasting framework for governmental decision-making and planning. *International Journal of Forecasting*, 29(2):311–321.
Smith, W. R. (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1):3–8.
Stimson, J. A. (1985). Regression in space and time: A statistical essay. *American Journal of Political Science*, pages 914–947.
Thomakos, D. and Nikolopoulos, K. (2012). Fathoming the theta method for a unit root process. *IMA Journal of Management Mathematics*, 25(1):105–124.
Vriens, M., Wedel, M., and Wilms, T. (1996). Metric conjoint segmentation methods: A monte carlo comparison. *Journal of Marketing Research*, pages 73–85.
Webby, R. and O'Connor, M. (1996). Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, 12(1):91–118.

Appendix. Forecasting methods

In the presentation of the following methods, X_t refers to the actual observation at time t, F_t represents the respective forecast, and h refers to the forecasting horizon.

Random Walk. All lead time forecasts are equal to the value of the last actual observation.

$$F_{t+h} = X_t \tag{15}$$

Drift method. This is a variation of the Random Walk method. It additionally adjusts the forecasts to increase or decrease over time, where the amount of change over time (called the drift) is equal to the average change observed in the historical observations.

$$F_{t+h} = X_t + \frac{h}{t-1}(X_t - X_1) \tag{16}$$

Exponential Smoothing. Exponential Smoothing gives more weight to the latest observations, as they are more relevant for extrapolating to the future. Single Exponential Smoothing assumes no trend or seasonal patterns and operates by averaging (smoothing) the past values of a time series, using exponentially decreasing weights, as observations get older.

$$F_{t+1} = \alpha X_t + (1-\alpha)F_t \tag{17}$$

where α is the exponential smoothing parameter.

Holt Exponential Smoothing. Holt Exponential Smoothing expands Single Exponential Smoothing by adding one additional parameter for smoothing the short-term trend (Holt, 2004). The equations are given as follows:

$$\begin{aligned} L_t &= \alpha X_t + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ F_{t+h} &= L_t + hT_t \end{aligned} \tag{18}$$

where β is the smoothing parameter for the trend, L_t refers to the forecast of the level for period t, and T_t is the forecast for the trend at time t.

Damped Exponential Smoothing introduces a dampening factor (ϕ) that is multiplied with the trend component of Holt's method in order to provide more control regarding the long-term extrapolation of the trend (Gardner et al., 1985). Forecasts for Damped method can be calculated as:

$$\begin{aligned} L_t &= \alpha X_t + (1-\alpha)(L_{t-1} + \phi T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)\phi T_{t-1} \\ F_{t+h} &= L_t + \sum_{i=1}^{h} \phi^h T_t s \end{aligned} \tag{19}$$

Theta model. The Theta model (Assimakopoulos and Nikolopoulos, 2000; Thomakos and Nikolopoulos, 2012) decomposes the time series into two periods that are described as "Theta lines". The first Theta-line represents the long-term

trend of the data. The second Theta-line is extrapolated based on Single Exponential Smoothing that focuses on recent change. In the last step, a combined point forecast is achieved by combining the respective point forecasts produced by the first and second Theta-line using equal weights.

MSKF. The MSKF is a univariate time series forecasting method and appropriate for short time series subject to no changes, transient effects, step changes and slope changes. A detailed description of this method is provided in Harrison and Stevens (1971).

The basic model is given as follows:

$$\begin{aligned} X_t &= T_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, V_\varepsilon) \\ T_t &= T_{t-1} + S_t + \gamma_t, \quad \gamma_t \sim N(0, V_\gamma) \\ S_t &= S_{t-1} + \rho_t, \quad \rho_t \sim N(0, V_\rho) \end{aligned} \tag{20}$$

ε_t represents observational disturbance,
γ_t represents trend disturbance,
ρ_t represents slope disturbance,

where X_t is the observation at time t; T_t is the current trend value of X_t; S_t refers to the current slope value of X_t; ε_t, γ_t, ρ_t are random disturbances of the process at time t and assumed to be independently normally-distributed with a mean of 0 and variances V_ε, V_γ, and V_ρ, respectively.

