Segmentation Approaches for the Identification of Analogies in a Forecasting Context

A thesis submitted to the University of Manchester

for the degree of Doctor of Philosophy

in the Faculty of Humanities

2018

By

Emiao Lu

Alliance Manchester Business School

Contents

Word Count: around 48,000

List of Tables

List of Figures

List of Abbreviations

AIC: Akaike Information Criterion

AR: Autoregressive Process

ARI: Adjusted Rand Index

ARIMA: Autoregressive Integrated Moving Average

ARMA: Autoregressive Moving Average

BagClust$_{CF}$: bagged Cross-Sectional Multi-State Kalman Filter model using a single-criterion clustering of causal factor data (perturbing the data during the segmentation stage)

BagClust$_{MC}$: bagged Cross-Sectional Multi-State Kalman Filter model using a multi-criteria clustering of causal factors and time series data (perturbing the data during the segmentation stage)

BagClust$_{TS}$: bagged Cross-Sectional Multi-State Kalman Filter model using a single-criterion clustering of time series data (perturbing the data during the segmentation stage)

BagFcst$_{CF}$: bagged Cross-Sectional Multi-State Kalman Filter model using a single-criterion clustering of causal factor data (perturbing the data during the forecasting stage)

BagFcst$_{MC}$: bagged Cross-Sectional Multi-State Kalman Filter model using a multicriteria clustering of causal factors and time series data (perturbing the data during the forecasting stage)

BagFcst$_{TS}$: bagged Cross-Sectional Multi-State Kalman Filter model using a single-criterion clustering of time series data (perturbing the data during the forecasting stage)

Bass: Bass Diffusion Model

CIHM: the Conditionally Independent Hierarchical Model

CF: single-criterion clustering of causal factors

C-MSKF: Cross-Sectional Multi-State Kalman Filter

Damped: exponential smoothing with damped trends

DFT: discrete Fourier transform

DLM: Dynamic Linear Model

DGP: data generation process

Drift: random walk or naïve approaches with the drift method

ETS: exponential smoothing methods

Eucl: Euclidean distance

Fig.: figure

IID: identical and independent distribution

MSE: Mean Absolute Error

MSKF: Multi-State Kalman Filter

MBB: moving block bootstrap

MC: multicriteria clustering approaches

MnMx: Min-Max method

PAM: Partitioning Around Medoids

Pear: Pearson

RW: random walk or naïve method

sMAPE: Symmetric Mean Absolute Percentage Error

Sil: Silhouette Width

STING: statistical information grid-based

TBB: transformation based bootstrap

TS: single-criterion clustering of time series data

Zsc: the z-score method

Abstract

This thesis considers the problem of analogy identification in the context of forecasting. We develop and test a range of segmentation approaches, with the aim of improving the accuracy of forecasting methods that employ analogies.

The first manuscript of the thesis outlines our core methodological framework. This framework describes a forecasting process that integrates a multicriteria segmentation approach using a weighted-sum method for the identification of analogies during the segmentation stage. This combines the information from past realizations of a set of time series with information about the factors that govern the patterns observed, at the level of the distance function. Using simulated and real-world data, we illustrate that a concurrent consideration of multiple criteria at the segmentation stage can help to achieve better clustering results, which feed forward to improved forecasting accuracy. This paper contributes to the first methodological framework for the forecasting of analogous time series. Mulcriterion segmentation approaches demonstrate a significant improvement in the forecasting performance compared to single-criterion segmentation methods.

The second manuscript focuses on discussing the model selection problem related to the use of multicriteria clustering approaches. Although multicriteria approaches to clustering are advantageous to the final increase of forecasting accuracy, the use of these approaches introduces the challenge of an additional model selection during the

segmentation stage. This is because even for the same number of clusters, multicriteria clustering approaches may return sets of clustering solutions that reflect different trade-offs between the conflicting criteria. Therefore, this thesis also includes work addressing the model selection problem for multicriteria clustering in a forecasting context. We demonstrate that the quality of clustering solutions is best assessed in the problem-specific (forecasting) context. Computationally, this is the most expensive approach, and we, therefore, describe a compromise, which uses a standard internal validation technique (the Silhouette Width measure) for the determination of clusters, but performs weight selection based on the best average (historical) forecasting performance of the forecasting algorithm.

Further, the third manuscript addresses instability issues stemming from the clustering procedures by integrating bagging techniques into the forecasting process. Segmentations of analogies have been reported to give rise to further increase in the final forecasting accuracy, but the application of clustering techniques in the segmentation stage may result in instabilities related to the model selection step. By combining the forecasts derived from multiple models, the aggregated forecast is expected to lower down the uncertainty of the results via the aggregation scheme. We, therefore, employ the bootstrap aggregation techniques to further improve the forecasting process, and this results in a further boost to the forecasting accuracy.

In the final manuscript, we consider the use of multicriteria approaches in time series clustering, where multiple criteria (*i.e.,* distance metrics and/or normalization techniques) are available, but where these relate to time series data alone. Different distance metrics / standardization techniques may emphasize different notions of similarity. In applications where we are not sure which notion of similarity is accurate or where several notions of similarity are relevant, we might benefit from combining multiple distance metrics / standardization techniques, to capture complementary notions

of similarity. Our findings suggest a continued advantage of multicriteria clustering approaches in this context.

Keywords: Analogies; Bootstrap aggregating; Time series forecasting; Model Selection; Multicriteria clustering

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or

Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses

Acknowledgements

I would like to thank Dr. Julia Handl and Prof Dong-ling Xu for their outstanding role as my supervisors. I express extra thanks and gratitude especially to Julia who contributed and inspired me with her passion, energy and care with regards to my research and Ph.D. studies. I would not be able to finish this work without her constructive advice, suggestions and critical feedback. I thank Dong-ling, who offered me various opportunities for collaboration with companies or organizations and I am also grateful for her kind advice concerning life in the UK. Because of both of you, my experience of staying in the UK has significantly broadened my horizons and understanding.

I would like to thank Prof Jian-bo Yang for his outstanding role in leading the Decision and Cognitive Sciences Research Centre, which offered me so many opportunities to disseminate and obtain feedback from colleagues and experienced supervisors. In addition, I thank Dr. Yu-wang Chen and Prof Konstantinos Nikolopoulos for their insightful and inspiring feedback regarding my Ph.D. thesis.

Finally, I would express my gratitude to my lovely family members for their support and trust. In addition, I want to express my gratitude for my lovely friends, especially Giselle, Katherine, Leon, Jade, and Jas, who always had my back when I needed help. I could not finish this work without all of their support and patient listening.

Chapter 1

Introduction

1.1 Research context

Time series forecasting is inherently a challenging research topic that has gleaned intense interest from both academics and practitioners. A variety of time series forecasting models have been proposed in the forecasting literature and further developed to tackle particular challenges of forecasting.

Statistical forecasts are dedicated for the development of models based on historical observations. The models are then utilized to extrapolate the time series into the future. These models are the most popular ones applied in the domain of time series forecasting, as they are usually simple to implement and inexpensive to operate. De Gooijer and Hyndman (2006) provided a comprehensive survey regarding time series forecasting methods by summarizing 380 journal papers available in the forecasting literature and classified these methods according to the models and various problems that they address. In general, the families of exponential smoothing and ARIMA are the most well-known tools for the addressing of the time series forecasting problems. Univariate time series models including the simple exponential smoothing and Box Jenkins

ARIMA are the basic forms proposed in previous work, and these can be further developed into multivariate versions. Univariate time series model regards lag values of itself as independent variables. Multivariate time series model extends the univariate case to incorporate two or more independent variables. It does not limit itself to its past information but it also contains the past of other variables. These models have been applied to circumstances where several related time series are observed simultaneously over time. Multivariate time series approaches model relationships among time series variables. These relationships are often investigated by considering the correlation structures among the component series. Other models such as combining forecasts has been proposed to address particular forecasting challenges.

The use of statistical forecasts is advantageous in scenarios where no satisfactory extraneous information is available, or the information can only be obtained at prohibitively high costs. These methods, however, can provide unsatisfactory forecasts for real-world problems, as they might impose severe restrictions on the form of the eventual forecast function. In fact, no single statistical forecasting mechanism can be guaranteed to reliably model the true mechanism of the way data are generated (Clemen, 1989).

On account of this, a bulk of research has been dedicated to developing related techniques or integrating new concepts for tackling the limitations of statistical forecasting methods. Most related methods developed for time series forecasting comprising the use of judgmental forecasts, combining forecasts, bootstrap aggregation, , machine learning techniques, and analogies.

A common practice used to tackle the limitations of the statistical forecast takes into account expert opinions for adjusting the final forecast by integrating information extrinsic to the time series data; judgmental forecasts. For example, this approach can be used in exceptional circumstances where a known event cannot be modeled by a physical model using the historical observations. Methodologically, statistical forecasting methods offer little opportunities to incorporate such contextual information mathematically without a post-hoc procedure.

Combining forecasts derived from various statistical models is advantageous because no single statistical prediction method can demonstrate optimal performance. In addition to this, the use of combined forecasts helps to overcome the limitations of the statistical methods where the statistical forecast cannot model the true mechanism of the way the data are generated. Combining forecasts methods may not necessarily improve the final forecasting accuracy but has been reported to give rise to a significant decrease in the risk of employing a wrong statistical forecasting method.

Another promising method known as the bootstrap aggregation technique can be utilized. This method perturbs the time series data; a forecast is then made for each bootstrapped sample, combining these individual forecasts provides an aggregated forecast. Use of this method might assist in the reduction of instabilities stemming from the data, and this might give rise to additional accuracy in point forecasts.

Over recent years, machine learning techniques have become a competitive option for forecasting applications. Numerous applications can be found in time series forecasting literature that employ machine learning techniques to make forecasts. Among these techniques, support vector machines, neural network, random forests techniques have gained intense research interest. These methods are promising due to their capability to model non-linearity that might not be modelled by a conventional statistical model *e.g., ARIMA*, but the improvements achieved in accuracy might compromise the clarity of the modelling process or increase the corresponding computational time.

Another way of improving the statistical forecasts is the pooling of information available from analogies, and this has received surprisingly little attention in previous work. Different from the methods that make use of expert judgment, the pooling of information can be integrated into the forecasting process using existing statistical forecasting methods such as the Cross-Sectional State Kalman Filter algorithm. The use of analogies is particularly useful when investigating problems where there is little prior knowledge available with regards to the target time series. This may be the case shortly after the launch of new products or where records are missing (Goodwin, Dyussekeneva, and Meeran, 2013). Under circumstances where little historical information is available, many statistical models become inapplicable. This might also affect the applicability of other approaches, such as those that combine forecasts from a set of different statistical modeling techniques. In such a setting, pooling information from a set of related time series data can provide a powerful mechanism to reduce the influence of outliers in individual time series or to complement the information for individual short time series.

As stressed by Stimson (1985), the homogeneity of analogies is important for pooling methods. Further evidence has been provided by Duncan, Gorr, and Szczypula (2001), that the use of a set of heterogeneous time series results in a poorer forecasting precision compared to grouped analogous time series. Subjective ways of deriving analogies based on expert opinions can be non-efficacious as the forecaster is required to recall similar cases (*e.g.,* the most similar promotion campaigns) from memory and judge their similarity to the target case. This can be prone to error. Moreover, limitations in human information processing capacity may mean that the forecaster relies on a single recalled case, limiting the robustness of their judgment. Considering the limitations of subjective methods, there is a strong motivation to develop objective methods that can draw information from analogies without relying on human inputs.

Nevertheless, there remains a lack of empirical evidence that supports the principled selection and judgment of similarity between analogies (Lee et al., 2007). This is also true regarding the assessment of the impact on the forecasting accuracy of methods which make use of such analogies.

The identification of suitable analogous time series / analogies is the first step in the application of forecasting methods that make use of analogies (Stimson, 1985). There is a shortage of previous work that has observed this and attempted to develop objective approaches for the selection of analogies, and to contrast the performance of different techniques. This is despite the fact that the accurate discernment and identification of similarities between time series (Lee et al., 2007) is critical to the successful use of forecasting approaches that employ analogies (Armstrong, 2001).

1.2 Research aims

- Given this gap in the literature, the primary aim of this thesis is to propose a solution for supporting the principled selection of analogies to achieve better forecasting results. We provide empirical evidence to verify the significance of analogies and to judge the impact of the segmentation performance on the forecasting stage in a setting where analogies are employed.
- We further aim to design objective methods that can be easily adapted and generalized to real case scenarios where masses of data must be analyzed.

1.3 Structure of the thesis

The remainder of this thesis is structured as follows.

To begin, **Chapter 2** presents a comprehensive review of related concepts and surveys the literature concerning the common practices for improving the forecasting performance of conventional forecasting methods. We further highlight the significant role that analogies can play in time series forecasting. This comprises the use of analogies in both subjective and objective forecasting methods. On account of the crucial role that analogies occupy in the boosting of forecasting accuracy, we introduce the multicriteria segmentation approach in the next section. This is to support the principled selection of analogies. We provide a review regarding the basics of clustering such as the choice of distance metrics and model selection problems in this context. The following part of the literature review covers the topic of bootstrap aggregating, which involves the resampling of data with replacement that helps to achieve additional gains in time series forecasting. Next, we discuss some of the limitations of previous work, and outline opportunities to address these. Specifically, this section begins with a discussion related to the problem of analogy identification, where multiple information sources are recommended for achieving better clustering results and therefore the improved forecasting accuracy of statistical models where analogies are fitted. Considering the additional challenges arising from the multicriteria segmentation stage, a discussion regarding model selection during clustering is provided. Although the use of analogies creates more opportunities for the improvement of the forecasting performance, it also raises additional issues of instabilities during the segmentation stage. Thus, bagging methods are proposed to tackle the instability issue with the aim of obtaining a further increase in the forecasting performance. Finally, the concept of multicriteria approaches is introduced in the context of time series clustering where the similarity between time series is hard to judge, as mixed patterns are commonly present in time series data. The last part of **Chapter 2** offers a justification of chosen research methodology for the present thesis.

Segmentation problems are inherently multicriteria problems (Liu et al., 2010). Fittingly, **Chapter 3** investigates the issues about the identification of analogies using multiple information sources. Time series information can be regarded as response variables that describe the performance of analogies. Causal factors underlying the time series, on the other hand, are explanatory variables that help to interpret the time-based patterns observed. The independent use of either information source could be sub-optimal regarding clustering performance. Therefore, a concurrent consideration of the two information sources is meaningful for the segmentation approach to yield more meaningful groupings. We propose the use of a post-hoc method (*i.e.,* multicriteria clustering approach) in the segmentation stage to combine multiple information sources using a weighting scheme. Using this strategy, we demonstrate the positive relation between the segmentation and the forecasting stages which identifies and utilizes analogies, respectively.

As discussed, the use of multicriteria clustering approaches can open up opportunities for the generation of more robust clustering solutions that can mean better forecasting results. However, the use of these techniques can present additional challenges to the model selection in the context of multicriteria clustering —weight selection that previously was not involved in the single-criterion clustering procedures. To further address the problem of model selection, we propose various objective methods in **Chapter 4**. Some of these have been adapted from popular internal validation techniques that have been described in the clustering literature, while others have been taken from the optimization literature. For example, the Silhouette Width metric was taken from the clustering literature, while the angle-based measure was taken from the multi-objective optimization literature. Following the applications of these techniques to a given circumstance, our results suggest that clustering solutions are best assessed in the problem-specific context, *i.e.,* forecasting. Hence, the model selection step should also take this factor into account. Finally, we go on to develop new methods that support the objective weight selection with consideration of forecasting performance, *i.e.,* a single best partitioning out of candidate clustering solutions for multicriteria clustering problems.

In **Chapter 5**, we consider the issue of instability that stems from the clustering procedure. This encompasses both the single-criterion as well as multicriteria clustering approaches. We propose two bagging procedures for improving the performance of the forecasting process that exploits information from analogies. These procedures function to perturb data in the clustering and forecasting stages sequentially to generate bootstrapped samples which help to derive a better aggregated forecast. We show that the integration of independent and identically distributed (IID) bootstrap in the clustering procedure can result in significant gains in the forecasting accuracy. Specifically, instead of directly bootstrapping the time series data and we focus on bootstrapping labels associated with these times series by converting this problem to a typical IID bootstrap problem.

In **Chapter 6**, we concentrate on exploring the potential of multiple criteria to time series clustering in a forecasting context. To define a cluster, the notion of similarity between time series is complex and dependent on the use of criteria. Different distance metrics / standardization techniques have been developed in previous works, but a single measure typically emphasizes on tackling a particular aspect of the problem. For example, one distance metric developing upon the linear correlation between pairs of time series may neglect the non-linear patterns present in the data. Additionally, some distance metrics may be sensitive to scale differences. Hence, different standardization techniques may yield quite different clustering results. In applications, there is no universally accepted notion of similarity between pairs of time series. The notion varies with the criteria considered as they emphasize on different aspects of the clustering. Specifically, we experiment with the use of multiple distance metrics / standardization techniques during the clustering procedure. We show that multicriteria approaches may be helpful in delivering better clustering and forecasting results when low correlated measures are used.

Finally **Chapter 7** summarizes our motivation of employing advanced segmentation approaches for the forecasting methods in the presence of analogies as presented in this thesis. We additionally discuss further ideas for work that could be conducted to expand the science described herein.

1.4 Contributions of the thesis

The main contributions of the thesis are summarized as follows:

- This thesis aims to develop objective computational methods that are suitable for the identification of analogies in the context of forecasting. We have contributed to the development of automated processes. This includes the development of different forecasting procedures that utilize multicriteria segmentation approaches, model selection as well as bagging techniques for improved forecasting.

- We propose a two-stage forecasting process that performs a forecasting task by pooling information from analogies. Specifically, the segmentation stage concerns the use of multicriteria clustering techniques that identify analogies using multiple information sources. The information drawn from analogies is then fed into the subsequent forecasting stage in the overall prediction process to improve the final forecasting accuracy. The details of the prediction processes developed are given in **Chapter 3**.

- Peer-reviewed studies are available in the literature that supports the positive impact that clustering quality of analogies can have on the forecasting stage. However, to the best of our knowledge, this has never been systematically studied. Using extensive experiments, we provide new insight into the relationship between the accuracy of segmentation stage and the performance of a forecasting algorithm that makes use of analogies (see **Chapter 3**).

- We provide empirical evidence to support the argument that the clustering quality is best evaluated in the context of an application. Therefore, our proposed model selection techniques take into account the forecasting performance at the weight selection step for the selection of a single best partitioning out of a set of candidate solutions (see **Chapter 4**).

- For forecasting methods that pool information from analogies, the use of segmentation approaches, either single- or multicriteria, inevitably face the challenge of instabilities that stem from the model selection step. To address the challenge of instabilities, we, therefore, integrate the IID bootstrap concept into the forecasting of analogous time series. By perturbing the data in the clustering stage, the use of bootstrap aggregating techniques in this context provides better forecasting results with respect to accuracy and robustness (see **Chapter 5**).

- We additionally extend the concept of multicriteria segmentation using multiple distance metrics or normalization techniques in the context of time series clustering. We aim to demonstrate the effectiveness of multicriteria approaches with regards to the added strength of defining the similarity between pairs of time series. This is especially in circumstances where there are no definitive recommendations for the definition of similarity or where almost all of the definitions are relevant. The results show that multicriteria approaches to time series clustering can provide better clustering results when combining low correlated dissimilarity matrices and then translate to improved forecasting performance (see **Chapter 6**).

1.5 Publications resulting from the thesis

The work described in this thesis has and is expected to yield several publications. All journal and conference contributions that have been submitted / published to date are listed below:

Refereed conference proceeding

Lu, E., Handl, J. (2015, June). multicriteria Segmentation of Demand Markets to Increase Forecasting Accuracy of Analogous Time Series: A First Investigation. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 379-388). Springer International Publishing.

Refereed journal papers

Lu, E., Handl, J., D.-L. Xu., (2017). Determining analogies based on the integration of multiple information sources. *International Journal of Forecasting*, Accepted for publication, 2018.

Conference abstract

Handl, J., Lu E (2017, August). Cluster validation in multicriteria data clustering. Conference of the International Federation of Classification Societies.

References

[1] J. S. Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Vol. 30. Springer Science & Business Media, 2001.

[2] R. T. Clemen. “Combining forecasts: A review and annotated bibliography”. In: *International Journal of Forecasting* 5.4 (1989), pp. 559–583.

[3] J. G. De Gooijer and R. J. Hyndman. “25 years of time series forecasting”. In: *International journal of forecasting* 22.3 (2006), pp. 443–473.

[4] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[5] P. Goodwin, K. Dyussekeneva, and S. Meeran. “The use of analogies in forecasting the annual sales of new electronics products”. In: *IMA Journal of Management Mathematics* 24.4 (2013), pp. 407–422.

[6] W. Y. Lee et al. “Providing support for the use of analogies in demand forecasting tasks”. In: *International Journal of Forecasting* 23.3 (2007), pp. 377–390.

[7] Y. Liu et al. “Multicriterion market segmentation: a new model, implementation, and evaluation”. In: *Marketing Science* 29.5 (2010), pp. 880–894.

[8] J. A. Stimson. “Regression in space and time: A statistical essay”. In: *American Journal of Political Science* (1985), pp. 914–947.

Chapter 2

Literature review

2.1 Improving the performance of statistical time series forecasting models

Time series forecasting is a significant research topic in the domain of forecasting. Previous work has divided methods of forecasting into two main categories - subjective and objective methods. Subjective methods encompass those which rely on expert judgment, while objective methods focus on the development of statistical forecasting methods which describe historical observations using mathematical models. The models derived from objective methods can be further employed to extrapolate the time series to allow for future predictions. In contrast to the subjective methods, statistical forecasts (objective methods) show advantages in terms of their applicability and scalability. Therefore, these models have received widespread use in practical applications across areas of economic, energy, finance, marketing, public budgeting and tourism.

Statistical time series forecasting methods have been developed for the addressing of particular challenges associated with data predictions. For example, one of the most popular types of univariate forecasting methods is the exponential smoothing method.

Simple exponential smoothing has been proposed to deal with time series, where there are only a small number of observations as a result of structural change in the time series. Analytically, the exponential smoothing method (Brown, 2004) assigns more weights on recent observations, and this discounts the importance of the past observations. The procedure executed is sensible but lacks statistical foundations. Another family of forecasting methods falls into ARIMA models. Box Jenkins popularized ARIMA process (Box et al., 2015) by developing a procedure that includes the identification, selection and determination of the model forms using linear stochastic equations. These methods are popular due to its flexibility in terms of its statistical attributes. Specifically, exponential smoothing can be transformed as a special case of the ARIMA process. However, the main limitation of this procedure is the prior restriction of the linear form posed on the data generation process. The form specified, prior to the application of the model, is unable to capture non-linear patterns presented in the data. Consequently, this might lead to unsatisfactory forecasting performance in complex real-applications. To model non-linear patterns associated with the time series data, models such as neural-networks and random forests, that are well studied in machine learning area, have been introduced in the forecasting field. For instance, Faruk (2010) proposed a hybrid method, that is comprised of ARIMA and neural-network methods, to model both linear and non-linear patterns in water quality time series. This is because neither of them in isolation can adequately model the patterns of the time series. In the area of short-term load forecasting, random forests have been employed (Dudek, 2015; Cheng, Chan, and Qiu, 2012), and these methods have shown better accuracy than conventional time series forecasting methods such as simple exponential smoothing methods.

In general, conventional time series forecasting methods are useful and relatively easy to apply and interpret. Nevertheless, almost all statistical forecasting models assume a prior mathematical form of the data generation process. In fact, no single model can recover the true mechanism of the data generation process (Harvey, 1990), and all of them attempt to approximate the data generation process. On account of this, there is a bulk of literature research that investigates various means of combining additional knowledge to support the forecasting analysis in either subjective or objective way. We divide these approaches into four major categories: the integration of expert judgment, the use of combining forecasts, the application of bootstrap aggregating techniques or the inclusion of analogies' information. Additional class of methods that we are going to discuss in this section is the machine learning technique.

The most common way of integrating additional information is to adjust the statistical forecasts based on expert judgment in a post-hoc manner. One would expect that the adjustment of the model with the assistance of domain knowledge, may be beneficial to bring information to the forecast that a statistical model could not. Blattberg and Hoch (1990) claimed that both judgment based on expert opinion as well as objective statistical methods have valuable and complementary contributions to the forecasting process. Specifically, statistical methods are adept at filtering regular time series patterns from noisy data, while judgmental forecasts tend to efficiently detect false patterns in noise and to overreact to random movements in series (O'Connor, Remus, and Griggs, 1993). Furthermore, when it is known that special events will occur in the future, expert judgment can be used to anticipate their effects. In these instances, statistical estimation are often limited in their effectiveness due to the scarcity or diverse nature of the data (Sanders and Ritzman, 2001). Lawrence et al. (2006) commented that judgmental forecasts can help motivate the respective forecasters to apply their expertise in applications and therefore generate a sense of ownership. Nevertheless, the application of expert judgment, without contextual information, can worsen the accuracy of the final forecasting due to anchoring bias and adjustment heuristic. The effectiveness of adjustment may also depend on the initial accuracy of the base statistical model; this is dependent on the series characteristics. Larrick and Soll (2006) showed that under certain conditions it is better not to combine the forecasts of experts. Besides this, the involvement of judgmental forecasts require the inputs of human experts and consequently incur the associated risk of bias. This can lead to problems with regards to reproducibility as well as the speed of throughput.

Alternatively, the effect of combining statistical forecasts methods is well studied (Bates and Granger, 1969). This allows for the summation forecasts derived from forecasts. There is a great deal of evidence to suggest that combining two or more independent statistical forecasts can lead to significant improvements in the final forecasting accuracy (Clemen, 1989). The combining forecasts is a valid alternative to the adjustment of statistical forecasts by expert judgment. Despite this, there is only a few systematic studies that compare these two common approaches. Lim (1993) compared these two methods and found that expert judgment with adjustment tended to be less accurate than combining forecasts methods. This was specifically when contextual information is not included in the expert judgment forecasting procedure. Other studies have also suggested that human experts tend to avoid combining estimates across sources, due to human limitations especially where used forecast models make different assumptions. This can increase errors since the act of averaging studies has the effect of reducing errors (Soll, 1999). As a result combining forecasts method is often regarded as a more successful alternative to the implementation of an individual forecasting method. In some situations, the analyst may have more than one possible forecasting method. This is because, on their own, no single statistical forecasting method can precisely model the true mechanism through which observations are generated. The analyst may therefore often selects to combine forecasts in some way,*e.g.,* average, median to reduce the risk of implementing a unsuitable statistical model. For more in detailed descriptions of combining forecasting methods, comprehensive reviews are given by Clemen (1989) and Granger and Newbold (1974). There also is also various works that questions whether one should always combine forecasts. For example, Hibon and Evgeniou (2005) presented empirical findings which showed that the combining forecast approach has the advantage in that it is less risky in practice to combine forecasts than to select an individual forecasting method.

Bagging or bootstrap aggregation methods (Efron, 1992) act as the most competitive form of solving predictive tasks. Bagging is a simple form of ensemble methods that consists of a large set of models for estimating the distribution of an estimator or test statistic. The predictive performance is expected to be boosted via the aggregation of component models. As the key step of bagging, bootstrapped samples are obtained by re-sampling from the the original data with replacement. From the bootstrap sampling, a Monte Carlo approximation of the bootstrap estimate is obtained. Numerous methods have been developed in order to determine the most effective means of implementing the bootstrap procedure depending on whether the data are a random sample from a distribution or a time series. The problem becomes more complex when the data are time series because bootstrap sampling must be carried out in a way that suitably captures the dependent structure of the data generation process. This is not difficult if one has a finite-dimensional parametric model, which reduces the data generation process to independent random sampling. In this case and under suitable regularity conditions, the bootstrap has properties that are essentially the same as they are when the data is a random sample from a distribution (see Bose, 1988; Bose, 1990). Such approaches are inconsistent if the model used for resampling is misspecified. However, these model-based approaches are straightforward: the dependent structure is modeled explicitly and the resampled data is drawn from the fitted model. This has been pursued in numerous examples and cases, *e.g.,* Bose (1988) and Freedman (1984) for autoregressive models, Kreiss and Franke (1992) for ARMA models and Rajarshi (1990) for Markov models. However, the problem becomes more problematic when time series data are bootstrapped which no longer follows the IID assumption. In contrast to resampling a single observation at a time, Kunsch (1989) and Liu and Singh (1992) independently formulated a substantially new resampling scheme, termed the moving block bootstrap (MBB). This is applicable to dependent data without any parametric model assumptions. Generally, the MBB resamples blocks of (consecutive) observations at a time. The method developed is reported to be suitable for stationary time series, where the dependent data structure can be preserved. The bootstrap technique can be promising, but its highly time consuming in terms of computation. The key success of the bootstrap aggregating method requires the presence of instabilities in the statistical model in order to obtain additional accuracy gains (Efron, 1992).

Another method that has received insufficient attention is to draw information from analogies. The concept of analogies is commonly employed in judgmental forecasting, *i.e.,* forecasting by analogy. Hence, it can be regarded as a combination of judgmental and statistical forecasting techniques. The integration of additional information drawn from analogies aims to boost the derived forecasting accuracy through the reduction of bias of the forecasters (Hyndman and Athanasopoulos, 2014). There are a few models to our knowledge that are capable of objectively integrating information from analogies. For example, the Bass model that is often used to forecast sales of unreleased products, which have yet to be launched, through the use of information available from similar products (Goodwin, Dyussekeneva, and Meeran, 2013). Nikolopoulos et al. (2007) referred this approach as "nearest neighbour analysis" for the forecasting of TV audience ratings. In addition, Bayesian pooling approaches, *e.g.,* the Cross-Sectional Multi-State Kalman Filter (C-MSKF: Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001) offers a possibility to integrate information from analogies directly into the modeling process. In essence, these models are more complicated than univariate time series forecasting methods. Moreover, they might extend the univariate time series forecasting methods such as exponential smoothing or Multi-State Kalman filter to incorporate analogies' information. However experience derived from the the common application of analogies, indicates that in challenging forecasting settings the use of analogies might become particularly useful and overcome the limitations of the previous methods. Such challenging setting include for example, the absence of data for a target series prior to the launch of a product, or where time series are short and volatile.

The use of machine learning algorithms in predictive modelling has gained increasing attention in the context of time series forecasting. Numerous applications can be found in the forecasting literature related to the application of support vector machines, artificial neural networks techniques. Increasing attention has paid to random forests, which are an ensemble learning method previously proposed for both classification and regression problems (Breiman, 2001). A review and simple interpretation of random forests can be found in Biau and Scornet (2016), Friedman, Hastie, and Tibshirani (2001), and Verikas, Gelzinis, and Bacauskiene (2011). In principle, random forests are a combination of a set of binary decision trees (Breiman et al., 1984), each of which is constructed using a bootstrap sample coming from the learning sample and a subset of features randomly selected at each node. Furthermore, trees in the forest are grown to maximum size and the there is no pruning step employed. Random forests has gained popularity over the recent decades due to its advantages in ease of employment, requirement of a few parameters (such as the number of trees (Oshiro, Perez, and Baranauskas, 2012) and the number of input variables at each split node (Verikas, Gelzinis, and Bacauskiene, 2011)). However, elements that also might impact on the forecasting performance comprise the number of possible directions for splitting at each node of each tree (Kuhn and Johnson, 2013) and the number of examples in each cell (Tyralis and Papacharalampous, 2017). Additionally, this method shows flexibility in accommodating the prediction tasks with small sample size, high-dimensional feature spaces, and complex data structures (Scornet, Biau, and Vert, 2015; Biau and Scornet, 2016). The inclusion of unimportant predictor variables does no seriously impact the predictive performance of random forests, as implied in Kuhn and Johnson (2013). Nevertheless, using random forests for time series forecasting is not identical to the simple regression case. The role of the predictor variables is taken by previous variables, *i.e.,* the selected lagged variables, inevitably results in reducing the length of the training set. Using fewer predictor variables instead may reduce the information gained by the available knowledge of the temporal dependence. The application of random forests in time series forecasting field has been reported in short-term load forecasting (Dudek, 2015; Cheng, Chan, and Qiu, 2012), stock market prices (Khaidem, Saha, and Dey, 2016), stock index movement (Kumar and Thenmozhi, 2006), water level forecasting (Yang, Cheng, and Chan, 2017). To implement random forests in the time series forecasting applications (without resorting to exogenous variables), different bagging procedures have been explored. For example, Tyralis and Papacharalampous (2017) proposed to use past observations as lagged variables for estimating the next time points. Precisely, Let g be the function obtained from the training of random forests, which will be used for forecasting x_{n+1}, given $x_1,\ldots,x_n$. If we use k lagged variables then the forecasted x_{n+1} is given by the following equation for $t = n+1$:

$$x_t = g(x_{t-1},\ldots,x_{t-k}), t = k+1,\ldots,n+1 \tag{2.1}$$

The function g is not in closed form, but can be obtained by training the random forest algorithm using a training set of size $n-k$. In each sample of the fitting set the dependent variable is x_t, for $t = k+1,\ldots,n+1$, while the predictor variables are $x_{t-},\ldots,x_{t-k}$. When the number of predictor variables k increases, the size of the training set $n-k$ decreases. Dudek (2015) applied random forest to model patterns of the time series seasonal cycles which simplifies the forecasting problem especially when a time series exhibits nonstationality, heteroscedasitcity, trend and multiple seasonal cycles. To model the nonstationary time series that involve with trend and multiple seasonal cycles often require complex models with many parameters to tune. Random forests used as a forecasting tool are relatively straightforward and easy to apply. They combine regression trees with only few parameters to estimate.

2.2 Multicriteria segmentation approaches

Segmentation has been extensively studied in areas across image processing, economics, finance, operational research, pattern recognition, and public budgeting. The use of segmentation aims to identify meaningful groupings that are homogeneous with respect to specific criteria considered (Bab-Hadiashar and Suter, 2012). In addition, Wedel and Kamakura (2012) declared that segmentation is a grouping task, where a large variety of methods are available and have been used. Broadly, segmentation can be categorized into *a-prior* and *post-hoc* (Wind, 1978). A segmentation approach is classified as *a-prior* when the type and number of segments are decided before data collection, while it is called *post-hoc* approach when the type and number of segments are decided based on the results of data analysis. Statistical methods have been commonly applied to perform post-hoc segmentation. Wedel and Kamakura (2012) categorized such methods and techniques into four categories: cluster analysis, mixture, mixture regression and mixture scaling methods. Among these, clustering methods are the most popular tools used for post-hoc segmentation (Wedel and Kamakura, 2012). More specifically, single-criterion, multicriteria clustering techniques have been studied in the clustering literature. Interestingly, segmentation is often modeled as a single-criterion clustering problem in the traditional marketing literature as well as in practice. Ideally however, multicriteria clustering problems should be modeled homogeneously with respect to explanatory as well as response variables (Liu et al., 2010; Myers, 1996; Smith, 1956). Systematic studies should therefore be conducted to investigate whether homogeneously modelling would contribute to improved clustering results and therefore forecasting accuracy.

2.2.1 Methodological framework for multicriteria segmentation methods

To accommodate for multiple criteria during the segmentation process, a range of methodological frameworks have been proposed in the clustering literature. These include the multistage segmentation, the transformation approach and direct multicriteria clustering approaches.

Krieger and Green (1996) proposed a multistage segmentation approach that allowed the consideration of one criterion at a time. Within this approach various criteria are combined in a sequential manner. The approach employed the K-means clustering method to partition observations into segments that are optimized for identifiability, allowing for more accurate inferences to be derived. In the second stage, a heuristic approach is employed to enhance the responsiveness of segments. Here a detection threshold is employed to increase the models sensitivity to increases in within-segment heterogeneity. However, Brusco, Cradit, and Stahl (2002) commented that the multistage segmentation approach developed by Krieger and Green (1996) is inherently a sub-optimal strategy as information is not optimally shared between stages.

On account of the limitation of the multistage segmentation approach, direct multicriteria segmentation has been identified as a competing alternative. These types of segmentation approaches group observations into sub-groups using multiple criteria concurrently during the segmentation stage. An exact approach to bicriterion data clustering was first proposed by Delattre and Hansen (1980). The approach was specific to a particular pair of clustering criteria. Ferligoj and Batagelj (1992) described an approach to account for clustering criteria defined in view of different information sources. In addition, multi-objective evolutionary algorithms have been developed. This allows for more flexible identification of full sets of Pareto optimal solutions for different choices of objectives (Handl and Knowles, 2007). Direct multicriteria clustering techniques show strengths in discovering more robust data structure, and have the potential to ultimately recover the full set of Pareto-optimal clustering solutions. However, the identification step may be time-consuming and further raises additional challenges related to the model selection. These added challenges arise because for the same number of clusters, there might exist sets of Pareto-optimal clustering solutions.

The transformation technique may be regarded as the most straightforward method of multicriteria segmentation. In the transformation technique, multiple criteria can be combined using a weighting scheme (Brusco, Cradit, and Stahl, 2002; Brusco, Cradit, and Tashchian, 2003). In essence, multiple criteria will be ultimately transformed into a single criterion. The combination can be processed at distance function or objective level. This technique is easy to apply and straightforward to understand but limits itself in terms of discovering all Pareto-optimal clustering solutions. This is because the use of weight intervals may have an important impact on the final set of solutions identified. Another evident limitation of the transformation approach, is that difficulties are often met when defining objective utility or select weights, which is necessary for their combination.

In summary, the choice of methodological framework implemented for the multi-criteria segmentation approach is dependent on the final purpose of the specific applications. For example, a transformation method may be preferred to facilitate the interpretation and facile implementation where full sets of clustering solutions are not required.

2.2.2 Basics to clustering techniques

The needs of segmentation raise a further requirement on the investigation of suitable techniques around the implementation of clustering techniques. Hence, we follow on from this to review popular clustering algorithms and discuss the basics with respect to this topic. Han, Pei, and Kamber (2011) defined that clustering is to partition a collection of data objects into groups, where objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be noted as a clustering.

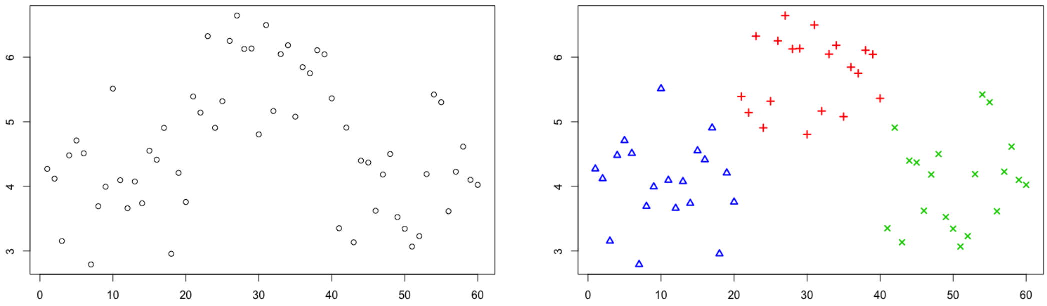
To construct a clustering process some decisions are required to be made regarding the choice relating to the selection of a suitable clustering algorithm, an appropriate distance function used for measuring the dissimilarity between objects and the determination of an optimal number of clusters. In the following sections, we aim to describe some of the basics in this regard.

2.2.2.1 Clustering algorithms

Clustering algorithms can easily be employed to operate on static data through the use of appropriate distance metrics that are dependent on the type of data as well as the selection of the optimal number of clusters if required. Data are denoted as static if the feature values do not vary significantly with time. An operational definition of the clustering procedure can be described as: given a set of N unlabeled data objects, a clustering method creates $K = \{1, 2, \ldots, N\}$ clusters of the data objects, where each cluster represents a group of objects. Therefore, the similarities between objects pertaining to the same cluster are high, while the similarities between objects in different clusters are low (see Fig. 2.1).

According to Liao (2005), five classes of clustering algorithms can be distinguished. These include partitioning approaches, hierarchical approaches, density based approaches, grid-based approaches, and model-based approaches.

Partitioning methods such as K-means or K-medoids are popular in practice. The partition is crisp if each object falls into only one cluster, or fuzzy if one object is allowed to be grouped into more than one cluster to a different degree. For example, K-means



a) Before clustering b) After PAM clustering

Figure 2.1: Illustration of the clustering process. The Partitioning Around Medoids (PAM) clustering approach clusters the objects into three distinctive groupings.

algorithm generates clusters that are represented by the mean value of the objects in a cluster. In terms of the K-medoids algorithm, each cluster is represented by the most centrally located object. These heuristic algorithms work well for finding spherical-shaped clusters and small to medium data sets. To find clusters with non-spherical or other complex shapes, specially designed algorithms such as density-based methods. K-means are reported as one of the most efficient algorithms but are known to be sensitive to outliers or noise. One variation of K-means is the K-medoid which minimizes the absolute distance between the objects and the selected centroids. A popular medoids-based algorithm is the Partitioning Around Medoids clustering algorithm (PAM: Kaufman and Rousseeuw, 2009). This offers additional flexibility allowing it to be used to operate on the dissimilarity matrix obtained by pair-wise comparison between objects. Using the dissimilarity matrix, PAM clustering can easily work on mix type of features.

The hierarchical clustering methods can be described as a nested sequence of partitions. Both agglomerative and divisive hierarchical clustering methods are widely used in practice. They work by merging two groups based on the optimization of linkage criteria at each stage of the algorithm. A popular linkage criterion is the sum of within-group sum of squares. Agglomerative methods start by placing an object in its own cluster. Then clusters are merged into increasingly larger clusters, until all objects are assigned in a single cluster or termination conditions such as the desired number of clusters are satisfied. Divisive methods work in an opposite manner. Basic hierarchical clustering methods (that do not incorporate heuristic methods or improvements) have the limitation of revoking an action. Ultimately this means these clustering algorithms can not adjust the partitioning once a merge or split decision has been executed. Another evident drawback of hierarchical clustering methods concerns the time-complexity property of the techniques. These methods directly work on the dissimilarity metric which requires the computation of pairwise comparisons across all pairs of objects. At least, the complexity of these methods is $O(N^2)$. A common practice of alleviating the issue of time-complexity is to regard the computation of the dissimilarity matrix as an independent step. Furthermore, there is a trend to integrate this method with other clustering techniques to improve the final clustering accuracy. A notable property of this method is that the hierarchical clustering algorithm does not require a predefined number of clusters. This is because it can return solutions for all possible numbers of clusters simultaneously. Further, the technique determines a dendrogram that can be cut at different specific heights to obtain the desired number of clusters.

Density-based methods are popular for the identification of clusters in a large multidimensional space. The general idea behind density-based methods is to identify distinctive groups, where the object space is contiguous and is associated with high object density. These methods are capable of discovering arbitrary shapes and handling noise. The idea is intrinsically different from the idea of generating a cluster explicitly. The method is advantageous as it does not require pre-defined parameters for a clustering algorithm. The selection of the number of clusters might be difficult but has a significant impact on clustering results. There may not exist any global parameters that could describe the internal data structure accurately. The determination of parameters may be dependent on the context of the specific application. OPTICS (Ankerst

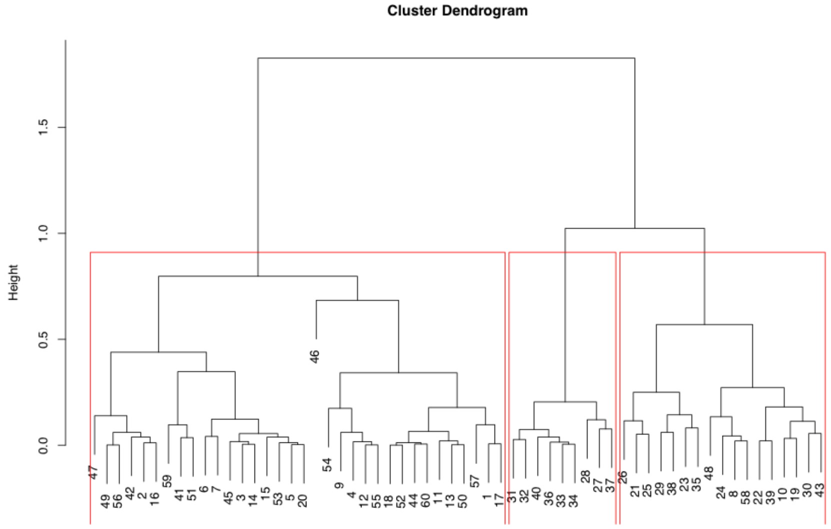


Figure 2.2: Illustration of dendrogram generated by the agglomerative hierarchical clustering algorithm using the average linkage criterion. Three clusters are highlighted in the red rectangles.

et al., 1999) is a well-established density-based clustering method. It computes an augmented cluster ordering for automatic and interactive cluster analysis. The ordering contains equivalent information comprised in density-based clustering. The information is obtained from a variety of parameter settings and in turn overcomes the difficulty of selecting parameter values.

In addition, grid-based methods have been proposed to efficiently handle large-data mining tasks. Grid-based methods quantitize the object space into a finite number of cells that form a grid structure without posing an assumption of the underlying distribution of the object. A well known method that falls into the grid-based approach is the statistical information grid-based (STING) algorithm (Wang, Yang, and Muntz, 1997). This method was proposed to be capable of dealing with large amounts of spatial data. Normally, the resulting computational complexity is at least linearly proportional to the number of objects to answer each spatial query. The general idea of this algorithm is to capture statistical information associated with spatial cells so that the whole classes of queries and clustering problems can be answered without recourse to the individual objects. In contrast, traditional methods might have to recourse all individual objects at least once.

Fraley and Raftery (2002) provided a detailed survey with respect to the model-based clustering methods. This series of clustering algorithms poses an assumption on the model for each of the clusters and attempts to best fit the data according to the hypothesized model. In general, there are two types of clustering methods fall into this category comprising statistical methods and neural network approaches, respectively. A well-known example of statistical approach is the AutoClass (Cheeseman et al., 1993), which estimates the number of clusters using Bayesian statistical analysis. In terms of the neural network approaches, popular clustering methods include the self-organizing maps (Kohonen, 1990). The self-organizing map technique can effectively create spatially organized internal representations of various features of input signals and their abstractions.

2.2.2.2 Choices of distance metrics to time series data

The data clustering particularly of static data has a much longer history compared to the clustering of non-static data, where observations are interdependent. Almost all the clustering algorithms when first proposed were motivated by applying to static data. Time series data are of recent interest because of its wide applicability in various areas ranging from biology, business, economic, finance, and health care. Given a set of unlabeled time series, it is often desirable to group time series into distinctive partitions. In terms of traditional data clustering, the determination of distance metrics has been extensively researched and is mainly dependent on the purpose of the application and type of the data. Relatively, the selection of distance metrics becomes more complex when it comes to non-static data, where observations are presented in an interdependent form.

In fact, most of clustering algorithms proposed have attempted to modify the existing clustering algorithms designed for static data. A time series that describes an object can be regarded as a feature/variable which comprises values that change with time. As a result, time series data can be handled or converted into the static form, so that existing clustering algorithms can be directly applied to. Three types of clustering algorithms are commonly applied to cluster time series data, including partitioning methods, hierarchical methods and model-based methods, either in a direct or modified manner. Precisely, the proposed clustering algorithms for handling time series data are working by the dissimilarity matrix level where distances between pairs of time series are computed. Hence, by using an existing or modified clustering algorithm, plenty of work has been dedicated to investigate different ways of measuring (dis)similarities between time series data where the features values vary with time. Thus, we further review important distance metrics suitable for dealing with time series data which supports clustering of time series. Liao (2005) categorized time series clustering approaches into three major streams depending upon whether they work directly on the raw data, indirectly with features that are extracted from the raw data, or with physical models that can used to describe the underlying time series patterns.

Raw-data-based approaches calculate the distance between pairs of time series by taking into account the time-based patterns of the series, either in the time or frequency domain. The two time series being compared are normally sampled at the same interval, but their length might or might not be the same. Clustering directly on raw data is straightforward. For example, correlation-based approaches are widely used to measure dissimilarities between time series. Yet, Granger and Newbold (1974) commented 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. Additionally, these types of methods are commonly limited to the computation complexity, particularly for high-dimensional data, where the observations are collected with fast sampling rates. Furthermore, high noise in the time series can present another issue to the clustering problem.

On account of this, an alternative method is to use the feature-based clustering techniques. We would generally expect this method to return more robust clustering results than raw-data-based methods by extracting features directly from the raw data. Although feature extraction methods are typically generic, the extracted features are essentially application dependent. Hence, the concept that one size fits all does not apply here. Since one set of features may performs well on one data set but may perform poorly on another. For example, Wang, Smith, and Hyndman (2006) clustered time series based on their structural characteristics on high dimensional data. This particular method does not cluster time series based on global features extracted from the time series. It can be further fit into arbitrary clustering algorithms. Essentially, the time series is represented using extracted features that can be regarded as a typical clustering problem based on static data.

The last popular group of methods concern the model-based approaches. The general idea behind these methods is the consideration of each time series through statistical models or by a mixture of underlying probability distributions. Time series are considered similar if the models characterizing individual time series are alike. In a similar manner, time series are regarded as similar, if the remaining residuals after fitting the model are similar. In general, model-based methods can be considered as a form of approximation of the data generation process that underlies the time series data. The approximation has a potential limitation in that they may be inadequate for accommodating other possible patterns underling the data. Kalpakis, Gada, and Puttagunta (2001) is an example of a model based approach which fits time series using Box Jenkins ARIMA models. The strengths of the method resides in its potential to differentiate overlaying time series by fitting the time series with ARIMA models, however it becomes insufficient for the modeling of short time series, where the estimation of model parameters becomes difficult when only a small number of observations present.

2.2.2.3 Cluster validation techniques

Cluster validation indexes can be sketched for two general purposes including (i) the evaluation of clustering solutions, (ii) the selection of the appropriate number of clusters. More specific, internal validation techniques are commonly employed to measure the quality of clustering solutions based on the intrinsic data structure. These techniques can also be applied to support the model selection step in the clustering procedure without involving external information. It is common to use external validation techniques for the evaluation of clustering solutions by comparing to the ground truth comprised in a synthetic data.

To measure the goodness of clustering solutions synthetic data sets, benchmark data sets, or real data sets have been widely applied to compare the performance of clustering solutions (Von Luxburg, 2010). based on the comparisons of the results. In circumstances where external information is not available, clustering quality scores can be computed to reflect the quality of a solution. These quality scores are designed based on two general types: external and internal validation measures. Cluster validation techniques for crisp clustering are reviewed below.

External measures that assess the clustering results based on the knowledge of the correct class labels are given in benchmark data or simulated data. These measures are useful in enabling an objective comparison of the clustering algorithms to cluster data, where true cluster structure is given. External measures can be further subdivided into unary and binary measures.

Unary measures are methods that take a single partitioning as the input, and then compare the assignment of the objects to a given set of class labels that are often regarded as the “ground truths” or “gold standards”. Conventionally, the gold standard will be complete and unique, in the sense that exactly one class label is provided for every data item, and that the label is clearly defined. By comparing the clustering solution to the ground truth, one can assess the degree of consensus between the two clustering solutions. A clustering solution can then be assessed in terms of purity and completeness of a partitioning. Here, purity denotes the fraction of objects in this predominant class that is grouped in the specific cluster. Clearly, both of these aspects provide a limited amount of information, and trivial solutions for both of them might be generated. A partitioning consists of singleton clusters can be highly scored under purity measure, while a one-cluster solution can be maximally scored under the criterion of completeness. In order to obtain an objective assessment of partition, accordance with the gold standard, it is therefore necessary to take into account the purity and completeness criteria. Measures such as the F-measure have been developed to take both factors into account. These metrics consider both purity and completeness of a partitioning and usually are more preferred in relation to simple techniques.

Alternatively, binary measures refer to measures that judge the agreement between two partitionings. These methods operate on contingency table of the pairwise assignment of object items. The ultimate aim of these methods is to assess the consensus between a partitioning and the ground truth. Most of these techniques are symmetric and therefore, well suited for the use of a binary measure for the evaluation the similarity between different clustering results. Perhaps, the most well-known methods is the Rand Index (Rand, 1971), which judges the similarity between two partitions as a function of positive and negative agreements in pair-wise partitioning assignments. The Adjusted Rand Index (Hubert and Arabie, 1985) is a correct-for-chance version of the Rand Index, which has the limitation where the expected value of the Rand index between two random partitions is not a constant. Another related index is the Jaccard coefficient (Niwattanakul et al., 2013), which employs a stricter definition of correspondence, where only positive agreements are rewarded.

In scenarios where class labels are not available or not satisfactorily defined, an evaluation may be appropriately undertaken using internal validation techniques. These techniques do not make use of extrinsic knowledge *i.e.,* class labels, but measure the clustering accuracy based on internal data structure alone. It means that these internal measures aim to measure how well a given clustering solution performs relative to the natural cluster structure discovered. Principally, most internal validation techniques can be used to assist the model selection step in addressing clustering problems. Generally, internal measures are proposed to assess the quality of clustering solutions from different perspectives which comprise the compactness, separation, connectivity and instability of the clustering quality.

Compactness of clustering solutions focuses on assessing the within-cluster homogeneity, where the intra-cluster variance is often measured to indicate the clustering quality. Alternative measures have been developed for the measurement of intra-cluster homogeneity. Examples of these methods include the sum-of-errors measure, the average or maximum pair-wise intra-cluster distances, as well as the use of graph-based approaches (Bezdek and Pal, 1998). Separation is used to quantify the distance between individual clusters. For instance, an overall rating of a partitioning can be defined as the average weighed inter-cluster distance, where distance between individual clusters can be computed as the distance between the cluster centroids, or as the minimum distance between data items belonging to different clusters. Alternatively, cluster separation in a partitioning may be assessed by the minimum separation observed between individual clusters in the partitioning. The connectivity of the clustering quality describes how well a given clustering solution agrees with the concept that to what degree a clustering solution observes local densities and groups data items together with their nearest neighbor in the data space. Instability-based techniques that assess the stability of a partitioning that categorized as a special class of internal validation measures. Instability measures behave different from traditional internal validation techniques as they require access to the clustering algorithm in order to assess the clustering quality. Measures of this type repeatedly re-sample from or perturb the original data set, and then re-cluster the resulting data. The consistency of the corresponding results provides an estimate of the significance of the clusters obtained from the original data set.

The literature provides a range of enhanced approaches that combine measures of the above measures. Combinations of compactness and separation are particularly common in practice, since these two measures function in opposing manners. This is because while intra-cluster homogeneity improves, the heterogeneity between clusters tends to deteriorate with an increasing numbers of clusters. Several techniques have therefore been proposed to assess both intra-cluster homogeneity and inter-cluster separation. These techniques compute the final quality score using a linear or non-linear combination of the two measures. An example of a linear combination is the SD-validity Index (Halkidi, Vazirgiannis, and Batistakis, 2000). Some well-known examples of non-linear combinations are the Dun Index (Dunn, 1974), Dunn-like Index (*e.g.,* Pal and Biswas, 1997), and the Silhouette Width (Rousseeuw, 1987).

2.2.3 Model selection in data clustering

Model selection in data clustering is still an open question, although a number of works have been produced regarding this topic. In principle, Liu et al. (2010) summarized that model selection can encompass the selection and standardization of clustering features, the determination of number of clusters, and the choice of clustering algorithms as well as the parameters concerning the non-determination of cluster algorithm. Generally, the major challenge in model selection problems can be reduced to the determination of the suitable number of clusters (Tibshirani, Walther, and Hastie, 2001), which is expected to have a major impact on the performance of clustering algorithms, where the number of clusters is required to be known. Note that in grid-based and density-based clustering approaches, there is no requirement for the determination of the number of clusters, but requires the pre-specification of density related parameters. This is because these techniques grow clusters based on density rather than the grouping of objects. In essence, the difficulty of model selection is associated with the challenge of defining the clustering. There does not exist any universally accepted definitions of clustering and one may be preferred than another in a certain problem context (Parsons, Haque, and Liu, 2004).

2.2.3.1 The existing validation techniques

A bulk of work has dedicated to developing automatic models for estimating the number of clusters in single-criterion clustering problems. Among these, most existing approaches devote to minimizing distance-based dissimilarity measures within clusters through the use of internal cluster validation techniques as discussed in Section 2.2.2.3. Other works such as Wang (2010) proposed an innovate technique to estimate the number of clusters. They attempted to select the number of clusters by minimizing the algorithm's instability via cross-validation techniques. Von Luxburg (2010) provided a comprehensive review on existing methods that utilize the cluster instability for the determination of number of clusters. In addition to these techniques, statistical approaches such as the Gap statistics has been proposed to deal with the challenge of model selection. However, to the best of our knowledge, there are no definitive recommendations regarding which model selection techniques would work best in practice. Consequently, traditional clustering methods (*e.g., K*-means) tend to utilize a subjective assessment to assist the selection of the appropriate number of clusters. The reader is refereed to a comprehensive review (Milligan and Cooper, 1985) concerning this topic for more information.

In general within the literature, the Elbow method, the Silhouette Width measure and the Gap statistics have received intensively applicability with success. Generically, internal cluster validation techniques can often be used to address the problem of K-determination. These techniques can be used to score the quality of clustering solutions, and the obtained quality scores are then used to determine the number of clusters. Normally, a higher quality score indicates a better clustering solution. The Elbow method and the Silhouette Width measure are developed on the internal cluster validation technique. These approaches tend to characterize the global characteristics of cluttering solutions.

Specifically, Elbow methods (Sugar, 1999; Sugar, Lenert, and Olshen, 1999) estimate the number of clusters by critically examining a graph of the percentage variance explained as a function of the number of clusters. The critical point where no further gains are achieved corresponds to the required number of clusters or partitionings for the optimal performance of the clustering procedure. This method is straightforward to apply, but practically it can be hard to determine the critical point where the variance plateaus.

The Silhouette Width measure (Rousseeuw, 1987) takes into account the separation and cohesion of the clustering solutions based on internal data structure alone. These methods are widely applied in single-criterion clustering problems. Specifically, they assume that data set contains N items and they can be partitioned into $k \in [2, N]$ clusters by employing a clustering algorithm. The Silhouette values can be computed for each cluster. The clustering solution returns the largest mean Silhouette value and the associated optimal cluster number K. The Silhouette Width technique takes value in the range [-1,1]. A higher value is indicative of a better clustering solution. Similar to the Elbow method, the Silhouette Width technique assesses the global characteristics of the entire partitioning.

Another statistical approach widely employed for the determination of K is known as the Gap statistic (Tibshirani, Walther, and Hastie, 2001). This approach normalizes the graph of log(W_k), where W_k is the pooled within-cluster sum of squares around the cluster means. Specifically, it compares the log(W_k) with the expected value derived from an appropriate null reference distribution of the data. (Gordon (1996) discussed the significance of the choice of an appropriate null model). The estimate of the optimal number of clusters is then the value of K, for which log(W_k) falls the farthest below the reference curve.

Handl and Knowles (2007) integrated the Gap statistics into the MOCK model for the identification of the single most promising solution from a particular set of trade-off solutions. This technique is based on an analysis of the location of solutions in objective space relative to a background of unstructured data. When applied to the multi-objective clustering technique MOCK, this approach has been shown to outperform more traditional techniques of model selection such as the Silhouette Width measure (Rousseeuw, 1987).

2.2.3.2 Adjustment of techniques from multi-objective optimization domain

Not limited to the clustering literature, there have been some methodologies and concepts developed in the domain of multi-objective optimization to address the selection of a single best solution.

Similar to multicriteria clustering problems, multi-objective optimization problems aim to optimize multiple criteria at the same time. These criteria are often conflicting, therefore it may not possible to find a single solution which is optimal with respect to all criteria. Instead, there exists a number of “Pareto-optimal” clustering solutions. These are characterized by the fact that an improvement in any one criteria can only be obtained at the expense of degradation of another. Without access to external information, none of the Pareto-optimal solutions can be identified as inferior compared to other solutions, this is because a single solution will not be universally better, *i.e.,* it may be only good with respect to specific criteria but not all. The idea of ”knees” have been investigated to reflect the user preferences which are of relevance to the decision maker. The most interesting solutions, or knees, of the Pareto-optimal front are those where a small improvement in one objective would lead to a large deterioration in another.

Knee points are well-recognized by multi-objective optimization researchers (Bechikh, Ben Said, and Ghédira, 2010; Branke et al., 2004; Deb and Sundar, 2006; Mattson, Mullur, and Messac, 2004; Schütze, Laumanns, and Coello, 2008; Rachmawati and Srinivasan, 2009). Owing to their advantages compared to other Pareto-optimal solutions, some evolutionary optimization methodologies have been designed to find knee point(s) (Bechikh, Ben Said, and Ghédira, 2010; Branke et al., 2004; Deb and Sundar, 2006; Rachmawati and Srinivasan, 2006a; Rachmawati and Srinivasan, 2006b; Rachmawati and Srinivasan, 2009; Schütze, Laumanns, and Coello, 2008). In some problems, instead of a single knee point, there may exist a sets of closely-packed trade-off solutions that altogether qualify as a knee region. Branke et al. (2004) proposed an enhanced angle-based approach that can identify a single best knee point through by looking at the different combinations of four nearest-point angels and the biggest angle is chosen as the final solution. Assuming a two-objective optimization problem, they claimed that a further movement in one direction might result in a significant degradation in another (see Fig. 2.3).

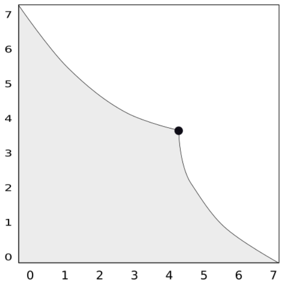


Figure 2.3: Illustration of a single knee point on the Pareto frontier based on two objectives

2.2.3.3 Evaluating the clustering performance in the domain of applications

Quality scores based on internal cluster validation techniques are useful at the level of algorithms. Here, they can be used as an objective function in the circumstance of optimization problems. A valid research question here concerns how different scores can be efficiently optimized. However, across different algorithms these scores provide only little information about the effectiveness of clustering solutions. This is because the preference of an accuracy measure over the other can vary dependent on the use of clustering algorithm.

Furthermore in situations where multiple information sources/feature spaces are used for clustering, quality scores will not valid without the consideration of Pareto-optimal clustering candidates. These reflect the different trade-offs between information sources. Overall, a unique, global, objective score for all clustering problems does not exist.

In exploratory data analysis, clustering can be used to discover aspects of the data which are either completely new, suspected to exist or which are hoped not to exist. For example, one can use clustering to define certain sub-categories of diseases in bio-medical science, or as a means for quality control to detect undesirable groupings that could suggest experimental artifacts in the data. It is unreasonable to expect a general evaluation procedure for clustering algorithms which is application-independent. In reality, a cluster is a subjective entity, the significance and interpretation of which requires domain knowledge. Indeed, there does not exist a universally accepted definition of clustering (Parsons, Haque, and Liu, 2004) and the structure identified from the data may vary with the final application purpose. Guyon, Von Luxburg, and Williamson (2009) argued that the success of clustering is best evaluated in the domain of the overall success of a particular application . This implies that a good clustering solution for forecasting should take into account forecasting performance rather than relying purely on the clustering quality without considering the particular application. Although previous works have proposed this idea, there is a lack of objective findings that support this argument. This suggests that the selection of the appropriate number of clusters based on the internal data structure could be insufficient for suitably enhancing the accuracy and precision of the final clustering results. Further work is needed to examine this.

2.3 Bootstrapping techniques in time series forecasting

Statistical problems are often plagued by three main sources of instabilities (Chatfield, 2000; Bergmeir, Hyndman, and Benítez, 2016). One of the main sources of instability is related to the choice of statistical models that describe the underlying data structure. Another source of instabilities stems from the model parameters, assuming the model structure is known. Furthermore, instabilities can arise from the data even when the choice of the statistical model and parameters involved are determined. More specifically, the instabilities associated with the data might encompass the unexplained random variation presented in the data observations. The random variation comprises errors related to measurement and/or recording.

The bootstrap aggregation (bagging) technique (Breiman, 1996) has been commonly applied to address the instability issues presented by modelling procedures. Through the reduction of instabilities, bagging techniques aim to additionally increase the accuracy of a predictive model.

Generically, a Bagging procedure can be formulated as follows. Let $X = (x_1, ..., x_n)$ be a random sample from a distribution characterized by a parameter θ. The inference about θ will be based on a statistic T. The basic bootstrap approach consists of generating bootstrapped samples with a size of m. The size is typically equal to the original sample size although it is not compulsory. Based on $(x_1, ...x_n)$, a bootstrapped sample $(x_1^b, ..., x_n^b)$ is formed each time. The procedure for the generation of data is repeated B times, where B is the number of bootstrap samples constructed. For each bootstrap sample, we compute the statistic T. We call this T^b. The distribution of T^b is known as the bootstrap distribution of T. We use this bootstrap distribution to make inferences about T. We use this bootstrap distribution to make inferences about θ. Under some circumstances, the bootstrap distribution enables us to make more accurate inferences than the asymptotic distribution of T. The bootstrap method described here is the basic procedure that is valid for IID observations.

By applying the model to different bootstrapped samples, multiple versions of a statistical forecasting model can be constructed. Then, an aggregated forecast is made using a suitable combination schemes such as the mean, median, trimmed mean or weighted mean. In principle, Bagging can yield substantial gains in forecasting accuracy if the perturbation of the learning results in significant changes in the constructed forecasting model (Breiman, 1996). The bootstrap method has often been reported to provide better approximations of distributions statistics than those of a first-order asymptotic theory (Härdle, Horowitz, and Kreiss, 2003). Numerous research studies have been conducted concerning the development of suitable bootstrapping techniques to deal with either independent (IID bootstrap) or dependent structure (time series). Here, we revisit previous works that have been utilized to bootstrap time series.

2.3.1 IID bootstrap for time series forecasting models

Aggarwal, Garg, and Gupta (2014) introduced the concept of non-parametric resampling that allows forecasters to carry out statistical inferences in a wide range of problems without imposing much structural assumptions on the underlying data generation process. We denote this scheme as the IID bootstrap, which re-samples independent and random variables with replacement.

For more details, Davison and Hinkley (1997), Efron and Tibshirani (1994), Shao and Tu (2012) provided discussions regarding the general bootstrap procedures. Numerous works have been conducted concerning the development of bootstrapping methods. The application of these methods is predominantly dependent on whether the data is a random sample from a distribution or is time series data with a dependent data structure. In general, Bootstrap techniques show better approximations than the first-order asymptotic theory.

If the data are a random sample, then the bootstrap can be implemented by sampling the data randomly with replacement or by sampling a parametric model of the distribution of the data. The distribution of a statistic is estimated by its empirical distribution under sampling from the data or parametric model. For example, Kushary (2000) provided a detailed discussion of bootstrap methods and their properties, and its applicability to data that have been randomly sampled from a distribution.

IID bootstrap methods described above are applicable either under the hypothesis of independence or under specific model assumptions for dependent data. The main idea in the latter case is to use the approximate independence of the residuals, and then apply the resampling scheme of IID bootstrap method to get the right approximation.

This is not difficult if one has a finite-dimensional parametric model (*e.g.,* a finite-order ARMA model) that reduces the data generation process to independent random sampling. In this case and under suitable conditions, the bootstrap has the same properties to a random sample from a distribution (see Bose, 1988; Bose, 1990). Such approaches are inconsistent if the model used for resampling is misspecified. However, these model-based approaches are straightforward because the dependent structure is modeled explicitly and the slightly different version of the original sample is drawn from the fitted model. This has been pursued in numerous cases, *e.g.,* (Bose, 1988) and (Freedman, 1984) for autoregressive models, (Kreiss and Franke, 1992; Fenga, 2017) for Autoregressive Moving Average (ARMA) models and (Rajarshi, 1990) for Markov models.

2.3.2 Bootstrapping time series data

Bootstrap and other resampling methods for dependent data still constitute an active field of research in statistics, even though monographs already exists that are especially devoted to bootstrapping for dependent data (Lahiri, 2003; Politis, Romano, and Wolf, 1999). General overviews of the variations of bootstrap methods have been published in the last decade (Berkowitz and Kilian, 2000; Bühlmann, 2002; Härdle, Horowitz, and Kreiss, 2003). Related review papers can be found in the area of econometrics. In particular Politis and Romano (1996) work mainly focus on the use of bootstrap methods for econometric models. Ruiz and Pascual (2002) and Paparoditis and Politis (2009) also investigates the problem of bootstrapping financial time series models.

Analytically, the situation is more complicated when the data set is a time series, this is because bootstrap sampling must be carried out in a way that suitably captures the time-dependent structure of the data generation process. Therefore, existing methods have been proposed to directly bootstrap time series data rather than reducing the data to independent random variables.

In situations where model-based approaches are not applicable, the standard bootstrap resampling method designed for independent and identically distributed errors is not applicable due to the violation of IID assumption. Correlated errors are not exchangeable, and lagged dependent variables create extra problems in pseudo data generation. Unit root and cointegration regression models create further complications in bootstrap data generation. Finally, to achieve an improvement over the asymptotic results, one needs to work with asymptotically pivotal statistics. This is usually not completed.

Bootstrapping time series data can be viewed as the simulation of a statistic or statistical procedure from an estimated distribution $\hat{T}_n$ of observed data $(x_1, ..., x_n)$. In time series data, the construction of $\hat{T}_n$ is more complicated due to the dependence structure presented and is far less "natural" than the seminal work proposed by Aggarwal, Garg, and Gupta (2014). Generally, previous methodologies fall into two classes: the time domain and frequency domain bootstrapping.

In contrast to the resampling of a single observation at a time, Kunsch (1989) and Liu and Singh (1992) independently formulated a substantially new resampling scheme, known as the moving block bootstrap (MBB). MBB is applicable to dependent data without any parametric model assumptions.

For dependent data, the most common approach to bootstrap time series is to resample "blocks" of sequential observations instead of resampling independent data observations. This preserves the dependence structure of the underlying process within the resampled blocks and is able to reproduce the effect of dependence at short lags. A relatively different approach to the problem was suggested by Zeger and Hurvich (1987). In their seminal work, Zeger and Hurvich (1987) considered the discrete Fourier transform (DFT of the data and rather than resampling the data values directly, they applied the IID bootstrap method of Efron (1992) to the DFT values. The transformation based bootstrap (TBB) described here is a generalization of Zeger and Hurvich (1987) idea. As a result, the time-dependent structure of the original observations is preserved within each block. Furthermore, the common length of the blocks increases with the sample size. As a result, when the data is generated by a weakly dependent process, the MBB reproduces the underlying dependence structure of the process asymptotically. Essentially the same principle was put forward by Hall (1985) in the context of bootstrapping spatial data and by Carlstein (1986) for estimating the variance of a statistic based on time series data. Limited work have been conducted to investigate the bootstrapping of non-stationary time series, where serial dependence and non-stationarity are present.

Cordeiro and Neves (2009) attempted to bootstrap time series by employing the sieve bootstrap technique that performs bagging with exponential smoothing models (ETS). They use ETS to decompose the data, then fit an AR model to the residuals, and generate new residuals from this AR process. Finally, they fit the ETS model that was used for the decomposition to all of the bootstrapped series. Overall, the results are not promising, although they achieved some success for quarterly and monthly data. A more promising method was proposed by Bergmeir, Hyndman, and Benítez (2016). They applied the Box-Cox transformation to decompose M-3 competition data into seasonal, trend, remainder components. The remaining component was bootstrapped using MBB technique. Finally, the trend and seasonal components were added back to the series. They applied the exponential smoothing model to each bootstrapped sample using the bias-corrected AIC to select the model. The bagged ETS shows consistent superiority in performance over basic ETS models. Particularly, in M-3 monthly data, the bagged exponential smoothing method performs the best among the contestant methods.

2.4 Research challenges and tasks in this thesis

2.4.1 Improving the analogy identification using multiple information sources

Considering the application of analogies in either subjective or objective methods, there might be a strong need to develop suitable analytical approaches for the identification of analogies. On account of this, some modeling approaches have been proposed to complete the task of analogy identification. It is evident that the identification of analogies typically involves the use of segmentation approaches to partition a collection of time series into a set of homogeneous groupings using clustering techniques (*e.g.,* Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001).

In the context of forecasting, some techniques have been explored for the segmentation of time series data into meaningful groups. Particularly, data-driven methods *i.e.,* clustering techniques have been explored to partition a collection of time series based on their similarities. These include the clustering of time series based on correlational co-movement, model-based approaches (Frühwirth-Schnatter and Kaufmann, 2008) and sets of causal variables associated with each time series (Duncan, Gorr, and Szczypula, 2001). The use of these techniques indicate that an independent consideration of information derived from either time series data or causal factors underlying the time series has been explored in the forecasting literature. However, it also implies that this might be inadequate for differentiating the analogous time series. This is particularly true when the information sources concern either time-based patterns or causal factors are noisy. It is evident that the characterization of analogies using either of the above approaches will often provide a partial or approximate picture at best. Ultimately this means that multiple information sources need to be considered during the segmentation stage to achieve a more meaningful partitioning of analogies.

Additionally, our understanding regarding the modelling of segmentation is in line with the work of Leitner and Leopold-Wildburger (2011) and Webby and O'Connor (1996). The authors claimed that time series data often comprises past realizations of the actual observations, as well as contextual information which includes factors that govern the behaviour of these time series. Both information data sources are crucial for a clear understanding of the causal relationship between factors and time-based patterns. In addition to this, various authors of clustering literature articles (such as Liu et al., 2010; Myers, 1996; Smith, 1956) argued that the segmentation problem is inherently a multicriteria problem. This is because clusters are typically preferred to be homogeneous with respect to a collection of explanatory as well as response variables. Consequently, the same idea can be applied to the forecasting analysis, where both past realizations of a given time series (response variables) and the associated causal factors (explanatory variables) should be considered for the identification of analogies.

At a theoretical level, multicriteria clustering techniques have been claimed to demonstrate a more robust recovery of the underlying data structure, as well as a more vigorous discovery of the data patterns that cannot be modelled by single-criterion clustering approaches (Handl and Knowles, 2007). This is because multicriteria clustering techniques are able to provide an objective assessment of the clustering quality from various (often conflicting) objective criteria, but the single-criterion clustering approaches offer little opportunities for this at the methodological level.

Considering the interpretability and superiority of the multicriteria clustering modelling approaches, such approaches may be more efficient at improving the identification of analogies for the segmentation of analogies. The use of multicriteria clustering approaches ensure that the analogies identified are homogeneous in terms of the underlying time-based patterns and causal factors that govern the patterns observed.

To the best of our knowledge, little work exists that has systematically explored and investigated the effectiveness of multicriteria clustering approaches in the context of forecasting, where analogies are required to be identified. Therefore, our first research task focuses on the development of suitable multicriteria clustering approaches, which are capable of integrating multiple information sources during in the clustering procedure. Using numerous experiments, we aim to (i) propose a methodological framework (forecasting process) which comprises procedures to allow the segmentation of analogies and the forecasting stage which pools information from the segmented analogies. The development of the segmentation stage should in particular, be able to accommodate for multiple incommensurable information sources; (ii) provide new insights regarding the relationship between the segmentation of analogies and the forecasting stage. Utilizing a statistical forecasting model in the forecasting stage, our framework could assess the impact of segmentation of analogies on the forecasting accuracy in an objective manner. This has not been covered in the previous literature; (iii) evaluate the relative performance of multicriteria segmentation approaches to the traditional single-criterion segmentation approaches.

2.4.2 Automatic model selection in the context of multicriteria clustering

Internal cluster validation metrics have been commonly employed to data clustering in order to address the challenge of model selection. This is particularly when one criterion is considered during the clustering procedure. However, limited work has been reported to investigate the effectiveness of these established measures for the exploration of new possibilities when multiple criteria (*e.g.,* information data sources or feature spaces, distance metrics, standardization techniques) are involved during the clustering procedure. A related literature study currently exists and is provided in Handl and Knowles (2013). Their work clusters objects using the multi-objective evolutionary algorithm and selects the single best partitioning out of sets of Pareto-optimal solutions using the Gap statistics. Matake et al. (2007) also provided some insights regarding the model selection process relevant to multicriteria market segmentation. However, they selected various regions of Pareto-optimal sets, which contain favorable trade-offs solutions based on subjective assessments. Subsequently, they identified one representative solution from the Pareto-optimal clustering solution. In other words, the authors tackled the problem of model selection through their subjective assessment opposed to automated procedures.

In the context of multi-objective optimization problems, a bulk of research has been conducted regarding the accurate determination of the knee point. However, the majority of the conducted studies make use of domain knowledge for determining the knee point. Promising objective methods include the angle-based measure (Branke et al., 2004). Instead of using only two nearest neighbours, they proposed an enhanced version of the angle-based method, by considering the four nearest neighbours as part of their determination of the largest angle. Individual points returning the largest angle were regarded as the knee point. This method has advantages in terms of its applicability and scalability. The idea behind this technique is that the most interesting solutions out of the Pareto-optimal candidates are those where a small improvement in one objective would result in a large deterioration in at least one other objective. Given the potential of this method, however, little work has been reported that specifically explored the possibility of applying or adapting this method from multi-objective optimization literature to the domain of multicriteria clustering problems.

Intrinsically, all approaches that we have discussed focus on the assessment of solution quality in terms of the procedure itself. As pointed by Guyon, Von Luxburg, and Williamson (2009), clustering might be a part of the whole chain of analysis in applications. Based on the same data, the structure discovered can differ as per the final application purpose. It can be misleading to derive a general evaluation procedure for clustering algorithms which is indeed application independent. Often, the evaluation of clustering solutions might require contextual information or domain knowledge. In line with Guyon, Von Luxburg, and Williamson (2009), the evaluation of clustering results is best assessed by taking into account a problem-specific context. Referring back to the model selection step, it might be a sub-optimal solution if we determine the single best clustering solution without the use of the forecasting results during the procedure.

Taking these factors into account, our second research challenge attempts to develop a set of automatic model selection techniques that serves to complement the multicriteria clustering procedure in the context of forecasting applications. We also attempt to compare the performance of the clustering-focused and forecasting-focused approaches that we have proposed to deal with the model selection problem.

2.4.3 Addressing the clustering instability using bootstrap aggregation techniques

For statistical models, different sources of instabilities might be involved in the modeling process. These include the instabilities originating from the input data, model parameters and the determination of model structure. Without using analogies, traditional statistical forecasting models may primarily inherit the instabilities from the above sources.

In contrast, for the forecasting process that makes use of analogies, additional instabilities might occur from the segmentation step where clustering procedure is applied to group analogies. Specifically, for clustering approaches, the determination of the number of clusters or the random initialization of the clustering algorithms might lead to clustering instability. For example, assuming the correct number of clusters is K, the incorrect determination of K might result in instabilities of the clustering results. $K+1$ clusters might lead to wrong split of the true clusters, while $K-1$ might yield wrong merge of the true clusters. This highlights that the estimation of the number of clusters is important for a clustering procedure, and an incorrect determination of K itself might deteriorate the clustering stability. In addition to this, multicriteria clustering approaches might expose to extra instabilities when determining the final partitions. This is because for the same K, multicriteria clustering approaches often return more than one clustering result. Furthermore, the random initialization required in a non-parametric clustering algorithm (*e.g., K*-means, PAM) can also result in instabilities. Taking the K-means as an example, this algorithm may produce different clustering solutions after each individual run, due to the algorithms use of random initialization.

In general, the clustering procedure can be unstable due to the model selection step, which comprises the selection of clustering variables, the determination of K, the specification of model parameter. Regarding the clustering stability, more details that are out of the scope of this review are provided by Von Luxburg (2010).

To address the instability issue, resampling methods have been widely applied in various fields. Our third manuscript demonstrates that non-parametric resampling methods such as the bootstrap aggregation technique can address this issue. As analyzed in Section 2.3, bootstrapping time series data is essentially challenging when both non-stationary and time-dependent structure are present in the data. Unfortunately, non-stationary time series are commonly present in practical forecasting applications. In spite of this, few successful applications have been reported from the forecasting literature that shows promising results regarding the bootstrapping of non-stationary time series (limited work is referred to Bergmeir, Hyndman, and Benítez, 2016).

In light of this, the third manuscript of our thesis focuses on the exploration of the potential of IID bootstrap as applied to forecasting models that make use of analogies. Instead of bootstrapping time series directly, this problem could be reduced to a typical problem of IID bootstrap by the resampling of a set of labels that are associated with the time series data. This set of labels are regarded as random variables that follow the identical and independent distribution (IID).

In a more mathematically detailed perspective, a bootstrapped sample is constructed by resampling a set of labels of size n, without perturbing the internal structure the time series data. The method proposed is straightforward and easy to apply. Specifically, we bootstrap on a set of labels **X** $= (X_1, X_2, \ldots, X_n)$, where labels X_i is associated with a specific time series i. The observed realizations for each $X_i = (x_1, x_2, \ldots, x_n)$. The bootstrapped samples can be represented as **X**$^b = (X_1^b, X_2^b, \ldots, X_n^b)$ and real realizations are unchanged for $X_i^b = (x_1, x_2, \ldots, x_n)$. To bootstrap non-stationary time series (only one promising paper Bergmeir, Hyndman, and Benítez, 2016), we expect that this innovative generation process could yield different groups of analogies that can be further utilized in the forecasting stage. By averaging point forecasts for each series across multiple bootstrapped samples, our ultimate goal is to generate aggregated point forecasts, that could be more reliable and accurate than individual forecasts.

2.4.4 Improving the performance of time series clustering using multicriteria approaches

Time series clustering is particularly useful and interesting. This is because the application of time series clustering can be easily applied in areas ranging from biology to finance and economics and even signal processing areas. The clustering of time series data is challenging, as there are no universally accepted notions of similarity among pairs of time series. The optimal definition of similarity may vary with the application context.

As reviewed in Section 2.2.2.2, conventionally, a single distance measure and standardization technique is employed during the clustering procedure for the grouping of a collection of time series. Each different distance metric / standardization technique may attempt to capture the notion of similarity, between pairs of time series, by emphasizing different aspects. However, it is common to observe that there are mixed types of patterns such as linearity and non-linearity simultaneously present in the time series data (Zhang, 2003). A distance measure may be able to capture the linear pattern underlying the data, yet fail to model the non-linearity of the data.

In fact, a single distance metric may perform well in approximating part of the whole picture of similarity. That is to say, the independent consideration of any one isolated distance metric may prove insufficient since various types of patterns can simultaneously present in the data. For example, the ARIMA model might be adequate for modelling linear patterns present in the US dollar exchange rate time series, yet fail to capture the non-linearity in these time series (Zhang, 2003). Similarly, Stoddard (1979) contended that any type of standardization can remove the between-cluster variation. The variation might be crucial for uncovering the underlying data structure. However, almost all existing clustering approaches employ a uniform normalization scheme over all data items on a set of variables. On account of this, it may be more appropriate to seek a suitable approach that combines the strengths of different metrics and automatically adjusts the importance of the considered criteria so as to satisfy the different application needs.

According to Handl and Knowles (2007) who employed multicriteria approaches to data clustering, where multiple clustering criteria are utilized to facilitate a more robust recovery of the data structure. Similarly, since there is neither a universally accepted notion of similarity, nor is there formal guideline of its use in different circumstances, we may benefit from combining multiple distance metrics / standardization techniques, to capture complementary information available from different metrics.

The last research task contained in this thesis is concerned with the development of advanced clustering approaches for addressing time series clustering problems. The criteria we consider here includes different distance metrics / standardization techniques. To help the automatic procedure of model selection, we measure the clustering quality based on the overall performance of the forecasting process, which employs analogous time series. This is because the success of a clustering solution is best assessed in context of the application where the solution is employed(Guyon, Von Luxburg, and Williamson, 2009). We aim to investigate the efficiency of multicriteria approaches to time series clustering.

2.5 Justification of chosen research methodology

2.5.1 Research philosophy

The term epistemology (what is known to be true) as opposed to doxology (what is believed to be true) encompasses the various philosophies of research approach. The purpose of science is the process of transforming things believed into things known: doxa to episteme. A research philosophy is a belief about the way where data about a phenomenon should be collected, analyzed and used. It works on the source, nature and development of knowledge (Cooper, Schindler, and Sun, 2006). Two major research philosophies have been identified in the Western tradition of science, namely positivist and interpretivist (Galliers, 1991). A positivists paradigm assumes a quantitative methodology while interpretivist assumes a qualitative methodology such as survey, questionnaires, interviews. In more details, positivists believe that reality is stable and can be observed and described from an objective viewpoint (Levin and Gaeth, 1988) / without interfering with the phenomena being studied. They contend that phenomena should be isolated and that observations should be repeatable. This often involves manipulation of reality with variations in only a single independent variable so as to identify regularities in,m and to form relationships between some of the constituent elements of the social world. Predictions can be made on the basis of the previously observed and explained realities and their inter-relationship. On the other hand, interpretivists contend that only through the subjective interpretation of and intervention in reality can that reality be fully understood. The study of phenomena in their natural environment is key to the interpretivists philosophy, together with the acknowledgement that scientist cannot avoid affecting those phenomena they study They admit that there may be many interpretations of reality, but maintain that these interpations are in themselves a part of the scientific knowledge they are pursuing. Interpretivism has a tradition that is no less glorious than that of positivism, nor is it shorter. In stead, we believe that both research methodologies are valuable if managed carefully. Our over-riding concern is that the research we undertake should be both relevant to our research challenges, as set out in Section 2.4, and rigorous in its operationalization. Overall, we believe that positivist philosophy is required for this purpose.

Positivist researchers remain detached from the participants of the research by creating a distance, which is important in remaining emotionally neutral to make clear distinctions between reason and feeling (Carson et al., 2001). They also maintain a clear distinction between science and personal experience and fact and value judgment. It is also important in positivist research to seek objectivity and use consistently rational and logical approaches to research (Carson et al., 2001). Statistical and mathematical techniques are central to positivist research, which adheres to specifically structured research techniques to uncover single and objective reality (Carson et al., 2001). The goal of positivist researchers is to make time and context free generalizations. They believe this is possible because human actions can be explained as a result of real causes that temporarily precedes their behaviour and the researcher and his research subjects are independent and do not influence each other (Hudson and Ozanne, 1988). Accordingly, positivist researchers also attempt to remain detached from the participants of the research by creating distance between themselves and the participants. Especially, this is an important step in remaining emotionally neutral to make clear distinctions between reason and feeling as well as between science and personal experience. Positivists also claim it is important to clearly distinguish between fact and value judgment. As positivist researchers they seek objectivity and use consistently rational and logical approaches to research (Carson et al., 2001; Hudson and Ozanne, 1988). However, positivism is associated with the following set of disadvantages: Positivism relies on experience as a valid source of knowledge. However, a wide range of basic and important concepts such as cause, time and spaces are not based on experience. Secondly, positivism assumes that all types of processes can be perceived as a certain variation of actions of individuals or relationship between individuals. Thirdly, adoption of positivism in business studies and other studies can be criticized for reliance on status quo. In other words, research findings in positivism studies are only descriptive, thus they lack insight into in-depth issues.

2.5.2 Research approach

Under positivism, research approaches fall into three major categories: the deductive research approach, inductive research approach and abductive research approach. The relevance of hypotheses to the study is the main distinction between deductive and inductive approaches. Deductive approaches test the validity of hypotheses, whereas inductive approaches aim to contribute to the emergence of new theories and generalizations. Abductive research begins with surprising facts or puzzles, and the research process is devoted to their interpretation.

The strategy adopted in this thesis aims to investigate the research challenges that were set out in Section 2.4. To improve the quality of analogies, segmentation approaches can be implemented. Segmentation methods can be broadly classified into *a-prior* and *post-hoc* (Wind, 1978) methods. *A-prior* methods refer to approaches with the type and numbers of clusters are decided before data collection, whereas *post-hoc* methods refer to approaches where the type and number of clusters are derived from data analysis. Post-hoc segmentation analysis generally involves the implementation of cluster analysis, mixture, mixture regression and mixture scaling techniques. Among these, clustering techniques are the most popular tools used for post-hoc segmentation (Wedel and Kamakura, 2012). Due to the reproductivity and scalability of clustering, such methods are the main concern for the proceeding of post-hoc segmentation analysis. Throughout the thesis, we mainly focus on addressing the segmentation of analogies using multicriteria clustering techniques. We aim to provide a systematic investigation to related to different clustering techniques (either single-criterion or multicriteria) for the improved segmentation of analogies in the context of forecasting.

More specifically, as discussed in Section 2.4, multicriteria clustering approaches are proposed here for the purpose of combining multiple criteria using a weighted-sum method. Given this, inductive research approach is carried out throughout the thesis. Following the literature, machine learning studies inductive as they might be carried out by algorithms. Hence, it might be more appropriate for conducting labs-based experiments via computational tool. The idea behind experiments is to investigate the impact and sensitivity of particular factors that might impact on forecasting accuracy of methods that exploit information from grouped analogies, which are homogeneous to group members.

References

[1] G. Aggarwal, S. Garg, and N. Gupta. “Combining clustering solutions with varying number of clusters”. In: *International Journal of Computer Science Issues (IJCSI)* 11.2 (2014), p. 240.

[2] M. Ankerst et al. “OPTICS: ordering points to identify the clustering structure”. In: *ACM SIGMOD Record*. Vol. 28. 2. ACM. 1999, pp. 49–60.

[3] A. Bab-Hadiashar and D. Suter. *Data segmentation and model selection for computer vision: a statistical approach*. Springer Science & Business Media, 2012.

[4] J. M. Bates and C. W. J. Granger. “The combination of forecasts”. In: *Or* (1969), pp. 451–468.

[5] S. Bechikh, L. Ben Said, and K. Ghédira. “Searching for knee regions in multi-objective optimization using mobile reference points”. In: *Proceedings of the 2010 ACM symposium on applied computing*. ACM. 2010, pp. 1118–1125.

[6] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. “Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation”. In: *International Journal of Forecasting* 32.2 (2016), pp. 303–312.

[7] J. Berkowitz and L. Kilian. “Recent developments in bootstrapping time series”. In: *Econometric Reviews* 19.1 (2000), pp. 1–48.

[8] J. C. Bezdek and N. R. Pal. “Some new indexes of cluster validity”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.3 (1998), pp. 301–315.

[9] G. Biau and E. Scornet. “A random forest guided tour”. In: *Test* 25.2 (2016), pp. 197–227.

[10] R. C. Blattberg and S. J. Hoch. “Database models and managerial intuition: 50% model+ 50% manager”. In: *Management Science* 36.8 (1990), pp. 887–899.

[11] A. Bose. “Bootstrap in moving average models”. In: *Annals of the Institute of Statistical Mathematics* 42.4 (1990), pp. 753–768.

[12] A. Bose. “Edgeworth correction by bootstrap in autoregressions”. In: *The Annals of Statistics* 16.4 (1988), pp. 1709–1722.

[13] G. E. Box et al. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[14] J. Branke et al. “Finding knees in multi-objective optimization”. In: *PPSN*. Vol. 3242. 2004, pp. 722–731.

[15] L. Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (1996), pp. 123–140.

[16] L. Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[17] L. Breiman et al. *Classification and regression trees*. CRC press, 1984.

[18] R. G. Brown. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation, 2004.

[19] M. J. Brusco, J. D. Cradit, and S. Stahl. “A simulated annealing heuristic for a bicriterion partitioning problem in market segmentation”. In: *Journal of Marketing Research* 39.1 (2002), pp. 99–109.

[20] M. J. Brusco, J. D. Cradit, and A. Tashchian. “Multicriterion clusterwise regression for joint segmentation settings: An application to customer value”. In: *Journal of Marketing Research* 40.2 (2003), pp. 225–234.

[21] P. Bühlmann. “Bootstraps for time series”. In: *Statistical Science* (2002), pp. 52–72.

[22] E. Carlstein. “The use of subseries values for estimating the variance of a general statistic from a stationary sequence”. In: *The Annals of Statistics* 14.3 (1986), pp. 1171–1179.

[23] D. Carson et al. *Qualitative marketing research*. Sage, 2001.

[24] C. Chatfield. *Time-series forecasting*. CRC Press, 2000.

[25] P. Cheeseman et al. “Autoclass: A Bayesian classification system”. In: *Readings in knowledge acquisition and learning*. Morgan Kaufmann Publishers Inc. 1993, pp. 431–441.

[26] Y.-Y. Cheng, P. P. Chan, and Z.-W. Qiu. “Random forest based ensemble system for short term load forecasting”. In: *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*. Vol. 1. IEEE. 2012, pp. 52–56.

[27] R. T. Clemen. “Combining forecasts: A review and annotated bibliography”. In: *International Journal of Forecasting* 5.4 (1989), pp. 559–583.

[28] D. R. Cooper, P. S. Schindler, and J. Sun. *Business research methods*. Vol. 9. McGraw-Hill Irwin New York, 2006.

[29] C. Cordeiro and M. M. Neves. “Forecasting time series with Boot. EXPOS procedure”. In: *Revstat* 7.2 (2009), pp. 135–149.

[30] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Vol. 1. Cambridge university press, 1997.

[31] K. Deb and J Sundar. “Reference point based multi-objective optimization using evolutionary algorithms”. In: *Proceedings of the 8th annual conference on Genetic and evolutionary computation*. ACM. 2006, pp. 635–642.

[32] M. Delattre and P. Hansen. “Bicriterion cluster analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4 (1980), pp. 277–291.

[33] G. Dudek. “Short-term load forecasting using random forests”. In: *Intelligent Systems' 2014*. Springer, 2015, pp. 821–828.

[34] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[35] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[36] J. C. Dunn. “Well-separated clusters and optimal fuzzy partitions”. In: *Journal of cybernetics* 4.1 (1974), pp. 95–104.

[37] B. Efron. “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics*. Springer, 1992, pp. 569–593.

[38] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[39] D. Ö. Faruk. “A hybrid neural network and ARIMA model for water quality time series prediction”. In: *Engineering Applications of Artificial Intelligence* 23.4 (2010), pp. 586–594.

[40] L. Fenga. “Bootstrap Order Determination for ARMA Models: A Comparison between Different Model Selection Criteria”. In: *Journal of Probability and Statistics* 2017 (2017).

[41] A. Ferligoj and V. Batagelj. “Direct multicriteria clustering algorithms”. In: *Journal of Classification* 9.1 (1992), pp. 43–61.

[42] C. Fraley and A. E. Raftery. “Model-based clustering, discriminant analysis, and density estimation”. In: *Journal of the American statistical Association* 97.458 (2002), pp. 611–631.

[43] D. Freedman. “On bootstrapping two-stage least-squares estimates in stationary linear models”. In: *The Annals of Statistics* 12.3 (1984), pp. 827–842.

[44] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York, 2001.

[45] S. Frühwirth-Schnatter and S. Kaufmann. "Model-based clustering of multiple time series". In: *Journal of Business & Economic Statistics* 26.1 (2008), pp. 78–89.

[46] R. D. Galliers. "Strategic information systems planning: myths, reality and guidelines for successful implementation". In: *European Journal of Information Systems* 1.1 (1991), pp. 55–64.

[47] P. Goodwin, K. Dyussekeneva, and S. Meeran. "The use of analogies in forecasting the annual sales of new electronics products". In: *IMA Journal of Management Mathematics* 24.4 (2013), pp. 407–422.

[48] A. D. Gordon. "Null models in cluster validation". In: *From data to knowledge*. Springer, 1996, pp. 32–44.

[49] C. W. J. Granger and P. Newbold. "Spurious regressions in econometrics". In: *Journal of econometrics* 2.2 (1974), pp. 111–120.

[50] I. Guyon, U. Von Luxburg, and R. C. Williamson. "Clustering: Science or art". In: *NIPS 2009 Workshop on Clustering Theory*. 2009, pp. 1–11.

[51] M. Halkidi, M. Vazirgiannis, and Y. Batistakis. "Quality scheme assessment in the clustering process". In: *Principles of Data Mining and Knowledge Discovery* (2000), pp. 265–276.

[52] P. Hall. "Resampling a coverage pattern". In: *Stochastic processes and their applications* 20.2 (1985), pp. 231–246.