In summary, the MSKF method can be implemented through five steps. The notation here is as follows: We work with the joint distribution of T_t and S_t, which jointly follow a bivariate normal distribution:

$$\begin{pmatrix} T_t \\ S_t \end{pmatrix} \sim N\left[\begin{pmatrix} m_t \\ b_t \end{pmatrix}, C_t = \begin{pmatrix} V_{11,t} & V_{12,t} \\ V_{12,t} & V_{22,t} \end{pmatrix}\right] \tag{21}$$

where C_t is the covariance matrix of (T_t, S_t) at time t; Φ_t refers to the entire set of moments that is used. Suffices and superscripts applied to Φ can be understood to be associated with each parameter in this set, *e.g.*,

$$\Phi_t^{(j)} = (m_t^{(j)}, b_t^{(j)}, C_t^{(j)})$$

Step 1. Suppose the posterior distribution $(T_{t-1}, S_{t-1}|X_{t-1})$ of observation X_{t-1} is a mixed bi-variate normal distribution:

$$(T_{t-1}, S_{t-1}|X_{t-1}) \sim \sum_{j=1}^{J} q_{t-1}^{(j)} N(\Phi_{t-1}^{(j)})$$

where the parameters of the distribution arise from state j at time $t-1$: $q_{t-1}^{(j)}$ is the posterior probability of being in state j at time $t-1$; the parameters $\Phi_{t-1}^{(j)}$ are known.

Step 2. The process is in one of four possible states ($j \in$ {no change, step change, slope change, transient}). At time t, the prior of the occurrence of X_t is given as:

π_j is the probability of state j

$V_\varepsilon^{(j)}$, $V_\gamma^{(j)}$, $V_\rho^{(j)}$ are the variances of the random disturbances $\varepsilon_t|j$, $\gamma_t|j$ and $\rho_t|j$ for state j at time t, respectively.

Step 3. From time $t-1$ to t, the Kalman Filter algorithm of Harrison and West (1999) is employed to update each component of the distribution:

$$(T_t, S_t|X_t) \sim \sum_{j=1}^{J} \sum_{k=1}^{J} p_t^{(j,k)} N(\Phi_{t-1}^{(j,k)})$$

where $p_t^{(j,k)}$ is the posterior probability with respect to observation X_t that the process was in state j in the period $t-1$ and is currently in state k.

The Kalman Filter recursive equations are employed to obtain the terms in the above equation:

$$\begin{aligned} m_t^{(j,k)} &= m_{t-1}^{(j)} + b_{t-1}^{(j)} + A_{1,t}^{(j,k)} e_t^{(j)} \\ b_t^{(j,k)} &= b_{t-1}^{(j)} + A_{2,t}^{(j,k)} e_t^{(j)} \\ V_{11,t}^{(j,k)} &= r_{11,t}^{(j,k)} - (A_{1,t}^{(j,k)})^2 V_e^{(k)} t \\ V_{12,t}^{(j,k)} &= r_{12,t}^{(j,k)} - A_{1,t}^{(j,k)} A_{2,t}^{(j,k)} V_e^{(k)} t \\ V_{22,t}^{(j,k)} &= r_{22,t}^{(j,k)} - (A_{2,t}^{(j,k)})^2 V_e^{(k)} t \\ p_t^{(j,k)} &= s(2\pi V_e^{(k)} t)^{-1/2} exp\Big\{ -(X_t - m_{t-1}^{(j)} - b_{t-1}^{(j)})^2 / 2V_e^{(k)} t \pi_j q_{t-1}^{(j)} \Big\} \end{aligned}$$

where each element of A_t acts similar to the "smoothing factor" in Exponential Smoothing methods; π_j refers to the probability of occurrence of state j; s is a probability normalization factor

$$\begin{aligned} e_t^{(j)} &= X_t - (m_{t-1}^{(j)} + b_{t-1}^{(j)}) \\ A_{1,t}^{(j,k)} &= r_{11,t}^{(j,k)} / V_e^{(k)} t \\ A_{2,t}^{(j,k)} &= r_{12,t}^{(j,k)} / V_e^{(k)} t \\ V_e^{(k)} t &= r_{11,t}^{(j,k)} + V_\varepsilon^{(k)} \\ r_{11,t}^{(k)} &= V_{11,t-1}^{(j)} + 2V_{12,t-1}^{(j)} + V_{22,t-1}^{(j)} + V_\gamma^{(k)} + V_\rho^{(k)} \\ r_{12,t}^{(k)} &= V_{12,t-1}^{(j)} + V_{22,t-1}^{(j)} + V_\rho^{(k)} \\ r_{22,t}^{(k)} &= V_{22,t-1}^{(j)} + V_\rho^{(k)} \end{aligned}$$