[53] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[54] J. Handl and J. Knowles. "An evolutionary approach to multiobjective clustering". In: *IEEE transactions on Evolutionary Computation* 11.1 (2007), pp. 56–76.

[55] J. Handl and J. Knowles. "Evidence accumulation in multiobjective data clustering". In: *International Conference on Evolutionary Multi-Criterion Optimization*. Springer. 2013, pp. 543–557.

[56] W. Härdle, J. Horowitz, and J.-P. Kreiss. "Bootstrap methods for time series". In: *International Statistical Review* 71.2 (2003), pp. 435–459.

[57] A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

[58] M. Hibon and T. Evgeniou. "To combine or not to combine: selecting among forecasts and their combinations". In: *International Journal of Forecasting* 21.1 (2005), pp. 15–24.

[59] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218.

[60] L. A. Hudson and J. L. Ozanne. "Alternative ways of seeking knowledge in consumer research". In: *Journal of consumer research* 14.4 (1988), pp. 508–521.

[61] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.

[62] K. Kalpakis, D. Gada, and V. Puttagunta. "Distance measures for effective clustering of ARIMA time-series". In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE. 2001, pp. 273–280.

[63] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[64] L. Khaidem, S. Saha, and S. R. Dey. "Predicting the direction of stock market prices using random forest". In: *arXiv preprint arXiv:1605.00003* (2016).

[65] T. Kohonen. “The self-organizing map”. In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480.

[66] J.-P. Kreiss and J. Franke. “BOOTSTRAPPING STATIONARY AUTOREGRESSIVE MOVING-AVERAGE MODELS”. In: *Journal of Time Series Analysis* 13.4 (1992), pp. 297–317.

[67] A. M. Krieger and P. E. Green. “Modifying cluster-based segments to enhance agreement with an exogenous response variable”. In: *Journal of Marketing Research* (1996), pp. 351–363.

[68] M. Kuhn and K. Johnson. *Applied predictive modeling*. Vol. 26. Springer, 2013.

[69] M. Kumar and M Thenmozhi. “Forecasting stock index movement: A comparison of support vector machines and random forest”. In: (2006).

[70] H. R. Kunsch. “The jackknife and the bootstrap for general stationary observations”. In: *The Annals of Statistics* 17.3 (1989), pp. 1217–1241.

[71] D. Kushary. *Bootstrap Methods and Their Application*. 2000.

[72] P. Lahiri et al. “On the impact of bootstrap in survey sampling and small-area estimation”. In: *Statistical Science* 18.2 (2003), pp. 199–210.

[73] R. P. Larrick and J. B. Soll. “Intuitions about combining opinions: Misappreciation of the averaging principle”. In: *Management Science* 52.1 (2006), pp. 111–127.

[74] M. Lawrence et al. “Judgmental forecasting: A review of progress over the last 25years”. In: *International Journal of Forecasting* 22.3 (2006), pp. 493–518.

[75] J. Leitner and U. Leopold-Wildburger. “Experiments on forecasting behavior with several sources of information–A review of the literature”. In: *European Journal of Operational Research* 213.3 (2011), pp. 459–469.

[76] I. P. Levin and G. J. Gaeth. “How consumers are affected by the framing of attribute information before and after consuming the product”. In: *Journal of consumer research* 15.3 (1988), pp. 374–378.

[77] T. W. Liao. “Clustering of time series data—a survey”. In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[78] J.-S. Lim. “An Empirical Investigation of the Effectiveness of Time Series Judgmental Adjustment Using Forecasting Support Systems”. PhD thesis. University of New South Wales, 1993.

[79] R. Y. Liu and K. Singh. “Moving blocks jackknife and bootstrap capture weak dependence”. In: *Exploring the limits of bootstrap* 225 (1992), p. 248.

[80] Y. Liu et al. “Multicriterion market segmentation: a new model, implementation, and evaluation”. In: *Marketing Science* 29.5 (2010), pp. 880–894.

[81] N. Matake et al. “Multiobjective clustering with automatic k-determination for large-scale data”. In: *Proceedings of the 9th annual conference on Genetic and evolutionary computation*. ACM. 2007, pp. 861–868.

[82] C. A. Mattson, A. A. Mullur, and A. Messac. “Smart Pareto filter: Obtaining a minimal representation of multiobjective design space”. In: *Engineering Optimization* 36.6 (2004), pp. 721–740.

[83] G. W. Milligan and M. C. Cooper. “An examination of procedures for determining the number of clusters in a data set”. In: *Psychometrika* 50.2 (1985), pp. 159–179.

[84] J. H. Myers. “Segmentation and positioning for strategic marketing decisions”. In: American Marketing Association. 1996.

[85] K. Nikolopoulos et al. “Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches”. In: *European Journal of Operational Research* 180.1 (2007), pp. 354–368.

[86] S. Niwattanakul et al. “Using of Jaccard coefficient for keywords similarity”. In: *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 6. 2013.

[87] M. O’Connor, W. Remus, and K. Griggs. “Judgemental forecasting in times of change”. In: *International Journal of Forecasting* 9.2 (1993), pp. 163–172.

[88] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas. “How many trees in a random forest?” In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2012, pp. 154–168.

[89] N. R. Pal and J Biswas. “Cluster validation using graph theoretic concepts”. In: *Pattern Recognition* 30.6 (1997), pp. 847–857.

[90] E. Paparoditis and D. N. Politis. “Resampling and subsampling for financial time series”. In: *Handbook of financial time series* (2009), pp. 983–999.

[91] L. Parsons, E. Haque, and H. Liu. “Subspace clustering for high dimensional data: a review”. In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 90–105.

[92] D. Politis, J. P. Romano, and M. Wolf. “Weak convergence of dependent empirical measures with application to subsampling in function spaces”. In: *Journal of statistical planning and inference* 79.2 (1999), pp. 179–190.

[93] D. N. Politis and J. P. Romano. “On flat-top kernel spectral density estimators for homogeneous random fields”. In: *Journal of Statistical Planning and Inference* 51.1 (1996), pp. 41–53.

[94] L. Rachmawati and D. Srinivasan. “A multi-objective evolutionary algorithm with weighted-sum niching for convergence on knee regions”. In: *Proceedings of the 8th annual conference on Genetic and evolutionary computation*. ACM. 2006, pp. 749–750.

[95] L. Rachmawati and D. Srinivasan. “Multiobjective evolutionary algorithm with controllable focus on the knees of the Pareto front”. In: *IEEE Transactions on Evolutionary Computation* 13.4 (2009), pp. 810–824.

[96] L. Rachmawati and D. Srinivasan. “Preference incorporation in multi-objective evolutionary algorithms: A survey”. In: *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*. IEEE. 2006, pp. 962–968.

[97] M. B. Rajarshi. “Bootstrap in Markov-sequences based on estimates of transition density”. In: *Annals of the Institute of Statistical Mathematics* 42.2 (1990), pp. 253–268.

[98] W. M. Rand. “Objective criteria for the evaluation of clustering methods”. In: *Journal of the American Statistical association* 66.336 (1971), pp. 846–850.

[99] P. J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[100] E. Ruiz and L. Pascual. “Bootstrapping financial time series”. In: *Journal of Economic Surveys* 16.3 (2002), pp. 271–300.

[101] N. R. Sanders and L. P. Ritzman. “Judgmental adjustment of statistical forecasts”. In: *Principles of Forecasting*. Springer, 2001, pp. 405–416.

[102] O. Schütze, M. Laumanns, and C. A. C. Coello. “Approximating the Knee of an MOP with Stochastic Search Algorithms.” In: *PPSN*. Springer. 2008, pp. 795–804.

[103] E. Scornet, G. Biau, J.-P. Vert, et al. “Consistency of random forests”. In: *The Annals of Statistics* 43.4 (2015), pp. 1716–1741.

[104] J. Shao and D. Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[105] W. R. Smith. “Product differentiation and market segmentation as alternative marketing strategies”. In: *Journal of marketing* 21.1 (1956), pp. 3–8.

[106] J. B. Soll. “Intuitive theories of information: Beliefs about the value of redundancy”. In: *Cognitive Psychology* 38.2 (1999), pp. 317–346.

[107] A. M. Stoddard. “Standardization of measures prior to cluster analysis”. In: *Biometrics* (1979), pp. 765–773.

[108] C. A. Sugar. “Techniques for clustering and classification with applications to medical problems.” In: (1999).

[109] C. A. Sugar, L. A. Lenert, and R. A. Olshen. “An application of cluster analysis to health services research: Empirically defined health states for depression from the sf-12”. In: (1999).

[110] R. Tibshirani, G. Walther, and T. Hastie. “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[111] H. Tyralis and G. Papacharalampous. “Variable selection in time series forecasting using random forests”. In: *Algorithms* 10.4 (2017), p. 114.

[112] A. Verikas, A. Gelzinis, and M. Bacauskiene. “Mining data with random forests: A survey and results of new tests”. In: *Pattern recognition* 44.2 (2011), pp. 330–349.

[113] U. Von Luxburg. “Clustering stability: an overview”. In: *Foundations and Trends® in Machine Learning* 2.3 (2010), pp. 235–274.

[114] J. Wang. “Consistent selection of the number of clusters via crossvalidation”. In: *Biometrika* 97.4 (2010), pp. 893–904.

[115] W. Wang, J. Yang, and R. Muntz. “STING: A statistical information grid approach to spatial data mining”. In: *VLDB*. Vol. 97. 1997, pp. 186–195.

[116] X. Wang, K. Smith, and R. Hyndman. “Characteristic-based clustering for time series data”. In: *Data mining and knowledge Discovery* 13.3 (2006), pp. 335–364.

[117] R. Webby and M. O’Connor. “Judgemental and statistical time series forecasting: a review of the literature”. In: *International Journal of Forecasting* 12.1 (1996), pp. 91–118.

[118] M. Wedel and W. A. Kamakura. *Market segmentation: Conceptual and methodological foundations*. Vol. 8. Springer Science & Business Media, 2012.

[119] Y. Wind. “Issues and advances in segmentation research”. In: *Journal of marketing research* (1978), pp. 317–337.

[120] J.-H. Yang, C.-H. Cheng, and C.-P. Chan. “A Time-Series Water Level Forecasting Model Based on Imputation and Variable Selection Method”. In: *Computational intelligence and neuroscience* 2017 (2017).

[121] S. L. Zeger and C. M. Hurvich. “A frequency-domain median time series”. In: *Journal of the American Statistical Association* 82.399 (1987), pp. 832–835.

[122] G. P. Zhang. “Time series forecasting using a hybrid ARIMA and neural network model”. In: *Neurocomputing* 50 (2003), pp. 159–175.

Chapter 3

Determining analogies based on the integration of multiple information sources (paper 1)

3.1 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

3.2 Introduction

Forecasting approaches such as the Cross-Sectional Multi-State Kalman Filter algorithm (C-MSKF: Duncan, Gorr, and Szczypula, 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, Gorr, and Szczypula, 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, Gorr, and Szczypula, 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 3.3 surveys related work, including pooling approaches and popular segmentation approaches in the literature. Section 3.4 proposes our three-stage prediction process. Section 3.5 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 3.6 summarizes results on a data set describing personal income tax liability data. Finally, Section 3.7 concludes.

3.3 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, Gorr, and Szczypula, 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, Dyussekeneva, and Meeran, 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, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 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, Gorr, and Szczypula, 1995) and tax revenue (Duncan, Gorr, and Szczypula, 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, Gorr, and Szczypula, 1994; Duncan, Gorr, and Szczypula, 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, Gorr, and Szczypula, 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, Gorr, and Szczypula, 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, Gada, and Puttagunta, 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, Wedel, and Wilms, 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, Cradit, and Stahl, 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, Cradit, and Stahl, 2002; Brusco, Cradit, and Tashchian, 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.

3.4 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.

3.4.1 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:

$$\delta^{TS}(i,j) = 1 - \frac{T(\sum_t x_{it}x_{jt}) - (\sum_t x_{it})(\sum_t x_{jt})}{\sqrt{(T(\sum_t x_{it}^2) - (\sum_t x_{it})^2)(T(\sum_t x_{jt}^2) - (\sum_t x_{jt})^2)}} \tag{3.1}$$

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:

$$\delta^{CF}(i,j) = \sum_{m} (a_{im} - a_{jm})^2 \tag{3.2}$$

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 $\delta^{CF}(i,j)$:

$$\delta^{CF}(i,j) = \#\{m : a_{im} \neq a_{jm}, m = 1, ..., M\} \tag{3.3}$$

3.4.2 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:

$$d_{ij}^{CF} \leftarrow \frac{d_{ij}^{CF} - \text{min}(\mathbf{D^{CF}})}{\text{max}(\mathbf{D^{CF}}) - \text{min}(\mathbf{D^{CF}})} \tag{3.4}$$

$$d_{ij}^{TS} \leftarrow \frac{d_{ij}^{TS} - \text{min}(\mathbf{D^{TS}})}{\text{max}(\mathbf{D^{TS}}) - \text{min}(\mathbf{D^{TS}})} \tag{3.5}$$

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 $\mathbf{D^{MC}_{\omega}}$ is obtained as follows:

$$d_{ij\omega}^{MC} = (1-\omega) \times d_{ij}^{CF} + \omega \times d_{ij}^{TS} \tag{3.6}$$

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.

[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.

3.4.3 Model selection

3.4.3.1 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.

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:

$$\text{Sil}(i) = \frac{b_i - c_i}{\text{max}(c_i, b_i)} \tag{3.7}$$

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.

3.4.3.2 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, Von Luxburg, and Williamson, 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:

$$\text{MSE} = \text{mean}(e_t^2) \tag{3.8}$$

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.

3.4.4 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, Gorr, and Szczypula, 1993).

A full description of the C-MSKF algorithm is available in the literature (Duncan, Gorr, and Szczypula, 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.

3.4.5 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

3.5 Empirical evaluation

3.5.1 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 3.6. 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:

$$f_1(t) = 0.8t + 2.8, \quad \text{if } 1 \leq t \leq q \tag{3.9}$$

$$f_2(t) = 4ln(t) + 2, \quad \text{if } 1 \leq t \leq q \tag{3.10}$$

$$f_3(t) = \begin{cases} 0.7t + 2.8, & \text{if } 1 \leq t \leq p \\ -0.9t + 25, & \text{if } p+1 \leq t \leq q \end{cases} \tag{3.11}$$

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:

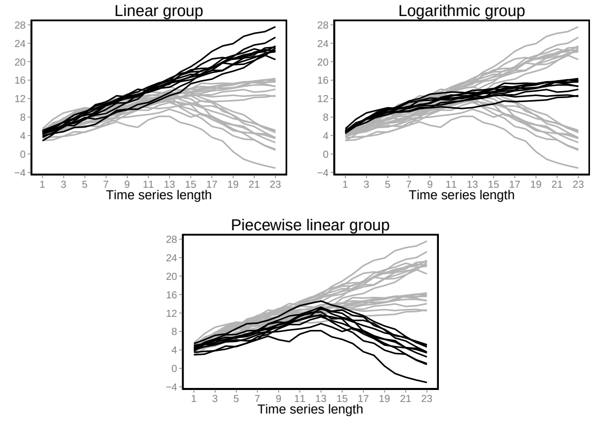
$$X_{it} = \begin{cases} f_g(0) + N(f_g(t+1) - f_g(t), \sigma^2_{TS}), & \text{if } t = 1 \\ X_{i(t-1)} + N(f_g(t+1) - f_g(t), \sigma^2_{TS}), & \text{if } 1 < t \leq q-1 \end{cases} \tag{3.12}$$

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 (3.12), 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 3.1, and it is evident that differentiation between these series is challenging for earlier time intervals. Following Duncan, Gorr, and Szczypula (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.

[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.

Figure 3.1: Illustration of raw time series data generated from a linear, logarithmic, and piecewise linear function.



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 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 $N(\mu_{CF}, \sigma_{CF}^2)$, 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 3.1). Specifically, σ_{CF} is fixed at 0.35 while σ_{TS} is increased from 0.35 to 1.15 in steps of 0.2.

All other parameters are kept constant in the experiments, and are summarized in

[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.

Table 3.1: 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

Table 3.2. 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 3.2). 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.

Table 3.2: 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

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.

3.5.2 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 3.4. 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 3.4.3.2, 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.

3.5.3 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 (3.8). Additionally, we also employ the Symmetric Mean Absolute Percentage Error, sMAPE (Bergmeir, Hyndman, and Benítez, 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:

$$\text{sMAPE} = \text{mean}(200\frac{|e_t|}{|X_t| + |F_t|}) \tag{3.13}$$

where all relevant variables have been defined previously (see Equation (3.8)). 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.

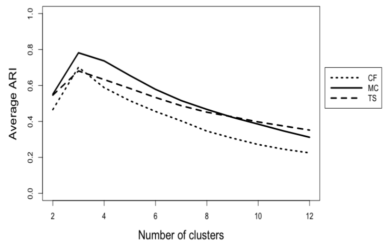
Using a representation based on the $L \times K$ contingency table defined by two partitions (of the same data) with L and K clusters, respectively, the Adjusted Rand Index between the two partitions is given as

$$\text{ARI} = \frac{\sum\limits_{l,m}\binom{N_{lm}}{2} - [\sum\limits_{l}\binom{N_{l.}}{2} \cdot \sum\limits_{k}\binom{N_{.m}}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum\limits_{l}\binom{N_{l.}}{2} + \sum\limits_{m}\binom{N_{.m}}{2}] - [\sum\limits_{l}\binom{N_{l\cdot}}{2} \cdot \sum\limits_{m}\binom{N_{.m}}{2}]/\binom{N}{2}} \tag{3.14}$$

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 $N_{l.}$ and $N_{.m}$ represent row and column totals for row l and column m of the table, respectively.

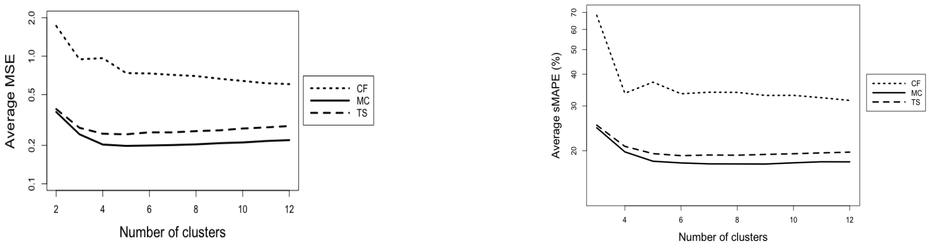
The ARI has been constructed so that the expected value of two random partitions is 0, with the generalized hypergeometric distribution as the model of randomness. The ARI takes a maximum value of 1 and an expected minimum value of 0, with higher values indicating a closer match between the partitions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

Figure 3.2: 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 3.3: 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)

3.5.4 Results

3.5.4.1 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 3.4.3), 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 3.2) 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 3.3), 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.

3.5.4.2 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 3.4.3 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

[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.

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.

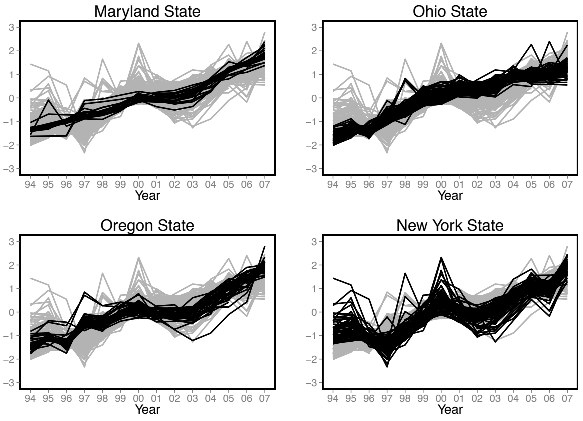
To explore the impact of automated weight selection in more detail, this section contrast the results obtained after the first model selection step (MC, which continues to select the weight for a given K by considering the best possible forecasting accuracy), with a fully automatic approach, MC$_{SilHist}$, that implements both of the model selection steps outlined in Section 3.4. 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 3.3 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 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), 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 3.4), we can confirm the consistent advantage of C-MSKF when employing partitions that have been generated based on multicriteria clustering (MC and MC$_{SilHist}$), as compared to TS or CF clustering. Only for the highest noise level is MC$_{SilHist}$ method 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.

Figure 3.4: Standardized time series of personal income tax in 208 counties in Maryland, New York, Ohio and Oregon State from 1994 to 2007.



3.6 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 type of forecasting task meets the conditions for the applicability of the C-MSKF algorithm, as summarized in Section 3.3.

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 3.4 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.

3.6.1 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 3.4 and Section 3.5.

3.6.2 Results

Table 3.5 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 3.11 in the Appendix.

In line with previous work (Duncan, Gorr, and Szczypula, 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.

3.7 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 multicriteria 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.

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 3.6, 3.7,..., 3.10. 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 MC$_{SilHist}$ and TS. Comparing the difference between CF and MC-based forecasting methods, including MC and MC$_{SilHist}$, the gap closes and eventually (for the highest noise setting) CF starts to outperform the 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 3.11 considers the significance of performance differences for the income tax liability data. For these data, weight selection in the 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 MC$_{SilHist}$ perform best among the contestant forecasting methods.

Table 3.3: 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

Table 3.4: 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	h=1	h=2	h=3	h=4	h=5	h=6
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 3.5: 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	2-year	3-year	1-year	2-year	3-year
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

Table 3.6: 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 3.7: 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 3.8: 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 3.9: 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 3.10: 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 3.11: 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.

References

[1] J. S. Armstrong. “Findings from evidence-based forecasting: Methods for reducing forecast error”. In: *International Journal of Forecasting* 22.3 (2006), pp. 583–598.

[2] J. S. Armstrong. *Principles of forecasting: a handbook for researchers and practitioners*. Vol. 30. Springer Science & Business Media, 2001.

[3] F. M. Bass. “A new product growth for model consumer durables”. In: *Management Science* 15.5 (1969), pp. 215–227.

[4] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. “Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation”. In: *International Journal of Forecasting* 32.2 (2016), pp. 303–312.

[5] R. G. Brown. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation, 2004.

[6] M. J. Brusco, J. D. Cradit, and S. Stahl. “A simulated annealing heuristic for a bicriterion partitioning problem in market segmentation”. In: *Journal of Marketing Research* 39.1 (2002), pp. 99–109.

[7] M. J. Brusco, J. D. Cradit, and A. Tashchian. “Multicriterion clusterwise regression for joint segmentation settings: An application to customer value”. In: *Journal of Marketing Research* 40.2 (2003), pp. 225–234.

[8] M. Delattre and P. Hansen. “Bicriterion cluster analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4 (1980), pp. 277–291.

[9] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[10] G. Duncan, W. Gorr, and J. Szczypula. *Comparative Study of Cross Sectional Methods for Time Series With Structural Changes*. Tech. rep. Carnegie Mellon University, 1994.

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

[12] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[13] A. Ferligoj and V. Batagelj. “Direct multicriteria clustering algorithms”. In: *Journal of Classification* 9.1 (1992), pp. 43–61.

[14] S. Frühwirth-Schnatter and S. Kaufmann. “Model-based clustering of multiple time series”. In: *Journal of Business & Economic Statistics* 26.1 (2008), pp. 78–89.

[15] P. Goodwin, K. Dyussekeneva, and S. Meeran. “The use of analogies in forecasting the annual sales of new electronics products”. In: *IMA Journal of Management Mathematics* 24.4 (2013), pp. 407–422.

[16] C. W. J. Granger and P. Newbold. “Spurious regressions in econometrics”. In: *Journal of econometrics* 2.2 (1974), pp. 111–120.

[17] K. C. Green and J. S. Armstrong. “Structured analogies for forecasting”. In: *International Journal of Forecasting* 23.3 (2007), pp. 365–376.

[18] N. P. Greis and C. Z. Gilstein. “Empirical Bayes methods for telecommunications forecasting”. In: *International Journal of Forecasting* 7.2 (1991), pp. 183–197.

[19] I. Guyon, U. Von Luxburg, and R. C. Williamson. “Clustering: Science or art”. In: *NIPS 2009 Workshop on Clustering Theory*. 2009, pp. 1–11.

[20] J. Handl and J. Knowles. “An evolutionary approach to multiobjective clustering”. In: *IEEE transactions on Evolutionary Computation* 11.1 (2007), pp. 56–76.

[21] P. J. Harrison and C. F. Stevens. “A Bayesian approach to short-term forecasting”. In: *Operational Research Quarterly* (1971), pp. 341–362.

[22] L. Hubert and P. Arabie. “Comparing partitions”. In: *Journal of Classification* 2.1 (1985), pp. 193–218.

[23] K. Kalpakis, D. Gada, and V. Puttagunta. “Distance measures for effective clustering of ARIMA time-series”. In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*. IEEE. 2001, pp. 273–280.

[24] R. E. Kass and D. Steffey. “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)”. In: *Journal of the American Statistical Association* 84.407 (1989), pp. 717–726.

[25] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[26] W. Y. Lee et al. “Providing support for the use of analogies in demand forecasting tasks”. In: *International Journal of Forecasting* 23.3 (2007), pp. 377–390.

[27] J. Leitner and U. Leopold-Wildburger. “Experiments on forecasting behavior with several sources of information–A review of the literature”. In: *European Journal of Operational Research* 213.3 (2011), pp. 459–469.

[28] T. W. Liao. “Clustering of time series data—a survey”. In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[29] Y. Liu et al. “Multicriterion market segmentation: a new model, implementation, and evaluation”. In: *Marketing Science* 29.5 (2010), pp. 880–894.

[30] S. Makridakis and M. Hibon. “The M3-Competition: results, conclusions and implications”. In: *International journal of forecasting* 16.4 (2000), pp. 451–476.

[31] J. H. Myers. “Segmentation and positioning for strategic marketing decisions”. In: American Marketing Association. 1996.

[32] K. Nikolopoulos et al. “Forecasting branded and generic pharmaceuticals”. In: *International Journal of Forecasting* 32.2 (2016), pp. 344–357.

[33] K. Nikolopoulos et al. “Relative performance of methods for forecasting special events”. In: *Journal of Business Research* 68.8 (2015), pp. 1785–1791.

[34] F. Petropoulos et al. “‘Horses for Courses’ in demand forecasting”. In: *European Journal of Operational Research* 237.1 (2014), pp. 152–163.

[35] M. I. Piecyk and A. C. McKinnon. “Forecasting the carbon footprint of road freight transport in 2020”. In: *International Journal of Production Economics* 128.1 (2010), pp. 31–42.

[36] P. J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[37] N. D. Savio and K. Nikolopoulos. “A strategic forecasting framework for governmental decision-making and planning”. In: *International Journal of Forecasting* 29.2 (2013), pp. 311–321.

[38] W. R. Smith. “Product differentiation and market segmentation as alternative marketing strategies”. In: *Journal of marketing* 21.1 (1956), pp. 3–8.

[39] J. A. Stimson. “Regression in space and time: A statistical essay”. In: *American Journal of Political Science* (1985), pp. 914–947.

[40] M. Vriens, M. Wedel, and T. Wilms. “Metric conjoint segmentation methods: A Monte Carlo comparison”. In: *Journal of Marketing Research* (1996), pp. 73–85.

[41] R. Webby and M. O'Connor. "Judgemental and statistical time series forecasting: a review of the literature". In: *International Journal of Forecasting* 12.1 (1996), pp. 91–118.

Chapter 4

Model selection in multicriteria clustering problems (paper 2)

4.1 Abstract

Multicriterion approaches partition objects into distinctive clusters by optimizing multiple criteria during the clustering procedure. The combination of various criteria potentially supports the discovery of complex data structure that may not be identified through single-criterion approaches. However, the use of multiple criteria raises additional issues related to model selection: even for a given number of clusters, multicriterion clustering approaches will potentially return a set of Pareto-optimal clustering solutions. These solutions reflect trade-offs between conflicting criteria and are said to be incomparable as an improvement in one criterion can only be obtained at the expense of another. Here, we develop various techniques to automatically determine a single partition from a set of Pareto optimal solutions, and test our approaches in the context of a forecasting problem that involves analogies. In particular, we propose a combination approach that employs the Silhouette Width technique for efficiently estimating the number of clusters, and determines the final partition based on the best historical forecasting results. Empirical analysis suggests the strong performance of this method and confirms that the performance of a clustering solution is best assessed in a problem-specific context, in our case forecasting of analogous time series.

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

4.2 Introduction

Multicriteria approaches to data clustering have received attention across a range of different areas, including bioinformatics (Handl and Kell, 2006), biomedical science (Saha and Bandyopadhyay, 2011), and market segmentation (Liu et al., 2010). These approaches have been developed to create clusters of objects by optimizing more than one criterion during the clustering procedure. These approaches are promising as they yield clustering solutions that trade off among multiple often conflicting criteria. Typically, individual criteria emphasize different aspects of the definition of clustering such as connectivity or compactness of clusters. Methodologically, a multicriteria clustering approach is capable of facilitating more robust discovery of data structure where this cannot be accommodated by a single clustering criterion.

The use of multicriteria approaches to data clustering is advantageous as they might facilitate a better recovery of the underlying data structure. However, they introduce additional challenges related to the model selection step during the clustering procedure. Model selection is a fundamental challenge in the field of data clustering. This is a generic term that encompasses the identification of a suitable number of clusters, the choice of model parameters, and the initialization of the clustering algorithm if required. In single-criterion clustering problems, the major issue associated with the model selection is to determine the appropriate number of clusters. In particular, dissimilarity-based clustering algorithms such as partitioning methods and hierarchical clustering methods all require the selection of the number groups in a single-criterion or multicriteria clustering context. The model selection problem is further complicated by the use of multiple criteria during the clustering procedure, as for the same number of clusters the clustering algorithm might return a set of Pareto-optimal clustering solutions. These solutions reflect trade-offs among various often conflicting criteria. Thus, the clustering Pareto front contains all the partitions where one criterion can not improved at the expense of another.

To our knowledge, a plenty of automatic methods have been developed in the clustering literature to tackle the challenge of model selection, such as the Elbow method, the Gap statistics and the Silhouette Width measure. However, little work has been reported to investigate the effectiveness of these established measures in the context of multicriteria clustering. A primary concern associated with the existing approaches is that model selection is often carried out independent of problem-specific context. Guyon, Von Luxburg, and Williamson, 2009 argued that the effectiveness of clustering solutions is best evaluated by taking into account the overall performance of the application. The idea behind this is in line with Parsons, Haque, and Liu, 2004, who declared that there is no universal definition of clustering, but one might be more suitable than another for a particular application. On account of this, it might be more meaningful and promising to determine the final partitions by taking into account the application's performance. In summary, there exists limited work that provides systematic investigation regarding the applicability of the existing model selection techniques in the context of multicriteria clustering problems. More importantly, further work should be completed to develop model selection methods that assess the performance of clustering solutions with consideration of the final performance of the application.

In light of this, we consider the application of multicriteria clustering approaches in a forecasting context. Specifically, multicriteria clustering approaches have been proposed to tackle the challenge of identifying analogies using multiple useful information sources, *i.e.,* time-based patterns as well as the causal factors that govern these patterns. The concurrent consideration of multiple criteria during the clustering of analogies has been reported to deliver more promising forecasting results. With the problem-specific context, we aim to propose and analyze the effectiveness of various model selection methods, which can take into account the application's performance during the process.

In summary, we divide our model selection techniques into two broad groups: application independent as well as application dependent methods. Application independent methods select the best performing clustering solution based on the highest quality score, which is measured by the Silhouette Width measure. The model selection step does not make use of the forecasting algorithm's accuracy. In contrast, application dependent techniques measure the quality of clustering solutions based on the overall performance of the forecasting model. This explicitly takes into account the accuracy of the forecasting method when determining the final partitions.

The rest of the paper is organized as follows. Section 4.3 revisits the literature concerning the topic of model selection. Section 4.4 describes the main components of the overall prediction process where multicriteria clustering approaches are employed. Section 4.5 presents details regarding the model selection approaches investigated in this paper. Section 4.6 discusses main findings derived from the experiments. Section 4.7 concludes.

4.3 Previous research

The proliferation of clustering methods has been witnessed across various disciplines including finance, marketing, operational research and pattern recognition.

Generically, the use of clustering approaches aims to help discover meaningful groups that reveal the underlying structure of the data. The identified clusters of data objects may further be employed to carry out a cluster-specific analysis.

Over the recent years, the development of multicriteria approaches to data clustering has been reported to create new opportunities for overcoming the limitation of traditional single-criterion clustering methods. These methods are useful for identifying complex data structure with the consideration of multiple conflicting criteria. multicriteria clustering approaches have been demonstrated to be powerful tools, which facilitate the recovery of more robust data structures and create more natural groupings than single-criterion clustering approaches (Handl and Knowles, 2004).

However, the application of multicriteria clustering approaches also cause additional issues related to model selection. Model selection is inherently a fundamental and complex problem in the clustering domain. It can comprise the tasks of variable standardization (Milligan and Cooper, 1988), variable selection (Steinley and Brusco, 2008), choices of the number of clusters (Dimitriadou, Dolničar, and Weingessel, 2002). Nevertheless, the major challenge related to model selection is the determination of the appropriate number of clusters. As multiple criteria are optimized during the clustering stage, for the same number of clusters, the clustering algorithm often returns a set of Pareto-optimal solutions that correspond to trade-offs between often conflicting criteria (Handl and Knowles, 2004). In fact, the quality of these Pareto-optimal solutions is not directly comparable, as an improvement in one criterion is obtained at the expense of at least one of the others. Practically, decision-makers are often required to select only one best solution out of the set of Pareto-optimal clustering solutions for targeting a cluster-specific problem. This is usually done by integrating domain knowledge from the decision makers' (*e.g.,* Liu et al., 2010). Consequently, this can lead to problems with regards to reproducibility and applicability of the model selection process.

In the clustering literature, some techniques have been developed to assist in the estimation of the number of clusters in traditional (single-criterion) clustering problems in an objective manner. Most commonly, the existing techniques devote to minimizing distance-based dissimilarity measures within groups based on the internal cluster validation techniques such as the Elbow methods (Sugar, 1999; Sugar, Lenert, and Olshen, 1999) and the Silhouette Width measure (Rousseeuw, 1987). The Elbow method determines the number of clusters by critically examining a graph of the percentage variance explained as a function of the number of clusters. This method is easy to apply, but the performance can be unsatisfactory as the number of clusters becomes hard to decide when the variance plateaus due to the weak clusterability of the data. The Silhouette Width measure combines the notion of separation and cohesion of the clustering solutions based on the internal data structure. This measure has been widely used to score the quality of a clustering solution. A higher value indicates a better clustering result. Similar to the Elbow method, the Silhouette Width technique also assesses the global characteristics of the entire partitioning. Another well-established method is the Gap statistic. This method standardizes the graph of log(W_k), where W_k is the pooled within-cluster sum of squares around the cluster means. Specifically, it makes a comparison between the log(W_k) and the expected value derived from an suitable null reference distribution of the data. The value of k is chosen as the number of clusters corresponding to the point, where log(W_k) falls the farthest below the reference curve. Later, Handl and Knowles, 2007 combined the Gap statistics with the MOCK model to identify the most interesting clustering solution, namely “knee point”, from a set of Pareto-optimal solutions. In essence, this technique analyzes the location of solutions in objective space relative to a background of unstructured data.

Considering the “knee” concept (Bechikh, Ben Said, and Ghédira, 2010; Branke et al., 2004; Das, 1999; Deb and Sundar, 2006; Mattson, Mullur, and Messac, 2004; Rachmawati and Srinivasan, 2006), there have been some techniques proposed in the multi-objective optimization literature. These optimization models are constructed by

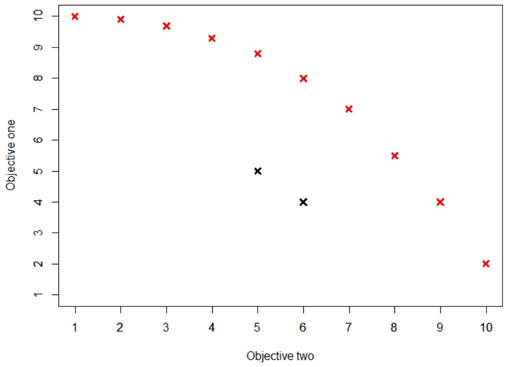
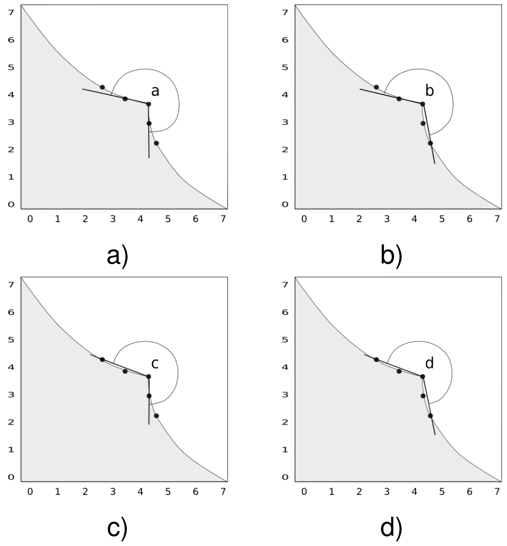


Figure 4.1: Illustration of the Pareto frontier where Pareto-optimal solutions are highlighted in red colour

optimizing multiple objectives at the same time. Often, multiple Pareto-optimal solutions are produced during the optimization process, and these solutions reflect user preferences concerning various criteria. The most interesting solutions of the Pareto-optimal frontier (see Fig. 4.1) are those where a small improvement in one objective might cause an evident degradation in another. Related to this topic, some evolutionary optimization methodologies have been developed to find knee point(s) (Branke et al., 2004; Deb and Sundar, 2006; Rachmawati and Srinivasan, 2006; Schütze, Laumanns, and Coello, 2008).

To identify the knee point, Branke et al., 2004 proposed the enhanced angle-based measure (a data-driven approach) in the context of multi-objective optimization problems. Their intensified version computes four angles between the individual x_i and its four nearest neighbors (x_{i-1}, x_i) and $(x_i,\ x_{i+1})$. These five individuals have to be pairwise linearly independent. If no neighbor to the left or right is available, a vertical or horizontal line is applied to calculate the angle, respectively. The largest of these four angles then assigned to the individual. Individuals with a larger angle-measure are preferred. The demonstration of the four angles is presented in Fig. 4.2. Potentially, the angle-based measure could also be suitable to address the model selection challenge presented in multicriteria clustering problems.

In essence, all the existing methods discussed are application-dependent. These techniques do not take into account the overall performance of the application. In fact, there is no universally accepted definition of clustering. Some definitions may be preferred than the others for certain problems (Parsons, Haque, and Liu, 2004). The groupings discovered become more meaningful when the clustering is appropriately defined from the perspective of the overall application. Again, supported by Guyon, Von Luxburg, and Williamson, 2009, the authors declared that the effectiveness of clustering is best assessed by taking into account the overall performance of the application.



a) b)

c) d)

Figure 4.2: Illustration of the intensified angle-based measure. The standard version just calculates a, the intensified version takes 4 neighbors into account and assigns the maximum angle among a, b, c, d to the individual investigated.

4.4 Forecasting analogous time series using multiple criteria

In this article, we analyze the performance of different model selection methods that address the problem of multicriteria approaches to data clustering. As discussed previously, the application's performance is the key to the evaluation of different clustering solutions. Here, we demonstrate our problems in forecasting circumstances where analogies are utilized during the forecasting stage. As discussed in (Lu and Handl, 2015), analogies are intrinsically better modeled using multiple information criteria. A suitable way of doing this is to base our analysis on the methods with the performance can be objectively evaluated. To satisfy the needs of the application, we locate our study in a forecasting context where clustering methods are widely employed to identify analogies. The concept of analogies has been widely applied in time series forecasting for improving the forecasting accuracy (Armstrong, 2006; Green and Armstrong, 2007; Piecyk and McKinnon, 2010). According to (Duncan, Gorr, and Szczypula, 2001), analogies are commonly used for judgmental approaches (such as forecasting by analogy) Hyndman and Athanasopoulos, 2014 to adjust statistical forecasts (Webby and O'Connor, 1996). The consideration of analogies may reduce biases caused by optimistic or wishful thoughts (Armstrong, 2001; Petropoulos et al., 2014). Moreover, statistical methods have also been proposed to exploit information available from analogies. The Bass model Bass, 1969 is a well-established method which forecasts sales of products which have yet to be launched, through the use of analogous products (Goodwin, Dyussekeneva, and Meeran, 2013). Also, Bayesian pooling approaches, *e.g.,* the Cross-sectional Multi-state Kalman Filter (Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001) have been developed to integrate the information from analogies directly into the stage of forecasting to improve the responsiveness of the algorithm after a structural change caused by external influence while increasing the accuracy of the point forecasts. As shown in previous work, this approach is promising as it requires a relatively small number of parameters and has been reported to show strengths in challenging scenarios such as forecasting of churn in telecommunication networks (Greis and Gilstein, 1991), infant mortality rates (Duncan, Gorr, and Szczypula, 2001) and tax revenue (Duncan, Gorr, and Szczypula, 1993).

This paper aims to analyze the model selection problem lies in multicriteria clustering in the context of forecasting. Hence, we base our analysis on the prediction method that exploits information from analogies. Typically, multicriteria clustering approaches can be employed to identify such analogies, and it becomes suitable for analyzing the model selection challenges in this context. Generally, the whole analytical process examined comprises three main elements: (i) the clustering of analogies using a multicriteria approach with a weighted-sum method; (ii) the implementation of a suitable forecasting algorithm that pools information from the previously identified analogies; (iii) a further step of selecting a most preferred partitioning out of sets of clustering candidates.

4.4.1 Combination of multiple information sources

According to (Lu and Handl, 2015), the optimal identification of analogies should consider the concurrent use of various information sources: the past realizations of time series as well as the factors that are associated with the patterns observed. Multicriteria approaches to the clustering of analogies have shown a promise for recovering

The improved homogeneity of analogies is reported to feed forward into improved forecasting accuracy. Specifically, we implement the idea of combining these two information sources using multicriteria clustering approaches with a weighted-sum method at the distance function level.

For the clustering of causal variables, the squared Euclidean distance is used to measure the distance between the sets of values associated with each pair of time series. Specifically, we denote this distance measure as $\delta^{CF}(i,j)$, i and j are two different time series, and the equation for calculating the distance is presented as follows:

$$\delta^{CF}(i,j) = \sum_{m}(a_{im} - a_{jm})^2 \tag{4.1}$$

where $d_{ij}^{CF} = \delta^{CF}(i,j)$; a_{im} and a_{jm} represent the value of causal variable m associated with time series i and j; $m = 1,2,...,M$, where M represents the number of causal factors, respectively; the dissimilarity matrix $\mathbf{D_{CF}} = (d_{ij}^{CF})$; To eliminate any scale differences, the z-score method is deployed to standardize the causal variable.

Additionally, the similarity between time series is measured based on raw observations, namely the Pearson correlation coefficients. Each object is described as a vector that the values varies over time, and the distance d_{ij}^{TS} between pairs of time series i and j is calculated based on the correlation between these vectors. The formulae of Pearson correlation coefficients are given as:

$$\delta^{TS}(ij) = 1 - \frac{T(\sum_t x_{it}x_{jt}) - (\sum_t x_{it})(\sum_t x_{jt})}{\sqrt{(T(\sum_t x_{it}^2) - (\sum_t x_{it})^2)(T(\sum_t x_{jt}^2) - (\sum_t x_{jt})^2)}} \tag{4.2}$$

where $d_{ij}^{CF} = \delta^{TS}(ij)$; d_{ij}^{TS} are elements of the dissimilarity matrix $\mathbf{D_{TS}}$; t is the index of time $(t = 1,2,...,T)$, and T is the number of time steps; x_{it} and x_{jt} describe the values of time series i and j over time, respectively.

For the multicriteria clustering approach, we integrate the two information sources at the distance function level using a weighted-sum method on the standardized distance values. The standardization technique (transforming data into the range [0,1]) implemented for updating each element of the dissimilarity matrices is presented as follows:

$$d_{ij}^{CF} \leftarrow \frac{d_{ij}^{CF} - min(\mathbf{D^{CF}})}{max(\mathbf{D^{CF}}) - min(\mathbf{D^{CF}})} \tag{4.3}$$

$$d_{ij}^{TS} \leftarrow \frac{d_{ij}^{TS} - min(\mathbf{D^{TS}})}{max(\mathbf{D^{TS}}) - min(\mathbf{D^{TS}})} \tag{4.4}$$

Ultimately, the distance function of $d_{ij\omega}^{MC}$ through a weighted-sum method is then given as:

$$d_{ij\omega}^{MC} = (1-\omega) \times d_{ij}^{CF} + \omega \times d_{ij}^{TS} \tag{4.5}$$

where the relative weight ω ranges from 0 to 1 in steps of 0.10; A new dissimilarity matrix is formed based on the integration of two information sources at the distance function level $\mathbf{D}_{\omega}^{\mathbf{MC}} = (d_{ij\omega}^{MC})$.

4.4.2 Clustering algorithm

Here, we employ the Partition Around Medoids (PAM) clustering algorithm (Kaufman and Rousseeuw, 2009) to partition the time series into distinctive groupings. PAM clustering is a standard clustering approach based on medoids. More important, this algorithm has been reported to be suitable for the clustering of analogies in the forecasting context, as PAM clustering tends to produce partitions consisting of equally-sized clusters (see (Lu and Handl, 2015)). This property is considered advantageous in the application of forecasting. To minimize the negative impacts of converging to local optima, we repeat the clustering procedure 30 times and select the clustering result with the minimum sum of within-cluster dissimilarities.

4.4.3 Determination of the number of clusters

For dissimilarity matrix-based clustering methods, a major challenge related to model selection is the determination of partitions. One main step involved in either single-criterion or multi-criterion clustering procedures is to determine the number of clusters. Here, we employ a popular internal validation technique to facilitate an automatic determination of the number of clusters, namely the Silhouette Width measure (Kaufman and Rousseeuw, 2009). This Silhouette Width has been widely applied in the clustering field to score clustering solutions based on the internal data structure. This metric evaluates the quality of a clustering solution by considering both the cluster cohesion and separation. The Silhouette Width takes value in the range [-1,1], with a larger value representing a better solution. The formulae are presented as follows:

$$Sil(i) = \frac{b_i - c_i}{max(c_i, b_i)} \tag{4.6}$$

where c_i is the average distance between item i and all data items in the same partition; b_i describes the average distance between i and all data items in the closest another cluster. This is defined as the clustering solution, which returns the minimum b_i. The Silhouette value of the entire partition is then calculated as the mean Silhouette value of all data items.

This measure is adequate to assist in the estimation of the number of clusters. However, it still leaves a question unresolved where a multicriteria clustering approach can return a set of Pareto-optimal clustering solutions for the same number of clusters. These Pareto-optimal solutions are essential incomparable as an improvement in one criterion may degrade the performance of another. On account of this, a range of techniques are proposed in Section 4.5.2 to complement the determination of the partitions.

4.4.4 Forecasting algorithm

To demonstrate the model selection problem in the context of forecasting, we proceed our analysis by employing a well-established forecasting algorithm: the Cross-Sectional State Kalman Filter algorithm (C-MSKF: Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001). The employment of a forecasting algorithms aims to provide objective assessment for model selection methods using the same multicriteria clustering procedure.

This algorithm is designed to exploit information from analogies that can be objectively identified using multicriteria clustering approaches. Analogies are commonly present in the context of forecasting. For instance, a set of products may fall into a group due to the same sphere of influence, similar consumer preferences, or local trends. These time series are typically co-vary and are thus positively correlated over time.

By integrating information available from analogies, the C-MSKF method is a Bayesian pooling method that has been proposed to forecast short and volatile time series. The C-MSKF method has shown to be a powerful tool in tackling challenging forecasting scenarios, such as churn on a telecommunications network (Greis and Gilstein, 1991), infant mortality rates (Duncan, Gorr, and Szczypula, 2012) and tax revenue (Duncan, Gorr, and Szczypula, 1993). Analytically, The C-MSKF algorithm is an extension of the Multi-State Kalman Filter (MSKF: Harrison and Stevens, 1971) with the Conditionally Independent Hierarchical Model (CIHM: Kass and Steffey, 1989) using the DGS shrinkage formula (DGS's shrinkage: Duncan, Gorr, and Szczypula, 1993). The use of additional information extrinsic to the time series data improves the responsiveness to the changes caused by an external influence (Duncan, Gorr, and Szczypula, 1994; Duncan, Gorr, and Szczypula, 2001), *e.g.,* such as the action of a competitor. 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 to outliers. For reference, a full interpretation of the C-MSKF algorithm is available in the literature (Duncan, Gorr, and Szczypula, 1993) and full syntax implemented in Fortran language for the algorithm refers to Duncan, Gorr, and Szczypula, 2012.