Step 4. The J^2-component distribution at the previous step is condensed into an approximately equivalent distribution:

$$(T_{t-1}, S_{t-1}|X_{t-1}) \sim \sum_{j=1}^{J} q_t^{(k)} N(\Phi_t^{(j)})$$

where $q_t^{(k)} = \sum_j p_t^{(j,k)}$ and the parameters $\Phi_t^{(k)}$ are given by:

$$m_t^{(k)} = \sum_i p_t^{(j,k)} m_t^{(j,k)} / q_t^{(k)}$$

$$b_t^{(k)} = \sum_i p_t^{(j,k)} b_t^{(j,k)} / q_t^{(k)}$$

$$V_{11,t}^{(k)} = \sum_j p_t^{(j,k)} (V_{11,t}^{(j,k)} + (m_t^{(j,k)} - m_t^{(k)})^2) / q_t^{(k)}$$

$$V_{12,t}^{(k)} = \sum_j p_t^{(j,k)} (V_{12,t}^{(j,k)} + (m_t^{(j,k)} - m_t^{(k)})(b_t^{(j,k)} - b_t^{(k)})) / q_t^{(k)}$$

$$V_{22,t}^{(k)} = \sum_j p_t^{(j,k)} (V_{22,t}^{(j,k)} + (b_t^{(j,k)} - b_t^{(k)})^2) / q_t^{(k)}$$

Step 5. The posterior distribution at the end of Step 4 is now in the same form as in Step 1. The updating procedure is repeated until all the historical observations are processed.

C-MSKF. The C-MSKF algorithm combines the capabilities of the MSKF (Harrison and Stevens, 1971) and the CIHM method (Kass and Steffey, 1989), which are both are standard, well-developed Bayesian approaches. The CIHM can be considered as a random effects method that pools information from analogous time series and boosts prediction accuracy and responsiveness. Here, the C-MSKF algorithm is summarized in six steps. Step one through five are repeated recursively for each series within a cluster. This method introduces the additional symbol i to indicate individual time series within a cluster, and additional steps are integrated to combine information available from clusters with that from a target series using the CIHM method. The algorithm syntax follows the definitions provided in previous work (Duncan et al., 1995). The C-MSKF algorithm employed for each cluster is presented as follows:

The models for four possible states ($j \in$ { no change, step change, slope change, transient }) are defined as:

$$X_{it} = T_{it} + \varepsilon_{it},\ \varepsilon_{it}|j \sim N(0, V_\varepsilon^{(j)} i)$$

$$T_{it} = T_{it-1} + S_{it} + \gamma_{it},\ \gamma_{it}|j \sim N(0, V_\gamma^{(j)} i)$$

$$S_{it} = S_{it-1} + \rho_{it},\ \rho_{it}|j \sim N(0, V_\rho^{(j)} i)$$

$$\text{Prior } (T_{i0}, S_{i0}|X_{i0}) \sim \sum_{j=1}^{J} q_{i0}^{(j)} N/((m_{i0}^{(j)}, b_{i0}^{(j)}), C_{i0}^{(j)})$$

where X_{it} is the observation for series i at time t; T_{it} is the current trend value X_{it}; and S_{it} is current slope value X_{it}.

$\varepsilon_{it}|j$, $\gamma_{it}|j$, $\rho_{it}|j$ are serially uncorrelated and mutually independent disturbance terms for each state j.

$$\begin{pmatrix} T_{it} \\ S_{it} \end{pmatrix} \sim N\left[\begin{pmatrix} m_{it} \\ b_{it} \end{pmatrix}, C_t = \begin{pmatrix} V_{11,it}^{(j)} & V_{12,it}^{(j)} \\ V_{12,it}^{(j)} & V_{22,it}^{(j)} \end{pmatrix} \right] \tag{22}$$

$m_{it,}^{(j)}, b_{it}^{(j)}$ are the means of T_{it} and S_{it} in state j

$C_{it}^{(j)}$ is the covariance matrix of (T_{it}, S_{it}) in state j for series i at time t, and

$q_{it}^{(j)}$ is the posterior probability of series i being in state j at time t.

The complete C-MSKF algorithm is presented by the following steps:

Step 1. Conditionally on X_{it-1} the joint distribution of (T_{it-1}, S_{it-1}) for series i at time $t-1$ is a mixture of bivariate normal distributions defined for each of the J states:

$$(T_{it-1}, S_{it-1}|X_{it-1}) \sim \sum_{j=1}^{J} q_{it-1}^{(j)} N((m_{it-1}^{(j)}, b_{it-1}^{(j)}), C_{it-1}^{(j)}).$$

Step 2. After the observation X_{it}, apply the Kalman Filter algorithm of Harrison and West (1999) to each of the J current components J times (since each of the current components at time $t-1$ can be in any state at time t). This operation creates J^2 (16 components since $J = 4$) normally-distributed components:

$$(T_{it}, S_{it}|X_{it}) \sim \sum_{k=1}^{J} \sum_{j=1}^{J} p_{it}^{(j,k)} N((m_{it}^{(j,k)}, b_{it}^{(j,k)}), C_{it}^{(j,k)})$$

where $p_{it}^{(j,k)}$ is the posterior probability with respect to observation X_{it} that the process was in state j in the period $t-1$ and is currently in state k.