4.5 Empirical evaluation

4.5.1 Simulated data

To demonstrate our ideas, we model two information sources (criteria) using simulated data: specifically, the time series patterns as well as the causal factors characterize the patterns observed.

The the first information is derived from time-based patterns. we use a couple of mathematical functions to describe the patterns of time series. Specifically, a linear, logarithmic and piece-wise linear function are applied to describe the trend changes as a function of time t. Ultimately, we aim to generate a set of time series that are correlated at an initial time point but later present different trend changes. We assume that the trend changes are caused by an external influence and shared across subsets of analogous time series. Principally, the linear function describes a time series that presents a stable increasing trend. The logarithmic model shows a time series with decreasing increasing rate in the trend. In essence, both functions do not capture sudden pattern changes. The piece-wise linear function can be interpreted as a time series showing a slope change from positive to negative due to an external influence occurring at time p. Specifically, $f_g(t)$ denotes the function used for simulating a time series and $g = 1, \ldots, 3$ indicates the choice made for using a linear, logarithmic and piece-wise linear function, respectively. The equations are given as follow:

$$f_1(t) = 0.8t + 2.8, \quad if\, 1 \leq t \leq q \tag{4.7}$$

$$f_2(t) = 4ln(t) + 2, \quad if\, 1 \leq t \leq q \tag{4.8}$$

$$f_3(t) = \begin{cases} 0.7t + 2.8, & if\, 1 \leq t \leq p \\ -0.9t + 25, & if\, p+1 \leq t \leq q \end{cases} \tag{4.9}$$

where q refers to the number of time points for a time series; p refers to the time of the trend change for the piece-wise linear function.

Based on the above functions, a group of analogous series are generated by adding normally-distributed noise at each time step. The normal distribution is considered in order to meet the assumption of Kalman Filters. Specifically, the noisy time series pattern X_{it} for time series i at time t, associated with generating model g, is obtained as follows:

$$X_{it} = \begin{cases} f_g(0) + N(f_g(t+1) - f_g(t), \sigma_{TS}^2), & if\, t = 1 \\ X_{i(t-1)} + N(f_g(t+1) - f_g(t), \sigma_{TS}^2), & if\, 1 < t \leq q-1 \end{cases} \tag{4.10}$$

where g is the choice of generating function. The notation $N(\mu_{TS}, \sigma_{TS}^2)$ describes a random variate drawn from a normal distribution with mean μ_{TS} and variance σ_{TS}^2. here σ_{TS}^2 is constant, 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 (4.10), 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 Fig.4.3, and it is evident that differentiation between these series is challenging for earlier time intervals.

Finally, all time series are standardized using the z-score method 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 behavior between the time series. 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 to derive suitable (informative but noisy) data for the causal factor. Specifically, the values of the causal factor for time

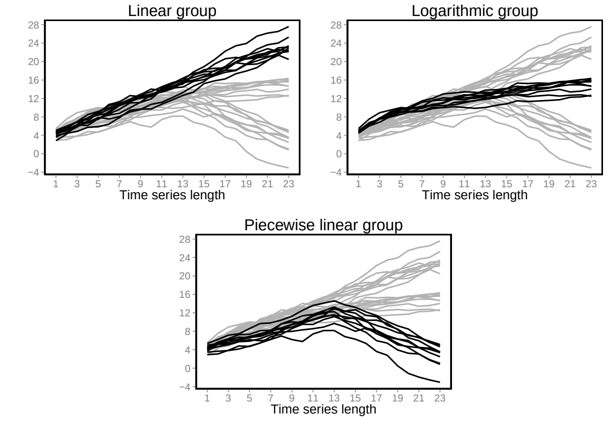


Figure 4.3: Illustration of simulated time series (raw data) generated from a linear, logarithmic, and piecewise linear function

series i is drawn from normal distributions $N(\mu_{CF}, \sigma^2_{CF})$, where μ_{CF} 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 adjusted the levels of σ_{CF} and σ_{TS} relative to each other (see Table 4.1). Specifically, σ_{CF} is fixed to 0.35 while σ_{TS} is increased from 0.35 to 1.15 in steps of 0.2.

Table 4.1: Standard deviation used to generate simulated causal variables and time series 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 below.Specifically, we fix the forecasting origin at $t = T$ throughout our analysis. The parameter T is chosen to allow for more than three observations after the trend change of the time series, thus meeting one of the key assumptions behind the C-MSKF algorithm. The parameter *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 causal factor and time series datasets).

Table 4.2: 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
Time of change	p=14

4.5.2 Compared model selection methods

In this section, we focus on detailing the model selection methods proposed in this article to pick up a single best clustering solution in the context of multicriteria clustering problems. We sub-divide the model selection methods into two categories: (i) The selection of the best partitioning based on internal data structure of the data, and this is assessed using internal validation index, the Silhouette Width (Sil); (ii) The selection of the best partitioning takes into account of the application context (forecasting accuracy). In the following, we present details of different approaches compared in the experiments. For contrasting purpose, we benchmark multicriteria clustering approaches on single-criterion clustering approaches and multicriteria clustering approaches with prior knowledge of model selection.

4.5.2.1 Benchmarks

In general, four models are used as benchmarks here. Specifically the single-criterion clustering of causal factor (CF), the single-criterion clustering of time series data (TS), and the multicriteria clustering on both information sources using a weighted-sum method (MC). For these three methods, the Silhouette Width measure is employed to determine the number of clusters. In addition to this, MC clustering additionally selects the weight interval using the smallest forecasting errors on the lead time period ($t = T+1,\ldots,H$). This assumes a *aprior* regarding the weight selection. Furthermore, we consider a situation, where both the optimal number of clusters (*K* $= 3$) and the weight interval (using the best performance on lead time period $t = T+1,\ldots,H$) are known, namely MC$_{ThreeMin}$ method.

4.5.2.2 Application-independent model selection approaches

- **Angle-based approaches, denoted by *MC*$_{Angles}$**

To score different clustering results, the Silhouette Width measure is employed by considering the specified *K* and weight interval ω. In a two-dimensional space, clustering solutions within the range considered are demonstrated in Fig. 4.4. This figure shows the Silhouette scores on clustering solutions, generated by multicriteria clustering approaches, that are projected to a single dissimilarity matrix derived from the CF (x-axis) and TS (y-axis) clustering, respectively. Specifically, the Silhouette Width measure determines a suitable number of clusters from a range of $k = 2,\ldots,6$ and ω takes value from 0 to 1 with an increment of 0.1. For example, partition 1 refers to a clustering solution that takes *K* $= 2$ and the $\omega = 0$. Clustering solutions plotted on the Pareto front are highlighted in black color. These partitionings are considered as Pareto-optimal solutions, as they show different trade-offs between the CF and TS information criteria. Further to select a single most promising solution out of the Pareto-optimal solutions.

The angle-based measure is applied to the pre-selected clustering candidates in the second step. Details regarding the angle-based method can be found in Section 4.3.

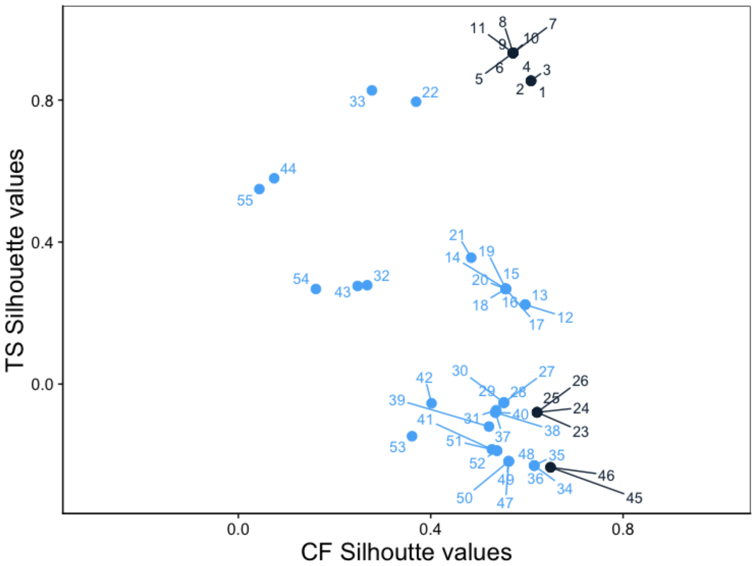


Figure 4.4: Illustration of Pareto-optimal clustering solutions on efficient frontier. $k = 2, \ldots, 6$ and ω takes value from 0 to 1 in steps of 0.1, and the clustering solutions are sequentially numbered from 1 to 55, correspondingly, for the purpose of illustration.

Here, similar to MC$_{ParetoHist}$ and MC$_{ParetoTest}$ methods (will be discussed later), MC$_{Angles}$ incorporates the concept of Pareto-optimality in the modelling process. As studied in optimization literature, one way to find good solutions is to find the Pareto optimal front (Baumgartner, Magele, and Renhart, 2004). By definition, Pareto-optimal solutions refer to those that cannot be improved in one objective at the expense of others. For stochastic clustering approaches, different choices of model parameters and random initialization might result in differing nondominated clustering solutions on Pareto-front. The resulting clustering solutions are considered Pareto-optimal in one run might become inferior in another run, and thus lead to different forecasting results. In contrast, clustering procedures (*e.g.,* Hierarchical clustering) that are deterministic may give rise to the same forecasting results in different run. Nevertheless, as stressed in Section 4.4.2, we implement PAM clustering in our experiments for the sake of equal-sized clustering, which is found to be more beneficial to the final forecasting performance.

- **Variates of quality scores, denoted by MC$_{MaxMax}$, MC$_{MaxMin}$, MC$_{MaxSum}$**

For this class of methods, we develop variations on the calculation of the quality scores using the Silhouette Width measure. Specifically, each clustering solution takes the number of clusters $k = 2, \ldots, N$ with ω from 0 to 1 in steps of 0.1. Partitionings using the two parameters are scored by Silhouette Width measure taking into account of the CF and TS information sources independently. To be specifical, the quality scores are computed on the dissimilarity matrices $\mathbf{D_{CF}}$ and $\mathbf{D_{TS}}$, respectively. Specifically, for each clustering solution, the maximum value measured on $\mathbf{D_{CF}}$ and $\mathbf{D_{TS}}$ is chosen first and proceed by taking the maximum Silhouette value across clustering solutions for MC$_{MaxMax}$. MC$_{MaxMin}$ takes the minimum quality score, Silhouette values, for each solution while MC$_{MaxSum}$ takes the sum of quality scores for each solution.

- **Clustering solutiones with the largest average Silhouette values, denoted by MC$_{SilSil}$**

MC$_{SilSil}$ describes a sequential procedure that determines the number of clusters using the largest mean Silhouette value in the first step and subsequently picks up a single best partitioning for the same number of groups with the largest mean Silhouette values across weight intervals, ω from 0 to 1 in steps of 0.1.

4.5.2.3 Application-dependent model selection approaches

For the second category of model selection methods, we choose the best weight interval using the best historical average forecasting performance. The questions raised by this strategy involves the choices of the number of data points used for weight selection; further question regarding whether the data points used in the weight selection should be included in the clustering. Therefore, two generic type of methods are proposed to answer the above questions.

To measure the forecasting results, the Mean Square Error (MSE) measure is applied throughout the paper to calculate the forecasting errors. The MSE is calculated as follows:

$$MSE = mean(e_t^2) \tag{4.11}$$

where t is the time step, $e_t = X_t - F_t$, X_t is the observation of the time series X at time t, and F_t is the respective forecast.

- **Clustering solutions with the best average historical forecasting performance, denoted by MC$_{SilHist}$, MC$_{SilTest}$**

MC$_{SilHist}$ determines the number of clusters K based on the largest mean Silhouette value in the first step. For the same number of clusters, a single best partitioning (ω^*) producing the best average historical forecasting results is chosen for the prediction of future data points in the weight selection step. More specifically, the forecasting origin $t = T$ is used to support model selection in this part of the analysis, and observations on time steps $t \leq T$ are used during the clustering step.

- **Clustering solutions with the best historical forecasting results, denoted by MC$_{SilTest}$**

MC$_{SilTest}$ approach again uses the Silhouette Width measure to determine the number of clusters in the first step. Subsequently, the average historical forecasting performance at time step $t = T$ is used for weight selection. Different from MC$_{SilHist}$, observations on $t < T$ period are used for clustering. The data points used for weight selection are excluded during the clustering stage.

- **Selecting the partitions based on the best average historical forecasting performance out of sets of Pareto-optimal solutions, MC$_{ParetoHist}$, MC$_{ParetoTest}$**

The idea of MC$_{ParetoHist}$ methods is to choose the best partitioning out of a set of Pareto-optimal clustering solutions based on average historical forecasting results. The Silhouette Width measures are employed by considering the specified K and weight interval ω. Specifically, the Silhouette Width measure determines a suitable number of clusters from a range of k considered and ω takes value from 0 to 1 with an increment of 0.1. Pareto-optimal clustering solutions are obtained in the first step, and the same procedure is applied to get the best performing clustering solution that produces the best average historical forecasting results. The second step is the same as described in MC$_{SilHist}$.

4.5.3 Performance evaluation

To measure the bias of various forecasting models, Mean Error (ME) is used to measure the forecasting results.

$$ME = mean(X_t - F_t) \tag{4.12}$$

where all variables retain the same meaning as Equation 4.11.

To measure the forecasting accuracy, two well-known accuracy measures are applied, including the Mean Absolute Scaled Error (MASE: Hyndman, 2006) and the Symmetric Mean Absolute Percentage Error (sMAPE: Bergmeir, Hyndman, and Benítez, 2016), respectively.

$$MASE = mean(\left|\frac{e_t}{\frac{1}{T-1}\sum_{i=2}^{T}|X_i - X_{i-1}|}\right|) \tag{4.13}$$

$$sMAPE = mean(200\frac{|e_t|}{|X_t| + |F_t|}) \tag{4.14}$$

where T refers to the forecasting origin and the rest variables retain the same meaning as Equation 4.11.

For the purpose of assessing the performance of clustering results, we use external criterion: the Adjusted Rand Index (ARI: Hubert and Arabie, 1985) to measure the agreement between the produced clustering results and respective "ground truth", as defined by the generating function for each time series. The ARI takes the largest value of 1 and an expected smallest value of 0, with larger values representing a better consistency between the ground truth and the clusters generated.

Based on the $L \times K$ contingency table, the ARI defines two clusters (of the same data) with L and K clusters respectively. The Adjusted Rand Index between the two clusters is computed as follows:

$$ARI = \frac{\sum\limits_{l,m}\binom{N_{lm}}{2} - [\sum\limits_{l}\binom{N_{l.}}{2} \cdot \sum\limits_{k}\binom{N_{.m}}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum\limits_{l}\binom{N_{l.}}{2} + \sum\limits_{m}\binom{N_{.m}}{2}] - [\sum\limits_{l}\binom{N_{l\cdot}}{2} \cdot \sum\limits_{m}\binom{N_{.m}}{2}]/\binom{N}{2}} \tag{4.15}$$

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

4.6 Results

4.6.1 Demonstration of the evaluation of the clustering performance in the context of application

In Fig. 4.6, we present figures derived from the experiments that illustrates the clustering quality measured by agreement between clusters and "ground truth" using the average ARI. Some clustering solutions (*e.g.,* Partitions 1, 23) show significant drops in clustering quality, although are identified as Pareto-optimal clustering solutions in Fig. 4.5. Considering the data structure alone, solutions generated from CF clustering might be highly scored by the Silhouette Width measure. However, CF information might contribute much less than TS information source with respect to the final forecasting results when implementing MC clustering. This figure indicates that a highly scored or Pareto-optimal clustering solutions do not necessarily give rise to satisfactory forecasting results since the statistical measure used can be insufficient and less informative in a forecasting context. Furthermore, we also present the illustrating of a model selection based on the clustering quality alone. Fig.4.7 illustrates the performance of clustering solutions that maximize the average Silhouette Width values at the stage of the selection of the number of clusters and weight selection, basically MC$_{SilSil}$. Comparing Fig.4.7 to Fig. 4.5, clustering solution 45 and 46 are identified as the best performing solutions in Fig.4.7, but translates to poorer forecasting results than those of partitionings 6-11.

In summary, Fig. 4.5, Fig. 4.6 and Fig. 4.7 illustrate the idea that the performance of clustering solutions are associated with the performance of an application. However, it should be best assessed in the context of the application.

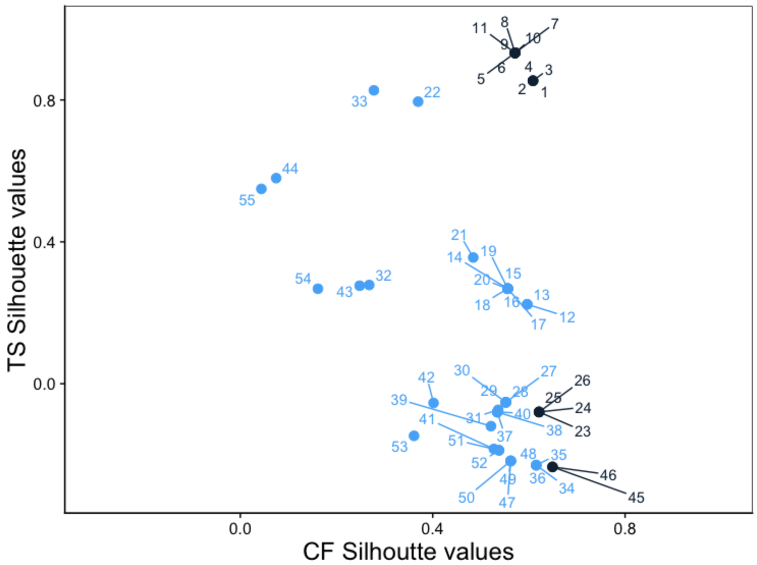


Figure 4.5: Partitionings integrate the information, using a weighted-sum method, derived from causal factors and time series patterns. Each partitioning is scored based on the Silhouette values projected on the dissimilarity matrix of an individual information source.

4.6.2 Assessing the historical forecasting performance of C-MSKF for model selection

Explicitly, there are two questions to be addressed for picking the weight using historical forecasting results: (i) How many data points should be used for model selection; (ii) Should the time steps used in the model selection step be included in the clustering procedure? Here, we address these two questions by varying e number of data points and change the strategy of utilizing these data points in the clustering procedure.

In summary, four model selection methods are compared and these are MC$_{ParetoHist}$, MC$_{ParetoTest}$, MC$_{SilHist}$ and MC$_{SilTest}$ approaches. Tables 4.3, 4.4, 4.5, 4.6 contrast forecasting accuracy of C-MSKF on model selection approaches that make use of historical forecasting accuracy in the model selection step. The forecasting accuracy is calculated by taking the average, minimum, maximum, median value across 30 sets of replicates, 6 time series lengths, and 6 forecasting horizons. Here, we compare two

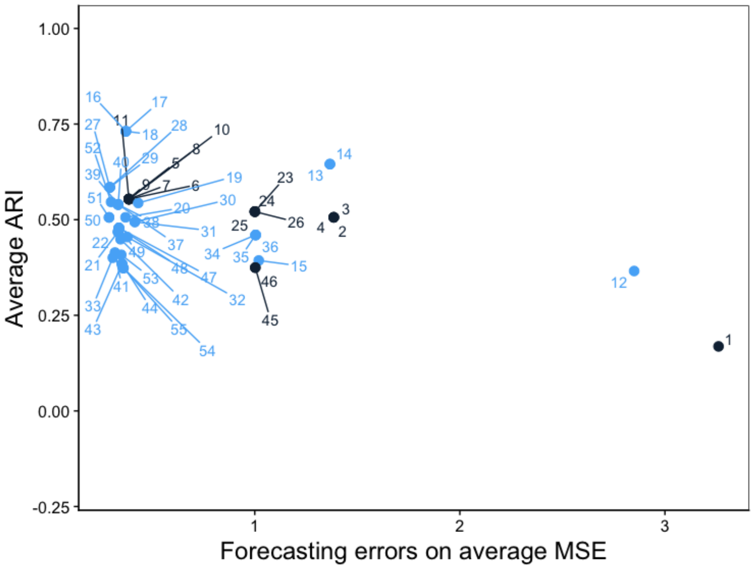


Figure 4.6: Partitionings are compared to the “ground truth” based on the average ARI. These partitionings are further scored based on the forecasting results measured by MSE.

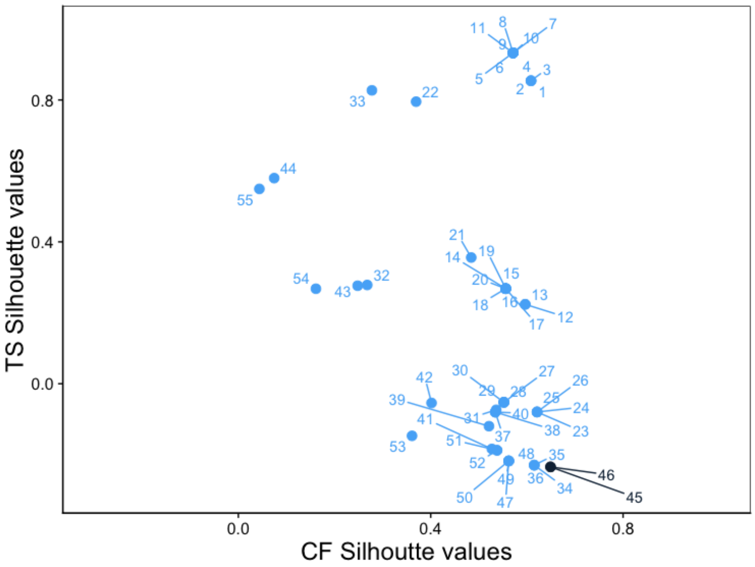


Figure 4.7: Partitionings are determined using the largest mean Silhouette values for estimating the number of clusters as well as the largest mean Silhouette values for the subsequent weight selection.

strategies for determining the number of data points in the clustering step. This aims to answer the second question proposed in this section. Considering the MC$_{SilHist}$ and MC$_{ParetoHist}$ approach, we systematically add a historical observation one at a time into the model selection as e (the number of successive time steps) increases from 1 ($t = T$) to 3 ($t = T-2, T-1, T$). Here, observations at time $t \leq T$ are consistently used during the clustering regardless of the change in e. While for MC$_{SilTest}$ and MC$_{ParetoTest}$, we vary e, the number of data points, used for model selection and where e varies from 1 to 3 ($t = T$) to 3 ($t = T-2, T-1, T$). Correspondingly, the data points used for model selection vary from $t < T$, $t < T-1$ to $t < T-2$,correspondingly.

As shown in Tables 4.3, 4.4, 4.5, 4.6, ME measure shows limited capability in differentiating the C-MSKF's accuracy across model selection methods compared. Based on average ME and median ME, we can see that negative results are reported in scenarios 1 and 3, while positive results are reported in scenarios 2, 4 and 5. In brief, a slightly higher proportion of the forecasting results shows positive bias in average or median value of the accuracy measures considered. For minimum ME and maximum ME, forecasts are found to be strongly negatively biased and positively biased from scenario 1 to 5. Overall, MC$SilHist$ methods show the consistenly best forecasting accuracy across different scenarios 1-5 and this is demonstrated in Tables 4.3, 4.4 and 4.6. In Table 4.5, Max ME shows unclear tendency to the best performing models whilst Max MSE presents a consistent better forecasting accuracy on the MC$_{ParetoHist}$ method. However, Max MASE and sMAPE present the first and second best performing results on methods of MC$_{SilHist}$.

Regarding the clustering accuracy, the MC$_{ParetoHist}$ methods demonstrate the best clustering accuracy in Tables 4.3, 4.4, 4.6. Table 4.5 presents ARI value of 1 across majority of scenarios and methods (50 out of 60). This implies that maximum ARI is not able to differentiate clustering performance among the methods discussed here.

In summary, based on ME methods, there is no clear conclusion can be drawn regarding the tendency of over- or under-forecast performance of MC$_{SilHist}$, MC$_{SilTest}$, MC$_{PartoHist}$ and MC$_{ParetoTest}$ methods across all scenarios. Min ME intends to yield negative bias, as the minimum forecasts across 30 replicates, 6 time series lengths, 6 forecasting horizons are consistently smaller than the actual observations. On the contrary, max ME presents consistently positive bias across all scenarios 1 to 5. In terms of the forecasting accuracy, the MC$_{SilHist}$ method helps to produce the best forecasting accuracy on C-MSKF's results. The overall satisfactory performance of this method indicates that the Silhouette Width measure performs effectively for the determination of the number of clusters. It also implies that the weight selection (clustering) is best assessed in the forecasting context. Our results confirm the conclusion of Von Luxburg (2010) that clustering is best evaluated in the overall application context. In consideration of clustering performance, the MC$_{ParetoHist}$ method shows the best performance over the contestant model selection methods here, and this is followed by MC$_{ParetoTest}$ approach. Generally speaking, model selection methods using the idea of Pareto-optimality generate better clustering results. This might be because that inferior clustering solutions were excluded in the weight selection step of the model selection. Pareto-optimal clustering solutions show significantly better clustering results that might have positive impacts on the following forecasting stage.

4.6.3 Performance comparison of model selection methods across different noise levels

In this section, we focus on contrasting the C-MSKF's accuracy based on different clustering strategies. Note that forecasting accuracy measured at time point $T = 17$ is used for model selection, and observations in the period $T \leq 17$ are included in the clustering procedure.

As shown in Tables 4.7, 4.8, 4.9, and 4.10, the ME measure is inadequate to differentiate the forecasting performance among the contestant approaches. In each scenario, almost all the compared approaches show the same forecasting results. However, considering the ME measure's interpretability in bias performance of forecasting methods, all compared models measured by minimum ME (see Table 4.8) report negative bias from scenario 1 to 5. The bias is positive for maximum ME for all contestant methods in these five scenarios. Average ME and median ME yield positive bias towards these contestant methods in scenarios 2, 4 and 5, while give rise to negative bias towards in scenarios 1 and 3.

Considering the forecasting accuracy of the contestant models, it can be observed that MASE, MSE and sMAPE measures (see Tables 4.7, 4.8, 4.9, and 4.10) in general rank the MC method as the best clustering approach that can assist in achieving the best forecasting accuracy of C-MSKF methods. Generally, the MC$_{ThreeMin}$ approach can be ranked as the second best performing model using the MASE, MSE and sMAPE metrics. However, this method makes use of prior knowledge and hindsight: the prior knowledge regarding the employment of underlying mathematical models, and the hindsight related to the weight selection, *i.e.,* the weight interval gives rise to the smallest forecasting errors is selected.

Finally, we can conclude that MC$_{ParetoHist}$ methods produce the best clustering results in scenarios where the clustering accuracy of the contestant methods is assessed using the average, minimum, median values of ARI. As demonstrated in Table 4.9, MC-based methods report maximum ARI of 1 from scenario 1 to 5. CF and TS methods show relatively low clustering accuracy in most cases. This indicates that ,considering the maximum clustering quality, the single-criterion clustering approaches show inferior performance to the multicriteria clustering approach.

In summary, MC$_{ParetoHist}$ methods are promising for the identification of good clustering quality of analogies using the concept of Pareto-optimality. However, the single clustering solution picked in the model selection step does not lead to the best forecasting results of C-MSKF methods. In stead, MC methods generally perform the best across scenarios and accuracy measures. Realistically, MC$_{SilHist}$ methods using objective model selection techniques perform promising with respect to forecasting accuracy of C-MSKF methods.

Furthermore, we break up the forecasting results by forecasting horizons from 1 to 6 using four different accuracy measures, each of these is summarized using four means of calculation, including the average, maximum, minimum, and median across 30 replicates, 6 time series lengths, and 6 forecasting horizons. Details refer to the following Tables presented in this section. Here, we contrast the performance of model selection methods that suitable for a real-world setting where there is no *a prior* or hindsight concerning the number of clusters or the best weight interval. We further benchmark these methods on forecasting performance of the MC clustering method. In summary, MC method consistently performs the best across six forecasting horizons and five noise levels as shown by MASE, MSE and sMAPE metrics (including the average, minimum, maximum values across 30 replicates, 6 time series lengths and 6 forecasting horizons). Generally, MC$_{SilHist}$ is the second best performing model in terms of forecasting accuracy of the C-MSKF method. This is particularly evident in shorter forecasting horizons when higher noise levels presented in scenario four and five on MSE measure. Since the MC method performs consistently the best across forecasting horizons based on (average, minimum, maximum, and median) MASE, MSE and sMAPE measures, this indicates that MC$_{SilHist}$ is sensitive to the increased noise in the time series data. By comparing the performance of application-based methods, MC$_{ParetoHist}$ methods show inferior forecasting performance to MC$_{SilHist}$ methods across the six forecasting horizons and the three accuracy measures.

Regarding the ME method, as shown in Tables 4.12, 4.16, 4.20, and 4.24, demonstrate almost the same performance across the compared methods at each forecasting horizon. In line with the experiments in the last section, ME measure does not highlight any MC-based methods that consistently outperform the rest with respect to the C-MSKF's forecasting accuracy. In terms of the bias performance, scenarios 1 and 3 give rise to negative bias while scenarios 2, 4 and 5 produce positive bias. It might be concluded that the bias of these MC-based methods are data-dependent as in each scenario different data sets are used, and all methods considered here show consistency towards the forecasting bias as they yield the same sign of the forecasting errors.

4.7 Conclusions

This paper investigates the challenge of model selection for a multicriteria clustering approach in the context of forecasting. We have proposed and adapted different techniques to support the automatic selection of a single best partitioning in multicriteria clustering problems.

In summary, MC$_{SilHist}$ is shown to be a promising method for the selection of a single best partitioning in the forecasting context. For the weight selection, it appears to be preferable to use a small number of data points during the model selection, and to include these data points during the clustering stage of the process. The fact that MC clustering methods consistently perform better than MC$_{SilHist}$ across forecasting horizons and noise levels (see Table **??**). This highlights the remaining limitations of the weight selection scheme. In particular, we observe a marked decrease in performance when the noise of the time series data increases. Our results also illustrate that the best clustering quality of partitionings does not necessarily give rise to the best forecasting results (see Table **??**). These findings provide additional empirical evidence to support the view that the quality of a clustering solution should be best assessed in the context of an application, as discussed in Guyon, Von Luxburg, and Williamson, 2009.

In this article, we propose and compare a range of model selection methods adapted from the existing techniques in the clustering literature or borrowing ideas from the multi-objective optimization domain. Most of these are application independent, and can be seen as generic contributions to the development of automatic model selection approaches for multicriteria clustering.

A key limitation of our current work is the evaluation on simulated data alone, and further work needs to consider real-world applications for ratifying our ideas. Also, future work could explore model selection techniques that are computationally more expensive. *E.g.,* one potential model selection method would construct models for all possible numbers of clusters, and then pick a preferred partitioning (across weight levels and numbers of clusters) based on the best average historical prediction performance alone. This approach is likely to deliver accurate predictions, but is computationally extremely expensive, and was therefore not considered in our current work.

Table 4.3: Comparison of model selection methods, which make use of average historical forecasting in the weight selection step, on C-MSKF's *average* forecasting accuracy by varying the number of observations e used in the clustering step. Final results are obtained by taking the *average* across 30 replicates, 6 time series lengths, and 6 forecasting horizons. e indicates the number of observations included (MC$_{SilHist}$, MC$_{ParetoHist}$) or excluded (MC$_{SilTest}$, MC$_{ParetoTest}$) in the clustering step.

		MC$_{SilHist}$			MC$_{SilTest}$			MC$_{ParetoHist}$			MC$_{ParetoTest}$		
	Scenarios	e =1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3
Average ARI	S_1	0.66	0.71	0.72	***0.78***	**0.80**	0.77	**0.80**	**0.80**	**0.80**	**0.80**	**0.80**	***0.78***
	S_2	0.57	0.56	0.57	0.66	0.61	0.59	**0.74**	**0.74**	**0.74**	***0.72***	0.70	0.70
	S_3	0.47	0.46	0.48	0.54	0.57	0.54	**0.74**	***0.73***	***0.73***	0.71	0.68	0.66
	S_4	0.41	0.42	0.41	0.44	0.52	0.50	**0.69**	***0.68***	**0.69**	0.67	0.66	0.67
	S_5	0.41	0.36	0.40	0.42	0.48	0.46	**0.74**	**0.74**	**0.74**	0.70	***0.72***	0.71
Average MASE	S_1	**0.94**	***0.95***	0.96	***0.95***	0.97	1.04	1.13	1.14	1.14	1.15	1.17	1.28
	S_2	**1.5**	***1.51***	***1.51***	1.54	1.63	1.7	1.73	1.72	1.72	1.74	1.79	1.84
	S_3	**1.82**	***1.85***	1.92	1.88	1.94	2.01	2.01	2.01	2.01	2.04	2.07	2.1
	S_4	**1.93**	***1.94***	2	2.02	2.11	2.17	2.14	2.4	2.39	2.15	2.21	2.2
	S_5	**2.06**	2.11	2.12	***2.08***	2.19	2.22	2.13	2.12	2.11	2.14	2.16	2.19
Average MSE	S_1	**0.16**	***0.17***	0.18	***0.17***	0.18	0.24	0.31	0.31	0.31	0.34	0.36	0.46
	S_2	**0.59**	***0.6***	***0.6***	0.62	0.73	0.82	0.78	0.77	0.77	0.8	0.86	0.91
	S_3	**1.02**	***1.07***	***1.07***	***1.07***	1.16	1.17	1.18	1.17	1.17	1.23	1.26	1.29
	S_4	**1.47**	***1.51***	1.57	1.54	1.67	1.7	1.58	1.67	1.66	1.62	1.69	1.7
	S_5	1.8	1.91	1.82	1.75	1.93	1.9	***1.65***	***1.65***	**1.64**	1.71	1.74	1.78
Average sMAPE (%)	S_1	**21.29**	***21.36***	21.77	21.68	22.16	23.91	25.78	25.96	25.96	26.26	27.25	29.97
	S_2	***38.08***	**37.86**	38.4	38.23	41.09	43.46	43.38	43.26	43.22	44.17	46.05	47.63
	S_3	**55.05**	***55.91***	59.01	57.4	60.13	62.61	60.73	60.58	60.52	62.79	65.27	66.91
	S_4	**60.93**	***61.28***	63.05	64.1	66.71	69.63	67.34	69.37	69.17	68.74	70.66	70.93
	S_5	**68.04**	70.04	69.98	***68.64***	73.94	74.85	71.09	70.85	70.76	72.77	73.38	74.39
Average ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	-**0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	**0.02**	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***
	S_3	***-0.02***	***-0.02***	**-0.03**	**-0.03**	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***
	S_4	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***	**0.04**	**0.04**	***0.08***	***0.08***	***0.08***
	S_5	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**

Table 4.4: Comparison of model selection methods, which make use of average historical forecasting in the weight selection step, on C-MSKF's *minimum* forecasting accuracy by varying the number of observations e used in the clustering step. Final results are obtained by taking the *minimum* value across 30 replicates, 6 time series lengths, and 6 forecasting horizons. e indicates the number of observations included (MC$_{SilHist}$, MC$_{ParetoHist}$) or excluded (MC$_{SilTest}$, MC$_{ParetoTest}$) in the clustering step.

		MC$_{SilHist}$			MC$_{SilTest}$			MC$_{ParetoHist}$			MC$_{ParetoTest}$		
	Scenarios	e =1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3
Min ARI	S_1	0.33	0.33	0.33	0.33	0.32	0.17	**0.42**	**0.42**	**0.42**	***0.39***	**0.42**	0.35
	S_2	0.30	0.30	0.30	0.28	0.21	0.17	**0.43**	***0.36***	***0.36***	0.35	0.21	0.34
	S_3	0.12	0.12	0.17	0.16	0.09	-0.02	**0.43**	0.25	0.25	***0.26***	***0.26***	***0.26***
	S_4	0.07	0.07	0.07	0.03	0.11	0.01	**0.37**	***0.26***	***0.26***	0.14	**0.37**	**0.37**
	S_5	0.02	0.04	0.02	0.07	-0.06	-0.03	**0.48**	***0.28***	***0.28***	0.21	0.14	0.01
Min MASE	S_1	**0.51**	**0.51**	**0.51**	***0.54***	0.55	**0.51**	0.58	0.56	0.56	0.58	0.58	0.58
	S_2	0.78	0.78	0.78	***0.75***	**0.71**	0.84	0.85	0.85	0.85	0.86	0.85	0.85
	S_3	**0.79**	***0.81***	0.89	0.98	1.03	1.04	1.14	1.05	1.14	1.03	1.05	1.21
	S_4	**0.92**	***0.95***	***0.95***	0.97	1.00	1.09	1.08	1.15	1.15	1.07	1.07	1.11
	S_5	**0.99**	**0.99**	**0.99**	***1.08***	1.27	1.28	1.36	1.36	1.36	1.26	1.36	1.36
Min MSE	S_1	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	***0.05***	***0.05***	***0.05***	**0.03**	**0.03**	**0.03**
	S_2	**0.07**	**0.07**	**0.07**	0.09	0.09	***0.08***	0.1	0.1	0.1	**0.07**	0.1	0.1
	S_3	**0.11**	0.13	0.14	***0.12***	0.13	0.15	0.17	0.15	0.17	0.17	0.15	0.28
	S_4	**0.14**	***0.16***	**0.14**	0.18	0.20	0.2	0.2	0.35	0.35	0.2	0.2	0.23
	S_5	**0.26**	**0.26**	**0.26**	***0.28***	0.30	0.33	0.33	0.33	0.33	0.29	0.38	0.38
Min sMAPE (%)	S_1	**10.30**	**10.30**	**10.30**	***12.06***	14.22	14.17	14.17	14.17	14.17	14.17	14.17	14.17
	S_2	21.86	21.86	21.86	***20.86***	**20.40**	21.97	22.07	22.07	22.07	24.37	24.37	24.37
	S_3	**26.79**	**26.79**	**26.79**	30.36	***29.30***	36.55	38.21	38.21	38.21	30.37	38.21	38.05
	S_4	***34.88***	***34.88***	36.71	**34.81**	44.24	45.18	47.05	38.93	38.93	39.64	47.16	48.71
	S_5	**40.56**	**40.56**	**40.56**	44.20	***43.36***	***43.36***	44.2	44.2	44.2	46.68	44.7	44.2
Min ME	S_1	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**
	S_2	***-0.57***	**-0.58**	**-0.58**	***-0.57***	**-0.58**	**-0.58**	***-0.57***	***-0.57***	***-0.57***	***-0.57***	***-0.57***	***-0.57***
	S_3	-0.54	**-0.56**	-0.54	-0.54	***-0.55***	-0.54	-0.54	-0.54	-0.54	-0.54	-0.54	-0.54
	S_4	**-0.74**	**-0.74**	**-0.74**	**-0.74**	**-0.74**	**-0.74**	**-0.74**	***-0.72***	***-0.72***	**-0.74**	**-0.74**	**-0.74**
	S_5	**-0.44**	**-0.44**	**-0.44**	***-0.43***	***-0.43***	-0.41	***-0.43***	***-0.43***	***-0.43***	***-0.43***	-0.42	-0.42

Table 4.5: Comparison of model selection methods, which make use of average historical forecasting in the weight selection step, on C-MSKF's *maximum* forecasting accuracy by varying the number of observations e used in the clustering step. Final results are obtained by taking the *maximum* value across 30 replicates, 6 time series lengths, and 6 forecasting horizons. e indicates the number of observations included (MC$_{SilHist}$, MC$_{ParetoHist}$) or excluded (MC$_{SilTest}$, MC$_{ParetoTest}$) in the clustering step.

		MC$_{SilHist}$			MC$_{SilTest}$			MC$_{ParetoHist}$			MC$_{ParetoTest}$		
	Scenarios	e =1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3
Max ARI	S_1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_3	**1.00**	***0.95***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_4	0.76	***0.90***	0.82	***0.90***	***0.90***	***0.90***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_5	**1.00**	0.90	**1.00**	**1.00**	0.90	***0.95***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Max MASE	S_1	1.81	1.81	**1.72**	1.78	***1.75***	2.72	2.26	2.26	2.26	2.46	2.54	2.79
	S_2	**2.66**	***2.67***	**2.66**	3.3	3.17	3.22	3.56	3.52	3.52	3.56	3.87	3.87
	S_3	**3.29**	**3.29**	**3.29**	3.56	3.45	4.11	***3.31***	***3.31***	***3.31***	***3.31***	***3.31***	***3.31***
	S_4	**3.35**	***3.51***	3.58	3.82	3.6	3.74	3.9	3.94	3.89	3.89	3.93	3.93
	S_5	3.87	4.7	3.87	3.69	3.77	4.3	**3.46**	**3.46**	**3.46**	**3.46**	***3.59***	4.35
Max MSE	S_1	**0.63**	0.78	1.7	0.78	***0.74***	2.21	1.7	1.7	1.7	1.7	2.04	2.58
	S_2	**5.19**	***5.31***	4.14	**5.19**	5.4	6.7	**5.19**	**5.19**	**5.19**	**5.19**	**5.19**	5.84
	S_3	10.28	11.33	4.69	10.98	9.99	**3.96**	4.39	4.89	4.89	***4.37***	***4.37***	4.68
	S_4	7.86	7.86	6.97	6.96	8.46	7.74	***5.68***	**5.25**	**5.25**	5.7	6.82	6.64
	S_5	10.96	13.47	12.07	9.23	12.3	10	**7.23**	**7.23**	**7.23**	**7.23**	***7.52***	7.23
Max sMAPE (%)	S_1	**43.24**	**43.24**	**43.24**	**43.24**	**43.24**	***49.6***	64.44	64.44	64.44	64.44	51.25	96.1
	S_2	**69.3**	**69.3**	**69.3**	***75.82***	78.92	85.06	78.94	78.94	78.94	82.49	106.33	110.64
	S_3	***97.56***	**92.76**	108.4	101.87	107.34	110.57	108.4	102.95	102.95	108.4	120.47	120.47
	S_4	**99.61**	***102.75***	106.96	111.22	99.44	131.17	123.31	107.69	107.69	118.59	128.45	131.22
	S_5	120.09	122.79	119.09	116.01	121.87	144.32	***114.8***	***114.8***	***114.8***	***114.8***	***114.8***	**113.87**
Max ME	S_1	***0.27***	***0.27***	0.28	**0.26**	***0.27***	**0.26**	***0.27***	***0.27***	***0.27***	***0.27***	**0.26**	***0.27***
	S_2	***0.48***	***0.48***	***0.48***	***0.48***	***0.48***	***0.48***	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**	**0.47**
	S_3	***0.42***	***0.42***	***0.42***	**0.41**	0.43	***0.42***	0.43	0.43	0.43	0.43	0.43	0.43
	S_4	***0.77***	***0.77***	***0.77***	***0.77***	***0.77***	**0.76**	***0.77***	0.81	0.81	***0.77***	***0.77***	***0.77***
	S_5	***0.68***	0.7	***0.68***	0.7	**0.66**	0.7	0.69	0.69	0.69	0.69	0.69	0.69

Table 4.6: Comparison of model selection methods, which make use of average historical forecasting in the weight selection step, on C-MSKF's *median* forecasting accuracy by varying the number of observations e used in the clustering step. Final results are obtained by taking the *median* value across 30 replicates, 6 time series lengths, and 6 forecasting horizons. e indicates the number of observations included (MC$_{SilHist}$, MC$_{ParetoHist}$) or excluded (MC$_{SilTest}$, MC$_{ParetoTest}$) in the clustering step.

		MC$_{SilHist}$			MC$_{SilTest}$			MC$_{ParetoHist}$			MC$_{ParetoTest}$		
	Scenarios	e =1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3	e=1	e=2	e=3
Median ARI	S_1	0.73	0.77	**0.81**	***0.80***	**0.81**	**0.81**	***0.80***	***0.80***	***0.80***	***0.80***	**0.81**	***0.80***
	S_2	0.54	0.54	0.55	0.62	0.57	0.59	**0.73**	**0.73**	**0.73**	**0.73**	0.66	***0.68***
	S_3	0.40	0.40	0.42	0.45	0.54	0.54	0.77	0.73	**0.80**	***0.79***	0.72	0.66
	S_4	0.40	0.39	0.38	0.44	0.47	0.49	0.70	***0.71***	**0.72**	0.70	0.67	0.67
	S_5	0.43	0.36	0.40	0.46	0.47	***0.54***	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**
Median MASE	S_1	**0.92**	**0.92**	***0.93***	***0.93***	0.95	1.00	1.09	1.09	1.09	1.10	1.11	1.21
	S_2	**1.46**	***1.48***	1.49	1.51	1.58	1.64	1.67	1.67	1.66	1.65	1.71	1.75
	S_3	**1.79**	***1.81***	1.89	1.83	1.90	1.97	1.97	1.97	1.96	2.00	2.03	2.05
	S_4	**1.89**	***1.90***	1.97	1.99	2.08	2.12	2.09	2.40	2.39	2.10	2.15	2.15
	S_5	**2.00**	2.06	2.08	***2.04***	2.14	2.18	2.10	2.09	2.09	2.11	2.12	2.14
Median MSE	S_1	**0.14**	**0.14**	***0.15***	***0.15***	0.16	0.18	0.24	0.24	0.24	0.25	0.26	0.35
	S_2	**0.46**	**0.46**	**0.46**	***0.49***	0.55	0.61	0.63	0.62	0.62	0.62	0.68	0.74
	S_3	**0.77**	***0.80***	0.88	0.82	0.92	1.02	1.01	1.00	1.00	1.07	1.10	1.12
	S_4	**1.18**	***1.20***	1.25	1.26	1.36	1.41	1.38	1.52	1.51	1.42	1.46	1.49
	S_5	1.44	1.52	1.52	1.44	1.61	1.57	1.42	***1.41***	**1.40**	1.46	1.47	1.49
Median sMAPE (%)	S_1	**20.96**	21.07	21.37	***21.06***	21.46	22.76	25.19	25.22	25.22	25.86	26.12	27.62
	S_2	**36.22**	***36.46***	36.80	36.95	39.19	41.59	40.27	40.52	40.47	40.50	42.66	43.98
	S_3	**53.05**	***54.00***	56.87	56.55	58.46	60.33	57.74	57.72	57.69	60.72	62.20	65.10
	S_4	**59.77**	***61.14***	62.28	63.09	64.65	68.59	64.45	68.98	68.77	66.66	68.92	68.81
	S_5	**64.76**	68.61	***66.57***	67.03	71.61	72.54	68.68	67.73	67.70	69.89	71.47	71.75
Median ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
	S_3	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**
	S_4	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***	**0.04**	**0.04**	***0.07***	***0.07***	***0.07***
	S_5	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**

Table 4.7: Summary of contestant methods for different noise levels of the time series data in the simulated data. Comparison of *average* forecasting accuracy on 6-step ahead forecast using CF, TS and MC-based clustering approaches. $t = 17$ is used to as the model selection period to support the weight selection process on MC$_{SilHist}$, MC$_{ParetoHist}$ methods, and $t \leq 17$ are used during the clustering stage.

	Scenarios	CF	TS	MC	MC$_{ThreeMin}$	MC$_{Angles}$	MC$_{MaxMax}$	MC$_{MaxMin}$	MC$_{MaxSum}$	MC$_{ParetoHist}$	MC$_{SilHist}$	MC$_{SilSil}$
Average MASE	S_1	1.36	0.99	**0.91**	***0.93***	1.12	1.06	1.01	0.98	1.13	0.94	1.06
	S_2	1.75	1.59	**1.42**	1.51	1.69	1.67	1.53	1.62	1.73	***1.50***	1.69
	S_3	2.05	2.05	**1.74**	1.86	1.95	2.09	1.85	1.93	2.01	***1.82***	2.08
	S_4	2.15	2.12	**1.86**	2.1	2.08	2.15	2.01	2.09	2.14	***1.93***	2.15
	S_5	2.16	2.29	**1.96**	2.08	2.12	2.18	2.1	2.12	2.13	***2.06***	2.17
Average MSE	S_1	0.51	***0.17***	**0.16**	**0.16**	0.32	0.24	0.21	0.19	0.31	**0.16**	0.23
	S_2	0.83	0.68	**0.47**	***0.49***	0.74	0.76	0.60	0.69	0.78	0.59	0.77
	S_3	1.18	1.25	**0.80**	***0.88***	1.10	1.23	1.03	1.08	1.18	1.02	1.22
	S_4	1.57	1.80	**1.16**	***1.22***	1.55	1.59	1.52	1.52	1.58	1.47	1.59
	S_5	1.67	2.24	**1.37**	***1.45***	1.74	1.70	1.79	1.68	1.65	1.80	1.68
Average sMAPE(%)	S_1	34.38	21.89	***20.08***	**21.06**	26.31	24.31	23.26	22.17	25.78	21.29	24.06
	S_2	46.61	39.09	**34.23**	***35.25***	43.22	42.27	37.31	40.43	43.38	38.08	42.95
	S_3	62.79	58.75	**49.77**	***51.24***	59.07	62.69	57.18	59.12	60.73	55.05	63.03
	S_4	68.76	62.57	**55.09**	***56.06***	64.84	68.14	64.18	65.86	67.34	60.93	68.19
	S_5	71.40	72.75	**61.09**	***62.57***	70.81	71.89	71.81	71.17	71.09	68.04	71.84
Average ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	***0.03***	**0.02**	**0.02**	***0.03***	***0.03***	**0.02**	***0.03***	***0.03***	***0.03***	***0.03***	**0.02**
	S_3	***-0.02***	**-0.03**	**-0.03**	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***	***-0.02***
	S_4	0.08	0.08	***0.07***	**0.04**	0.08	0.08	0.08	0.08	0.08	0.08	0.08
	S_5	***0.04***	***0.04***	**0.03**	***0.04***	***0.04***	***0.04***	***0.04***	***0.04***	***0.04***	***0.04***	***0.04***
Average ARI	S_1	0.53	0.66	0.75	0.75	***0.80***	0.63	**0.85**	0.78	***0.80***	0.69	0.64
	S_2	0.55	0.48	***0.72***	0.69	0.67	0.51	0.70	0.70	**0.74**	0.57	0.50
	S_3	0.46	0.32	0.61	0.60	***0.67***	0.44	0.60	***0.67***	**0.74**	0.47	0.40
	S_4	0.51	0.26	0.55	0.54	0.60	0.47	0.50	***0.64***	**0.69**	0.41	0.45
	S_5	0.56	0.21	0.62	0.59	0.59	0.55	0.48	***0.66***	**0.74**	0.43	0.52

Table 4.8: Summary of contestant methods for different noise levels of the time series data in the simulated data. Comparison of *minimum* forecasting accuracy on 6-step ahead forecast using CF, TS and MC-based clustering approaches. $t = 17$ is used to as the model selection period to support the weight selection process on MC$_{SilHist}$, MC$_{ParetoHist}$ methods, and $t \leq 17$ are used during the clustering stage.

	Scenarios	CF	TS	MC	MC$_{ThreeMin}$	MC$_{Angles}$	MC$_{MaxMax}$	MC$_{MaxMin}$	MC$_{MaxSum}$	MC$_{ParetoHist}$	MC$_{SilHist}$	MC$_{SilSil}$
Min MASE	S_1	0.52	***0.51***	**0.49**	***0.51***	0.58	***0.51***	***0.51***	***0.51***	0.58	***0.51***	***0.51***
	S_2	**0.76**	***0.78***	**0.76**	***0.78***	***0.78***	**0.76**	***0.78***	0.85	0.85	***0.78***	**0.76**
	S_3	1	0.81	**0.76**	***0.79***	1.01	0.95	0.89	0.89	1.14	***0.79***	0.95
	S_4	1.17	0.95	**0.8**	0.96	1.01	1.2	1	1.07	1.08	***0.92***	1.2
	S_5	1.1	1.12	***1.05***	1.17	1.14	1.33	***1.05***	1.36	1.36	**0.99**	1.28
Min MSE	S_1	0.05	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	***0.04***	**0.03**	0.05	**0.03**	**0.03**
	S_2	0.08	0.08	**0.06**	**0.06**	**0.06**	0.08	**0.06**	0.10	0.10	***0.07***	0.08
	S_3	0.15	0.15	**0.11**	0.15	0.16	0.15	***0.14***	0.17	0.17	**0.11**	0.15
	S_4	0.29	**0.14**	**0.14**	**0.14**	0.18	0.29	***0.16***	0.19	0.20	**0.14**	0.29
	S_5	0.30	***0.29***	**0.26**	***0.29***	0.33	0.30	0.34	0.33	0.33	**0.26**	0.30
Min sMAPE(%)	S_1	14.17	**10.30**	**10.30**	**10.30**	***14.16***	**10.30**	14.39	**10.30**	14.17	**10.30**	**10.30**
	S_2	23.49	21.86	**18.56**	**18.56**	22.07	21.91	***19.10***	22.07	22.07	21.86	21.91
	S_3	37.70	27.11	***26.79***	**26.25**	29.38	27.11	28.00	**26.25**	38.21	***26.79***	27.11
	S_4	45.18	***34.88***	37.66	**34.31**	36.71	37.87	38.21	35.36	47.05	***34.88***	37.87
	S_5	44.14	44.07	43.67	**39.95**	43.62	43.07	40.70	40.70	44.20	***40.56***	44.14
Min ME	S_1	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**	**-0.27**
	S_2	***-0.57***	**-0.58**	**-0.58**	***-0.57***	***-0.57***	**-0.58**	-***-0.57***	***-0.57***	***-0.57***	***-0.57***	**-0.58**
	S_3	-0.54	**-0.56**	**-0.56**	-0.54	-0.54	**-0.56**	***-0.55***	-0.54	-0.54	-0.54	**-0.56**
	S_4	-0.74	***-0.77***	**-0.78**	-0.71	-0.73	-0.74	-0.74	-0.73	-0.74	-0.74	-0.74
	S_5	-0.42	**-0.44**	**-0.44**	**-0.44**	-0.41	***-0.43***	**-0.44**	-0.41	***-0.43***	**-0.44**	***-0.43***
Min ARI	S_1	0.23	0.39	0.33	***0.42***	0.32	0.23	**0.44**	0.41	***0.42***	0.33	0.23
	S_2	0.19	0.13	***0.36***	0.34	0.24	0.16	0.33	0.33	**0.43**	0.30	0.16
	S_3	0.22	0.04	***0.27***	0.23	0.26	0.04	***0.27***	***0.27***	**0.43**	0.12	0.04
	S_4	***0.21***	-0.02	0.16	0.13	0.19	0.07	0.13	0.13	**0.37**	0.07	0.07
	S_5	***0.26***	-0.04	0.19	0.10	0.10	0.01	0.18	0.10	**0.48**	0.02	0.01

Table 4.9: Summary of contestant methods for different noise levels of the time series data in the simulated data. Comparison of *maximum* forecasting accuracy on 6-step ahead forecast using CF, TS and MC-based clustering approaches. $t = 17$ is used to as the model selection period to support the weight selection process on MC$_{SilHist}$, MC$_{ParetoHist}$ methods, and $t \leq 17$ are used during the clustering stage.

	Scenarios	CF	TS	MC	MC$_{ThreeMin}$	MC$_{Angles}$	MC$_{MaxMax}$	MC$_{MaxMin}$	MC$_{MaxSum}$	MC$_{ParetoHist}$	MC$_{SilHist}$	MC$_{SilSil}$
Max MASE	S_1	2.80	***1.81***	**1.64**	**1.64**	2.40	2.69	1.93	1.82	2.26	***1.81***	2.69
	S_2	3.00	2.90	**2.47**	2.69	3.54	2.96	2.91	3.56	3.56	***2.66***	2.96
	S_3	3.45	4.38	**3.08**	3.19	3.37	3.56	***3.14***	3.28	3.31	3.29	3.56
	S_4	3.67	4.22	**3.2**	3.55	3.7	3.67	3.7	3.7	3.9	***3.35***	3.67
	S_5	3.52	4.7	**3.13**	***3.26***	3.59	3.52	3.44	3.46	3.46	3.87	3.52
Max MSE	S_1	2.58	0.78	**0.60**	**0.60**	1.88	2.27	1.08	1.70	1.70	***0.63***	2.27
	S_2	6.00	7.94	**2.58**	**2.58**	***4.05***	7.94	***4.05***	4.26	5.19	5.19	7.94
	S_3	3.83	13.23	***3.47***	**3.44**	3.96	4.26	9.13	3.92	4.39	10.28	3.92
	S_4	***6.35***	7.14	**5.68**	**5.68**	7.06	***6.35***	7.90	6.82	**5.68**	7.86	***6.35***
	S_5	***6.37***	15.76	**6.19**	**6.19**	9.56	7.31	11.60	6.64	7.23	10.96	7.31
Max sMAPE(%)	S_1	73.66	**43.24**	**43.24**	**43.24**	64.44	58.82	73.85	***47.42***	64.44	**43.24**	58.82
	S_2	87.39	67.76	65.65	**60.70**	90.40	89.67	***65.34***	90.40	78.94	69.30	87.39
	S_3	98.02	***95.49***	108.40	**78.70**	100.96	106.29	100.42	108.40	108.40	97.56	101.05
	S_4	107.30	106.96	***90.34***	**84.93**	105.11	105.02	107.37	105.23	123.31	99.61	107.30
	S_5	***103.28***	122.79	107.95	**95.36**	113.87	***103.28***	120.09	113.87	114.80	120.09	***103.28***
Max ME	S_1	***0.27***	***0.27***	**0.26**	***0.27***	***0.27***	***0.27***	**0.26**	**0.26**	***0.27***	***0.27***	***0.27***
	S_2	**0.47**	***0.48***	**0.47**	***0.48***	***0.48***	***0.48***	***0.48***	***0.48***	**0.47**	***0.48***	***0.48***
	S_3	0.43	0.42	**0.39**	0.43	0.42	0.43	***0.41***	0.42	0.43	0.42	0.43
	S_4	***0.77***	***0.77***	**0.76**	0.8	***0.77***	***0.77***	***0.77***	***0.77***	***0.77***	***0.77***	***0.77***
	S_5	***0.69***	0.73	**0.68**	***0.69***	**0.68**	***0.69***	**0.68**	***0.69***	***0.69***	**0.68**	***0.69***
Max ARI	S_1	***0.90***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_2	**1.00**	***0.90***	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	S_3	0.80	0.64	**1.00**	**1.00**	**1.00**	***0.90***	**1.00**	**1.00**	**1.00**	**1.00**	0.73
	S_4	**1.00**	0.55	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	***0.76***	**1.00**
	S_5	**1.00**	0.50	**1.00**	**1.00**	**1.00**	**1.00**	***0.95***	**1.00**	**1.00**	**1.00**	**1.00**

Table 4.10: Summary of contestant methods for different noise levels of the time series data in the simulated data. Comparison of *median* of the forecasting accuracy on 6-step ahead forecast using CF, TS and MC-based clustering approaches. $t = 17$ is used to as the model selection period to support the weight selection process on MC$_{SilHist}$, MC$_{ParetoHist}$ methods, and $t \leq 17$ are used during the clustering stage.

	Scenarios	CF	TS	MC	MC$_{ThreeMin}$	MC$_{Angles}$	MC$_{MaxMax}$	MC$_{MaxMin}$	MC$_{MaxSum}$	MC$_{ParetoHist}$	MC$_{SilHist}$	MC$_{SilSil}$
Median MASE	S_1	1.31	0.97	**0.90**	***0.92***	1.07	1.02	0.99	0.95	1.09	***0.92***	1.01
	S_2	1.70	1.57	**1.40**	1.47	1.62	1.65	1.49	1.57	1.67	***1.46***	1.66
	S_3	2.01	1.99	**1.7**	1.82	1.91	2.04	1.82	1.9	1.97	***1.79***	2.04
	S_4	2.11	2.06	**1.84**	2.06	2.04	2.13	1.99	2.05	2.09	***1.89***	2.12
	S_5	2.12	2.26	**1.93**	2.05	2.09	2.15	2.05	2.09	2.1	***2***	2.13
Median MSE	S_1	0.40	***0.15***	**0.14**	**0.14**	0.23	0.17	0.18	***0.15***	0.24	**0.14**	0.17
	S_2	0.68	0.52	**0.39**	***0.42***	0.57	0.60	0.47	0.54	0.63	0.46	0.61
	S_3	1.02	0.97	**0.67**	***0.74***	0.94	1.05	0.83	0.93	1.01	0.77	1.04
	S_4	1.34	1.41	**0.96**	***1.03***	1.28	1.34	1.24	1.29	1.38	1.18	1.33
	S_5	1.44	1.83	**1.16**	***1.25***	1.48	1.46	1.52	1.42	1.42	1.44	1.43
Median sMAPE (%)	S_1	31.75	21.82	**20.49**	20.54	24.68	22.69	21.99	21.73	25.19	***20.96***	22.55
	S_2	45.17	38.01	***34.21***	**33.86**	41.11	39.34	36.15	39.22	40.27	36.22	40.47
	S_3	62.92	57.92	***52.80***	**50.38**	57.72	62.59	54.37	56.47	57.74	53.05	63.08
	S_4	66.81	62.62	60.80	**56.36**	61.91	66.62	61.89	64.07	64.45	***59.77***	66.38
	S_5	69.84	71.27	64.91	**61.18**	68.36	70.90	70.33	68.48	68.68	***64.76***	70.68
Median ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	***0.03***	***0.03***	**0.02**	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***
	S_3	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**	**-0.03**
	S_4	***0.07***	***0.07***	***0.07***	**0.05**	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***	***0.07***
	S_5	***0.03***	***0.03***	**0.02**	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***	***0.03***
Median ARI	S_1	0.48	0.62	***0.81***	***0.81***	***0.81***	0.55	**0.90**	**0.90**	0.80	0.73	0.57
	S_2	0.51	0.48	**0.73**	**0.73**	0.62	0.50	***0.72***	**0.73**	**0.73**	0.54	0.50
	S_3	0.40	0.33	0.62	0.57	***0.72***	0.40	0.61	***0.72***	**0.77**	0.40	0.38
	S_4	0.46	0.27	0.54	0.52	0.57	0.44	0.49	***0.68***	**0.70**	0.40	0.44
	S_5	0.51	0.20	***0.62***	0.57	0.59	0.52	0.45	**0.72**	**0.72**	0.43	0.50

Table 4.11: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy based on *average MASE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are average across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.69**	**0.79**	**0.87**	**0.96**	**1.04**	**1.13**
	MC$_{Angles}$	0.86	0.97	1.07	1.17	1.27	1.37
	MC$_{MaxMax}$	0.8	0.92	1.01	1.12	1.22	1.32
	MC$_{MaxMin}$	0.78	0.88	0.97	1.06	1.15	1.24
	MC$_{MaxSum}$	0.75	0.85	0.93	1.02	1.11	1.2
	MC$_{ParetoHist}$	0.87	0.98	1.08	1.18	1.29	1.39
	MC$_{SilHist}$	***0.71***	***0.81***	***0.90***	***0.99***	***1.08***	***1.18***
	MC$_{SilSil}$	0.8	0.91	1.01	1.11	1.21	1.31
S_2	MC	**1.05**	**1.20**	**1.35**	**1.50**	**1.65**	**1.78**
	MC$_{Angles}$	1.3	1.46	1.61	1.77	1.92	2.08
	MC$_{MaxMax}$	1.27	1.43	1.58	1.76	1.92	2.09
	MC$_{MaxMin}$	1.15	1.29	1.45	1.6	1.76	1.91
	MC$_{MaxSum}$	1.25	1.4	1.54	1.7	1.85	1.99
	MC$_{ParetoHist}$	1.36	1.51	1.66	1.81	1.96	2.11
	MC$_{SilHist}$	***1.10***	***1.25***	***1.41***	***1.58***	***1.74***	***1.90***
	MC$_{SilSil}$	1.28	1.44	1.6	1.77	1.94	2.1
S_3	MC	**1.27**	**1.47**	**1.66**	**1.84**	**2.02**	**2.19**
	MC$_{Angles}$	1.5	1.69	1.87	2.04	2.21	2.38
	MC$_{MaxMax}$	1.55	1.78	1.99	2.19	2.4	2.61
	MC$_{MaxMin}$	1.37	1.57	1.77	1.96	2.14	***2.33***
	MC$_{MaxSum}$	1.47	1.67	1.86	2.03	2.2	2.38
	MC$_{ParetoHist}$	1.57	1.76	1.94	2.10	2.27	2.44
	MC$_{SilHist}$	***1.30***	***1.52***	***1.73***	***1.93***	***2.13***	2.34
	MC$_{SilSil}$	1.55	1.77	1.98	2.18	2.4	2.61
S_4	MC	**1.35**	**1.56**	**1.78**	**1.98**	**2.16**	**2.35**
	MC$_{Angles}$	1.6	1.78	1.98	2.18	2.38	2.57
	MC$_{MaxMax}$	1.64	1.84	2.06	2.26	2.47	2.67
	MC$_{MaxMin}$	1.5	1.69	1.9	2.11	2.32	2.53
	MC$_{M}axSum$	1.62	1.81	2	2.18	2.38	2.57
	MC$_{ParetoHist}$	1.68	1.85	2.04	2.22	2.42	2.61
	MC$_{SilHist}$	***1.38***	***1.59***	***1.82***	***2.04***	***2.26***	***2.49***
	MC$_{SilSil}$	1.63	1.84	2.06	2.25	2.46	2.66
S_5	MC	**1.45**	**1.67**	**1.86**	**2.06**	**2.27**	**2.46**
	MC$_{Angles}$	1.65	1.85	2.03	2.21	2.4	2.6
	MC$_{MaxMax}$	1.72	1.9	2.08	2.26	2.47	2.68
	MC$_{M}axMin$	1.57	1.79	1.99	2.2	2.41	2.63
	MC$_{M}axSum$	1.66	1.85	2.02	2.2	***2.39***	***2.59***
	MC$_{ParetoHist}$	1.68	1.86	2.03	2.21	2.40	***2.59***
	MC$_{SilHist}$	***1.49***	***1.73***	***1.95***	***2.17***	2.40	2.62
	MC$_{SilSil}$	1.71	1.89	2.06	2.25	2.46	2.67

Table 4.12: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *average ME* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are average across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{Angles}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{MaxMax}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{MaxMin}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{MaxSum}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{ParetoHist}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{SilHist}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	MC$_{SilSil}$	**0**	***0***	**-0.01**	**-0.01**	**-0.02**	**-0.02**
S_2	MC	**0.01**	**0.01**	**0.02**	**0.03**	**0.03**	**0.04**
	MC$_{Angles}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
	MC$_{MaxMax}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
	MC$_{MaxMin}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
	MC$_{MaxSum}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
	MC$_{ParetoHist}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	***0.05***
	MC$_{SilHist}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
	MC$_{SilSil}$	**0.01**	**0.01**	**0.02**	**0.03**	***0.04***	**0.04**
S_3	MC	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**	**-0.05**
	MC$_{Angles}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	-0.03
	MC$_{MaxMax}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	-0.03
	MC$_{MaxMin}$	**-0.02**	**-0.02**	***-0.02***	**-0.03**	***-0.03***	***-0.04***
	MC$_{MaxSum}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	***-0.04***
	MC$_{ParetoHist}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	-0.03
	MC$_{SilHist}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	***-0.04***
	MC$_{SilSil}$	**-0.02**	**-0.02**	***-0.02***	***-0.02***	***-0.03***	-0.03
S_4	MC	**0.03**	**0.05**	**0.07**	**0.08**	**0.09**	**0.1**
	MC$_{Angles}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
	MC$_{MaxMax}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
	MC$_{MaxMin}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
	MC$_{MaxSum}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
	MC$_{ParetoHist}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
	MC$_{SilHist}$	***0.04***	**0.05**	***0.08***	***0.09***	***0.1***	***0.11***
	MC$_{SilSil}$	***0.04***	**0.05**	**0.07**	***0.09***	***0.1***	***0.11***
S_5	MC	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**	**0.05**
	MC$_{Angles}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{MaxMax}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{MaxMin}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{M}axSum$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{ParetoHist}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{SilHist}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***
	MC$_{SilSil}$	**0.01**	***0.03***	***0.04***	***0.05***	***0.05***	***0.06***

Table 4.13: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *average MSE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6.Shown are average across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.08**	**0.11**	**0.14**	**0.17**	**0.20**	**0.24**
	MC$_{Angles}$	0.18	0.23	0.28	0.35	0.42	0.49
	MC$_{MaxMax}$	0.13	0.17	0.21	0.25	0.3	0.36
	MC$_{MaxMin}$	0.12	0.15	0.18	0.22	0.26	0.31
	MC$_{MaxSum}$	0.11	***0.14***	***0.17***	***0.21***	0.25	0.3
	MC$_{ParetoHist}$	0.17	0.22	0.27	0.33	0.40	0.47
	MC$_{SilHist}$	***0.09***	**0.11**	**0.14**	**0.17**	***0.21***	***0.25***
	MC$_{SilSil}$	0.12	0.16	0.2	0.24	0.29	0.35
S_2	MC	**0.23**	**0.31**	**0.40**	**0.52**	**0.63**	**0.75**
	MC$_{Angles}$	0.4	0.51	0.64	0.8	0.96	1.14
	MC$_{MaxMax}$	0.37	0.49	0.63	0.82	1.02	1.23
	MC$_{M}axMin$	0.29	0.38	0.5	0.65	0.81	***0.98***
	MC$_{M}axSum$	0.36	0.47	0.59	0.75	0.91	1.08
	MC$_{ParetoHist}$	0.44	0.55	0.68	0.84	1.00	1.17
	MC$_{SilHist}$	***0.26***	***0.35***	***0.47***	***0.64***	***0.82***	1.01
	MC$_{SilSil}$	0.37	0.5	0.64	0.84	1.04	1.25
S_3	MC	**0.37**	**0.52**	**0.70**	**0.87**	**1.06**	**1.27**
	MC$_{Angles}$	0.61	0.78	0.97	1.17	***1.4***	***1.66***
	MC$_{MaxMax}$	0.63	0.83	1.07	1.31	1.6	1.92
	MC$_{MaxMin}$	0.48	0.66	0.88	1.12	1.38	1.67
	MC$_{MaxSum}$	0.58	0.75	0.95	1.16	***1.4***	1.66
	MC$_{ParetoHist}$	0.68	0.86	1.06	1.25	1.48	1.73
	MC$_{SilHist}$	***0.41***	***0.59***	***0.83***	***1.11***	1.41	1.76
	MC$_{SilSil}$	0.62	0.82	1.05	1.3	1.59	1.92
S_4	MC	**0.51**	**0.73**	**1.00**	**1.28**	**1.57**	**1.88**
	MC$_{Angles}$	0.79	1.02	1.33	1.67	2.04	2.45
	MC$_{MaxMax}$	0.81	1.05	1.36	1.7	2.1	2.51
	MC$_{MaxMin}$	0.71	0.96	1.28	1.64	2.05	2.49
	MC$_{MaxSum}$	0.8	1.03	1.31	1.62	***1.98***	2.38
	MC$_{ParetoHist}$	0.88	1.10	1.37	1.68	2.03	***2.41***
	MC$_{SilHist}$	***0.59***	***0.85***	***1.20***	***1.61***	2.05	2.53
	MC$_{SilSil}$	0.81	1.05	1.36	1.71	2.1	2.52
S_5	MC	**0.66**	**0.91**	**1.17**	**1.48**	**1.82**	**2.17**
	MC$_{Angles}$	0.94	1.22	1.51	1.86	2.26	2.69
	MC$_{MaxMax}$	0.97	1.21	1.47	1.8	2.18	2.58
	MC$_{MaxMin}$	0.89	1.19	1.51	1.91	2.37	2.86
	MC$_{MaxSum}$	0.96	1.21	1.46	1.77	2.13	2.52
	MC$_{ParetoHist}$	0.97	1.21	***1.45***	***1.75***	***2.09***	***2.45***
	MC$_{SilHist}$	***0.76***	***1.11***	1.48	1.93	2.46	3.04
	MC$_{SilSil}$	0.95	1.19	***1.45***	1.78	2.16	2.56

Table 4.14: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *average sMAPE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6.Shown are average across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	***23.68***	**22.05**	**20.92**	**20.26**	**19.87**	**19.69**
	MC$_{Angles}$	29.04	27.33	26.16	25.48	25.06	24.82
	MC$_{MaxMax}$	26.56	25.1	24.11	23.59	23.3	23.2
	MC$_{MaxMin}$	26.11	24.29	23.09	22.39	21.96	21.72
	MC$_{MaxSum}$	24.75	23.12	21.99	21.35	20.99	20.83
	MC$_{ParetoHist}$	28.67	26.85	25.63	24.89	24.46	24.21
	MC$_{SilHist}$	**23.67**	***22.11***	***21.1***	***20.54***	***20.22***	***20.11***
	MC$_{SilSil}$	26.34	24.87	23.86	23.33	23.03	22.92
S_2	MC	**37.93**	**36.67**	**35.74**	**35.26**	**34.97**	**34.77**
	MC$_{Angles}$	45.02	43.9	43.02	42.67	42.42	42.27
	MC$_{MaxMax}$	43.79	42.78	41.99	41.76	41.68	41.63
	MC$_{MaxMin}$	***38.34***	***37.46***	***37.08***	***37.02***	***37.03***	***36.95***
	MC$_{MaxSum}$	42.15	41.07	40.24	39.9	39.71	39.54
	MC$_{ParetoHist}$	45.76	44.32	43.23	42.68	42.29	41.98
	MC$_{SilHist}$	38.58	37.98	37.74	37.87	38.09	38.25
	MC$_{SilSil}$	44.33	43.45	42.73	42.51	42.4	42.31
S_3	MC	***56.21***	**55.11**	**53.96**	**52.9**	**52.48**	**52.24**
	MC$_{Angles}$	61.67	60.55	59.25	58.18	57.58	57.19
	MC$_{MaxMax}$	62.9	63.02	62.91	62.51	62.37	62.44
	MC$_{MaxMin}$	57.04	57.5	57.43	57.19	57.01	56.9
	MC$_{MaxSum}$	60.77	60.26	59.43	58.51	58.03	57.71
	MC$_{ParetoHist}$	64.26	62.65	61.02	59.5	58.74	58.22
	MC$_{SilHist}$	**54.54**	***55.19***	***55.16***	***55.05***	***55.13***	***55.26***
	MC$_{SilSil}$	62.92	63.1	63.13	62.89	62.97	63.18
S_4	MC	***62.56***	***62.07***	***61.98***	**61.45**	**61.28**	**61.2**
	MC$_{Angles}$	65.55	64.78	64.8	64.6	64.65	64.66
	MC$_{MaxMax}$	68.06	67.94	68.41	68.17	68.16	68.09
	MC$_{MaxMin}$	63.95	63.88	64.13	64.1	64.42	64.63
	MC$_{MaxSum}$	67.14	66.13	65.77	65.33	65.41	65.36
	MC$_{ParetoHist}$	68.79	67.85	67.34	66.82	66.7	66.54
	MC$_{SilHist}$	**58.55**	**59.66**	**60.99**	***61.59***	***62.18***	***62.62***
	MC$_{SilSil}$	68.23	68	68.48	68.2	68.17	68.05
S_5	MC	***67.81***	**67.3**	**66.1**	**65.66**	**65.64**	**65.79**
	MC$_{Angles}$	71.8	71.52	71.01	70.34	70.12	70.05
	MC$_{MaxMax}$	73.55	72.63	71.69	71.11	71.12	71.26
	MC$_{MaxMin}$	70.74	71.72	71.88	72.03	72.13	72.33
	MC$_{MaxSum}$	72.21	71.84	71.28	70.66	70.52	70.53
	MC$_{ParetoHist}$	72.81	71.96	71.2	70.44	70.13	70
	MC$_{SilHist}$	**65.96**	***67.52***	***68.06***	***68.5***	***68.92***	***69.25***
	MC$_{SilSil}$	73.65	72.46	71.48	71.03	71.1	71.32

Table 4.15: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *minimum MASE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6.Shown are minimum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.49**	**0.6**	**0.61**	**0.65**	**0.68**	**0.73**
	MC$_{Angles}$	0.58	0.67	0.71	0.78	0.82	0.86
	MC$_{MaxMax}$	***0.51***	**0.6**	***0.62***	***0.66***	***0.69***	***0.74***
	MC$_{MaxMin}$	***0.51***	0.64	0.65	0.7	0.73	0.78
	MC$_{MaxSum}$	***0.51***	**0.6**	***0.62***	***0.66***	***0.69***	***0.74***
	MC$_{ParetoHist}$	0.58	0.66	0.73	0.81	0.84	0.87
	MC$_{SilHist}$	***0.51***	***0.61***	**0.61**	**0.65**	**0.68**	**0.73**
	MC$_{SilSil}$	***0.51***	**0.6**	***0.62***	***0.66***	***0.69***	***0.74***
S_2	MC	**0.76**	**0.87**	**0.89**	**1.01**	**1.1**	**1.19**
	MC$_{Angles}$	***0.78***	**0.87**	**0.89**	**1.01**	**1.1**	**1.19**
	MC$_{MaxMax}$	**0.76**	0.95	1.04	1.16	1.24	1.3
	MC$_{MaxMin}$	***0.78***	**0.87**	**0.89**	**1.01**	**1.1**	**1.19**
	MC$_{MaxSum}$	0.85	0.95	***1.01***	***1.11***	1.18	***1.27***
	MC$_{ParetoHist}$	0.85	1.07	1.14	1.23	1.27	1.3
	MC$_{SilHist}$	***0.78***	***0.88***	1.02	1.12	***1.17***	**1.19**
	MC$_{SilSil}$	**0.76**	0.95	1.04	1.16	1.24	1.3
S_3	MC	**0.76**	**1.03**	***1.22***	1.35	**1.48**	**1.62**
	MC$_{Angles}$	1.01	***1.12***	**1.21**	**1.33**	***1.49***	1.67
	MC$_{MaxMax}$	0.95	1.17	1.4	1.53	1.63	1.72
	MC$_{MaxMin}$	0.89	***1.12***	1.33	1.46	1.56	***1.66***
	MC$_{MaxSum}$	0.89	***1.12***	1.27	1.42	1.61	1.76
	MC$_{ParetoHist}$	1.14	1.3	1.45	1.58	1.69	1.79
	MC$_{SilHist}$	***0.79***	1.14	1.23	***1.34***	1.51	***1.66***
	MC$_{SilSil}$	0.95	1.17	1.4	1.53	1.67	1.78
S_4	MC	**0.8**	**1.03**	**1.1**	**1.14**	**1.27**	**1.46**
	MC$_{Angles}$	1.01	1.19	1.24	1.32	1.4	1.52
	MC$_{MaxMax}$	1.2	1.3	1.45	1.61	1.73	1.86
	MC$_{MaxMin}$	1	1.17	1.23	1.29	1.37	***1.5***
	MC$_{MaxSum}$	1.07	1.25	1.37	1.45	1.55	1.66
	MC$_{ParetoHist}$	1.08	1.2	1.37	1.45	1.55	1.66
	MC$_{SilHist}$	***0.92***	***1.05***	***1.14***	***1.2***	***1.34***	1.52
	MC$_{SilSil}$	1.2	1.3	1.45	1.61	1.73	1.86
S_5	MC	***1.05***	***1.12***	***1.28***	***1.5***	***1.71***	***1.81***
	MC$_{Angles}$	1.14	1.3	1.45	1.6	***1.71***	1.82
	MC$_{MaxMax}$	1.33	1.38	1.5	1.67	1.81	1.92
	MC$_{MaxMin}$	1.05	1.24	1.33	***1.5***	1.66	1.78
	MC$_{MaxSum}$	1.36	1.38	1.5	1.67	1.76	1.85
	MC$_{ParetoHist}$	1.36	1.38	1.5	1.67	1.76	1.85
	MC$_{SilHist}$	**0.99**	**1.02**	**1.15**	**1.35**	**1.55**	**1.74**
	MC$_{SilSil}$	1.28	1.38	1.48	1.67	1.81	1.92

Table 4.16: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *minimum ME* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are minimum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**-0.09**	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	MC$_{Angles}$	**-0.09**	**-0.14**	***-0.17***	***-0.2***	**-0.24**	**-0.27**
	MC$_{MaxMax}$	**-0.09**	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
	MC$_{MaxMin}$	**-0.09**	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
	MC$_{MaxSum}$	**-0.09**	**-0.14**	***-0.17***	***-0.2***	**-0.24**	**-0.27**
	MC$_{ParetoHist}$	**-0.09**	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
	MC$_{SilHist}$	**-0.09**	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	MC$_{SilSil}$	**-0.09**	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
S_2	MC	**-0.15**	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	MC$_{Angles}$	**-0.15**	**-0.28**	**-0.35**	**-0.45**	**-0.52**	***-0.57***
	MC$_{MaxMax}$	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	MC$_{MaxMin}$	**-0.15**	**-0.28**	**-0.35**	**-0.45**	**-0.52**	***-0.57***
	MC$_{MaxSum}$	**-0.15**	**-0.28**	**-0.35**	**-0.45**	**-0.52**	***-0.57***
	MC$_{ParetoHist}$	***-0.14***	**-0.28**	***-0.34***	**-0.45**	**-0.52**	***-0.57***
	MC$_{SilHist}$	***-0.14***	**-0.28**	***-0.34***	**-0.45**	**-0.52**	***-0.57***
	MC$_{SilSil}$	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
S_3	MC	**-0.23**	**-0.34**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	MC$_{Angles}$	**-0.23**	***-0.33***	***-0.36***	***-0.42***	***-0.47***	-0.54
	MC$_{MaxMax}$	**-0.23**	***-0.33***	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	MC$_{MaxMin}$	**-0.23**	***-0.33***	***-0.36***	**-0.43**	***-0.47***	***-0.55***
	MC$_{MaxSum}$	**-0.23**	***-0.33***	***-0.36***	***-0.42***	***-0.47***	-0.54
	MC$_{ParetoHist}$	**-0.23**	***-0.33***	***-0.36***	***-0.42***	***-0.47***	-0.54
	MC$_{SilHist}$	**-0.23**	***-0.33***	***-0.36***	***-0.42***	***-0.47***	-0.54
	MC$_{SilSil}$	**-0.23**	***-0.33***	**-0.37**	**-0.43**	**-0.48**	**-0.56**
S_4	MC	**-0.23**	**-0.26**	**-0.39**	**-0.53**	**-0.65**	**-0.78**
	MC$_{Angles}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	-0.62	-0.73
	MC$_{MaxMax}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	-0.62	***-0.74***
	MC$_{MaxMin}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	***-0.63***	***-0.74***
	MC$_{MaxSum}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	-0.62	-0.73
	MC$_{ParetoHist}$	**-0.23**	**-0.26**	***-0.37***	***-0.5***	-0.62	***-0.74***
	MC$_{SilHist}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	-0.62	***-0.74***
	MC$_{SilSil}$	***-0.22***	***-0.25***	***-0.37***	***-0.5***	-0.62	***-0.74***
S_5	MC	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	MC$_{Angles}$	**-0.16**	***-0.24***	***-0.25***	-0.28	***-0.34***	-0.41
	MC$_{MaxMax}$	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	***-0.43***
	MC$_{MaxMin}$	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	MC$_{MaxSum}$	**-0.16**	***-0.24***	***-0.25***	-0.28	***-0.34***	-0.41
	MC$_{ParetoHist}$	**-0.16**	**-0.25**	**-0.26**	***-0.29***	-0.36	***-0.43***
	MC$_{SilHist}$	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	MC$_{SilSil}$	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	***-0.43***

Table 4.17: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *minimum MSE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are minimum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.03**	**0.03**	**0.04**	**0.04**	**0.04**	**0.05**
	MC$_{Angles}$	**0.03**	0.05	***0.05***	0.06	0.07	0.08
	MC$_{MaxMax}$	**0.03**	***0.04***	**0.04**	**0.04**	**0.04**	**0.05**
	MC$_{MaxMin}$	***0.04***	0.05	***0.05***	***0.05***	***0.06***	***0.06***
	MC$_{MaxSum}$	**0.03**	***0.04***	**0.04**	**0.04**	**0.04**	**0.05**
	MC$_{ParetoHist}$	0.05	0.05	0.06	0.07	0.08	0.09
	MC$_{SilHist}$	**0.03**	**0.03**	**0.04**	**0.04**	**0.04**	**0.05**
	MC$_{SilSil}$	**0.03**	***0.04***	**0.04**	**0.04**	**0.04**	**0.05**
S_2	MC	**0.06**	**0.09**	**0.1**	**0.13**	**0.15**	**0.15**
	MC$_{Angles}$	**0.06**	**0.09**	**0.1**	**0.13**	**0.15**	**0.15**
	MC$_{MaxMax}$	0.08	0.12	0.15	0.17	0.19	0.21
	MC$_{MaxMin}$	**0.06**	**0.09**	**0.1**	**0.13**	**0.15**	**0.15**
	MC$_{MaxSum}$	0.1	0.12	***0.12***	***0.14***	***0.16***	***0.18***
	MC$_{ParetoHist}$	0.1	0.13	0.16	0.17	0.19	0.21
	MC$_{SilHist}$	***0.07***	***0.1***	***0.12***	***0.14***	**0.15**	**0.15**
	MC$_{SilSil}$	0.08	0.12	0.15	0.17	0.19	0.21
S_3	MC	**0.11**	**0.16**	**0.21**	**0.26**	**0.32**	**0.38**
	MC$_{Angles}$	0.16	0.23	0.3	0.36	0.41	0.46
	MC$_{MaxMax}$	0.15	0.24	0.32	0.37	0.43	0.47
	MC$_{MaxMin}$	***0.14***	***0.19***	***0.26***	***0.3***	***0.35***	***0.39***
	MC$_{MaxSum}$	0.17	0.24	0.32	0.38	0.44	0.48
	MC$_{ParetoHist}$	0.17	0.25	0.33	0.39	0.45	0.49
	MC$_{SilHist}$	**0.11**	**0.16**	**0.21**	**0.26**	**0.32**	**0.38**
	MC$_{SilSil}$	0.15	0.24	0.31	0.36	0.43	0.5
S_4	MC	**0.14**	**0.17**	**0.18**	**0.23**	**0.27**	***0.34***
	MC$_{Angles}$	0.18	***0.19***	***0.2***	***0.24***	**0.27**	**0.32**
	MC$_{MaxMax}$	0.29	0.34	0.46	0.52	0.61	0.69
	MC$_{MaxMin}$	***0.16***	**0.17**	**0.18**	**0.23**	**0.2**7	***0.34***
	MC$_{MaxSum}$	0.19	0.28	0.37	0.42	0.54	0.65
	MC$_{ParetoHist}$	0.2	0.29	0.43	0.49	0.57	0.65
	MC$_{SilHist}$	**0.14**	**0.17**	**0.18**	**0.23**	***0.3***	0.37
	MC$_{SilSil}$	0.29	0.34	0.46	0.52	0.61	0.69
S_5	MC	**0.26**	**0.29**	**0.41**	**0.44**	**0.47**	**0.54**
	MC$_{Angles}$	0.33	0.37	***0.42***	**0.44**	**0.47**	**0.54**
	MC$_{MaxMax}$	***0.3***	***0.35***	0.43	0.48	0.53	***0.59***
	MC$_{MaxMin}$	0.34	0.42	0.5	0.58	0.74	0.86
	MC$_{MaxSum}$	0.33	0.37	***0.42***	**0.44**	**0.47**	**0.54**
	MC$_{ParetoHist}$	0.33	0.37	***0.42***	**0.44**	**0.47**	**0.54**
	MC$_{SilHist}$	**0.26**	**0.29**	**0.41**	***0.45***	***0.48***	**0.54**
	MC$_{SilSil}$	***0.3***	***0.35***	0.43	0.48	0.53	***0.59***

Table 4.18: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *minimum sMAPE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are minimum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**10.3**	**11.37**	**1.54**	**11.62**	**11.77**	**12.53**
	MC$_{Angles}$	14.88	16.2	16.03	15.35	14.55	***14.16***
	MC$_{MaxMax}$	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	MC$_{MaxMin}$	***14.39***	***14.97***	***15.22***	15.71	14.99	14.63
	MC$_{MaxSum}$	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	MC$_{ParetoHist}$	14.55	***14.97***	***15.22***	***15.19***	***14.52***	14.17
	MC$_{SilHist}$	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	MC$_{SilSil}$	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
S_2	MC	**18.56**	**19.42**	**20.72**	**21.67**	**22.76**	**23.26**
	MC$_{Angles}$	22.07	23.73	24.97	25.65	25.16	24.92
	MC$_{MaxMax}$	21.91	***22.63***	***24.18***	***25.64***	26.35	26.77
	MC$_{MaxMin}$	***19.1***	**19.42**	**20.72**	**21.67**	**22.76**	**23.26**
	MC$_{MaxSum}$	22.07	***22.63***	***24.18***	***25.64***	26.41	26.17
	MC$_{ParetoHist}$	22.07	24.93	25.01	25.97	26.56	26.17
	MC$_{SilHist}$	21.86	***22.63***	***24.18***	24.7	***24.48***	***24.16***
	MC$_{SilSil}$	21.91	***22.63***	***24.18***	***25.64***	26.35	26.77
S_3	MC	26.79	**30.1**	**33.64**	**35.18**	**35.68**	***37.11***
	MC$_{Angles}$	29.38	32.21	35.59	36.17	***36.16***	**36.47**
	MC$_{MaxMax}$	***27.11***	***30.78***	***34.68***	***35.77***	36.37	37.75
	MC$_{MaxMin}$	28	31.47	35.11	36.01	36.57	37.58
	MC$_{MaxSum}$	**26.25**	***30.78***	***34.68***	***35.77***	36.37	37.75
	MC$_{ParetoHist}$	38.21	40.7	41.82	41.91	41.91	41.78
	MC$_{SilHist}$	26.79	**30.1**	**33.64**	**35.18**	**35.68**	***37.11***
	MC$_{SilSil}$	***27.11***	***30.78***	***34.68***	***35.77***	36.37	37.75
S_4	MC	38.95	37.66	38.54	40.23	***39.37***	**40.32**
	MC$_{Angles}$	39.08	***36.71***	***37.33***	***37.95***	**39.21**	**40.32**
	MC$_{MaxMax}$	37.87	40.4	42.99	44.92	45.17	45.88
	MC$_{MaxMin}$	38.92	38.21	39.36	40.62	41.54	42.42
	MC$_{MaxSum}$	***35.36***	39.1	40.7	42.54	43.82	44.64
	MC$_{ParetoHist}$	48.19	49.73	50.86	48.3	47.87	47.05
	MC$_{SilHist}$	**34.88**	**35.52**	**36.39**	**37.83**	**39.21**	***40.92***
	MC$_{SilSil}$	37.87	40.4	42.99	44.92	45.17	45.88
S_5	MC	47.93	46.01	44.38	***44.64***	44.48	***43.67***
	MC$_{Angles}$	47.72	46.01	44.93	45.23	***44.45***	**43.62**
	MC$_{MaxMax}$	43.07	43.88	44.14	45	44.95	44.38
	MC$_{MaxMin}$	**41.27**	**40.7**	***42.34***	45.23	***44.45***	**43.62**
	MC$_{MaxSum}$	**41.27**	**40.7**	***42.34***	45.32	45.41	44.81
	MC$_{ParetoHist}$	49.48	46.5	44.2	45.32	45.41	44.81
	MC$_{SilHist}$	***42.32***	***41.43***	**40.56**	**41.75**	**43.28**	***43.67***
	MC$_{SilSil}$	51.56	46.05	44.14	45	44.95	44.38

Table 4.19: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *maximum MASE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are maximum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**1.02**	**1.15**	**1.24**	**1.39**	**1.52**	**1.64**
	MC$_{Angles}$	1.38	1.61	1.8	2.02	2.23	2.4
	MC$_{MaxMax}$	1.65	1.88	2.1	2.27	2.46	2.69
	MC$_{MaxMin}$	1.12	1.29	1.43	1.61	1.78	1.93
	MC$_{MaxSum}$	1.14	***1.23***	***1.36***	***1.53***	***1.69***	1.82
	MC$_{ParetoHist}$	1.35	1.56	1.73	1.94	2.11	2.26
	MC$_{SilHist}$	***1.07***	1.25	1.37	***1.53***	***1.69***	***1.81***
	MC$_{SilSil}$	1.65	1.88	2.1	2.27	2.46	2.69
S_2	MC	**1.4**	**1.66**	**1.9**	**2.13**	**2.31**	**2.47**
	MC$_{Angles}$	2.3	2.43	2.71	2.98	3.27	3.54
	MC$_{MaxMax}$	1.94	2.15	2.26	2.5	2.74	2.96
	MC$_{MaxMin}$	1.74	1.86	2.09	2.37	2.65	2.91
	MC$_{MaxSum}$	2.31	2.44	2.72	2.99	3.28	3.56
	MC$_{ParetoHist}$	2.31	2.44	2.72	2.99	3.28	3.56
	MC$_{SilHist}$	***1.6***	***1.71***	***1.91***	***2.19***	***2.42***	***2.66***
	MC$_{SilSil}$	1.94	2.15	2.26	2.5	2.74	2.96
S_3	MC	1.92	***2.16***	***2.38***	**2.59**	**2.85**	**3.08**
	MC$_{Angles}$	2.18	2.39	2.62	2.84	3.07	3.37
	MC$_{MaxMax}$	2.2	2.44	2.69	2.94	3.24	3.56
	MC$_{MaxMin}$	**1.88**	**2.11**	**2.37**	***2.62***	***2.87***	***3.14***
	MC$_{MaxSum}$	2.04	2.27	2.47	2.68	2.98	3.28
	MC$_{ParetoHist}$	2.18	2.39	2.62	2.84	3.07	3.31
	MC$_{SilHist}$	***1.91***	2.23	2.4	2.71	3	3.29
	MC$_{SilSil}$	2.2	2.44	2.69	2.94	3.24	3.56
S_4	MC	**1.84**	**2.08**	**2.36**	**2.68**	**2.91**	**3.2**
	MC$_{Angles}$	2.16	***2.41***	2.74	3.08	3.42	3.7
	MC$_{MaxMax}$	2.22	2.44	2.67	3.01	3.37	3.67
	MC$_{MaxMin}$	2.06	***2.41***	2.74	3.08	3.42	3.7
	MC$_{MaxSum}$	2.16	***2.41***	2.74	3.08	3.42	3.7
	MC$_{ParetoHist}$	2.3	2.53	2.74	3.12	3.5	3.9
	MC$_{SilHist}$	***1.91***	2.28	***2.64***	***2.94***	***3.14***	***3.35***
	MC$_{SilSil}$	2.22	2.44	2.67	3.01	3.37	3.67
S_5	MC	**1.86**	**2.06**	**2.29**	**2.59**	**2.87**	**3.13**
	MC$_{Angles}$	2.3	2.59	2.9	3.2	3.42	3.59
	MC$_{MaxMax}$	2.14	2.41	2.63	2.92	3.23	3.52
	MC$_{MaxMin}$	2.09	***2.26***	***2.54***	***2.75***	***3.09***	***3.44***
	MC$_{MaxSum}$	2.24	2.39	2.61	2.92	3.23	3.46
	MC$_{ParetoHist}$	2.24	2.39	2.61	2.92	3.23	3.46
	MC$_{SilHist}$	***1.93***	2.42	2.65	3.11	3.49	3.87
	MC$_{SilSil}$	2.2	2.48	2.68	2.96	3.23	3.52