The Kalman Filter recursive equations for the terms in the above formulae are:

$$m_{it}^{(j,k)} = m_{it-1}^{(j)} + b_{it-1}^{(j)} + A_{1,it}^{(j,k)} e_{it}^{(j)}$$

$$b_{it}^{(j,k)} = b_{it-1}^{(j)} + A_{2,it}^{(j,k)} e_{it}^{(j)}$$

$$V_{11,it}^{(j,k)} = r_{11,it}^{(j,k)} - (A_{1,it}^{(j,k)})^2 V_e^{(j,k)} it$$

$$V_{12,it}^{(j,k)} = r_{12,it}^{(j,k)} - A_{1,it}^{(j,k)} A_{2,it}^{(j,k)} V_e^{(j,k)} it$$

$$V_{22,it}^{(j,k)} = r_{22,it}^{(j,k)} - (A_{2,it}^{(j,k)})^2 V_e^{(j,k)} it$$

$$p_{it}^{(j,k)} = s(2\pi V_e^{(j,k)} it)^{-1/2} exp\Big\{ - (X_{it} - m_{it-1}^{(j)} - b_{it-1}^{(j)})^2 / 2V_e^{(j,k)} it \pi_j q_{it-1}^{(j)} \Big\}$$

where each element of A_{it} acts similar to "smoothing factor" in Exponential Smoothing methods; π_j is the probability of occurrence of state j (constant for each state j); s is a probability normalization factor.

$$e_{it}^{(j)} = X_{it} - (m_{it-1}^{(j)} + b_{it-1}^{(j)})$$

$$A_{1,it}^{(j,k)} = r_{11,it}^{(j,k)} / V_e^{(j,k)} it$$

$$A_{2,it}^{(j,k)} = r_{12,it}^{(j,k)} / V_e^{(j,k)} it, \text{ and where}$$

$$V_e^{(j,k)} it = r_{11,it}^{(j,k)} + V_\varepsilon^{(k)} i$$

$$r_{11,it}^{(j,k)} = V_{11,it-1}^{(j)} + 2V_{12,it-1}^{(j)} + V_{22,it-1}^{(j)} + V_\gamma^{(k)} i + V_\rho^{(k)} i,$$

$$r_{12,it}^{(j,k)} = V_{12,it-1}^{(j)} + V_{22,it-1}^{(j)} + V_{\rho}^{(k)} i$$

$$r_{22,it}^{(j,k)} = V_{22,it-1}^{(j)} + V_{\rho}^{(k)} i$$

Step 3. To achieve the form required in Step 1, collapse J^2 into a J component normal distribution:

$$(T_{it}, S_{it}|X_{it}) \sim \sum_{k=1}^{J} q_{it}^{(k)} N((m_{it}^{(k)}, b_{it}^{(k)}), C_{it}^{(k)})$$

Equations for collapsing densities are (see Kool (1983)):

$$q_{it}^{(k)} = \sum_{j} p_{it}^{(j,k)},$$
$$m_{it}^{(k)} = \sum_{j} p_{it}^{(j,k)} m_{it}^{(j,k)} / q_{it}^{(k)},$$
$$b_{it}^{(k)} = \sum_{j} p_{it}^{(j,k)} b_{it}^{(j,k)} / q_{it}^{(k)},$$
$$V_{11,it}^{(k)} = \sum_{j} p_{it}^{(j,k)} (V_{11,it}^{(j,k)} + (m_{it}^{(j,k)} - m_{it}^{(k)})^2) / q_{it}^{(k)},$$
$$V_{12,it}^{(k)} = \sum_{j} p_{it}^{(j,k)} (V_{12,it}^{(j,k)} + (m_{it}^{(j,k)} - m_{it}^{(k)})(b_{it}^{(j,k)} - b_{it}^{(k)})) / q_{it}^{(k)},$$
$$V_{22,it}^{(k)} = \sum_{j} p_{it}^{(j,k)} (V_{22,it}^{(j,k)} + (b_{it}^{(j,k)} - b_{it}^{(k)})^2) / q_{it}^{(k)}$$

Step 4. Repeat Steps 1 to 3 for each series given a cluster.

Step 5. Given the distribution for each analogous time series i, use the CIHM method to adjust means and variances for every series. The adjusted means of trends T_{it} are given by

$$E(m_{it}^{(j)}|T_{it}, \mu_0, \tau_0^2) = (\mu_0 V_{11,it}^{(j)} + T_{it}\tau_0^2)/(V_{11,it}^{(j)} + \tau_0^2)$$

where μ_0 and τ_0 are the MLEs of the hyperparameters μ and τ^2, they are the sample mean and the sample variance of $m_{1t}^{(j)}, m_{2t}^{(j)}, ...m_{lt}^{(j)}$,respectively. The adjusted variances of the trends T_{it} are given by

$$E(V_{11,it}^{(j)}|T_{it}, \vartheta_0, \nu_0) = (\vartheta_0 + (T_{it} - m_{it}^{(j)})^2)/(\nu_0 - 1)$$

where ϑ_0 and ν_0 are the MLEs of the hyperparameters ϑ and ν found by solving the likelihood equations

$$\vartheta = I\nu / \Big\{ \sum_{i=1}^{I} 1/V_{11,it}^{(j)} \Big\}$$

$$\Gamma'(\nu/2)/\Gamma(\nu/2) = (1/2)\Big\{ \log \vartheta - \log 2 - (1/I) \sum_{i=1}^{I} \log V_{11,it}^{(j)} \Big\}$$

where I refers to number of series in a cluster.

Step 6. Repeat the five steps above until all the historical observations are processed.

When Step 6 is completed, the final distributions prepared are utilized to forecast each series i individually.

Appendix. Paired comparison of approaches

To confirm the statistical significance of performance differences on the simulated data, we break up the forecasting results by differences in the forecasting horizon (h-step forecast with $h = 1, \ldots, 6$) and time series lengths $l = 12, \ldots, 17$. Every two forecasting methods are paired and the mean and standard error of the difference across the replicates are presented in Table 6, 7,..., 10. In conclusion, the MC method generally performs the best from scenario 1 to scenario 5, as measured by average MSE and sMAPE, except for scenario 1 where MSKF outperforms MC method as measured by average sMAPE. Additionally, as σ_{TS} increases from 0.35 to 1.15, the performance gap between MC's forecasting accura and that of TS increases, and the same conclusion also applies to $\text{MC}_{SilHist}$ and TS. Comparing the difference between CF and MC-based forecasting methods, including MC and $\text{MC}_{SilHist}$, the gap closes and eventually (for the highest noise setting) CF starts to outperform the $\text{MC}_{SilHist}$ clustering method, although it continues to perform worse than the MC method. This reflects the fact that the noise levels of time series information sources has a negative impact on MC's model selection step which relies on the noisy time series data. From a theoretical perspective, the MC approach with optimal model selection should always be able to meet or outperform the better performer amongst the CF and TS approaches.

Table 11 considers the significance of performance differences for the income tax liability data. For these data, weight selection in the $\text{MC}_{SilHist}$ performs well in picking up the final partitioning based on historical forecasting accuracy at time $t = 11$. Aggregating results for different horizons, we can identify that MC and $\text{MC}_{SilHist}$ perform best among the contestant forecasting methods.

Table 6: Scenario 1: $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.35$. The mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 30 sets of time series data, 30 series, 6 lengths and 6 prediction horizons. The standard error is calculated by breaking up the data across 6 lengths, 6 forecasting horizons and 30 replicates.

Methods	Average MSE									Average sMAPE (%)								
	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta
Damped	0.31									12.64								
	(0.01)									(0.35)								
Drift	-0.37	-0.68								-22.39	-35.02							
	(0.02)	(0.02)								(0.45)	(0.26)							
ETS	0.07	-0.25	0.44							5.11	-7.53	27.49						
	(0.01)	(0.01)	(0.01)							(0.44)	(0.26)	(0.45)						
MC	0.35	0.04	0.72	0.29						13.58	0.95	35.97	8.48					
	(0.01)	(0)	(0.02)	(0.01)						(0.31)	(0.17)	(0.34)	(0.31)					
$MC_{SilHist}$	0.35	0.03	0.72	0.28	-0.01					13.16	0.52	35.55	8.05	-0.42				
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)					(0.33)	(0.26)	(0.38)	(0.36)	(0.19)				
MSKF	0.34	0.02	0.71	0.27	-0.02	-0.01				14.79	2.15	37.17	9.68	1.2	1.63			
	(0.01)	(0)	(0.02)	(0.01)	(0)	(0.01)				(0.35)	(0.10)	(0.31)	(0.24)	(0.16)	(0.26)			
RW	-0.29	-0.6	0.08	-0.35	-0.64	-0.63	-0.62			-23.35	-35.98	-0.96	-28.45	-36.93	-36.51	-38.13		
	(0.01)	(0.01)	(0)	(0.01)	(0.02)	(0.02)	(0.02)			(0.48)	(0.27)	(0.14)	(0.43)	(0.37)	(0.42)	(0.31)		
Theta	-0.32	-0.64	0.05	-0.39	-0.68	-0.67	-0.66	-0.04		-22.41	-35.04	-0.02	-27.51	-35.99	-35.57	-37.19	0.94	
	(0.02)	(0.01)	(0)	(0.01)	(0.02)	(0.02)	(0.02)	(0)		(0.45)	(0.25)	(0.06)	(0.42)	(0.33)	(0.37)	(0.29)	(0.11)	
TS	0.34	0.02	0.7	0.27	-0.02	-0.01	0	0.62	0.66	12.49	-0.15	34.88	7.38	-1.09	-0.67	-2.3	35.84	34.9
	(0.01)	(0)	(0.02)	(0.01)	(0)	(0.01)	(0)	(0.02)	(0.02)	(0.32)	(0.17)	(0.34)	(0.31)	(0.04)	(0.19)	(0.16)	(0.38)	(0.33)