Table 4.20: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *maximum ME* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are maximum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.11**	**0.14**	**0.17**	**0.21**	**0.23**	**0.26**
	MC$_{Angles}$	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	MC$_{MaxMax}$	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	MC$_{MaxMin}$	**0.11**	**0.14**	**0.17**	**0.21**	**0.23**	**0.26**
	MC$_{MaxSum}$	**0.11**	**0.14**	**0.17**	**0.21**	**0.23**	**0.26**
	MC$_{ParetoHist}$	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	MC$_{SilHist}$	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	MC$_{SilSil}$	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
S_2	MC	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	**0.47**
	MC$_{Angles}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
	MC$_{MaxMax}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
	MC$_{MaxMin}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
	MC$_{MaxSum}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
	MC$_{ParetoHist}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	**0.47**
	MC$_{SilHist}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
	MC$_{SilSil}$	**0.14**	**0.24**	**0.28**	**0.34**	**0.38**	0.48
S_3	MC	**0.18**	**0.26**	**0.27**	**0.3**	**0.34**	**0.39**
	MC$_{Angles}$	***0.19***	***0.27***	0.29	0.33	0.37	0.42
	MC$_{MaxMax}$	***0.19***	***0.27***	0.29	0.33	0.37	0.43
	MC$_{MaxMin}$	***0.19***	***0.27***	***0.28***	***0.32***	***0.36***	***0.41***
	MC$_{MaxSum}$	0.2	***0.27***	0.29	0.33	0.37	0.42
	MC$_{ParetoHist}$	***0.19***	***0.27***	0.29	0.33	0.37	0.43
	MC$_{SilHist}$	***0.19***	***0.27***	0.29	0.33	0.37	0.42
	MC$_{SilSil}$	***0.19***	***0.27***	0.29	0.33	0.37	0.43
S_4	MC	**0.49**	**0.58**	**0.72**	**0.76**	**0.74**	**0.74**
	MC$_{Angles}$	**0.49**	***0.59***	***0.73***	***0.77***	0.76	0.76
	MC$_{MaxMax}$	**0.49**	***0.59***	**0.72**	***0.77***	0.76	***0.75***
	MC$_{MaxMin}$	**0.49**	***0.59***	**0.72**	***0.77***	***0.75***	***0.75***
	MC$_{MaxSum}$	**0.49**	***0.59***	***0.73***	***0.77***	0.76	0.76
	MC$_{ParetoHist}$	**0.49**	***0.59***	***0.73***	***0.77***	0.76	0.76
	MC$_{SilHist}$	**0.49**	***0.59***	**0.72**	***0.77***	***0.75***	***0.75***
	MC$_{SilSil}$	**0.49**	***0.59***	**0.72**	***0.77***	0.76	***0.75***
S_5	MC	**0.27**	**0.36**	**0.39**	**0.38**	**0.54**	**0.68**
	MC$_{Angles}$	***0.28***	0.38	0.41	***0.4***	**0.54**	**0.68**
	MC$_{MaxMax}$	**0.27**	**0.36**	**0.39**	**0.38**	**0.54**	***0.69***
	MC$_{MaxMin}$	**0.27**	***0.37***	0.41	***0.4***	**0.54**	**0.68**
	MC$_{MaxSum}$	**0.27**	**0.36**	**0.39**	**0.38**	**0.54**	***0.69***
	MC$_{ParetoHist}$	**0.27**	***0.37***	***0.4***	**0.38**	**0.54**	***0.69***
	MC$_{SilHist}$	**0.27**	***0.37***	***0.4***	***0.4***	**0.54**	**0.68**
	MC$_{SilSil}$	**0.27**	**0.36**	**0.39**	**0.38**	**0.54**	***0.69***

Table 4.21: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *maximum MSE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are maximum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.19**	**0.24**	**0.34**	**0.41**	**0.48**	**0.6**
	MC$_{Angles}$	0.63	0.79	0.99	1.25	1.57	1.88
	MC$_{MaxMax}$	0.72	0.95	1.26	1.55	1.91	2.27
	MC$_{MaxMin}$	0.26	0.36	0.48	***0.64***	***0.87***	1.08
	MC$_{MaxSum}$	0.49	0.62	0.77	0.98	1.36	1.7
	MC$_{ParetoHist}$	0.53	0.67	0.84	0.99	1.36	1.7
	MC$_{SilHist}$	***0.2***	***0.25***	***0.35***	**0.41**	**0.48**	***0.63***
	MC$_{SilSil}$	0.72	0.95	1.26	1.55	1.91	2.27
S_2	MC	**0.57**	**0.96**	**1.47**	**1.97**	**2.26**	**2.58**
	MC$_{Angles}$	1.99	2.11	2.35	***3.12***	***3.72***	***4.05***
	MC$_{MaxMax}$	1.25	1.58	2.48	4.61	6.76	7.94
	MC$_{MaxMin}$	0.85	1.5	***2.12***	***3.12***	***3.72***	***4.05***
	MC$_{MaxSum}$	1.03	1.5	***2.12***	***3.12***	3.73	4.26
	MC$_{ParetoHist}$	1.62	1.7	2.76	3.92	4.7	5.19
	MC$_{SilHist}$	***0.74***	***1.01***	2.76	3.92	4.7	5.19
	MC$_{SilSil}$	1.25	1.58	2.48	4.61	6.76	7.94
S_3	MC	**1.15**	1.66	**2.03**	***2.63***	***3.14***	**3.47**
	MC$_{Angles}$	1.66	2.05	2.37	2.74	3.44	3.96
	MC$_{MaxMax}$	1.51	1.87	2.28	2.75	3.37	4.26
	MC$_{MaxMin}$	***1.23***	***1.76***	3.36	5.14	6.83	9.13
	MC$_{MaxSum}$	1.37	**1.64**	***2.06***	**2.6**	**3.13**	***3.92***
	MC$_{ParetoHist}$	1.66	1.79	2.5	3.2	3.94	4.39
	MC$_{SilHist}$	1.25	1.9	3.46	5.5	7.52	10.28
	MC$_{SilSil}$	1.53	1.87	2.28	2.66	3.34	***3.92***
S_4	MC	**1.27**	**1.99**	**3.28**	**4.46**	**4.99**	**5.68**
	MC$_{Angles}$	1.98	2.97	3.92	5.3	6.18	7.06
	MC$_{MaxMax}$	1.83	2.52	3.36	***4.47***	***5.61***	***6.35***
	MC$_{MaxMin}$	1.98	2.97	4.51	6.11	6.88	7.9
	MC$_{MaxSum}$	1.98	2.97	3.92	5.03	6.17	6.82
	MC$_{ParetoHist}$	***1.57***	***2.36***	***3.31***	4.48	5	**5.68**
	MC$_{SilHist}$	1.69	2.48	4.34	5.98	6.82	7.86
	MC$_{SilSil}$	1.83	2.52	3.36	***4.47***	***5.61***	***6.35***
S_5	MC	**1.4**	**2.37**	**3.2**	**4.09**	**4.86**	**6.19**
	MC$_{Angles}$	2.05	***2.79***	***3.29***	4.47	6.81	9.56
	MC$_{MaxMax}$	2.25	3.05	3.3	***4.37***	5.72	7.31
	MC$_{MaxMin}$	2.54	3.08	4.22	5.77	8.4	11.6
	MC$_{MaxSum}$	2.64	3.05	3.64	4.49	***5.57***	***6.64***
	MC$_{ParetoHist}$	2.25	3.19	3.54	4.49	5.63	7.23
	MC$_{SilHist}$	***1.76***	3.18	4.77	6.45	7.9	10.96
	MC$_{SilSil}$	2.25	3.05	3.3	***4.37***	5.72	7.31

Table 4.22: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *maximum sMAPE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are maximum values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**43.24**	**36.69**	**32.03**	**29.68**	**29.57**	**29.22**
	MC$_{Angles}$	64.44	57.88	53.94	52.14	51.02	49.91
	MC$_{MaxMax}$	***54.49***	55.55	56.3	56.64	57.53	58.82
	MC$_{MaxMin}$	73.85	59.61	***51.44***	47.1	44.38	42.55
	MC$_{MaxSum}$	47.42	***39.95***	35.88	***33.35***	***31.62***	***32.65***
	MC$_{ParetoHist}$	64.44	56.42	51.55	49.25	47.68	46.41
	MC$_{SilHist}$	**43.24**	**36.69**	**32.03**	**29.68**	**29.57**	**29.22**
	MC$_{SilSil}$	***54.49***	55.55	56.3	56.64	**29.57**	58.82
S_2	MC	***65.65***	**62.13**	***60.21***	**58.87**	**58.15**	**56.94**
	MC$_{Angles}$	90.4	88.37	82.3	77.82	77.97	80.09
	MC$_{MaxMax}$	89.67	87.39	81.52	79.52	79.99	81.22
	MC$_{MaxMin}$	**61.79**	**62.13**	**59.45**	***61.54***	***63.73***	***65.34***
	MC$_{MaxSum}$	90.4	74.17	73.51	74.49	75.54	75.82
	MC$_{ParetoHist}$	78.94	76.96	75.86	74.58	74.06	74.16
	MC$_{SilHist}$	67.76	***63.64***	65.01	67.05	68.1	69.3
	MC$_{SilSil}$	87.13	87.39	81.52	79.52	79.99	81.22
S_3	MC	108.4	99.74	94.79	91.53	***89.82***	***89.12***
	MC$_{Angles}$	100.96	99.65	95.42	92.74	90.33	90.2
	MC$_{MaxMax}$	106.29	99.15	***93.57***	**89.24**	**88.43**	**88.64**
	MC$_{MaxMin}$	***100.42***	98.52	97.98	96.04	94.72	94.97
	MC$_{MaxSum}$	108.4	99.74	94.86	93.06	94.08	94.97
	MC$_{ParetoHist}$	108.4	99.74	94.86	91.53	89.82	***89.12***
	MC$_{SilHist}$	**86.44**	***97.56***	97.22	94.86	91.8	90.58
	MC$_{SilSil}$	101.05	**93.15**	**91.16**	***90.01***	89.84	90.78
S_4	MC	**89.62**	**90.34**	**88.66**	**88.07**	**88.38**	**88.4**
	MC$_{Angles}$	105.11	101.7	103.42	104.67	105.06	104.31
	MC$_{MaxMax}$	99.27	101.7	103.42	104.64	105.02	104.71
	MC$_{MaxMin}$	107.37	100.25	97.76	***98.05***	***97.69***	***96.35***
	MC$_{MaxSum}$	105.23	104.09	103.42	104.67	105.18	104.91
	MC$_{ParetoHist}$	119.87	123.31	116.52	115.96	115.51	114.93
	MC$_{SilHist}$	***96.57***	***94.91***	***95.84***	98.42	99.24	99.61
	MC$_{SilSil}$	107.3	102.49	103.42	104.64	105.02	104.71
S_5	MC	***107.95***	**101.2**	***99.72***	***97.34***	**95.57**	**94.6**
	MC$_{Angles}$	113.87	113.59	110.69	107.16	104.01	102.85
	MC$_{MaxMax}$	**103.28**	**101.2**	**99.65**	**96.74**	***96.82***	97.46
	MC$_{MaxMin}$	120.09	113.59	110.69	108.87	111.15	114.06
	MC$_{MaxSum}$	113.87	108.28	107.66	102.85	100.38	99.24
	MC$_{ParetoHist}$	114.8	***105.64***	105.69	101.57	99.86	***97.4***
	MC$_{SilHist}$	120.09	118.14	116.66	117.02	119.23	118.05
	MC$_{SilSil}$	**103.28**	**101.2**	**99.65**	**96.74**	***96.82***	97.46

Table 4.23: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *median MASE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are median values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.68**	**0.79**	**0.87**	**0.95**	**1.04**	**1.11**
	MC$_{Angles}$	0.82	0.92	1.03	1.15	1.26	1.35
	MC$_{MaxMax}$	0.77	0.88	0.96	1.06	1.17	1.28
	MC$_{MaxMin}$	0.77	0.87	0.96	1.04	1.13	1.23
	MC$_{MaxSum}$	0.74	0.84	0.92	1.01	1.09	1.19
	MC$_{ParetoHist}$	0.83	0.94	1.04	1.14	1.25	1.35
	MC$_{SilHist}$	***0.71***	***0.81***	***0.9***	***0.99***	***1.07***	***1.17***
	MC$_{SilSil}$	0.76	0.87	0.96	1.05	1.15	1.25
S_2	MC	**1.06**	**1.18**	**1.33**	**1.49**	**1.65**	**1.8**
	MC$_{Angles}$	1.24	1.42	1.58	1.76	1.9	2.03
	MC$_{MaxMax}$	1.23	1.37	1.54	1.72	1.88	2.03
	MC$_{MaxMin}$	1.14	1.27	1.43	1.58	1.74	1.9
	MC$_{MaxSum}$	1.2	1.34	1.5	1.65	1.8	1.94
	MC$_{ParetoHist}$	1.33	1.49	1.63	1.77	1.91	2.06
	MC$_{SilHist}$	***1.08***	***1.24***	***1.4***	***1.57***	***1.73***	***1.89***
	MC$_{SilSil}$	1.26	1.4	1.55	1.72	1.89	2.04
S_3	MC	**1.27**	**1.45**	**1.63**	**1.81**	**1.97**	**2.12**
	MC$_{Angles}$	1.48	1.67	1.83	1.96	2.13	2.35
	MC$_{MaxMax}$	1.57	1.74	1.95	2.12	2.33	2.53
	MC$_{MaxMin}$	1.37	1.55	1.78	1.97	2.14	2.33
	MC$_{MaxSum}$	1.45	1.64	1.83	1.98	2.15	2.39
	MC$_{ParetoHist}$	1.57	1.7	1.86	2.01	2.19	2.39
	MC$_{SilHist}$	***1.28***	***1.49***	***1.68***	***1.88***	***2.12***	***2.34***
	MC$_{SilSil}$	1.56	1.76	1.97	2.16	2.34	2.55
S_4	MC	**1.33**	**1.59**	**1.8**	**1.98**	**2.17**	**2.36**
	MC$_{Angles}$	1.59	1.76	1.98	2.17	2.37	2.55
	MC$_{MaxMax}$	1.65	1.82	2.02	2.22	2.42	2.6
	MC$_{MaxMin}$	1.47	1.67	1.89	2.09	2.3	2.48
	MC$_{MaxSum}$	1.6	1.77	1.98	2.14	2.34	2.51
	MC$_{ParetoHist}$	1.68	1.85	2.02	2.21	2.41	2.55
	MC$_{SilHist}$	***1.38***	***1.59***	***1.83***	***2.05***	***2.26***	***2.45***
	MC$_{SilSil}$	1.63	1.81	2.01	2.21	2.39	2.58
S_5	MC	**1.47**	**1.67**	**1.86**	**2.07**	**2.29**	**2.48**
	MC$_{Angles}$	1.64	1.84	2.03	2.22	***2.4***	***2.59***
	MC$_{MaxMax}$	1.72	1.92	2.11	2.28	2.49	2.73
	MC$_{MaxMin}$	1.54	1.79	1.97	2.2	2.44	2.65
	MC$_{MaxSum}$	1.66	1.86	2.06	2.22	2.42	2.6
	MC$_{ParetoHist}$	1.68	1.88	2.07	2.23	2.43	2.61
	MC$_{SilHist}$	***1.5***	***1.74***	***1.95***	***2.16***	2.42	2.67
	MC$_{SilSil}$	1.74	1.92	2.1	2.27	2.47	2.69

Table 4.24: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *median ME* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are median values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{Angles}$	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{MaxMax}$	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{MaxMin}$	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{MaxSum}$	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{ParetoHist}$	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{SilHist}$	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	MC$_{SilSil}$	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
S_2	MC	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	MC$_{Angles}$	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	***0.05***
	MC$_{MaxMax}$	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	MC$_{MaxMin}$	**0.01**	***0.02***	**0.03**	**0.03**	**0.04**	***0.05***
	MC$_{MaxSum}$	**0.01**	***0.02***	**0.03**	**0.03**	**0.04**	***0.05***
	MC$_{ParetoHist}$	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	***0.05***
	MC$_{SilHist}$	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	MC$_{SilSil}$	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
S_3	MC	**-0.03**	**-0.02**	**-0.03**	**-0.05**	**-0.05**	**-0.06**
	MC$_{Angles}$	**-0.03**	***-0.01***	**-0.03**	***-0.04***	**-0.05**	***-0.05***
	MC$_{MaxMax}$	**-0.03**	***-0.01***	**-0.03**	***-0.04***	**-0.05**	***-0.05***
	MC$_{MaxMin}$	***-0.02***	**-0.02**	**-0.03**	***-0.04***	**-0.05**	***-0.05***
	MC$_{MaxSum}$	***-0.02***	***-0.01***	**-0.03**	***-0.04***	**-0.05**	***-0.05***
	MC$_{ParetoHist}$	**-0.03**	***-0.01***	**-0.03**	***-0.04***	**-0.05**	***-0.05***
	MC$_{SilHist}$	***-0.02***	***-0.01***	**-0.03**	***-0.04***	**-0.05**	**-0.06**
	MC$_{SilSil}$	**-0.03**	***-0.01***	**-0.03**	***-0.04***	**-0.05**	***-0.05***
S_4	MC	**0.02**	**0.05**	**0.07**	**0.09**	**0.11**	**0.11**
	MC$_{Angles}$	**0.02**	**0.05**	***0.08***	***0.1***	***0.12***	***0.12***
	MC$_{MaxMax}$	**0.02**	**0.05**	***0.08***	***0.1***	**0.11**	***0.12***
	MC$_{MaxMin}$	**0.02**	***0.06***	***0.08***	***0.1***	**0.11**	***0.12***
	MC$_{MaxSum}$	**0.02**	**0.05**	***0.08***	**0.09**	**0.11**	***0.12***
	MC$_{ParetoHist}$	**0.02**	**0.05**	***0.08***	**0.09**	**0.11**	***0.12***
	MC$_{SilHist}$	***0.03***	***0.06***	0.09	***0.1***	***0.12***	***0.12***
	MC$_{SilSil}$	**0.02**	**0.05**	***0.08***	***0.1***	**0.11**	***0.12***
S_5	MC	**-0.01**	**0.03**	**0.03**	**0.04**	**0.05**	**0.06**
	MC$_{Angles}$	**-0.01**	**0.03**	**0.03**	0.06	0.07	***0.07***
	MC$_{MaxMax}$	**-0.01**	**0.03**	**0.03**	0.06	0.07	***0.07***
	MC$_{MaxMin}$	**-0.01**	**0.03**	**0.03**	0.06	***0.06***	**0.06**
	MC$_{MaxSum}$	***0***	**0.03**	**0.03**	0.06	***0.06***	**0.06**
	MC$_{ParetoHist}$	**-0.01**	**0.03**	**0.03**	0.06	***0.06***	***0.07***
	MC$_{SilHist}$	***0***	**0.03**	***0.04***	***0.05***	***0.06***	**0.06**
	MC$_{SilSil}$	**-0.01**	**0.03**	**0.03**	0.06	0.07	***0.07***

Table 4.25: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *median MSE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are median values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	**0.08**	**0.1**	**0.13**	**0.16**	**0.19**	**0.22**
	MC$_{Angles}$	0.14	0.18	0.21	0.26	0.31	0.37
	MC$_{MaxMax}$	***0.09***	0.12	0.15	0.2	0.23	0.28
	MC$_{MaxMin}$	0.11	0.14	0.17	0.21	0.24	0.29
	MC$_{MaxSum}$	***0.09***	0.12	***0.14***	***0.17***	***0.2***	0.25
	MC$_{ParetoHist}$	0.14	0.18	0.23	0.28	0.33	0.39
	MC$_{SilHist}$	**0.08**	***0.11***	**0.13**	**0.16**	**0.19**	***0.24***
	MC$_{SilSil}$	***0.09***	0.12	0.15	0.19	0.23	0.28
S_2	MC	**0.21**	**0.27**	**0.35**	**0.46**	**0.55**	**0.65**
	MC$_{Angles}$	0.35	0.43	0.53	0.66	0.79	0.93
	MC$_{MaxMax}$	0.32	0.43	0.55	0.69	0.85	1.03
	MC$_{MaxMin}$	0.26	0.34	0.44	***0.56***	***0.69***	***0.84***
	MC$_{MaxSum}$	0.32	0.42	0.48	0.59	0.76	0.92
	MC$_{ParetoHist}$	0.39	0.5	0.58	0.72	0.85	0.99
	MC$_{SilHist}$	***0.24***	***0.32***	***0.42***	0.57	0.7	0.88
	MC$_{SilSil}$	0.32	0.43	0.55	0.7	0.86	1.03
S_3	MC	**0.34**	**0.49**	**0.63**	**0.75**	**0.95**	**1.15**
	MC$_{Angles}$	0.56	0.71	0.87	1.05	1.25	1.51
	MC$_{MaxMax}$	0.6	0.76	0.98	1.21	1.45	1.81
	MC$_{MaxMin}$	0.44	0.58	0.76	0.94	1.18	1.41
	MC$_{MaxSum}$	0.53	0.7	0.87	1.05	1.27	1.51
	MC$_{ParetoHist}$	0.64	0.8	0.96	1.11	1.32	1.54
	MC$_{SilHist}$	***0.38***	***0.53***	***0.71***	***0.9***	***1.13***	***1.43***
	MC$_{SilSil}$	0.59	0.76	0.96	1.21	1.46	1.78
S_4	MC	**0.48**	**0.7**	**0.93**	**1.2**	**1.47**	**1.75**
	MC$_{Angles}$	0.75	0.95	1.23	1.49	***1.82***	***2.19***
	MC$_{MaxMax}$	0.78	0.98	1.28	1.58	1.94	2.32
	MC$_{MaxMin}$	0.71	0.92	1.18	1.48	1.84	2.24
	MC$_{MaxSum}$	0.79	0.98	1.25	1.51	1.86	2.23
	MC$_{ParetoHist}$	0.85	1.08	1.31	1.64	1.98	2.35
	MC$_{SilHist}$	***0.56***	***0.79***	***1.07***	***1.47***	1.88	2.25
	MC$_{SilSil}$	0.77	0.98	1.28	1.59	1.94	2.31
S_5	MC	**0.64**	**0.88**	**1.12**	**1.38**	**1.69**	**2.01**
	MC$_{Angles}$	0.87	1.14	1.37	1.65	1.98	2.35
	MC$_{MaxMax}$	0.91	1.16	1.4	1.63	1.95	2.29
	MC$_{MaxMin}$	0.81	1.09	1.35	1.71	2.09	2.53
	MC$_{MaxSum}$	0.88	1.12	***1.31***	***1.55***	1.86	2.18
	MC$_{ParetoHist}$	0.9	1.11	***1.31***	***1.55***	***1.83***	***2.11***
	MC$_{SilHist}$	***0.72***	***1.01***	1.35	1.73	2.22	2.77
	MC$_{SilSil}$	0.91	1.15	1.35	1.62	1.96	2.3

Table 4.26: In-depth comparison of different model selection methods on C-MSKF's forecasting accuracy *median sMAPE* on the simulated data, broken up by noise levels and forecasting horizons, from 1 to 6. Shown are median values across 30 replicates and 6 different time series length. The best performance obtained for each setting is highlighted in bold faces, with the second best performance highlighted in italic bold face.

Scenarios	Methods	1-period	2-period	3-period	4-period	5-period	6-period
S_1	MC	***22.72***	***21.79***	**20.91**	**19.99**	**19.92**	**19.66**
	MC$_{Angles}$	27.06	26.17	24.64	24.02	23.45	23.11
	MC$_{MaxMax}$	25.07	23.99	22.92	22.31	22.13	22.13
	MC$_{MaxMin}$	24.21	23.17	22.17	21.68	21.02	20.85
	MC$_{MaxSum}$	24.12	22.88	21.83	21.45	20.99	20.7
	MC$_{ParetoHist}$	27.24	26.18	25.26	24.6	24.27	24.22
	MC$_{SilHist}$	**22.55**	**21.6**	***21.12***	***20.7***	***20.45***	***20.27***
	MC$_{SilSil}$	24.66	23.74	22.86	22.26	21.95	21.95
S_2	MC	**36.95**	**35.22**	**34.03**	**33.72**	**33.62**	**33.05**
	MC$_{Angles}$	43.48	42.88	40.91	40.15	39.56	39.24
	MC$_{MaxMax}$	41.52	39.89	39.3	38.97	39.06	39.13
	MC$_{MaxMin}$	37.43	37.51	36.14	***35.62***	***34.6***	***34.59***
	MC$_{MaxSum}$	41.32	40.91	39.87	39.04	37.84	37.7
	MC$_{ParetoHist}$	43.58	42.7	40.72	39.37	38.58	38.43
	MC$_{SilHist}$	***37.84***	***35.76***	***35.92***	36.18	36.15	36.35
	MC$_{SilSil}$	42.82	41.29	39.97	40.07	39.89	40.17
S_3	MC	55.35	***53.77***	***53.14***	**52.19**	**51.75**	**51**
	MC$_{Angles}$	60.86	59.06	58.31	56.83	55.24	54.34
	MC$_{MaxMax}$	63	63.28	63.5	62.57	62.07	61.54
	MC$_{MaxMin}$	***55.23***	55.58	54.83	54	53.42	53.18
	MC$_{MaxSum}$	58.42	57.46	56.82	55.81	54.69	54
	MC$_{ParetoHist}$	61.01	59.27	58.26	56.38	55.16	54.36
	MC$_{SilHist}$	**52.64**	**53.22**	**53**	***53.28***	***53.15***	***52.79***
	MC$_{SilSil}$	63.87	64.04	63.5	62.75	62.71	62.23
S_4	MC	61.53	***61.05***	***61.24***	**59.77**	**60.67**	**60.23**
	MC$_{Angles}$	63.28	62.2	61.89	61.32	***61.38***	***61.56***
	MC$_{MaxMax}$	67.77	65.94	67.17	66.72	66.18	65.2
	MC$_{MaxMin}$	***61.34***	61.47	61.97	61.88	62.17	62.36
	MC$_{MaxSum}$	66.27	64.33	64.18	63.77	63.99	63.55
	MC$_{ParetoHist}$	66.8	66.01	64.7	64.11	64.04	63.6
	MC$_{SilHist}$	**56.43**	**57.84**	**59.64**	60.72	61.61	62.22
	MC$_{SilSil}$	67.08	65.71	67.1	66.49	65.87	65.17
S_5	MC	***65.63***	***65.46***	**64.37**	**63.96**	**64.74**	**64.93**
	MC$_{Angles}$	69.04	69.08	68.91	68.16	67.71	67.46
	MC$_{MaxMax}$	73.57	71.84	70.05	69.63	69.78	70.38
	MC$_{MaxMin}$	68.06	70.17	70.96	71.24	70.86	71.17
	MC$_{MaxSum}$	70.42	69.45	68.8	67.74	67.6	67.98
	MC$_{ParetoHist}$	70.62	69.45	68.7	67.77	67.51	67.39
	MC$_{SilHist}$	**62.89**	**64.68**	***65.22***	***64.56***	***66.33***	***67.23***
	MC$_{SilSil}$	72.55	71.39	69.78	69.35	69.76	70.32

References

[1] J. S. Armstrong. “Findings from evidence-based forecasting: Methods for reducing forecast error”. In: *International Journal of Forecasting* 22.3 (2006), pp. 583–598.

[2] J. S. Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Vol. 30. Springer Science & Business Media, 2001.

[3] F. M. Bass. “A new product growth for model consumer durables”. In: *Management Science* 15.5 (1969), pp. 215–227.

[4] U Baumgartner, C. Magele, and W Renhart. “Pareto optimality and particle swarm optimization”. In: *IEEE Transactions on magnetics* 40.2 (2004), pp. 1172–1175.

[5] S. Bechikh, L. Ben Said, and K. Ghédira. “Searching for knee regions in multi-objective optimization using mobile reference points”. In: *Proceedings of the 2010 ACM symposium on applied computing*. ACM. 2010, pp. 1118–1125.

[6] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. “Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation”. In: *International Journal of Forecasting* 32.2 (2016), pp. 303–312.

[7] J. Branke et al. “Finding knees in multi-objective optimization”. In: *PPSN*. Vol. 3242. 2004, pp. 722–731.

[8] I. Das. “On characterizing the “knee” of the Pareto curve based on normal-boundary intersection”. In: *Structural and Multidisciplinary Optimization* 18.2 (1999), pp. 107–115.

[9] K. Deb and J Sundar. “Reference point based multi-objective optimization using evolutionary algorithms”. In: *Proceedings of the 8th annual conference on Genetic and evolutionary computation*. ACM. 2006, pp. 635–642.

[10] E. Dimitriadou, S. Dolničar, and A. Weingessel. “An examination of indexes for determining the number of clusters in binary data sets”. In: *Psychometrika* 67.1 (2002), pp. 137–159.

[11] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[12] G. Duncan, W. Gorr, and J. Szczypula. *Comparative Study of Cross Sectional Methods for Time Series With Structural Changes*. Tech. rep. Carnegie Mellon University, 1994.

[13] G. T. Duncan, W. Gorr, and J. Szczypula. “14 Bayesian Hierarchical Forecasts for Dynamic Systems: Case Study on Backcasting School District Income Tax”. In: *New Directions in Spatial Econometrics* (2012), p. 322.

[14] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[15] P. Goodwin, K. Dyussekeneva, and S. Meeran. “The use of analogies in forecasting the annual sales of new electronics products”. In: *IMA Journal of Management Mathematics* 24.4 (2013), pp. 407–422.

[16] K. C. Green and J. S. Armstrong. “Structured analogies for forecasting”. In: *International Journal of Forecasting* 23.3 (2007), pp. 365–376.

[17] N. P. Greis and C. Z. Gilstein. “Empirical Bayes methods for telecommunications forecasting”. In: *International Journal of Forecasting* 7.2 (1991), pp. 183–197.

[18] I. Guyon, U. Von Luxburg, and R. C. Williamson. “Clustering: Science or art”. In: *NIPS 2009 Workshop on Clustering Theory*. 2009, pp. 1–11.

[19] J. Handl and J. Knowles. “An evolutionary approach to multiobjective clustering”. In: *IEEE transactions on Evolutionary Computation* 11.1 (2007), pp. 56–76.

[20] J. Handl and J. Knowles. "Multiobjective clustering with automatic determination of the number of clusters". In: *Technical Report* (2004).

[21] J. K. Handl and Kell. "Multiobjective approaches to the data-driven analysis of biological systems". PhD thesis. University of Manchester, 2006.

[22] P. J. Harrison and C. F. Stevens. "A Bayesian approach to short-term forecasting". In: *Operational Research Quarterly* (1971), pp. 341–362.

[23] L. Hubert and P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218.

[24] R. J. Hyndman. "Another look at forecast-accuracy metrics for intermittent demand". In: *Foresight: The International Journal of Applied Forecasting* 4.4 (2006), pp. 43–46.

[25] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.

[26] R. E. Kass and D. Steffey. "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)". In: *Journal of the American Statistical Association* 84.407 (1989), pp. 717–726.

[27] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[28] Y. Liu et al. "Multicriterion market segmentation: a new model, implementation, and evaluation". In: *Marketing Science* 29.5 (2010), pp. 880–894.

[29] E. Lu and J. Handl. "Multicriterion Segmentation of Demand Markets to Increase Forecasting Accuracy of Analogous Time Series: A First Investigation". In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 379–388.

[30] C. A. Mattson, A. A. Mullur, and A. Messac. “Smart Pareto filter: Obtaining a minimal representation of multiobjective design space”. In: *Engineering Optimization* 36.6 (2004), pp. 721–740.

[31] G. W. Milligan and M. C. Cooper. “A study of standardization of variables in cluster analysis”. In: *Journal of Classification* 5.2 (1988), pp. 181–204.

[32] L. Parsons, E. Haque, and H. Liu. “Subspace clustering for high dimensional data: a review”. In: *ACM SIGKDD Explorations Newsletter* 6.1 (2004), pp. 90–105.

[33] F. Petropoulos et al. “‘Horses for Courses’ in demand forecasting”. In: *European Journal of Operational Research* 237.1 (2014), pp. 152–163.

[34] M. I. Piecyk and A. C. McKinnon. “Forecasting the carbon footprint of road freight transport in 2020”. In: *International Journal of Production Economics* 128.1 (2010), pp. 31–42.

[35] L. Rachmawati and D. Srinivasan. “Preference incorporation in multi-objective evolutionary algorithms: A survey”. In: *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*. IEEE. 2006, pp. 962–968.

[36] P. J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[37] S. Saha and S. Bandyopadhyay. “Automatic MR brain image segmentation using a multiseed based multiobjective clustering approach”. In: *Applied Intelligence* 35.3 (2011), pp. 411–427.

[38] O. Schütze, M. Laumanns, and C. A. C. Coello. “Approximating the Knee of an MOP with Stochastic Search Algorithms.” In: *PPSN*. Springer. 2008, pp. 795–804.

[39] D. Steinley and M. J. Brusco. “A new variable weighting and selection procedure for K-means cluster analysis”. In: *Multivariate Behavioral Research* 43.1 (2008), pp. 77–108.

[40] C. A. Sugar. “Techniques for clustering and classification with applications to medical problems.” In: (1999).

[41] C. A. Sugar, L. A. Lenert, and R. A. Olshen. “An application of cluster analysis to health services research: Empirically defined health states for depression from the sf-12”. In: (1999).

[42] U. Von Luxburg. “Clustering stability: an overview”. In: *Foundations and Trends® in Machine Learning* 2.3 (2010), pp. 235–274.

[43] R. Webby and M. O’Connor. “Judgemental and statistical time series forecasting: a review of the literature”. In: *International Journal of Forecasting* 12.1 (1996), pp. 91–118.

Chapter 5

Bagging approaches to the forecasting of analogies (paper 3)

5.1 Abstract

Analogies have been well recognized in the area of time series forecasting. They have been widely applied to help improve the performance of forecasting processes in situations where analogies are present and can be identified. Within the previous literature, clustering approaches have shown promise in supporting the identification of meaningful groupings of analogies and thus lead to improved forecasting results. Nevertheless, the grouping of analogies during the segmentation stage introduces additional instabilities associated with the clustering procedure. In part, these instabilities stem from the model selection step of the clustering process. For example, clustering techniques such as non-parametric clustering or hierarchical clustering approaches all require the identification of an appropriate number of clusters, and inaccuracies in the estimation of the number of clusters will cause the incorrect splitting or merging of true clusters. Additionally, the random initialization step required by certain clustering algorithms potentially introduces further noise. On account of this, we integrate the notion of bootstrap aggregation into the forecasting process for handling the issue of instabilities. By perturbing input data at the segmentation stage, we demonstrate that bootstrap aggregation techniques can give rise to significant gains in the forecasting accuracy of the prediction process.

Keywords: Bayesian pooling; Bootstrap aggregating; Clustering; Kalman filter;

5.2 Introduction

The importance of analogies has been well recognized in the field of time series forecasting (*e.g.,* Glantz, 1991; Murawski, 1993; Goodwin, Dyussekeneva, and Meeran, 2013). Judgmental forecasting approaches utilize analogies to adjust the final forecasts which are subject to forecasters' over-optimistic views or wishful thinking (Hyndman and Athanasopoulos, 2014). Statistical forecasting models make use of analogies to boost the accuracy of point forecasts. For example, the Bass model (Bass, 1969; Nikolopoulos et al., 2016) was employed to forecast sales of products shortly after the launch of new products by integrating information available from similar products (Goodwin, Dyussekeneva, and Meeran, 2013).

Analogies play a crucial role for the forecasting approaches that employ such analogies. However, the use of heterogeneous time series tends to yield poorer forecasting results than those of homogeneous analogies (Duncan, Gorr, and Szczypula, 2001). The homogeneity of analogies is important for the effectiveness of forecasting processes where analogies are needed (Stimson, 1985). This indicates that the appropriate identification of analogies can be essential to the success of forecasting processes, where analogies are required.

In general, within the literature, clustering techniques have been proposed to tackle the analogy identification challenge. This could potentially comprise the application of single-criterion and multi-criteria approaches (see **Chapter 3**). As demonstrated in previous work (Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001), a proper grouping of analogies might help to obtain additional gains in the final accuracy of forecasts that cannot be achieved by traditional statistical forecasting models. This is because conventional statistical models are methodologically limited to take into account additional information that is not present within the time series patterns.

Principally, almost all statistical models are plagued by different sources of instabilities that can stem from the input data, the model parameters and/or the inaccurate assumption of model structures. Unfortunately, the clustering of analogies during the segmentation stage of the overall prediction process might introduce additional instabilities. These additional instabilities mainly stem from the model selection step during the clustering procedure. For instance, the incorrect determination of a number of clusters can lead to either wrong split or merge of partitions; the random initialization of a clustering algorithm might yield different clustering results each time.

The bootstrap aggregation (bagging) technique proposed by Breiman (1996) has been commonly applied to address the instability issues presented by statistical models. Through the reduction of instabilities, bagging techniques aim to additionally increase the accuracy of the forecasting process. The main advantage of this technique is to lower down the variability with the final statistical forecast through the application of the combination scheme such as the mean, median, trimmed mean or weighted mean. By combining the forecasts derived from multiple models, the aggregated forecast is expected to increase the accuracy of the results via the instability reduction. In principle, Bagging can obtain an improvement in the performance of forecasting if the perturbation of the learning sets leads to significant changes in the constructed statistical models (Breiman, 1996).

In light of this, we experiment with integrating the bootstrap aggregation technique into the forecasting process that makes use of analogies. Instead of directly bootstrapping time series data, which is challenging for non-stationary time series, we reduce the problem to a typical problem of IID bootstrap by resampling a set of labels associated the time series data, and the model is constructed on each bootstrapped sample. Specifically, we regard this set of labels as random variables that follow identical and independent distribution (IID). We assume a time series data set with size of n and **X** $= \{X_1, X_2, \ldots, X_n\}$, where X_i is a time series with measurements $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$. A bootstrapped sample **X**$^b = \{X_1^b, X_2^b, \ldots, X_n^b\}$ of size n is generated by resampling the set of labels associated with the time series without perturbing the internal structure of the time series. Thus, $X_i^b = (x_{i1}, x_{i2}, \ldots, x_{in})$, where measures of X_i^b remain the same as X_i. We apply a simple mean method to obtain aggregated forecasts derived from different bootstrapped samples.

Here, bootstrapped samples are obtained at different stages of the prediction process that employs analogies: (i) The first class of methods resample from the original order with replacement during the stage of segmentation. Subsequently, a suitable model selection method is applied to the generated bootstrapped samples, where the forecasting algorithm is then used to forecast the resultant clusters of analogous time series; (ii) The second class of methods bootstraps directly on analogies, which were identified at the segmentation stage, and further combines point forecasts across bootstrapped samples. Based on objective forecasting methods, *i.e.,* the Cross-Sectional State Kalman Filter algorithm, our findings demonstrate that by perturbing the data at different stages of the process, bagging methods are capable of boosting the forecasting accuracy of forecasting methods make use of analogies.

The rest of the paper is organized as follows: Section 5.3 surveys previous work related to the concept of bagging and multicriteria clustering. Section 5.4 presents the details of the prediction process that makes use of analogies. Section 5.5 describes the overall bagging procedures. Section 5.6 provides details regarding the design of the experiments based on simulated data. Section 5.7 analyzes results of the experiments, and Section 5.8 concludes the paper.

5.3 Previous work

Analogies have been widely employed in the forecasting field for boosting the forecasting accuracy (Easingwood, 1989; Armstrong, 2006; Green and Armstrong, 2007; Piecyk and McKinnon, 2010). Most commonly, analogies have been utilized in judgmental forecasting, where forecasting by analogy approaches make use of analogies to adjust the point forecasts derived from statistical forecasts (Hyndman and Athanasopoulos, 2014; Webby and O'Connor, 1996). The idea behind this is to reduce biases caused by forecasters' over-optimistic view or wishful thinking (Armstrong, 2001; Petropoulos et al., 2014).

According to Duncan, Gorr, and Szczypula (2001), analogies can be defined as time series that show similarity in terms of time-based patterns that are correlated over time. Recent work of Lu and Handl (2015) further refined the definition of analogies as a set of time series show similarity in term of time-based patterns and hypothesized factors that govern the behavior of the observed patterns. Within the forecasting literature, a few research studies have been conducted to help identify analogies in an objective manner.

Specifically, clustering techniques have been applied during the segmentation stage of the overall prediction process where analogies are involved. These clustering approaches have focused on partitioning a set of time series into distinctive groups where each group contains homogeneous time series. Time series is homogeneous if the time-based patterns are similar and/or the causal factors that drive these patterns are similar (Duncan, Gorr, and Szczypula, 2001; Lu and Handl, 2015). Previous work shows that segmentation approaches to the group of analogies can provide better clustering results and then translate to improved forecasting results. However, a major issue associated with segmentation approaches lies in the model selection step of the clustering procedures, where clustering results can be unstable and vary from time to time.

In more details, forecasting processes utilize analogies are inevitably facing the challenge of additional instabilities primarily stemming from the model selection process. Specifically, theses processes are involved with determining the final partitions of the analogies. Illustrated by (Guyon, Von Luxburg, and Williamson, 2009), clustering methods that require the determination of the number of clusters might exhibit instabilities as the incorrect choose of the number of clusters can lead to either wrong split or merge of the “true” clusters. In addition, the random initialization of clustering algorithms (if needed) can also return unstable clustering results at each run. Similar to all statistical forecasts, there might be different sources of instabilities stemming from the input data, model parameters and/or model structures. However, the identification of analogies which was proved to be crucial for the success of such forecasting process might arise additional issues of instabilities.

To handle the instabilities, bagging methods have been extensively studied in the forecasting field with the aim of boosting the final accuracy (*e.g.,* Breiman, 1996). Bagging methods are declared to be able to yield substantial gains in the model performance if the perturbation of learning sets can lead to different model results (Breiman, 1996). The bootstrap of time series data is not difficult if one has a finite-dimensional parametric model (*e.g.,* a finite-order ARMA model) that reduces the data generation process to independent random sampling. In this case and under suitable conditions, the bootstrap has the same properties to a random sample from a distribution (see Bose, 1988; Bose, 1990). Such approaches are inconsistent if the model used for resampling is mis-specified.

In terms of direct time series bootstrap, the most common approach to bootstrap time series is to resample “blocks” of sequential observations instead of resampling independent data observations. This preserves the dependence structure of the underlying process within the resampled blocks and can reproduce the effect of dependence at short lags. A typical example is the local moving block method (MBB) (Kunsch, 1989). These methods have shown particular successes for stationary time series. The bootstrap of non-stationary time series is more challenging as the notion of ”block” cannot be directly applied because the resampling of blocks can interrupt the dependent structure. One recent work related to this topic is referred to Bergmeir, Hyndman, and Benítez (2016), who firstly transformed the non-stationary time series using a Box-Cox transformation by decomposing the time series into trend, seasonal and remainder components. Later, MBB methods were applied to generate bootstrapped residuals. The final bootstrapped samples were constructed by transforming components back to time series. This method demonstrates better forecasting results of the exponential smoothing method limited in the monthly series of M-3 competition data. One earlier research study is provided in Cordeiro and Neves (2009), who employed a sieve bootstrap to perform bagging with ETS models. Unfortunately, the overall results are not promising. In fact, the bagged forecasts are often not as good as the original forecasts applied to the original time series.

5.4 The forecasting process makes use of analogies

This section details main components that are utilized in the prediction process where analogies are required. Ultimately, we aim to integrate the bootstrap aggregation concept into the overall prediction process discussed here to obtain a further gain in forecasting accuracy. The gain is expected by constructing an aggregated forecast that reduces the bias and variability resulted from the clustering procedure.

In general, there are three steps involved in the prediction process: (i) The segmentation of analogies partitions a set of time series into clusters based on criteria considered. The clustering process can make use of either single criterion or multiple criteria. The choice is dependent on the final application purpose and the availability of information data sources. (ii) The second step of the prediction process employs a pooling approach, namely the Cross-Sectional State Kalman Filter (C-MSKF method). This algorithm is capable of utilizing information available from analogies. We use the C-MSKF method to demonstrate the performance of various bagging procedures developed in this article. (iii) The notion of bagging based on IID bootstrap is integrated into the overall prediction process that comprises Step (i) and Step (ii). The overall bagging procedures are described later in this section.

5.4.1 Segmentation of analogies

In the context of forecasting, analogies have been widely employed to help the improvement of forecasting results derived from the statistical forecasts. To identify analogies, both single-criterion and multicriteria segmentation approaches have been investigated. These approaches make use of information from past realizations of time series and/or the causal factors that drive the patterns observed. The motivation of using these information sources is referred to our previous work Lu and Handl (2015). Here, we focus more on the development of methodologies that could generate better forecasting results via the instability reduction.

Specifically, three clustering approaches are considered during the segmentation step for the grouping of analogies: single-criterion clustering of time series data (TS clustering), single-criterion clustering of causal factors (CF clustering) and multicriteria clustering approaches that make use of both information sources (MC).

5.4.1.1 The choice of distance metric

CF clustering. The first information source is derived from causal factors that cause the time-based patterns underlying the analogies. We apply a squared Euclidean distance metric to measure the hypothesized causal factor by assuming a single numeric value is associated with each analogous time series. Specifically, $\delta_{CF}(i,j)$ denotes time series i and j. The equation applied to calculate the distance between the two-time series is given as:

$$\delta^{CF}(i,j) = \sum_{m}(a_{im} - a_{jm})^2 \tag{5.1}$$

where $\delta^{CF}(i,j) = d_{ij}^{CF}$ and d_{ij}^{CF} are elements of the dissimilarity matrix $\mathbf{D^{CF}}$; a_{im} and a_{jm} represent the values of causal variable m associated with time series i and j, respectively. Further, the z-score method is used to eliminate the scale differences among time series.

TS clustering. In terms of time series information, we measure the dissimilarity between time series based on Pearson correlation coefficients. Assume time series i and j, the distance between them is represented as d_{ij}^{TS} and calculated as follows:

$$\delta^{TS}(i,j) = 1 - \frac{T(\sum_{t} x_{it}x_{jt}) - (\sum_{t} x_{it})(\sum_{t} x_{jt})}{\sqrt{(T(\sum_{t} x_{it}^2) - (\sum_{t} x_{it})^2)(T(\sum_{t} x_{jt}^2) - (\sum_{t} x_{jt})^2)}} \tag{5.2}$$

where a dissimilarity matrix based on time series information defines as $\mathbf{D^{TS}} = (d_{ij}^{TS})$; $d_{ij}^{TS} = \delta^{TS}(i,j)$; t is the index of time $t = 1,2,...,T$; T is the number of time steps used for measuring correlation; x_{it} and x_{jt} describe the values of time series i and j over time, respectively.

MC clustering. With respect to MC clustering, the dissimilarity matrices derived from causal factors and time series patterns are combined at the distance function level through a weighting scheme. To reduce scale differences, we update each element of the dissimilarity matrices d_{ij}^{CF} and d_{ij}^{TS} using the Min-max method (see Equations (5.3) and (5.4)):

$$d_{ij}^{TS} \leftarrow \frac{d_{ij}^{TS} - min(\mathbf{D^{TS}})}{max(\mathbf{D^{TS}}) - min(\mathbf{D^{TS}})} \tag{5.3}$$

$$d_{ij}^{CF} \leftarrow \frac{d_{ij}^{CF} - min(\mathbf{D^{CF}})}{max(\mathbf{D^{CF}}) - min(\mathbf{D^{CF}})} \tag{5.4}$$

To combine the two information sources, the distance matrix $\mathbf{D}_{\omega}^{\mathbf{MC}} = (d_{ij\omega}^{MC})$ is obtained through a weighted-sum method, where the relative weight ω varied from 0 to 1 with increments of 0.10;

$$d_{ij\omega}^{MC} = (1-\omega) \times d_{ij}^{CF} + \omega \times d_{ij}^{TS} \tag{5.5}$$

Further, we employ a standard clustering technique, namely PAM clustering (Kaufman and Rousseeuw, 1990), to the clustering procedure. The main rationale of using this technique is due to its capability of combining incommensurable variables that can be easily realized using dissimilarity matrices[1]. In our context, equally-sized partitionings are considered advantageous for the forecasting analysis. PAM clustering approaches tend to yield such clusters. To reduce chances of generating local optima, we repeat the clustering procedure 30 times. Among these, the clustering solution with the minimal sum of within-cluster dissimilarities is chosen for the further analysis.

5.4.1.2 Model selection

- **Selection of the number of clusters**

For an automatic determination of the number of clusters, the Silhouette Width measure (Rousseeuw, 1987) is applied. The Silhouette Width measure has been widely employed to determine the number of clusters using the internal data structure alone.

[1] Clustering methods operate on dissimilarity matrix are applicable

This measure focuses on the compromise between the cluster cohesion and separation. Assume a data set with n objects, these objects can be grouped into k clusters, $k = 1, 2, \ldots, n$, through a suitable clustering approach. The Silhouette values will be calculated for each clustering solution. The clustering solution returns the largest mean Silhouette value is selected. The Silhouette value Sil takes values from [-1,1], and a larger value means a better clustering result. The calculation is presented as follows:

$$Sil(i) = \frac{b_i - c_i}{max(c_i, b_i)} \tag{5.6}$$

where c_i represents the average distance between object i and all data objects in the same partition; b_i denotes the average distance between i and all data objects in the closest other partition; The closest partition is defined as the one with the minimum b_i. The Silhouette Width of the entire partition is then calculated as the mean Silhouette Width of all data objects.

- **Selection of weight for multicriteria approaches**

In addition, MC clustering requires a further step in model selection, as for the same number of k, the clustering approach might return multiple solutions. Each weight interval might correspond to an individual clustering solution. In consistent with **Chapter 3** and **4**, the best partitioning is determined by calculating the historical average performance of the forecasting algorithm implemented in the prediction process. The main rationale behind this is the success of the clustering is best reflected by the overall success of the application (forecasting) (Guyon, Von Luxburg, and Williamson, 2009).

Specifically, the prediction origin $t = T$ is used to support this part of the analysis. Historical data points ($t \leq T$) are used during the clustering procedure. To measure the historical average forecasting performance, the choice of historical period and whether the data points should be included during the clustering procedure might have impacts on the forecasting accuracy. We employ the best method reported by the previous work, and more details are systematically analyzed in **Chapter 4**.

The accuracy measure used for the weight selection is the Mean Square Error (MSE) and given as:

$$MSE = mean(e_t^2) \tag{5.7}$$

where t is the time index; $e_t = X_t - F_t$, X_t is the actual observation of the time series X at time t; F_t is the respective forecast made at time t.

5.4.2 The C-MSKF forecasting algorithm

Here, we employ the C-MSKF algorithm during the forecasting stage. This approach is able to pool information from analogies and has shown promising results in dealing short and volatile time series. C-MSKF methods have demonstrated some successes in applications, *e.g.,* forecasting of churn in telecommunication networks (Greis and Gilstein, 1991), infant mortality rates (Duncan, Gorr, and Szczypula, 2012) and tax revenue (Duncan, Gorr, and Szczypula, 1993). In principle, the C-MSKF method was as a representative example, but the applicability of other pooling approaches in this framework is expected. Particularly, the idea of bagging lies in the perturbation of the labels of the time series data without interrupting the mechanisms of the forecasting algorithm.

In general, C-MSKF methods combine of the capability of the Multi-State Kalman Filter (MSKF: Harrison and Stevens, 1971) with the Conditionally Independent Hierarchical Model (CIHM: Kass and Steffey, 1989) using DGS's shrinkage formula (DGS's shrinkage: Duncan, Gorr, and Szczypula, 1993). The more detailed description regarding the C-MSKF algorithm is available in the literature (Duncan, Gorr, and Szczypula, 1993) and the Appendix. To handle the structural change presented in the time series patterns, C-MSKF methods were claimed to show better results when allowing at least three observations after the change caused by the impact of an external influence. Furthermore, a key assumption of the C-MSKF is that time series that are regarded as analogous during the investigation period do not frequently diverge in the forecasting period (Duncan, Gorr, and Szczypula, 1994; Duncan, Gorr, and Szczypula, 2001).

5.5 The overall bagging procedures

Based on the prediction process described in the previous sections, we further detail the procedure of bagging to the forecasting of analogies. Generally, the prediction process comprises three major components: (i) The clustering of analogies using a suitable clustering approach which can be a single-criterion or multicriteria clustering approach; (ii) the employment of the C-MSKF algorithm, which utilizes information from analogies that have been identified in Step (i); (iii) Bootstrapped samples are generated at different stages of such prediction process. At the aggregation step, point forecasts derived from different bootstrapped samples are combined using the mean method. In most cases, simple combination methods often work reasonably well relative to more complex combinations (Clemen, 1989).

Firstly, a word about notations: generically assume that $L = (l_1, l_2, \ldots, l_n)$ refers to a set (vector) of labels associated with each time series, respectively in a data set or a cluster. L collects a set of random variables that follow IID distribution. The time series data is denoted as $\mathbf{X}^{TS} = (x_{it}^{TS})$, where i refers to i^{th} time series and $t = 1, 2, \ldots, T$ and T denotes the total number of time steps. Additionally, the causal factor data $\mathbf{X}^{CF} = (x_{im}^{CF})$, where $m = 1, 2, \ldots, M$ and M denotes the number of causal variables, share the same set of labels with the time series data $\mathbf{X}^{TS}$. The bagging procedures proposed are dedicated to randomly draw the labels with a replacement rather than perturbing the internal structure or values of the time series and causal factor data.

5.5.1 BagClust

This procedure is a generic framework that deals with single-criterion of causal factor, time series data and multicriteria approach to segmentation of analogies. Specifically, we perturb the set of time series labels (the learning set) that associated with a collection of time series during the clustering process. By doing this, each bootstrap sample might give rise to different clustering results, which will be further utilized to benefit the forecasting stage.

(1) Construct B data samples based on the bootstrapped samples of the labels L. Specifically, we resample the set of labels $L = (l_1, l_2, \ldots, l_n)$ with replacement. A bootstrapped sample of labels refers to $L^b = (l_1^b, l_2^b, \ldots, l_n^b)$. A new data sample of causal factors and time series is $\mathbf{X}_b^{CF}$ and $\mathbf{X}_b^{TS}$, respectively. The procedure is performed B times. This procedure can be applied to both single-criterion and multicriteria clustering approaches.

(2) Apply the clustering procedure P to the bootstrapped samples $\mathbf{X}_b$. The new data sample used dependent on the criterion used during the clustering procedure P.

(3) Determine the final partitions. Determine the number of clusters based on the largest mean Silhouette Width measure. For MC clustering, we further proceed the model selection step by applying the C-MSKF algorithm to the each partition determined in Step (2). The partitioning returns the best average (historical) forecasting results of C-MSKF is chosen as the final solution for MC clustering (details refer to Section 5.4.1.2). This approach is denoted as MC$_{SilHist}$.

(4) Apply the C-MSKF algorithm to the final partitioning identified in Step (3).

(5) Obtain the aggregated point forecast by averaging across the bootstrapped samples $\mathbf{X}_b$.

5.5.2 BagFcst

The BagFcst-based framework operates at the forecasting stage of the overall prediction process. Particularly, clusters of analogies produced from the segmentation stage are labeled systematically. The bootstrapping procedure is applied at the cluster level. Instead of resampling the learning sets, BagFcst randomly draws labels within an individual cluster with replacement and generates a number of I time series that equals to the original cluster.

(1) Apply the clustering procedure P to the original learning set **X** $= (x_{it})$. Based on the criteria (CF, MC and TS clustering) considered, **X**^{CF} and/or **X**^{TS} are used.

(2) Determine the final partitions. Determine the number of clusters based on the largest mean Silhouette Width measure. For MC clustering, we further proceed by applying the C-MSKF algorithm to the each partition determined in Step (2). The partitioning returns the best average (historical) forecasting results of C-MSKF is regarded as the final solution for MC clustering (details refer to 5.4.1.2). The multicriteria approach takes in account the automated model selection step is denoted as MC$_{SilHist}$.

(3) Construct B bootstrapped samples **X**^{b} based on the resampled labels of $L^b = (l_1^b, l_2^b, \ldots, l_I^b)$, where I is the number of time series in a cluster, and it can vary from cluster to cluster. The set of labels are regarded as random variables following IID distribution within each cluster.

(4) Apply the C-MSKF algorithm to each bootstrapped sample independently.

(5) Obtain the aggregated point forecast by averaging point forecasts for individual time series across the bootstrapped samples.

5.6 Empirical validation

5.6.1 Simulated data

To enable comparison between different segmentation approaches, we generated simulated data corresponding to (i) the information derived from the time series data and (ii) the information obtained from the causal factors that are associated with each time series.

In terms of the time series data, we aim to generate a collection of time series that are correlated across an initial time interval but later display different trend changes, due to an external influence. Particularly, we use linear, logarithmic and piecewise linear functions to characterize these trend change as a function of time t. Conceptually, the linear model describes a time series that shows stable increasing trend. The logarithmic model describes a decreasing growth rate in the slope of the trend. An evident structural change has been introduced into the time-based pattern of a series by the piece-wise linear model. It reflects a trend change from a positive to a negative slope. The specific models used for these three generating functions $f_g(t)$, $g = 1, 2, \ldots, 3$, are defined as follows:

$$f_1(t) = 0.8t + 2.8, \quad if\, 1 \leq t \leq q, \tag{5.8}$$

$$f_2(t) = 4ln(t) + 2, \quad if\, 1 \leq t \leq q, \tag{5.9}$$

$$f_3(t) = \begin{cases} 0.7t + 2.8, & if\, 1 \leq t \leq p \\ -0.9t + 25, & if\, p + 1 \leq t \leq q \end{cases} \tag{5.10}$$

where p defines the time point where a change occurs in the pattern based on the piece-wise linear model; q denotes the number of time steps of the time series.

To further generate a set of analogous time series from a given model provided in

Equations (5.8), (5.9) and (5.10), we added normally-distributed noise to the trend at each time step. Specifically, the noisy time series pattern X_{it} for time series i at time t, associated with the choice of generating function g, is obtained by:

$$x_{it} = \begin{cases} f_g^i(1) + N(f_g^i(t+1) - f_g^i(t), \sigma_{TS}^2), & if\, t = 1 \\ x_{i(t-1)} + N(f_g^i(t+1) - f_g^i(t), \sigma_{TS}^2), & if\, 1 < t \leq q \end{cases} \tag{5.11}$$

where g denotes the choice of generating model; the notation $N(\mu_{TS}, \sigma_{TS}^2)$ describes a random variate drawn from a normal distribution with mean μ_{TS} and variance σ_{TS}^2; Here σ_{TS}^2 is static, but μ_{TS} varies over time and, for each time step t, is defined by the slope of the generating function $f_g(t+1) - f_g(t)$.

Based on Equation (5.11), each model is utilized to generate a set of analogous time series with size I and length of $q-1$ [2]. Noise is introduced to each set through the addition of additive noise, as described in Equation.(5.11). One of the resulting time series data sets is shown in Fig. 5.1.

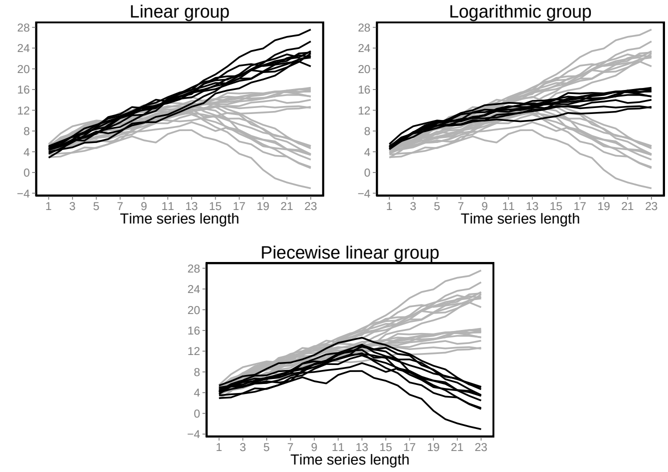


Figure 5.1: Illustration of the simulated time series produced from a linear, logarithmic, and piecewise linear functions

[2] Due to the *differencing* step in Equation.(5.11)

In addition, the z-score method is applied to normalize the time series in order to eliminate the magnitude differences and therefore improve the performance of CIHM cross-sectional adjustment.

With respect to the generation of causal factors, we assume a single causal variable that governs the pattern behavior of the time series. In principle, the methodology described above generalities to a feature space of arbitrary dimensions, as long as a suitable distance measure d_{ij}^{CF} can be defined. The core property modeled here is simply the availability of two different, incommensurable and noisy feature spaces.

In our simulated data, the ground truth (*i.e.,* the nature of the generating model for each time series) is known; this information could be therefore employed to derive suitable but noisy causal factor information. Specifically, the values of the causal factor for time series i is drawn from normal distributions $N(\mu_{CF}, \sigma_{CF}^2)$, where μ_{CF} corresponds to the index g of the generating function $f_g(t)$, associated with time series i (*i.e.,* it takes value in $1, \ldots, 3$).

As reported by previous work (see **Chapter 3**), the increase of noise levels presented in the time series might have negative impacts on C-MSKF's performance. Generally, as the noise level increases, the weight selection for multicriteria clustering approaches becomes more challenging and less reliable. To further understand the noise impact on bagging procedures, we adjust the levels of σ_{CF} and σ_{TS} relative to each other. In specific, the $\sigma_{CF} = 0.35$ and σ_{TS} increases from 0.35 to 0.75 in steps of 0.2, respectively. Across noise levels (scenarios), the parameters used to generate the data are constant and shown in Table 5.1.

Here, all parameters are constant in the experiments, and the forecasting origin is fixed to $t = T$ throughout our analysis. The parameter T is used to allow for more than three observations after the slope change. This makes sure that we satisfy the key assumption of the C-MSKF approach (refers to Section 5.4.2). The parameter *Length selection* reflects that we systematically drop the earliest historical points one at a time,

Table 5.1: Constant parameters used to generate simulate data sets across noise levels, from scenarios 1 to 3

Parameter	Values
Forecasting horizon	$h = 1, 2, \ldots, 6$
Forecasting origin	$T = 17$
Length selection	$l = 12, 13, \ldots, 17$
No. of bootstrapped samples	$B = 50$
No. of replicates	$R = 30$
No. of time series in a group	$I = 10$
Total No. of data points	$q = 24$
Time of trend change	$p = 14$

while fixing the forecasting origin at $t = T$, to take into account the effect of shorter time series.

We generate $R = 30$ replicates (including CF and TS data sets) using the same set of the noise level. Further, for each set of replicate, $B = 50$ bootstrapped samples are generated. A larger number does not give evident improvements regarding the forecasting results so that we keep $B = 50$.

5.6.2 Contestant forecasting techniques

Our main focus in this article is to analyze and compare the effectiveness of bagging approaches in the context of forecasting analogous time series. We aim to investigate accuracy performance of the forecasting forecasting process, which makes use of analogies, after the integration of bagging procedures.

At the segmentation stage of the forecasting process, we allow the comparison between a single-criterion approach of causal factors (CF), a single-criterion approach of time series (TS), and a multicriteria clustering approach (MC) to accommodate for both CF and TS information data sources. Thus, BagClust$_{CF}$, BagClust$_{MC}$, BagClust$_{TS}$, BagFcst$_{CF}$, BagFcst$_{MC}$ and BagFcst$_{TS}$ are compared in the experiments.

In addition, we compare our bagged C-MSKF methods to the unbagged C-MSKF methods including CF, TS, MC, MC$_{SilHist}$, TS methods as well as the baseline model MSKF. Furthermore, popular univariate forecasting methods are included: Damped Exponential Smoothing (Damped), Drift, Exponential Smoothing (ETS), Random Walk (RW), and the Theta method. Here ETS method refers to the automated process provided in the *forecast* R package. Details regarding all the methods considered here are given in the Appendix.

5.6.3 Performance evaluation

In addition to the MSE measure given in Equation (5.6). We also measure the forecasting accuracy of a method using the Symmetric Mean Absolute Percentage Error measure (sMAPE: Bergmeir, Hyndman, and Benítez, 2016).

$$sMAPE = mean(200\frac{|e_t|}{|X_t| + |F_t|}) \tag{5.12}$$

where t refers to the time steps; $e_t = X_t - F_t$ and X_t is the observation at t; F_t represents the respective forecast made at time t.

We assess the accuracy of the point forecasts by computing the average MSE and sMAPE across different forecasting horizons, replicates, bootstrapped samples, time series, and time series lengths. In order to provide further insight, some of our results are broken up by key aspects that are observed to influence the forecasting accuracy. These encompass the three noise scenarios and six forecasting horizons.

5.7 Results

Table 5.2 shows that BagClust$_{MC}$ consistently performs the best among the contestant methods from scenarios 1 to 3. BagClust$_{TS}$ is ranked as the second-best performing method in 4 out of 6 scenarios. We do a pair-wise comparison; we can conclude that bagged C-MSKF methods outperform all the unbagged C-MSKF methods in all scenarios. For example, BagClust$_{CF}$ and BagFcst$_{CF}$ both significantly improve the CF method from scenario 1 to 3 based on average MSE and sMAPE. This indicates that by perturbing time data at either the clustering or forecasting stage, the aggregated point forecasts obtained by averaging the 50 bootstrapped samples has successfully demonstrated the improved forecasting results.

Comparing the MC with MC$_{SilHist}$ method, both approaches employ the Silhouette Width measure to determine the number of cluster. MC then applies the prior knowledge to determine the optimal weight interval, while MC$_{SilHist}$ selects the weight interval based on the best historical average forecasting results of the C-MSKF method. (Particularly, $t = 17$ is used to assess the forecasting performance, and $t \leq T$ are included during the clustering procedure). From scenarios 1 to 3, MC shows equivalent performance to MC$_{SilHist}$ in scenario 1, but consistently outperform the latter in scenario 2 and 3. This has been agreed by both MSE and sMAPE measure. This highlights the limitation of the model selection method implemented in MC$_{SilHist}$ approaches; the weight selection is limited in dealing robustly with the increasingly noisy nature of the time series data. The results here were in line with the previous conclusion provided in **Chapter 3**.

Finally, BagClust$_{MC}$ and BagFcst$_{MC}$ both outperform the MC$_{SilHist}$ in S_1, S_2 and S_3. The IID bootstrap developed based on the MC$_{MC}$ method demonstrates the potential of overcoming the model selection difficulty associated with the robustness related to noisy nature of the time series data. In summary, IID bootstrap employed during the clustering step of the overall prediction process shows better forecasting results than its application in the forecasting stage.

In addition, we break up the average forecasting results by six forecasting horizons to investigate the performance of the compared techniques. As reported by previous work (see **Chapter 3**), weight select in multicriteria clustering approaches is dependent

Table 5.2: Results of forecasting methods across noise levels from scenario 1 to 3. The results are averaged across 30 replicates, 50 bootstrapped samples, 6 time series lengths and 6 prediction horizons

	Average MSE			Average sMAPE		
Methods	S_1	S_2	S_3	S_1	S_2	S_3
CF	0.51	0.83	1.18	34.38	46.61	62.79
TS	0.17	0.68	1.25	21.89	39.09	58.75
MC	0.16	0.47	0.80	20.80	34.23	49.77
MC$_{SilHist}$	0.16	0.59	1.00	21.22	37.67	54.97
BagClust$_{CF}$	0.29	0.57	0.84	28.31	40.40	53.61
BagClust$_{TS}$	***0.12***	0.49	0.84	***18.35***	***32.79***	***48.28***
BagClust$_{MC}$	**0.10**	**0.42**	**0.68**	**16.50**	**31.05**	**44.90**
BagFcst$_{CF}$	0.42	0.72	1.04	31.48	43.68	58.98
BagFcst$_{TS}$	0.16	0.64	1.18	21.13	38.10	57.50
BagFcst$_{MC}$	0.14	***0.46***	***0.76***	20.28	34.70	51.56
Damped	0.20	0.64	0.98	21.75	38.09	52.43
Drift	0.88	1.07	1.24	56.77	59.73	64.34
ETS	0.44	0.96	1.16	29.28	51.03	61.58
MSKF	0.17	0.72	1.14	19.60	39.26	50.62
RW	0.80	1.04	1.20	57.73	61.15	65.22
Theta	0.83	1.06	1.26	56.79	60.53	65.20

on the noise level in the time series data, and MC was generally demonstrated as better performing method, although it cannot be directly employed in a real application setting due the assumption of prior knowledge on the best weight selection.

In summary, BagClust$_{MC}$ presents the best performance in all situations (across noise levels and forecasting horizons) measured by average MSE and sMAPE. By assuming the existence of prior knowledge for the weight selection, MC outperforms all the benchmark models including MC$_{SilHist}$. Comparing MC to BagClust$_{MC}$, the latter forecasting method outperforms MC in all noise levels across forecasting horizons, from 1 to 6. Based on MSE, BagFcst$_{MC}$ consistently outperform MC methods across noise levels and forecasting horizons form 1 to 6. However, this conclusion is not confirmed on sMAPE, and we see opposite results.,

5.8 Conclusions

In this article, we have introduced the notion of bagging, based on IID bootstrap, to a forecasting process that makes use of analogies. The idea of the IID bootstrap is straightforward and easy to implement in the forecasting process. This method shows significant improvements to the final forecasting results through the reduction of clustering instability. By bootstrapping time series labels at the segmentation stage, BagClust-based methods consistently improve upon BagFcst-based approaches.

The main contributions of this manuscript are as follows: (i) We analyze the performance of bagging approaches in the context of forecasting analogous time series. To the best of our knowledge, little work has been reported to investigate the effectiveness of bagging in circumstances where analogies are extensively employed. The IID bootstrap procedures proposed can be easily transferred to different forecasting approaches where analogies are applicable. (ii) Our bagged C-MSKF methods show superior performance over the original C-MSKF methods. Furthermore, these methods improve the robustness of the forecasting process based on MC$_{SilHist}$. This was demonstrated by the comparison of forecasting accuracy between BagClust$_{MC}$ and MC$_{SilHist}$.

Specifically, the bagging procedures have found to be powerful in lowering down the process instability through the aggregation of numerous point forecasts. Surprisingly, the BagClust$_{MC}$ method outperforms the MC method across 6 forecasting horizons and 3 noise levels (see Tables 5.3, 5.4,5.5 and 5.6). This indicates that the employment of IID bootstrap at the segmentation stage helps address the shortcoming of our model selection method, where the weight is determined based on best historical forecasting accuracy of C-MSKF. The process was reported to be unstable and sensitive to the increase of noise associated with time series, particularly for longer forecasting horizons (see **Chapter 4**).

Another option of reducing the clustering instability is to construct a consistent partition through the ensemble method. Ensemble clustering might be able to produce better clustering results, but the procedure does not consider forecasting accuracy performance when determining the final partitions. According to Guyon, Von Luxburg, and Williamson (2009), the goodness of a clustering solution is best measured by taking into account the overall aim of an application. Hence, we retain our problem-specific approach to model selection, which works on the historical forecasting results of C-MSKF algorithms.

In future work, more research could be conducted concerning the determination of parameters used for the bagging procedures. This includes decisions related to the number of bootstrapped samples, and the sample size of bootstrapped samples. To further evaluate the model performance, it will be desirable to apply our bagged C-MSKF methods to real-world applications.

Table 5.3: In-depth contrast of the forecasting errors on MSE where C-MSKF algorithm is used as the baseline model for pooling methods, broken up by noise level and forecasting horizon ranging from 1 to 6. Shown are averages across 30 replicates and 6 different time series lengths. The best performing methods are highlighted in bold face, with the second best performance highlighted in italic bold face.

Scenarios	Methods	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	BagClust$_{CF}$	0.15	0.2	0.25	0.31	0.38	0.45
	BagClust$_{MC}$	**0.04**	**0.06**	**0.08**	**0.11**	**0.14**	**0.18**
	BagClust$_{TS}$	***0.06***	***0.08***	***0.1***	***0.13***	***0.16***	***0.2***
	BagFcst$_{CF}$	0.22	0.29	0.37	0.45	0.54	0.64
	BagFcst$_{MC}$	0.08	0.1	0.13	0.15	0.18	0.22
	BagFcst$_{TS}$	0.08	0.11	0.14	0.17	0.21	0.25
	CF	0.27	0.36	0.45	0.55	0.66	0.78
	Damped	**0.04**	***0.08***	0.13	0.20	0.30	0.44
	Drift	0.14	0.33	0.60	0.95	1.38	1.89
	ETS	0.08	0.17	0.30	0.46	0.68	0.96
	MC	0.09	0.11	0.14	0.17	0.20	0.24
	MC$_{SilHist}$	0.09	0.11	0.14	0.17	0.21	0.25
	MSKF	**0.04**	***0.08***	0.12	0.19	0.26	0.34
	RW	0.13	0.30	0.55	0.86	1.24	1.69
	Theta	0.14	0.32	0.57	0.90	1.30	1.78
	TS	0.09	0.12	0.15	0.19	0.22	0.27
S_2	BagClust$_{CF}$	0.28	0.37	0.47	0.62	0.76	0.93
	BagClust$_{MC}$	**0.14**	**0.22**	**0.32**	**0.47**	***0.61***	0.78
	BagClust$_{TS}$	0.17	***0.26***	***0.37***	0.54	0.71	0.89
	BagFcst$_{CF}$	0.38	0.49	0.61	0.78	0.94	1.12
	BagFcst$_{MC}$	0.25	0.32	0.4	***0.5***	**0.6**	**0.71**
	BagFcst$_{TS}$	0.26	0.37	0.51	0.7	0.9	1.12
	CF	0.45	0.57	0.71	0.89	1.08	1.27
	Damped	***0.15***	***0.26***	0.42	0.68	0.99	1.36
	Drift	0.23	0.46	0.76	1.17	1.64	2.17
	ETS	0.21	0.39	0.65	1.01	1.48	2.04
	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
	MSKF	0.17	0.32	0.51	0.79	1.08	1.43
	RW	0.22	0.45	0.74	1.14	1.59	2.10
	Theta	0.23	0.45	0.75	1.16	1.62	2.14
	TS	0.29	0.40	0.54	0.74	0.95	1.18

Table 5.4: In-depth contrast of the forecasting errors on MSE where C-MSKF algorithm is used as the baseline model for pooling methods, broken up by noise level and forecasting horizon ranging from 1 to 6. Shown are averages across 30 replicates and 6 different time series lengths. The best performing methods are highlighted in bold face, with the second best performance highlighted in italic bold face.

Scenarios 6-period	Methods	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_3	BagClust$_{CF}$	0.4	0.55	0.72	0.9	1.12	1.36
	BagClust$_{MC}$	**0.23**	**0.35**	**0.53**	**0.73**	**0.97**	**1.24**
	BagClust$_{TS}$	0.29	***0.44***	***0.65***	0.91	1.2	1.54
	BagFcst$_{CF}$	0.51	0.69	0.89	1.11	1.38	1.68
	BagFcst$_{MC}$	0.4	0.52	0.66	***0.81***	***0.99***	***1.19***
	BagFcst$_{TS}$	0.48	0.68	0.96	1.28	1.63	2.03
	CF	0.61	0.80	1.02	1.26	1.54	1.85
	Damped	0.29	0.47	0.74	1.07	1.44	1.86
	Drift	0.29	0.56	0.92	1.36	1.87	2.46
	ETS	0.29	0.54	0.87	1.26	1.72	2.25
	MC	0.37	0.52	0.70	0.87	1.06	1.27
	MC$_{SilHist}$	0.4	0.58	0.82	1.08	1.38	1.72
	MSKF	***0.26***	0.49	0.81	1.23	1.72	2.32
	RW	0.28	0.55	0.90	1.32	1.80	2.36
	Theta	0.31	0.59	0.95	1.38	1.87	2.44
	TS	0.51	0.73	1.02	1.36	1.73	2.15

Table 5.5: In-depth contrast of the forecasting errors on sMAPE where C-MSKF algorithm is used as the baseline model for pooling methods, broken up by noise level and forecasting horizon ranging from 1 to 6. Shown are averages across 30 replicates and 6 different time series lengths. The best performing methods are highlighted in bold face, with the second best performance highlighted in italic bold face.

Scenarios	Methods	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	BagClust$_{CF}$	30.14	28.95	28.11	27.7	27.51	27.47
	BagClust$_{MC}$	***17.47***	**16.62**	**16.21**	**16.14**	**16.2**	**16.36**
	BagClust$_{TS}$	19.88	18.76	***18.12***	***17.84***	***17.72***	***17.77***
	BagFcst$_{CF}$	33.61	32.24	31.28	30.78	30.54	30.45
	BagFcst$_{MC}$	22.76	21.18	20.11	19.5	19.15	19
	BagFcst$_{TS}$	23.32	21.84	20.92	20.43	20.18	20.1
	CF	36.67	35.19	34.18	33.64	33.37	33.24
	Damped	18.73	19.57	20.83	22.28	23.77	25.30
	Drift	45.29	51.90	56.37	59.78	62.50	64.78
	ETS	24.33	26.48	28.55	30.44	32.14	33.72
	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
	MSKF	**17.27**	***17.93***	19.03	20.15	21.14	22.06
	RW	42.57	50.29	56.39	61.50	65.88	69.73
	Theta	44.87	51.62	56.28	59.92	62.81	65.24
	TS	24.19	22.66	21.69	21.16	20.87	20.77
S_2	BagClust$_{CF}$	41.52	40.7	40.14	40.02	39.97	40.06
	BagClust$_{MC}$	**29.66**	**30.22**	**30.66**	**31.34**	**31.92**	**32.47**
	BagClust$_{TS}$	***31.34***	***31.85***	***32.34***	***33.09***	***33.77***	34.33
	BagFcst$_{CF}$	45.19	44.07	43.29	43.16	43.13	43.21
	BagFcst$_{MC}$	36.56	35.43	34.52	34.12	33.87	33.72
	BagFcst$_{TS}$	38.26	37.84	37.74	37.95	38.28	38.5
	CF	48.32	47.20	46.35	46.07	45.90	45.82
	Damped	32.90	35.53	37.42	39.30	40.96	42.42
	Drift	48.02	54.58	59.23	62.89	65.71	67.94
	ETS	40.94	46.40	50.29	53.57	56.31	58.68
	MC	35.22	34.55	34.12	33.96	33.83	***33.67***
	MC$_{SilHist}$	38.35	37.62	37.3	37.39	37.59	37.75
	MSKF	33.17	36.45	38.64	40.81	42.49	44.00
	RW	45.51	53.62	59.78	65.08	69.56	73.36
	Theta	47.85	54.84	59.81	63.94	67.10	69.64
	TS	39.56	38.92	38.72	38.87	39.15	39.34

Table 5.6: In-depth contrast of the forecasting errors on sMAPE where C-MSKF algorithm is used as the baseline model for pooling methods, broken up by noise level and forecasting horizon ranging from 1 to 6.Shown are averages across 30 replicates and 6 different time series lengths. The best performing methods are highlighted in bold face, with the second best performance highlighted in italic bold face.

Scenarios	Methods	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_3	BagClust$_{CF}$	53.15	53.27	53.65	53.64	53.87	54.09
	BagClust$_{MC}$	**41.22**	**43.51**	**45.03**	**45.76**	**46.61**	**47.25**
	BagClust$_{TS}$	44.69	47.24	***48.36***	***49.07***	49.83	50.5
	BagFcst$_{CF}$	58.31	58.57	59.05	59.05	59.29	59.63
	BagFcst$_{MC}$	53.38	52.39	51.58	50.79	50.62	50.59
	BagFcst$_{TS}$	55.8	57.08	57.65	57.84	58.16	58.45
	CF	62.68	62.82	62.84	62.65	62.76	63.01
	Damped	44.17	48.56	52.20	54.70	56.67	58.26
	Drift	50.49	58.40	64.01	68.10	71.27	73.75
	ETS	48.21	55.29	60.89	65.16	68.60	71.36
	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
	MSKF	***42.31***	***47.22***	50.43	52.64	54.69	56.42
	RW	48.81	57.14	64.08	69.44	74.03	77.80
	Theta	50.76	58.87	64.73	69.08	72.53	75.22
	TS	57.24	58.43	58.88	59.04	59.33	59.60

References

[1] J. S. Armstrong. “Findings from evidence-based forecasting: Methods for reducing forecast error”. In: *International Journal of Forecasting* 22.3 (2006), pp. 583–598.

[2] J. S. Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Vol. 30. Springer Science & Business Media, 2001.

[3] F. M. Bass. “A new product growth for model consumer durables”. In: *Management Science* 15.5 (1969), pp. 215–227.

[4] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. “Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation”. In: *International Journal of Forecasting* 32.2 (2016), pp. 303–312.

[5] A. Bose. “Bootstrap in moving average models”. In: *Annals of the Institute of Statistical Mathematics* 42.4 (1990), pp. 753–768.

[6] A. Bose. “Edgeworth correction by bootstrap in autoregressions”. In: *The Annals of Statistics* 16.4 (1988), pp. 1709–1722.

[7] L. Breiman. “Bagging predictors”. In: *Machine Learning* 24.2 (1996), pp. 123–140.

[8] R. T. Clemen. “Combining forecasts: A review and annotated bibliography”. In: *International Journal of Forecasting* 5.4 (1989), pp. 559–583.

[9] C. Cordeiro and M. M. Neves. “Forecasting time series with Boot. EXPOS procedure”. In: *Revstat* 7.2 (2009), pp. 135–149.

[10] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[11] G. Duncan, W. Gorr, and J. Szczypula. *Comparative Study of Cross Sectional Methods for Time Series With Structural Changes*. Tech. rep. Carnegie Mellon University, 1994.

[12] G. T. Duncan, W. Gorr, and J. Szczypula. “14 Bayesian Hierarchical Forecasts for Dynamic Systems: Case Study on Backcasting School District Income Tax”. In: *New Directions in Spatial Econometrics* (2012), p. 322.

[13] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[14] C. J. Easingwood. “An analogical approach to the long term forecasting of major new product sales”. In: *International Journal of Forecasting* 5.1 (1989), pp. 69–82.

[15] M. H. Glantz. “The use of analogies: in forecasting ecological and societal responses to global warming”. In: *Environment: Science and Policy for Sustainable Development* 33.5 (1991), pp. 10–33.

[16] P. Goodwin, K. Dyussekeneva, and S. Meeran. “The use of analogies in forecasting the annual sales of new electronics products”. In: *IMA Journal of Management Mathematics* 24.4 (2013), pp. 407–422.

[17] K. C. Green and J. S. Armstrong. “Structured analogies for forecasting”. In: *International Journal of Forecasting* 23.3 (2007), pp. 365–376.

[18] N. P. Greis and C. Z. Gilstein. “Empirical Bayes methods for telecommunications forecasting”. In: *International Journal of Forecasting* 7.2 (1991), pp. 183–197.

[19] I. Guyon, U. Von Luxburg, and R. C. Williamson. “Clustering: Science or art”. In: *NIPS 2009 Workshop on Clustering Theory*. 2009, pp. 1–11.

[20] P. J. Harrison and C. F. Stevens. “A Bayesian approach to short-term forecasting”. In: *Operational Research Quarterly* (1971), pp. 341–362.

[21] R. J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.

[22] R. E. Kass and D. Steffey. “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)”. In: *Journal of the American Statistical Association* 84.407 (1989), pp. 717–726.

[23] L. Kaufman and P. J. Rousseeuw. *Partitioning around medoids (program pam)*. 1990.

[24] H. R. Kunsch. “The jackknife and the bootstrap for general stationary observations”. In: *The Annals of Statistics* 17.3 (1989), pp. 1217–1241.

[25] E. Lu and J. Handl. “Multicriterion Segmentation of Demand Markets to Increase Forecasting Accuracy of Analogous Time Series: A First Investigation”. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 379–388.

[26] S. Murawski. “Climate change and marine fish distributions: forecasting from historical analogy”. In: *Transactions of the American Fisheries Society* 122.5 (1993), pp. 647–658.

[27] K. Nikolopoulos et al. “Forecasting branded and generic pharmaceuticals”. In: *International Journal of Forecasting* 32.2 (2016), pp. 344–357.

[28] F. Petropoulos et al. “‘Horses for Courses’ in demand forecasting”. In: *European Journal of Operational Research* 237.1 (2014), pp. 152–163.

[29] M. I. Piecyk and A. C. McKinnon. “Forecasting the carbon footprint of road freight transport in 2020”. In: *International Journal of Production Economics* 128.1 (2010), pp. 31–42.

[30] P. J. Rousseeuw. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[31] J. A. Stimson. “Regression in space and time: A statistical essay”. In: *American Journal of Political Science* (1985), pp. 914–947.

[32] R. Webby and M. O’Connor. “Judgemental and statistical time series forecasting: a review of the literature”. In: *International Journal of Forecasting* 12.1 (1996), pp. 91–118.

Chapter 6

Improving time series clustering using multiple criteria (paper 4)

6.1 Abstract

Time series clustering is of significant interest in research areas ranging from bioinformatics, economics, finance, and forecasting to signal processing. The clustering of time series data is challenging due to the difficulty of defining similarity between pairs of time series. In fact, there are no universally accepted definitions of similarity between time series, and the best notion of similarity might vary with the application context. Traditionally, a single distance metric and standardization technique is employed during the clustering procedure to partition a set of time series into distinctive groupings. However, different distance metrics / standardization techniques may emphasize different notions of similarity, *e.g.,* one distance measure may capture the linear pattern underlying the data, yet fail to model the non-linearity present in the time series. In applications where we are not sure which notion of similarity is accurate or where several notions of similarity are relevant, we might benefit from combining multiple distance metrics / standardization techniques, to capture complementary notions of similarity. Here, we describe a multicriteria clustering approach that optimizes a partition with respect to pairs of different distance metrics / standardization techniques. Using simulated data, we demonstrate that such a multicriteria clustering approach consistently outperforms single-criterion approaches in situations where the selected distance measures / standardization measures show a low correlation with each other.

Keywords: Analogy; Bayesian pooling; Distance metrics; Kalman filter; Standardization techniques

6.2 Introduction

Time series clustering has gleaned extensive interest across demand forecasting, health care, pattern recognition, public budgeting, signal processing. A time series can be regarded as a feature that comprises values varied with time. The similarity between pairs of time series is always considered as a whole due to its numerical and continuous nature.

Traditionally, time series clustering problems are modeled as single-criterion problems that partitions a set of time series into homogeneous groupings, which share similarity with regards to time-based patterns. Model selection, as the main step in clustering procedures, often encompasses the choices related to the determination of a proper distance metric, the choice of a suitable normalization technique.

These distance metrics can generally be classified into three classes. Straightforwardly, raw-data-based methods have been investigated to directly work with raw data, but these methods are often computational expensive due to the high dimensionality of time series data. Additionally, feature-based methods have been developed to overcome this limitation. These methods approximate the time series patterns by extracting key features that optimally describe the underlying patterns. Time series are considered similar if the features used are similar. The third class of methods refers to model-based approaches, where time series are fitted by statistical models or by a mix of underlying probability distributions. Time series are regarded as similar if the models that characterize individual series are similar, or the remaining residuals after fitting the model are similar. In fact, the isolated consideration of any one isolated distance metric may be inadequate since various types of patterns can present the data at the same time. For example, the ARIMA model might be sufficient for modeling linear patterns present in the US dollar exchange rate time series, yet fail to capture the non-linearity in these series (Zhang, 2003).

Furthermore, Stoddard (1979) argued that any type of standardization can remove the between-cluster variation. The variation might be crucial for identifying the underlying data structure. However, almost all clustering procedures used a uniform normalization scheme across all data items on the variable. However, time series data sets can contain a mixture of patterns for either an individual time series or entire data sets. We argue that a uniform application of standardization technique can be inadequate and lead to a hidden data structure. Indeed, there has been no universal agreement on the choice of distance metrics or standardization schemes that could always perform the best across different applications. Different distance metrics / standardization techniques may of different focus related to the notion of similarity among pairs of time series. For instance, one distance measure may capture the linear pattern underlying the data, yet fail to model the non-linearity present in the time series.

In applications where we are not sure which notion of similarity is accurate or where several notions of similarity are relevant, we might benefit from combining multiple distance metrics / standardization techniques, to capture complementary notions of similarity. On account of this, it may be more promising to use multiple criteria during the clustering procedure in order to accommodate for various views regarding the similarity. Here, we introduce the notion of multicriteria approaches to time series clustering. These approaches have been explored in the previous literature studies (*e.g.,* Handl and Knowles, 2007) to trade off various often conflicting criteria during the clustering process. Multicriteria clustering approaches have demonstrated the potential of facilitating a more robust discovery of the data structure than single-criterion clustering approaches.

In this paper, we explore the potential of multicriteria approaches to aid in the forecast of analogous time series, where analogous series are identified using clustering approaches. In particular, we are concerned with circumstances where only time series data are available, and this eliminates the possibility of using multiple features spaces as criteria. This is often true as extraneous information associated time series data may not be available or can only be obtained with prohibitively high cost. To our knowledge, little work has been conducted that investigates the possibilities of improving the time series clustering by employing multiple criteria. Based on simulate data, we explore the further potential of multicriteria approaches in the problem-specific context, *i.e.,* time series forecasting.

The remainder of the paper is structured as follows: Section 6.3 mainly surveys previous work concerning time series clustering. Section 6.4 details the main components of the prediction process that utilizes clustering approaches for the identification of analogies. Section 6.5 presents details involved in the experiments in order to evaluate the performance of multicriteria approaches. Section 6.6 analyzes the results. Section 6.7 discusses limitations of the present manuscript and future work.

6.3 Previous research

Over the recent decades, the amount of research conducted regarding the clustering of time series data has increased significantly. This is true across various disciplines, including data mining (Last, Kandel, and Bunke, 2004), energy time series (Alvarez et al., 2011),empirical finance (Franses and Van Dijk, 2000), multimedia (Alon et al., 2003), and marketing areas. Most commonly, clustering algorithms developed for time series data have been previously proposed to handle the static data, which are random variables. In essence, the current clustering algorithms or specific dissimilarity metrics proposed in the clustering or data mining literature mainly attempt to transform the time series data into static forms so that the time series data can be easily handled (Liao, 2005).

Time series can be regarded as a feature changed with time. The difficulty related to time series clustering primarily originates from internal properties of the data. Time series data show time-dependent structure, which is continuous as well as a mixture of patterns. It might simultaneously encompass the patterns of linearity, non-linearity, seasonality, structure changes. To partition a set of time series into homogeneous clusters, numerous approaches have been investigated in the literature, where some surveys with respect to time series clustering are provided in Esling and Agon (2012), Fu (2011), Keogh and Kasetty (2003), Liao (2005),and Rani and Sikka (2012). In summary, the majority of the previous work strives to recover the structure underlying the time series data from different perspectives, which are reflected by the development of various techniques, including distance functions and standardization schemes.

To satisfy different needs of applications, many distance metrics have been developed to measure the (dis)similarity between time series. Liao (2005) segmented these techniques into three main categories: raw-data-based, feature-based and model-based approaches. Raw-data-based approaches directly work with the raw data in a way all measurements are considered to obtain the similarity between series. Typical examples include the applications of correlation-based approaches (Ernst, Nau, and Bar-Joseph, 2005). These methods are straightforward to apply and easy to interpret but can be sensitive to outliers and require much effort in computation. To overcome the limitation particularly lies in the time-complexity of the existing methods, feature-based and model-based approaches have been developed. These methods approximate time series patterns by using key features of the time series or assuming a statistical model that characterizes the individual series, respectively. However, the features obtained might be application dependent while an individual model identified can be inadequate to allow the mixture of the time series patterns, *e.g.,* linearity and non-linearity. Also, distance metrics are sensitive to the differences in scales of the variables (Milligan and Cooper, 1985).

Additionally, the uniform application of one standardization technique across all items on differing variables is often observed in practice. Standardization is used to equalize the magnitude and the variability of the input variables. Stoddard (1979) declared that any standardization could remove the between-cluster variability, which is important for the discovery of the underlying data structure. Due to the possible mixing of variables that exhibit a wide range of cluster structures, Steinley (2004) argued that the routine standardization of all variables during the clustering procedure may be unwise. The routine normalization of the data may lead to poor clustering performance. In addition to this, practically different distance metrics can be sensitive to the scale differences or means of standardization of the time series data.

Commonly observed in previous work, differing standardization and distance metrics have been proposed to assist in the recovery of the data structure. Nevertheless, these work typical model time series clustering as a single-criterion problem where a distance metric and standardization metric are used. However, the performance of these measures are often application or even data-dependent. This is because various distance metrics / standardization techniques may focus on different aspects of the notion of similarity. No metric can capture all information underlying the notion of similarity. As a result, some metrics may be preferred than the others from in different applications. For example, the ARIMA model might be adequate for modeling linear patterns present in the water quality time series, yet fail to capture the non-linearity in these series (Faruk, 2010). Again, supported by Guyon, Von Luxburg, and Williamson (2009), the goodness of clustering solutions is best judged by the overall success of the application. This means the judge of similarity among time series can be inadequate without taking into account the aim of the application.

On account of this, the independent consideration of any one isolated distance metric / normalization measure may prove insufficient since various types of patterns can simultaneously present in the data. Thus, a combination different measures might complement the information that these measures neglect when defining the notion of similarity. In circumstances where we are not sure which notion of similarity for time series is accurate or where all notions of similarity are relevant, we might benefit from combining multiple often conflicting criteria to capture various aspects of the notion.

One way of doing this is to integrate additional information data sources that extrinsic to the time series data during the clustering procedure. The combination of data features or information sources is importance to facilitate a better recover of the underlying data structure (Brusco and Cradit, 2001; Dash and Liu, 2000; Lu and Handl, 2015). This implies that the same set of objects should be clustered with the consideration of the relative importance of features or information sources, and different representations of the cluster might reveal of a specific part of the overall clustering definition. Additionally, Handl and Knowles (2007) considered multiple clustering criteria concerning the quality of a clustering solution and their results demonstrated the potential of multicriteria approaches with regards to more natural and robust groupings. Additionally, Law, Topchy, and Jain (2004) developed a two-step clustering approach, where each clustering algorithm, in parallel, generates a clustering solution and then external criteria based on the stability of the clustering results was used to determine the final partitions.

6.4 Multicriteria approaches to time series clustering in the context of forecasting

Here, we demonstrate our ideas of multicriteria approaches to time series clustering in the context of forecasting. A highly related work could be referred to **Chapter 3**, where multiple information sources were used for boosting the performance of analogy identification. A potential challenge associate with the previous work lies in the fact that it requires the presence of information, which is extrinsic to the time series information. In many scenarios, the causal factors that describe the time-based patterns observed are simply absent or that can only be obtained at prohibitively high cost. Alternatively, we could further explore the potential of multicriteria approaches to time series clustering, where analogies are identified using time series information alone.

To ratify our concepts, we attempt to identify analogous time series using multiple criteria during the clustering stage in order to cater for broader application needs. Specifically, all major elements used for constructing the prediction process are detailed as follows. Particularly, we focus on the choice of criteria for the clustering of analogous time series.

6.4.1 The choice of criteria

6.4.1.1 Standardization techniques

For clustering of the time series data, an early step is to standardize all time series to a proper format for eliminating magnitude differences. By applying the distance metric, large values might dominate the final clustering solution and further leads to an unsatisfactory recovery of the data structure.

Regarding standardization, we apply two standard techniques the z-score and range or Mon-max methods. The z-score measure has been extensively studied and has been proposed by numerous authors, for example Dubes and Jain (1980), Lorr (1983), and Lance and Williams (1966). Specifically, the standardized variable has been transformed to have zero mean and unity variance. This method may not perform properly if there are substantial differences among the within-cluster standard deviations. The Min-max method is another option we consider to apply for standardizing time series. This method was reported to give the best recovery of the clusters based on extensive Monte Carlo simulation (Milligan and Cooper, 1988).