Mean values are not placed in parentheses.
Standard errors are placed in parentheses.

Table 7: Scenario 2: $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.55$. The mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 30 sets of time series data, 30 series, 6 lengths and 6 prediction horizons. The standard error is calculated by breaking up the data across 6 lengths, 6 forecasting horizons and 30 replicates.

Methods	Average MSE									Average sMAPE (%)								
	CF	Damped	Drift	ETS	MC	$MC_{silHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	MC	$MC_{silHist}$	MSKF	RW	Theta
CF																		
Damped	0.18									8.52								
	(0.02)									(0.41)								
Drift	-0.24	-0.43								-13.12	-21.64							
	(0.02)	(0.01)								(0.47)	(0.27)							
ETS	-0.13	-0.32	0.11							-4.42	-12.94	8.7						
	(0.02)	(0.01)	(0.01)							(0.48)	(0.20)	(0.26)						
MC	0.35	0.17	0.6	0.49						12.39	3.86	25.5	16.81					
	(0.01)	(0.01)	(0.02)	(0.02)						(0.37)	(0.25)	(0.35)	(0.33)					
$MC_{silHist}$	0.24	0.06	0.48	0.37	-0.11					8.95	0.42	22.06	13.37	-3.44				
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)					(0.24)	(0.39)	(0.44)	(0.47)	(0.32)				
MSKF	0.11	-0.07	0.35	0.24	-0.24	-0.13				7.35	-1.17	20.47	11.77	-5.03	-1.59			
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)				(0.41)	(0.19)	(0.26)	(0.24)	(0.24)	(0.39)			
RW	-0.21	-0.4	0.03	-0.08	-0.57	-0.45	-0.32			-14.54	-23.06	-1.42	-10.12	-26.92	-23.48	-21.89		
	(0.02)	(0.01)	(0.00)	(0.01)	(0.02)	(0.02)	(0.01)			(0.51)	(0.27)	(0.17)	(0.21)	(0.40)	(0.50)	(0.28)		
Theta	-0.23	-0.42	0.01	-0.1	-0.59	-0.47	-0.34	-0.02		-13.92	-22.44	-0.8	-9.5	-26.3	-22.86	-21.27	0.62	
	(0.02)	(0.01)	(0.00)	(0.01)	(0.02)	(0.02)	(0.01)	(0.00)		(0.48)	(0.25)	(0.08)	(0.22)	(0.36)	(0.46)	(0.26)	(0.12)	
TS	0.14	-0.04	0.39	0.28	-0.21	-0.1	0.03	0.36	0.38	7.52	-1	20.64	11.94	-4.87	-1.43	0.17	22.06	21.44
	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.35)	(0.27)	(0.35)	(0.35)	(0.16)	(0.30)	(0.25)	(0.40)	(0.36)

Mean values are not placed in parentheses.
Standard errors are placed in parentheses.

Table 8: Scenario 3: $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.75$. The mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 30 sets of time series data, 30 series, 6 lengths and 6 prediction horizons. The standard error is calculated by breaking up the data across 6 lengths, 6 forecasting horizons and 30 replicates.

Methods	Average MSE									Average sMAPE (%)								
	CF	Damped	Drift	ETS	MC	$MC_{silHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	MC	$MC_{silHist}$	MSKF	RW	Theta
CF																		
Damped	0.2									10.37								
	(0.01)									(0.41)								
Drift	-0.06	-0.26								-1.54	-11.91							
	(0.02)	(0.01)								(0.43)	(0.22)							
ETS	0.02	-0.18	0.09							1.21	-9.16	2.75						
	(0.01)	(0.01)	(0.01)							(0.45)	(0.20)	(0.18)						
MC	0.38	0.18	0.44	0.36						13.03	2.66	14.57	11.82					
	(0.02)	(0.02)	(0.02)	(0.02)						(0.33)	(0.34)	(0.40)	(0.41)					
$MC_{silHist}$	0.18	-0.02	0.24	0.16	-0.2					7.82	-2.54	9.37	6.61	-5.2				
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)					(0.26)	(0.42)	(0.44)	(0.46)	(0.34)				
MSKF	0.04	-0.16	0.11	0.02	-0.34	-0.14				12.17	1.81	13.72	10.97	-0.85	4.35			
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)				(0.38)	(0.21)	(0.30)	(0.29)	(0.31)	(0.40)			
RW	-0.02	-0.22	0.04	-0.05	-0.4	-0.2	-0.07			-2.42	-12.79	-0.88	-3.63	-15.45	-10.25	-14.6		
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)			(0.45)	(0.23)	(0.15)	(0.14)	(0.43)	(0.46)	(0.28)		
Theta	-0.08	-0.28	-0.01	-0.1	-0.46	-0.26	-0.12	-0.05		-2.41	-12.77	-0.87	-3.62	-15.44	-10.23	-14.58	0.02	
	(0.02)	(0.01)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)	(0.00)		(0.44)	(0.21)	(0.09)	(0.14)	(0.41)	(0.45)	(0.29)	(0.11)	
TS	-0.07	-0.27	-0.01	-0.09	-0.45	-0.25	-0.11	-0.05	0.01	4.04	-6.33	5.58	2.83	-8.99	-3.78	-8.13	6.47	6.45
	(0.03)	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.41)	(0.37)	(0.42)	(0.43)	(0.29)	(0.39)	(0.37)	(0.44)	(0.43)

Mean values are not placed in parentheses.
Standard errors are placed in parentheses.

Table 9: Scenario 4: $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 0.95$. The mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 30 sets of time series data, 30 series, 6 lengths and 6 prediction horizons. The standard error is calculated by breaking up the data across 6 lengths, 6 forecasting horizons and 30 replicates.

Methods	Average MSE									Average sMAPE (%)								
	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta
Damped	0.17									10.65								
	(0.02)									(0.42)								
Drift	0.2	0.03								5.07	-5.58							
	(0.01)	(0.02)								(0.43)	(0.19)							
ETS	0.05	-0.12	-0.15							3.12	-7.53	-1.95						
	(0.02)	(0.02)	(0.01)							(0.47)	(0.17)	(0.18)						
MC	0.41	0.23	0.21	0.36						13.66	3.01	8.59	10.54					
	(0.02)	(0.02)	(0.02)	(0.02)						(0.40)	(0.37)	(0.38)	(0.42)					
$MC_{SilHist}$	0.1	-0.08	-0.11	0.05	-0.31					7.85	-2.8	2.78	4.73	-5.81				
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)					(0.13)	(0.42)	(0.43)	(0.47)	(0.39)				
MSKF	-0.25	-0.43	-0.45	-0.3	-0.66	-0.35				10.67	0.02	5.6	7.55	-2.99	2.82			
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)				(0.42)	(0.23)	(0.23)	(0.22)	(0.36)	(0.42)			
RW	0.18	0	-0.02	0.13	-0.23	0.08	0.43			3.27	-7.38	-1.8	0.15	-10.39	-4.58	-7.4		
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)			(0.46)	(0.20)	(0.14)	(0.13)	(0.41)	(0.46)	(0.21)		
Theta	0.1	-0.08	-0.1	0.05	-0.31	0	0.35	-0.08		3.1	-7.56	-1.97	-0.02	-10.57	-4.76	-7.58	-0.18	
	(0.02)	(0.02)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.01)		(0.44)	(0.17)	(0.11)	(0.13)	(0.40)	(0.44)	(0.23)	(0.12)	
TS	-0.23	-0.4	-0.43	-0.28	-0.63	-0.32	0.03	-0.4	-0.32	6.19	-4.46	1.12	3.07	-7.47	-1.66	-4.48	2.92	3.09
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.42)	(0.34)	(0.37)	(0.39)	(0.28)	(0.41)	(0.34)	(0.38)	(0.38)

Mean values are not placed in parentheses.
Standard errors are placed in parentheses.

Table 10: Scenario 5: $\sigma_{CF} = 0.35$ and $\sigma_{TS} = 1.15$. The mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 30 sets of time series data, 30 series, 6 lengths and 6 prediction horizons. The standard error is calculated by breaking up the data across 6 lengths, 6 forecasting horizons and 30 replicates.

Methods	Average MSE									Average sMAPE (%)								
	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	MC	$MC_{SilHist}$	MSKF	RW	Theta
Damped	-0.23									6.26								
	(0.04)									(0.38)								
Drift	0.11	0.34								4.57	-1.68							
	(0.02)	(0.03)								(0.37)	(0.20)							
ETS	-0.07	0.16	-0.18							0.7	-5.56	-3.88						
	(0.03)	(0.03)	(0.02)							(0.43)	(0.16)	(0.18)						
MC	0.3	0.53	0.2	0.38						10.21	3.95	5.63	9.51					
	(0.02)	(0.04)	(0.02)	(0.03)						(0.40)	(0.42)	(0.45)	(0.46)					
$MC_{SilHist}$	-0.1	0.13	-0.21	-0.03	-0.41					3.81	-2.45	-0.77	3.11	-6.4				
	(0.01)	(0.04)	(0.02)	(0.03)	(0.02)					(0.14)	(0.38)	(0.38)	(0.44)	(0.41)				
MSKF	-0.79	-0.56	-0.89	-0.71	-1.09	-0.68				4.98	-1.28	0.4	4.28	-5.23	1.17			
	(0.05)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)				(0.44)	(0.25)	(0.27)	(0.27)	(0.42)	(0.44)			
RW	0.11	0.34	0	0.18	-0.2	0.21	0.89			2.15	-4.1	-2.42	1.45	-8.05	-1.65	-2.82		
	(0.02)	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.04)			(0.43)	(0.19)	(0.14)	(0.13)	(0.45)	(0.44)	(0.23)		
Theta	0.03	0.26	-0.08	0.1	-0.28	0.13	0.81	-0.08		1.88	-4.38	-2.7	1.18	-8.33	-1.93	-3.1	-0.27	
	(0.02)	(0.03)	(0.01)	(0.02)	(0.02)	(0.02)	(0.05)	(0.01)		(0.41)	(0.17)	(0.11)	(0.13)	(0.47)	(0.42)	(0.28)	(0.13)	
TS	-0.57	-0.34	-0.68	-0.5	-0.88	-0.47	0.21	-0.68	-0.6	-1.35	-7.61	-5.93	-2.05	-11.56	-5.16	-6.33	-3.51	-3.23
	(0.03)	(0.04)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.46)	(0.39)	(0.43)	(0.43)	(0.35)	(0.46)	(0.40)	(0.43)	(0.44)

Mean values are not placed in parentheses.
Standard errors are placed in parentheses.

Table 11: Income tax liability data: the mean and standard error of the difference between the column and row. The mean is obtained by taking the average across 3 forecasting horizons. The standard error is calculated by breaking up the data across 3 horizons and 208 time series.

Methods	Average MSE								Average sMAPE (%)							
	CF	Damped	Drift	ETS	$MC_{SilHist}$	MSKF	RW	Theta	CF	Damped	Drift	ETS	$MC_{SilHist}$	MSKF	RW	Theta
Damped	-0.40								-9.16							
	(0.15)								(1.55)							
Drift	-0.12	0.28							-3.48	5.68						
	(0.13)	(0.04)							(1.39)	(0.73)						
ETS	-0.64	-0.24	-0.52						-15.71	-6.55	-12.23					
	(0.17)	(0.04)	(0.06)						(1.88)	(0.91)	(1.13)					
$MC_{SilHist}$	0.04	0.44	0.16	0.68					1.02	10.18	4.50	16.73				
	(0.11)	(0.15)	(0.13)	(0.17)					(0.82)	(1.58)	(1.41)	(1.92)				
MSKF	-0.09	0.31	0.03	0.55	-0.13				0.05	9.21	3.53	15.76	-0.97			
	(0.11)	(0.15)	(0.13)	(0.17)	(0.11)				(1.21)	(1.83)	(1.66)	(2.14)	(1.19)			
RW	-0.49	-0.10	-0.37	0.15	-0.53	-0.40			-9.66	-0.50	-6.18	6.05	-10.68	-9.71		
	(0.15)	(0.03)	(0.03)	(0.03)	(0.15)	(0.15)			(1.53)	(0.84)	(0.77)	(0.90)	(1.54)	(1.82)		
Theta	-0.51	-0.12	-0.39	0.13	-0.55	-0.42	-0.02		-9.74	-0.58	-6.26	5.97	-10.76	-9.79	-0.08	
	(0.17)	(0.03)	(0.06)	(0.04)	(0.16)	(0.17)	(0.04)		(1.69)	(0.70)	(0.87)	(0.84)	(1.72)	(1.97)	(0.71)	
TS	-0.10	0.30	0.02	0.54	-0.14	-0.01	0.40	0.42	-2.87	6.29	0.61	12.84	-3.89	-2.92	6.79	6.87
	(0.11)	(0.14)	(0.12)	(0.16)	(0.11)	(0.11)	(0.14)	(0.16)	(1.02)	(1.57)	(1.43)	(1.91)	(1.00)	(1.26)	(1.57)	(1.72)

Here, MC has the same performance as $MC_{SilHist}$ method.