Assuming a data set $\mathbf{X} = \{x_{ij}, \ldots, x_{NJ}\}$, and N represents the number of samples in the data set; Each sample is a vector of values that describe the feature varies over time. The transformation of the i^{th} time series on the t observation is given as:

The formula of z-score method is given as:

$$Z^{(1)} = \frac{x_{it}}{max(x_i) - min(x_i)} \tag{6.1}$$

The formula of Min-max method is given as:

$$Z^{(2)} = \frac{x_{it} - mean(x_j)}{\sigma(x_i)} \tag{6.2}$$

6.4.1.2 Distance measures

Subsequently, distance metrics are applied on the standardized time series to measure the distance between pairs of time series, and two simple distance metrics are considered here. A time series can be described as a vector that comprises a feature varying with time. The distance value d_{ij} between pairs of time series i and j is computed on the point-to-point distance using square Euclidean distance or the correlation (Pearson correlations' coefficients) between these vectors. Euclidean distance is perhaps the most popular distance metric, as it maintains the original scales of the variables (Gnanadesikan, Kettenring, and Tsao, 1995). Give no missing values exist, Euclidean distance has no influence on the subsequent cluster recovery process. Correlation methods have been widely applied to measure the similarity between time series (*e.g.,* Frühwirth-Schnatter and Kaufmann, 2008).

The square Euclidean distance between time series i and j are presented as follows:

$$\delta(i,j)^{(1)} = \sum_t (a_{it} - a_{jt})^2 \tag{6.3}$$

where the dissimilarity matrix derived from the time series data is defined as $\mathbf{D}^{(1)} = (d_{ij}^{(1)})$, and each element $d_{ij}^{(1)}$ is calculated as $d_{ij}^{(1)} = \delta(i,j)^{(1)}$; a_{it} and a_{jt} represent the values at time t associated with time series i and j, respectively; t is the time index. This method is simple and commonly used to measure the distance between series, but it becomes time-consuming for data with larger dimension.

A common way of measure similarity between variables is to use the Pearson correlation's coefficient, and the equation is given as:

$$\delta(i,j)^{(2)} = 1 - \frac{T(\sum_t x_{it}x_{jt}) - (\sum_t x_{it})(\sum_t x_{jt})}{\sqrt{(T(\sum_t x_{it}^2) - (\sum_t x_{it})^2)(T(\sum_t x_{jt}^2) - (\sum_t x_{jt})^2)}} \tag{6.4}$$

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

6.4.1.3 The combination of various criteria

To combine multiple criteria, we apply the weighted-sum method at the distance function, which allows trade-off between various complementary criteria. The weighted-sum method is easy to apply and interpret. Then, dissimilarity matrix-based clustering algorithms can be used to handle the time series data (see Liao, 2005).

Principally, different distance metrics / standardization techniques may emphasize different aspects of the notion of similarity among pairs of time series. The combination of different metrics may be able to provide complementary information regarding the notion of similarity.

Specifically, we formulate a combined dissimilarity matrix using multiple criteria derived from the mix of distance metric / standardization technique. For initial testing of our ideas, the following combinations are considered in this article. We focus on comparing the scenarios, in which the one clustering procedure shares a at least common distance metric / standardization technique with another.

$$d_{ij\omega}^{MC} = \begin{cases} (1-\omega) \times d_{ij}^{(1)} + \omega \times d_{ij}^{(2)},\ given\ Z^{(1)} \\ (1-\omega) \times d_{ij}^{(1)} + \omega \times d_{ij}^{(2)},\ given\ Z^{(2)} \\ (1-\omega) \times d_{ij}^{(1)} + \omega \times d_{ij}^{(1)},\ given\ Z^{(1)}\ and\ Z^{(2)},\ respectively \\ (1-\omega) \times d_{ij}^{(2)} + \omega \times d_{ij}^{(2)},\ given\ Z^{(1)}\ and\ Z^{(2)},\ respectively \end{cases} \tag{6.5}$$

Where ω varies from 0 to 1 with increments of 0.1; $\mathbf{D}_{w}^{\mathbf{MC}} = (d_{ijw})$ and d_{ijw}^{MC} [1] denotes the element of the combined dissimilarity matrix using weight ω.

[1] Note that two dissimilarity matrices, derived from different clustering methods, are combined without transformation into the same range as been done in **Chapters 3**, **5** and **4**. This is because the use of same range on the same data source is found to eliminate the differences between the clustering methods.

6.4.2 Clustering algorithm

Further, we apply a standard clustering algorithm for the clustering analysis, namely the Partitioning Around Medoids (PAM) clustering algorithm (Kaufman and Rousseeuw, 2009). This method tends to yield clustering solutions with equal size that are considered advantageous in this forecasting application. Moreover, PAM clustering operates on the dissimilarity matrix that offers flexibility for tackling time series data. The PAM clustering is repeated 30 times, where the partition with the smallest sum of within-cluster dissimilarities is selected for further analysis.

6.4.3 Model selection

To determine the clusters, we proceed by applying one of the most popular metrics (the Silhouette Width Kaufman and Rousseeuw, 2009) for determining the appropriate number of clusters. For the same number of clusters, multicriteria clustering approaches often return multiple clustering solutions that reflect trade-offs among various criteria. For the following weight selection step, we select the most suitable weight interval based on the average best historical forecasting results of the forecasting algorithm applied. More specifically, time point $t = T$ is used to support this part of analysis. $t \leq T$ are used during the clustering step. This model selection method is the best performing method reported in the previous work (see **Chapter 4**) that takes into account the overall performance of an application.

The Silhouette Width measure is a comprehensive measure that evaluates the quality of clustering by considering both the cluster cohesion and separation given the data structure alone. It has a minimum value of -1 and a maximum value of 1. A larger value indicates a better clustering solution. The Silhouette value for an individual data object i is given as:

$$Sil(i) = \frac{c_i - b_i}{max(c_i, b_i)} \tag{6.6}$$

where c_i indicates the average distance between i and all objects in the same group; b_i indicates the average distance between i and all objects in the closest another group that is referred to the cluster returning the minimum b_i. The Silhouette Width of the entire partitioning is then computed as the mean Silhouette Width of all data objects.

After the determination of the number of clusters, the best weight is selected based on the best average historical forecasting results at $t = T$.The Mean Square Error (MSE) measure is used to support the weight selection step:

$$MSE = mean(e_t^2) \tag{6.7}$$

where t refers to time step; $e_t = X_t - F_t$, X_t is the observation of the time series X at time t; F_t is the respective point forecast.

6.4.4 Forecasting algorithm

During the forecasting stage, we experiment with the Cross-Sectional Multistate Kalman Filter algorithm (C-MSKF: Duncan, Gorr, and Szczypula, 1993; Duncan, Gorr, and Szczypula, 2001) to extrapolate the time series into the future. According to to Stimson (1985), the homogeneity of grouping is essential for the effectiveness of pooling methods. Here, we propose multicriteria approaches to the clustering of time series data and illustrate our ideas using the C-MSKF forecasting method. However, our clustering approaches would be expected to generalize to different pooling methods that make use of analogous time series.

The C-MSKF method is a Bayesian pooling method developed from conventional time series forecasting methods, *i.e.,* the Multi-State Kalman Filter algorithm (Harrison and Stevens, 1971). The application of C-MSKF is considered suitable in this application because of its capability of drawing information from analogies. In brief, the C-MSKF method combines the strengths of the Conditionally Independent Hierarchical Model (CIHM: Kass and Steffey, 1989) and MSKF algorithms, which utlizes the Kalman Filter (Harvey and Forecasting, 1989). The Kalman filter is advantageous in quickly reaching a reliable prediction and thus lends efficiency to the C-MSKF methods. Details of C-MSKF and MSKF algorithm are referred to the Appendix. In addition, the use of MSKF can be principally replaced by the Single Exponential Smoothing method, and a counterpart approach was proposed in (Duncan, Gorr, and Szczypula, 1994).

In this article, C-MSKF is used to make forecasts for a set of prediction horizons. Specifically, for a given forecasting origin T, the h-step ahead forecast (for $h \geq 2$) is made by iteratively updating C-MSKF using the forecasts obtained for the $T+1, \ldots, h-1^{th}$ time steps, and predicting the succeeding time point.

6.5 Experimental design

6.5.1 Simulated data

We use simulated data to demonstrate our ideas concerning multicriteria approaches to address time series clustering problems in the context of forecasting. We generate a set of time series, which are correlated across an initial time step and further subject to different slope changes governed by an external influence. Particularly, a linear, logarithmic and piece-wise linear model is employed to produce the time series data. The models utilized are given in Equations (6.8), (6.9) and (6.10). The time series generated by the linear model shows a stable increasing trend as a function of time. The logarithmic model indicates a decreasing rate of growth in the time series. The time series generated by the piece-wise linear function can be interpreted as a series with pattern change in the slop, from the positive to negative value after the external influence occurring at the p time point. The specific models used for these three generating functions $f_g(t)$, $g = 1,\ldots,3$, are given as follows:

$$f_1(t) = 0.8t + 2.8, \quad if\, 1 \leq t \leq q, \tag{6.8}$$

$$f_2(t) = 4ln(t) + 2, \quad if\, 1 \leq t \leq q, \tag{6.9}$$

$$f_3(t) = \begin{cases} 0.7t + 2.8, & if\, 1 \leq t \leq p \\ -0.9t + 25, & if\, p+1 \leq t \leq q \end{cases} \tag{6.10}$$

where p refers to the time of slope change for Equation (6.10); q is the number of time points.

To generate a group of analogous time series from a given model, we perturb each time point with normally-distributed noise. The noisy time series pattern X_{it} for time series i at time t, associated with generating model g, is presented as follows:

$$x_{it} = \begin{cases} f_g^i(1) + N(f_g^i(t+1) - f_g^i(t), \sigma^2), & if\, t = 1 \\ x_{i(t-1)} + N(f_g^i(t+1) - f_g^i(t), \sigma^2), & if\, 1 < t \leq q \end{cases} \tag{6.11}$$

where for each $g = 1, 2$ or 3 generates a set of time series of size I; x_{it} denotes the value of series i observed at time t considering diversity (adding normally-distributed noise); **X** $= (x_{it})$ forms a data set with size $3I$; $f_g^i(t)$ is the value generated by function g for i time series at time t; $N(\mu_{TS}, \sigma_{TS}^2)$ describes a random variate drawn from a normal distribution with mean μ_{TS} and variance σ_{TS}^2.

Each model is assumed to be the true physical process underlying a group of I "known" analogous time series, where each time series has time steps of q. Noise was introduced to each group through the addition of additive noise, as described in Equation (6.5.1). The illustration of simulated time series data generated is shown in Fig. 6.1.

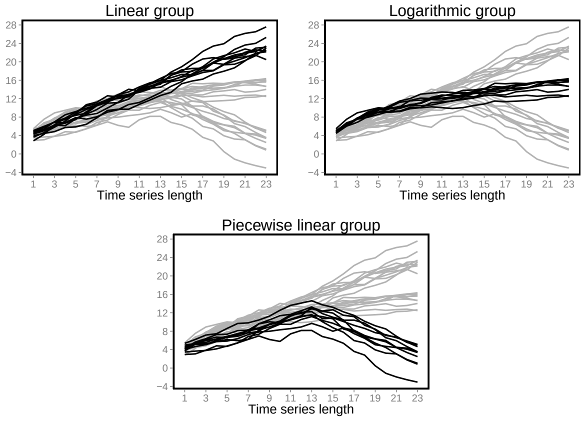


Figure 6.1: Illustration of simulated time series generated from a linear, logarithmic, and piecewise linear function

To provide insights about the impact of noise in the time series data on C-MSKF's precision, various noise levels are considered by varying the standard deviation during the data generation process. Specifically, σ varies from 0.35 to 1.15 in steps of 0.2. Under each noise level, we keep all other parameters constant and the details are provided in Table 6.1.

Table 6.1: Constant parameters for the generation of time series data across scenarios.

Parameter	Values
Prediction horizon	h=1, 2,. . . ,6
Prediction origin	T=17
Length selection	l=12, 13,...,17
No. of time series in a group	I=10
Time of trend change	p=14
Total No. of time steps	q=24

Throughout the experiments, we keep the prediction origin at $t = T$. The parameter

T is used to allow more than three observations after the time of slope change at time p (see Equation 6.10). This ensures that the generating time series satisfies the working condition of C-MSKF's algorithm (Duncan, Gorr, and Szczypula, 1994). The parameter *Length selection* indicates that we systematically drop the earliest historical points one at a time, while fixing the forecasting origin T in order to consider the effect of shorter time series. Based on the previous settings, we generate 30 data sets for each noise level so as to support our statistical analysis.

6.5.2 Compared methods

The main focus of the article is to investigate the potential of multicriteria approaches to time series clustering. Hence, we contrast different combinations of distance / standardization schemes considered in this article, and these combinations are referred to Equation 6.5. In addition, we benchmark multicriteria clustering approaches on single-criterion methods that make use of a distance metric and standardization measure.

6.5.3 Performance assessment

In analyzing our results, we measure the performance of the models by taking into account both the clustering and forecasting accuracy.

The accuracy with which analogies are identified is expected to have an influence on the accuracy of the forecasting algorithm. To evaluate the quality of clustering solutions, the Adjusted Rand Index (ARI: Hubert and Arabie, 1985) is employed throughout the experiments. The ARI measure is a popular cluster validation metric. This metric measures the agreement between two partitionings: resultant clustering solution and the ground truth, which was derived from the data generating models, *i.e.,* three mathematical models are used during the data generation process.

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

$$ARI = \frac{\sum\limits_{l,m} \binom{N_{lm}}{2} - [\sum\limits_{l} \binom{N_{l.}}{2} \cdot \sum\limits_{k} \binom{N_{.m}}{2}] / \binom{N}{2}}{\frac{1}{2}[\sum\limits_{l} \binom{N_{l.}}{2} + \sum\limits_{m} \binom{N_{.m}}{2}] - [\sum\limits_{l} \binom{N_{l\cdot}}{2} \cdot \sum\limits_{m} \binom{N_{.m}}{2}] / \binom{N}{2}} \tag{6.12}$$

where N is the total number of data items, N_{lm} represents 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 $N_{l.}$ and $N_{.m}$ represent row and column totals for row l and column m of the table, respectively.

To measure the bias of various forecasting models, Mean Error (ME) is used to measure the forecasting results.

$$ME = mean(X_t - F_t) \tag{6.13}$$

where all variables retain the same meaning as Equation 4.11.

To measure the forecasting accuracy, two well-known accuracy measures are applied, including the Mean Absolute Scaled Error (MASE: Hyndman, 2006) and the Symmetric Mean Absolute Percentage Error (sMAPE: Bergmeir, Hyndman, and Benítez, 2016), respectively.

$$MASE = mean(\left| \frac{e_t}{\frac{1}{T-1} \sum\limits_{i=2}^{T} |X_i - X_{i-1}|} \right|) \tag{6.14}$$

$$sMAPE = mean(200 \frac{|e_t|}{|X_t| + |F_t|}) \tag{6.15}$$

where T refers to the forecasting origin and the rest variables retain the same meaning as Equation 4.11.

To provide in-depth insights regarding the forecasting performance, we assess the forecasting results using the average, maximum, minimum and median values of the accuracy measures presented above. Specifically, the forecasting results are calculated by taking the average / maximum / median / minimum across 6 forecasting horizons, 30 replicates, 30 time series. Further, some of our results are analyzed by breaking up in terms of key factors that are found to impact the forecasting accuracy of C-MSKF methods, particularly the noise level and the forecasting horizon.

6.6 Results

6.6.1 Forecasting accuracy comparison across different noise levels

Tables 6.2, 6.4, 6.3, and 6.5 present comparisons on the C-MSKF's forecasting accuracy after the implementation of different clustering approaches for the identification of analogous time series. The ME measure employed here primarily aims to evaluate the bias of the forecasting approaches. Generally speaking, this measure yields almost the same forecasting errors across clustering methods at a specific noise level. Consequently, it can be meaningless to judge the forecasting performance of the contestant approaches based on the results derived from the ME measure. In terms of the model bias, Tables 6.2 and 6.5 produce negative bias in scenarios 1 and 3, whereas positive bias in scenarios 2, 4 and 5.

By observing forecasting results shown by MASE, MSE and sMAPE measures (see Tables 6.2, 6.4, 6.3, and 6.5), it is evident that the Eucl_MnMx_Zsc method consistently shows the highest forecasting accuracy from scenario 1 to 5. This means that the multicriteria clustering approach outperforms the contestant methods by clustering time series data using the Euclidean distance, with the consideration of two criteria, *i.e.,* the min-max method and the z-score method for the normalization of time series data. In addition, there is no agreement on the second best performing method among the compared methods.

Considering the quality of time series clustering alone, the Eucl_MnMx_Zsc (multicriteria clustering approaches using multiple standardization techniques) returns the best clustering results in all scenarios considered. The rest competitive clustering methods perform as the second best across all scenarios. This indicates the robustness of the Eucl_MnMx_Zsc method concerning the recovery of a high-quality clustering solution. Additionally, as the time series becomes noisier, from scenario 1 to 5, the clustering quality generally decreases, and this applies to all methods compared here. Correspondingly, the good performance of the Eucl_MnMx_Zsc method on clustering results translates to satisfactory forecasting accuracy of C-MSKF methods in the forecasting stage from scenario 1 to 5.

To summarize, multicriteria clustering approaches Eucl_Pear_MnMx and Eucl_MnMx_Zsc demonstrate better clustering results in general (see Tables 6.2, 6.4, 6.5, and 6.3), and the improvement on clustering performance further translates to better forecasting precision of C-MSKF methods such as the Eucl_MnMx_Zsc method. However, the Eucl_Pear_Zsc and Pear_MnMx_Zsc approaches show almost no improvements over single-criterion clustering methods by using the time series data alone.

As the criteria are combined at the distance function level, we further investigate the correlation between pairs of dissimilarity matrices derived from the clustering methods in order to give more information regarding the effectiveness of different combinations. In general, the correlation for each pair of methods increases as the noise levels increases. Particularly in scenario 1, Eucl_MnMx is uncorrelated to Pear_MnMx, Eucl_Zsc and Pear_Zsc. Table 6.6 shows that the correlation between Pear_MinMax and Pear_Zsc methods gives the largest values among different combinations across scenario 1 to 5.

This implies that multicriteria clustering methods Pear_MinMax_Zsc utilizes the highest correlated standardization techniques when considering the Pearson correlation's coefficients for measuring the similarity between time series. Among these values, Eucl_MinMax and Eucl_Zsc give the least correlation values across noise levels. This might imply that Eucl_MinMax_Zsc, which provides the highest forecasting accuracy, is positively influenced by the concurrent consideration of least correlated standardization techniques on the final prediction results.

6.6.2 Performance comparison across noise levels, broken up by forecasting horizons

To gain more insights regarding the performance, we further break up our results by 6 forecasting horizons to investigate the changes on this. Tables presented below are concerned with the break-up forecasting results across 30 data sets, 6 time series lengths, broken up by 6 prediction horizons across 5 noise levels. Following the previous section, we measure the forecasting accuracy of the models using the MASE, ME, MSE and sMAPE metrics. Regarding the ME measure, clustering approaches show the same forecasting accuracy of the C-MSKF method at a particular forecasting horizon. This agrees with the previous conclusions drawn in Section 6.6.1. Again, average ME and median ME show negative bias in scenarios 1 and 3. On the contrast, positive bias can be found in scenarios 2, 4 and 5. This indicates that the contestant methods employed in this experiment result in over-forecast bias in 90 out of 150 cases (5 noise levels and 6 forecasting horizons) for each accuracy measure.

In terms of the forecasting accuracy, across noise levels, the Eucl_MnMx_Zsc is almost the best performing method among the contestant methods across 5 scenarios and 6 forecasting horizons. This has been confirmed by MASE, MSE and sMAPE measures (based on the average, maximum, median, and minimum values across 30 replicates, 6 time series lengths).

6.7 Conclusion

Building upon our previous work, this paper further explores the potential of integrating multiple criteria in the context of time series clustering. Our results are expected to have particular relevance in the context of forecasting of short and volatile time series. We provide empirical evidence to support the implementation of multicriteria approaches for time series clustering, which defines criteria at a variety of levels. Here, we consider scenarios where only two criteria are considered: distance metrics / standardization techniques. Primarily, the effectiveness of different clustering approaches is contrasted based on the C-MSKF's forecasting accuracy. Overall, the Eucl_Pear_MnMx, Eucl_Pear_Zsc, Eucl_MnMx_Zsc methods generally perform better than single-criterion clustering approaches: Eucl_MnMx, Pear_MnMx, Eucl_Zsc, and Pear_Zsc, across different 5 noise levels, as measured by MASE, MSE and sMAPE. However, the Pear_MnMx_Zsc approach shows no improvement on C-MSKF's accuracy after the application of single-criterion clustering methods. Particularly, the Eucl_MnMx_Zsc method consistently shows the best clustering and forecasting accuracy across noise levels.

According to an analysis of the correlation between criteria, the least correlated criteria tend to give better clustering and forecasting results in multicriteria clustering approaches. This might be because the additional value added through the integration of complementary criteria. As discussed before, there is no universally accepted definition of similarities among time series. By capturing capturing different notions of similarity, the use of multiple criteria during the clustering of analogous time series might give rise to better clustering results and thus the improved forecasting accuracy.

Regarding our current experiments, we refrained from experimenting with different model selection methods, but instead base our analysis on the most promising model selection methods developed in **Chapter 4**, in which the Silhouette Width measure for the determination of the number of clusters, and the best average historical forecasting performance to determine the best weight interval. Particularly, $T = 17$ was used for the weight selection, and the data points on $t \leq T$ period were included in the clustering step. From our findings, we show that multicriteria clustering approaches are capable of improving the clustering quality compared to the single-criterion clustering of time series data. However, one possible limitation underlying the present work is the that forecasting results may be affected by implementing different model selection models.

In our future work, it would be meaningful to further validate our multicriteria approaches in different real-world applications. It is likely that the advantage of multicriteria clustering approaches will decrease in situations where the noise is highly correlated, or where the reliability of individual criteria is poor. An increase in the noise level of the time series data may also introduce challenges: As pointed out in **Chapter 4**, the weight selection method used in the model selection step might cause the noticeable decrease in the clustering accuracy in such settings. Currently, we refrain from experimenting with multiple distance metrics and various standardization techniques at the same time to avoid complicating the problem. However, it might be useful to investigate further possible combinations of techniques. For example, our current work used two raw-data-based methods to define the notion of similarity among time series, but more advanced distance metrics such as the model-based technique could be employed. Such specialized distance metric might be more powerful at capturing specific characteristics of the time series data, and thus a combination of various types of distance metrics is a valuable avenue for future research.

Table 6.2: Summary of clustering and C-MSKF's performance for simulated time series data with multiple criteria across noise levels, from scenario 1 to 5. The results are the *average* across 30 replicates, 6 time series lengths. The best performing method is highlighted in bold face, and the second-best performing method is highlighted in italic, bold face.

	Scenarios	Eucl	Pear	Eucl	Pear	Eucl_Pear	Eucl_Pear	Eucl	Pear
		MnMx	MnMx	Zsc	Zsc	MnMx	Zsc	MnMx_Zsc	MnMx_Zsc
Average ARI	S_1	***0.71***	0.66	0.69	0.66	0.7	0.67	**0.72**	0.66
	S_2	0.39	0.48	0.48	0.48	***0.54***	0.48	**0.55**	0.48
	S_3	0.21	0.32	0.34	0.32	***0.36***	0.34	**0.38**	0.32
	S_4	0.15	0.26	***0.29***	0.26	0.28	0.28	**0.3**	0.26
	S_5	0.09	***0.21***	***0.21***	***0.21***	**0.22**	***0.21***	**0.22**	***0.21***
Average MASE	S_1	1.56	0.99	***0.98***	0.99	**0.93**	***0.98***	**0.93**	0.99
	S_2	3.32	1.59	1.56	1.59	***1.52***	1.55	**1.48**	1.59
	S_3	3.39	2.05	1.97	2.05	1.96	***1.94***	**1.87**	2.05
	S_4	3.25	2.12	2.06	2.12	2.09	***2.05***	**2.02**	2.12
	S_5	3.21	2.29	***2.24***	2.29	2.27	***2.24***	**2.21**	2.29
Average MSE	S_1	0.72	***0.17***	0.18	***0.17***	**0.16**	***0.17***	**0.16**	***0.17***
	S_2	2.35	0.68	0.66	0.68	***0.65***	***0.65***	**0.6**	0.68
	S_3	2.8	1.25	1.21	1.25	***1.16***	1.17	**1.06**	1.25
	S_4	3.11	1.8	1.76	1.8	***1.72***	1.73	**1.62**	1.8
	S_5	3.5	2.24	***2.13***	2.24	2.2	***2.13***	**2.06**	2.24
Average sMAPE	S_1	38.32	21.89	21.9	21.89	***21.12***	21.75	**21.09**	21.89
	S_2	97.72	39.09	39.39	39.09	***38.38***	38.93	**37.55**	39.09
	S_3	108.75	58.75	58.15	58.75	57.54	***57.41***	**55.86**	58.75
	S_4	109.86	***62.6***	63.17	***62.6***	63.51	62.62	**62.29**	***62.6***
	S_5	116	72.75	72.74	72.75	***72.32***	72.48	**71.51**	72.75
Average ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	**0.02**	**0.02**	***0.03***	**0.02**	***0.03***	***0.03***	***0.03***	**0.02**
	S_3	**-0.03**	**-0.03**	***-0.02***	**-0.03**	**-0.03**	**-0.03**	***-0.02***	**-0.03**
	S_4	**0.07**	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***	***0.08***
	S_5	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**

Table 6.3: Summary of clustering and C-MSKF's performance for simulated time series data with multiple criteria across noise levels, from scenario 1 to 5. The results are the *minimum* value across 30 replicates, six time series lengths. The best performing method is highlighted in bold face, and the second-best performing method is highlighted in italic, bold face.

	Scenarios	Eucl	Pear	Eucl	Pear	Eucl_Pear	Eucl_Pear	Eucl	Pear
		MnMx	MnMx	Zsc	Zsc	MnMx	Zsc	MnMx_Zsc	MnMx_Zsc
Min ARI	S_1	***0.36***	**0.39**	**0.39**	**0.39**	**0.39**	**0.39**	**0.39**	**0.39**
	S_2	0.05	***0.13***	**0.29**	***0.13***	***0.13***	**0.29**	**0.29**	***0.13***
	S_3	0.01	0.04	0.04	0.04	***0.08***	0.04	**0.11**	0.04
	S_4	-0.03	-0.02	**0.06**	-0.02	-0.01	**0.06**	***0.04***	-0.02
	S_5	-0.04	-0.04	-0.04	-0.04	***0***	-0.04	**0.02**	-0.04
Min MASE	S_1	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**
	S_2	0.29	0.19	***0.13***	0.19	**0.12**	***0.13***	**0.12**	0.19
	S_3	0.65	0.36	**0.24**	0.36	***0.28***	**0.24**	**0.24**	0.36
	S_4	0.52	**0.32**	**0.32**	**0.32**	**0.32**	**0.32**	***0.38***	**0.32**
	S_5	0.68	0.74	0.66	0.74	***0.46***	0.72	**0.39**	0.74
Min MSE	S_1	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**
	S_2	0.29	0.19	***0.13***	0.19	**0.12**	***0.13***	**0.12**	0.19
	S_3	0.65	0.36	**0.24**	0.36	***0.28***	**0.24**	**0.24**	0.36
	S_4	0.52	**0.32**	**0.32**	**0.32**	**0.32**	**0.32**	***0.38***	**0.32**
	S_5	0.68	0.74	0.66	0.74	***0.46***	0.72	**0.39**	0.74
Min sMAPE	S_1	***13.04***	**11.52**	**11.52**	**11.52**	**11.52**	**11.52**	**11.52**	**11.52**
	S_2	31.75	24.57	24.57	24.57	***20.97***	24.57	**20.46**	24.57
	S_3	50.97	33.74	33.74	33.74	***33.39***	33.74	**31.27**	33.74
	S_4	50.32	39.11	**37.77**	39.11	***38.78***	39.12	38.79	39.11
	S_5	45.02	44.9	***42.98***	44.9	44.92	43.22	**42.63**	44.9
Min ME	S_1	**-0.19**	**-0.19**	**-0.19**	**-0.19**	**-0.19**	**-0.19**	**-0.19**	**-0.19**
	S_2	**-0.39**	**-0.39**	**-0.39**	**-0.39**	***-0.38***	**-0.39**	**-0.39**	**-0.39**
	S_3	***-0.39***	**-0.4**	**-0.4**	**-0.4**	***-0.39*** ‘	**-0.4**	***-0.39***	**-0.4**
	S_4	**-0.47**	**-0.47**	**-0.47**	**-0.47**	**-0.47**	**-0.47**	**-0.47**	**-0.47**
	S_5	**-0.29**	**-0.29**	**-0.29**	**-0.29**	**-0.29**	**-0.29**	**-0.29**	**-0.29**

Table 6.4: Summary of clustering and C-MSKF's performance for simulated time series data with multiple criteria across noise levels, from scenario 1 to 5. The results are the *maximum* value across 30 replicates, six time series lengths. The best performing method is highlighted in bold face, and the second-best performing method is highlighted in italic, bold face.

	Scenarios	Eucl	Pear	Eucl	Pear	Eucl_Pear	Eucl_Pear	Eucl	Pear
		MnMx	MnMx	Zsc	Zsc	MnMx	Zsc	MnMx_Zsc	MnMx_Zsc
Max ARI	S_1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
	S_2	***0.9***	***0.9***	***0.9***	***0.9***	**1**	***0.9***	**1**	***0.9***
	S_3	0.61	0.64	0.67	0.64	***0.72***	0.67	**0.85**	0.64
	S_4	0.5	0.55	0.53	0.55	***0.62***	0.53	**0.66**	0.55
	S_5	0.55	0.5	0.5	0.5	***0.56***	0.52	**0.58**	0.5
Max MASE	S_1	4.54	1.46	1.4	1.46	***1.34***	1.46	**1.33**	1.46
	S_2	4.91	2.43	***2.1***	2.43	2.43	***2.1***	**2.03**	2.43
	S_3	4.61	3.33	***2.81***	3.33	3.19	***2.81***	**2.67**	3.33
	S_4	4.45	3.37	***2.8***	3.37	3.19	***2.84***	2.88	3.37
	S_5	4.37	3.67	3.67	3.67	**3.4**	3.67	***3.45***	3.67
Max MSE	S_1	3.41	0.51	**0.48**	0.51	***0.49***	**0.48**	***0.49***	0.51
	S_2	8.27	***4.04***	***4.04***	***4.04***	4.06	***4.04***	**4.02**	***4.04***
	S_3	7.68	6.65	***5.91***	6.65	6.65	***5.91***	**4.64**	6.65
	S_4	6.76	4.63	4.72	4.63	***4.5***	***4.5***	**4.29**	4.63
	S_5	8.11	7.47	7.43	7.47	***6.29***	7.43	**6.16**	7.47
Max sMAPE	S_1	126.02	33.4	33.6	33.4	**32.43**	33.4	***33.11***	33.4
	S_2	152.8	***64.1***	66.85	***64.1***	***64.1***	66.85	**62.8**	***64.1***
	S_3	146.1	***90.34***	90.42	***90.34***	95.16	90.42	**87.86**	***90.34***
	S_4	167.32	100.19	***100.14***	100.19	107.92	**99.02**	101.17	100.19
	S_5	170.47	**118.48**	121.69	**118.48**	131.52	121.69	***121.48***	**118.48**
Max ME	S_1	**0.19**	**0.19**	**0.19**	**0.19**	**0.19**	**0.19**	**0.19**	**0.19**
	S_2	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**
	S_3	***0.33***	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**	**0.31**
	S_4	**0.68**	**0.68**	**0.68**	**0.68**	**0.68**	**0.68**	**0.68**	**0.68**
	S_5	**0.44**	***0.46***	***0.46***	***f0.46***	**0.44**	***0.46***	**0.44**	***0.46***

Table 6.5: Summary of clustering and C-MSKF's performance for simulated time series data with multiple criteria across noise levels, from scenario 1 to 5. The results are the *median* value across 30 replicates, six time series lengths. The best performing method is highlighted in bold face, and the second-best performing method is highlighted in italic, bold face.

	Scenarios	Eucl	Pear	Eucl	Pear	Eucl_Pear	Eucl_Pear	Eucl	Pear
		MnMx	MnMx	Zsc	Zsc	MnMx	Zsc	MnMx_Zsc	MnMx_Zsc
Median ARI	S_1	0.69	0.62	***0.7***	0.62	0.68	0.67	**0.76**	0.62
	S_2	0.37	0.48	0.48	0.48	***0.52***	0.47	**0.53**	0.48
	S_3	0.19	0.33	0.33	0.33	***0.36***	0.34	**0.37**	0.33
	S_4	0.14	0.27	**0.29**	0.27	**0.29**	***0.28***	***0.28***	0.27
	S_5	0.08	***0.2***	***0.2***	***0.2***	**0.21**	***0.2***	***0.2***	***0.2***
Median MASE	S_1	1.33	0.98	0.98	0.98	**0.93**	***0.97***	**0.93**	0.98
	S_2	3.49	1.54	1.54	1.54	***1.51***	1.52	**1.48**	1.54
	S_3	3.45	1.99	1.95	1.99	***1.92***	1.93	**1.88**	1.99
	S_4	3.33	2.07	2.04	2.07	2.05	***2.02***	**1.98**	2.07
	S_5	3.23	2.3	***2.22***	2.3	2.25	2.23	**2.18**	2.3
Median MSE	S_1	0.46	***0.16***	***0.16***	***0.16***	**0.15**	***0.16***	**0.15**	***0.16***
	S_2	2.18	0.57	0.54	0.57	0.54	***0.53***	**0.48**	0.57
	S_3	2.79	1.07	1.02	1.07	***0.97***	0.98	**0.9**	1.07
	S_4	2.89	1.63	***1.58***	1.63	1.61	1.59	**1.5**	1.63
	S_5	3.21	2.01	***1.88***	2.01	1.99	1.93	**1.87**	2.01
Median sMAPE	S_1	32.54	22.01	21.95	22.01	***20.76***	21.78	**20.7**	22.01
	S_2	105.08	38.03	38.3	38.03	***37.03***	37.59	**36.06**	38.03
	S_3	111.15	58.05	58.64	58.05	***56.69***	57.23	**55.38**	58.05
	S_4	110.73	62.39	62.68	62.39	62.57	***62.02***	**61.42**	62.39
	S_5	117.65	71.25	70.87	71.25	70.11	***70.71***	**69.16**	71.25
Median ME	S_1	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	S_2	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**	**0.03**
	S_3	**-0.04**	***-0.03***	***-0.03***	***-0.03***	***-0.03***	***-0.03***	***-0.03***	***-0.03***
	S_4	**0.08**	***0.09***	**0.08**	***0.09***	***0.09***	**0.08**	***0.09***	***0.09***
	S_5	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**	**0.04**

Table 6.6: The correlation of dissimilarity matrices derived from single-criterion clustering methods across noise levels, from scenario 1 to 5. The expected values are averaged across 30 replicates and 6 time series lengths.

Scenarios		Eucl_MnMx	Pear_MnMx	Eucl_Zsc
S_1	Pear_MnMx	-0.05		
	Eucl_Zsc	-0.04	0.16	
	Pear_Zsc	-0.05	0.19	0.16
S_2	Pear_MnMx	0.11		
	Eucl_Zsc	0.09	0.27	
	Pear_Zsc	0.11	0.32	0.27
S_3	Pear_MnMx	0.18		
	Eucl_Zsc	0.16	0.4	
	Pear_Zsc	0.18	0.47	0.4
S_4	Pear_MnMx	0.17		
	Eucl_Zsc	0.16	0.38	
	Pear_Zsc	0.17	0.44	0.38
S_5	Pear_MnMx	0.21		
	Eucl_Zsc	0.19	0.4	
	Pear_Zsc	0.21	0.46	0.4

Table 6.7: Summary of forecasting accuracy on C-MSKF by *average* MASE using different clustering methods based on time series data. Results are obtained by taking the *average* across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	1.19	1.35	1.49	1.64	1.78	1.93
	Pear_MnMx	0.74	0.85	0.94	1.04	1.13	1.23
	Eucl_Zsc	0.74	0.84	0.93	1.03	1.12	1.22
	Pear_Zsc	0.74	0.85	0.94	1.04	1.13	1.23
	Eucl_Pear_MnMx	***0.7***	***0.8***	***0.89***	***0.98***	***1.07***	***1.17***
	Pear_Pear_Zsc	0.73	0.84	0.93	1.02	1.12	1.22
	Eucl_MnMx_Zsc	**0.69**	**0.79**	**0.88**	**0.97**	**1.06**	**1.16**
	Pear_MnMx_Zsc	0.74	0.85	0.94	1.04	1.13	1.23
S_2	Eucl_MnMx	2.5	2.83	3.15	3.49	3.81	4.13
	Pear_MnMx	1.16	1.33	1.5	1.68	1.86	2.03
	Eucl_Zsc	1.13	1.3	1.47	1.65	1.83	2
	Pear_Zsc	1.16	1.33	1.5	1.68	1.86	2.03
	Eucl_Pear_MnMx	***1.12***	***1.27***	***1.43***	***1.6***	***1.77***	***1.94***
	Pear_Pear_Zsc	***1.12***	1.28	1.45	1.63	1.81	1.99
	Eucl_MnMx_Zsc	**1.06**	**1.23**	**1.39**	**1.56**	**1.73**	**1.89**
	Pear_MnMx_Zsc	1.16	1.33	1.5	1.68	1.86	2.03
S_3	Eucl_MnMx	2.53	2.89	3.23	3.56	3.89	4.22
	Pear_MnMx	1.47	1.71	1.94	2.17	2.39	2.61
	Eucl_Zsc	1.41	1.64	1.87	2.08	2.3	2.51
	Pear_Zsc	1.47	1.71	1.94	2.17	2.39	2.61
	Eucl_Pear_MnMx	***1.39***	1.63	1.86	2.08	2.3	2.52
	Pear_Pear_Zsc	***1.39***	***1.62***	***1.84***	***2.05***	***2.27***	***2.48***
	Eucl_MnMx_Zsc	**1.33**	**1.55**	**1.77**	**1.98**	**2.19**	**2.4**
	Pear_MnMx_Zsc	1.47	1.71	1.94	2.17	2.39	2.61
S_4	Eucl_MnMx	2.47	2.78	3.1	3.41	3.72	4.02
	Pear_MnMx	1.52	1.76	2.01	2.24	2.48	2.72
	Eucl_Zsc	1.47	1.7	1.95	***2.18***	2.42	***2.65***
	Pear_Zsc	1.52	1.76	2.01	2.24	2.48	2.72
	Eucl_Pear_MnMx	1.48	1.72	1.97	2.21	2.45	2.7
	Pear_Pear_Zsc	***1.46***	***1.69***	***1.94***	***2.18***	***2.41***	***2.65***
	Eucl_MnMx_Zsc	**1.42**	**1.65**	**1.9**	**2.14**	**2.38**	**2.61**
	Pear_MnMx_Zsc	1.52	1.76	2.01	2.24	2.48	2.72
S_5	Eucl_MnMx	2.44	2.75	3.06	3.37	3.67	3.97
	Pear_MnMx	1.64	1.93	2.17	2.42	2.67	2.92
	Eucl_Zsc	***1.59***	***1.87***	***2.12***	***2.37***	***2.62***	***2.87***
	Pear_Zsc	1.64	1.93	2.17	2.42	2.67	2.92
	Eucl_Pear_MnMx	1.6	1.89	2.15	2.4	2.66	2.91
	Pear_Pear_Zsc	***1.59***	***1.87***	***2.12***	***2.37***	***2.62***	***2.87***
	Eucl_MnMx_Zsc	**1.54**	**1.83**	**2.08**	**2.34**	**2.6**	**2.85**
	Pear_MnMx_Zsc	1.64	1.93	2.17	2.42	2.67	2.92

Table 6.8: Summary of forecasting accuracy on C-MSKF by *average* ME using different clustering methods based on time series data. Results are obtained by taking the *average* across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Pear_MnMx	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Eucl_Zsc	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Pear_Zsc	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Eucl_Pear_MnMx	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Pear_Pear_Zsc	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Eucl_MnMx_Zsc	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
	Pear_MnMx_Zsc	**0**	**0**	**-0.01**	**-0.01**	**-0.02**	**-0.02**
S_2	Eucl_MnMx	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Pear_MnMx	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Eucl_Zsc	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Pear_Zsc	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Eucl_Pear_MnMx	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Pear_Pear_Zsc	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Eucl_MnMx_Zsc	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
	Pear_MnMx_Zsc	**0.01**	**0.01**	**0.02**	**0.03**	**0.04**	**0.04**
S_3	Eucl_MnMx	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Pear_MnMx	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Eucl_Zsc	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Pear_Zsc	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Eucl_Pear_MnMx	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Pear_Pear_Zsc	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Eucl_MnMx_Zsc	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
	Pear_MnMx_Zsc	**-0.02**	**-0.02**	**-0.02**	**-0.03**	**-0.03**	**-0.04**
S_4	Eucl_MnMx	**0.04**	**0.05**	**0.07**	**0.08**	**0.1**	**0.11**
	Pear_MnMx	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
	Eucl_Zsc	**0.04**	**0.05**	***0.08***	***0.09***	**0.1**	**0.11**
	Pear_Zsc	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
	Eucl_Pear_MnMx	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
	Pear_Pear_Zsc	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
	Eucl_MnMx_Zsc	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
	Pear_MnMx_Zsc	**0.04**	***0.06***	***0.08***	***0.09***	**0.1**	**0.11**
S_5	Eucl_MnMx	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Pear_MnMx	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Eucl_Zsc	**0.01**	**0.03**	**0.04**	**0.04**	**0.05**	**0.06**
	Pear_Zsc	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Eucl_Pear_MnMx	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Pear_Pear_Zsc	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Eucl_MnMx_Zsc	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**
	Pear_MnMx_Zsc	**0.01**	**0.03**	**0.04**	***0.05***	**0.05**	**0.06**

Table 6.9: Summary of forecasting accuracy on C-MSKF by *average* MSE using different clustering methods based on time series data. Results are obtained by taking the *average* across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	0.39	0.51	0.63	0.77	0.93	1.11
	Pear_MnMx	***0.09***	***0.12***	***0.15***	0.19	***0.22***	***0.27***
	Eucl_Zsc	***0.09***	***0.12***	***0.15***	0.19	0.23	***0.27***
	Pear_Zsc	***0.09***	***0.12***	***0.15***	0.19	***0.22***	***0.27***
	Eucl_Pear_MnMx	**0.08**	**0.11**	**0.14**	**0.17**	**0.21**	**0.25**
	Pear_Pear_Zsc	***0.09***	***0.12***	***0.15***	***0.18***	***0.22***	***0.27***
	Eucl_MnMx_Zsc	***0.09***	**0.11**	**0.14**	**0.17**	**0.21**	**0.25**
	Pear_MnMx_Zsc	***0.09***	***0.12***	***0.15***	0.19	***0.22***	***0.27***
S_2	Eucl_MnMx	1.24	1.61	2.02	2.53	3.06	3.64
	Pear_MnMx	0.29	0.4	0.54	0.74	0.95	1.18
	Eucl_Zsc	***0.27***	***0.38***	0.52	0.72	0.92	1.14
	Pear_Zsc	0.29	0.4	0.54	0.74	0.95	1.18
	Eucl_Pear_MnMx	***0.27***	***0.38***	***0.51***	***0.7***	***0.9***	***1.12***
	Pear_Pear_Zsc	***0.27***	***0.38***	***0.51***	0.71	0.91	1.13
	Eucl_MnMx_Zsc	**0.24**	**0.35**	**0.47**	**0.65**	**0.84**	**1.04**
	Pear_MnMx_Zsc	0.29	0.4	0.54	0.74	0.95	1.18
S_3	Eucl_MnMx	1.44	1.9	2.43	3.01	3.64	4.37
	Pear_MnMx	0.51	0.73	1.02	1.36	1.73	2.15
	Eucl_Zsc	0.49	0.71	0.99	1.31	1.67	2.08
	Pear_Zsc	0.51	0.73	1.02	1.36	1.73	2.15
	Eucl_Pear_MnMx	***0.46***	***0.67***	***0.94***	***1.26***	***1.61***	***2.02***
	Pear_Pear_Zsc	0.47	***0.67***	0.95	***1.26***	1.62	***2.02***
	Eucl_MnMx_Zsc	**0.43**	**0.62**	**0.87**	**1.15**	**1.47**	**1.83**
	Pear_MnMx_Zsc	0.51	0.73	1.02	1.36	1.73	2.15
S_4	Eucl_MnMx	1.62	2.1	2.69	3.34	4.05	4.84
	Pear_MnMx	0.71	1.04	1.46	1.96	2.51	3.12
	Eucl_Zsc	0.69	1.02	1.42	1.92	2.46	3.06
	Pear_Zsc	0.71	1.04	1.46	1.96	2.51	3.12
	Eucl_Pear_MnMx	***0.67***	***0.99***	***1.39***	***1.87***	***2.4***	***2.99***
	Pear_Pear_Zsc	***0.67***	***0.99***	***1.39***	1.89	2.43	3.02
	Eucl_MnMx_Zsc	**0.62**	**0.91**	**1.3**	**1.77**	**2.27**	**2.83**
	Pear_MnMx_Zsc	0.71	1.04	1.46	1.96	2.51	3.12
S_5	Eucl_MnMx	1.82	2.39	3.02	3.75	4.56	5.45
	Pear_MnMx	0.92	1.38	1.86	2.43	3.08	3.81
	Eucl_Zsc	0.87	1.3	***1.75***	***2.29***	***2.92***	***3.62***
	Pear_Zsc	0.92	1.38	1.86	2.43	3.08	3.81
	Eucl_Pear_MnMx	0.87	1.32	1.81	2.39	3.04	3.76
	Pear_Pear_Zsc	***0.86***	***1.29***	***1.75***	2.3	2.93	3.63
	Eucl_MnMx_Zsc	**0.81**	**1.23**	**1.69**	**2.23**	**2.85**	**3.54**
	Pear_MnMx_Zsc	0.92	1.38	1.86	2.43	3.08	3.81

Table 6.10: Summary of forecasting accuracy on C-MSKF by *average* sMAPE using different clustering methods based on time series data. Results are obtained by taking the *average* across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	40.94	39.13	38.11	37.53	37.2	37
	Pear_MnMx	24.19	22.66	21.69	21.16	20.87	20.77
	Eucl_Zsc	24.21	22.7	21.7	21.18	20.86	20.74
	Pear_Zsc	24.19	22.66	21.69	21.16	20.87	20.77
	Eucl_Pear_MnMx	***23.4***	***21.82***	***20.87***	***20.4***	***20.15***	***20.07***
	Pear_Pear_Zsc	24.05	22.52	21.54	21.02	20.72	20.62
	Eucl_MnMx_Zsc	**23.38**	**21.81**	**20.85**	**20.37**	**20.1**	**20**
	Pear_MnMx_Zsc	24.19	22.66	21.69	21.16	20.87	20.77
S_2	Eucl_MnMx	95.83	96.68	97.42	98.21	98.83	99.35
	Pear_MnMx	39.56	38.92	38.72	38.87	39.15	39.34
	Eucl_Zsc	39.39	39.11	39.05	39.29	39.61	39.87
	Pear_Zsc	39.56	38.92	38.72	38.87	39.15	39.34
	Eucl_Pear_MnMx	***38.84***	***38.21***	***38.02***	***38.17***	***38.41***	***38.64***
	Pear_Pear_Zsc	38.95	38.56	38.54	38.82	39.2	39.49
	Eucl_MnMx_Zsc	**37.59**	**37.22**	**37.15**	**37.43**	**37.79**	**38.11**
	Pear_MnMx_Zsc	39.56	38.92	38.72	38.87	39.15	39.34
S_3	Eucl_MnMx	100.22	105.01	108.59	111.11	112.99	114.56
	Pear_MnMx	57.24	58.43	58.88	59.04	59.33	59.6
	Eucl_Zsc	57.48	58.1	58.26	58.21	58.35	58.52
	Pear_Zsc	57.24	58.43	58.88	59.04	59.33	59.6
	Eucl_Pear_MnMx	***55.49***	***56.99***	57.68	57.96	58.38	58.72
	Pear_Pear_Zsc	56.83	57.49	***57.51***	***57.43***	***57.53***	***57.69***
	Eucl_MnMx_Zsc	**55.18**	**55.9**	**55.93**	**55.85**	**56.04**	**56.26**
	Pear_MnMx_Zsc	57.24	58.43	58.88	59.04	59.33	59.6
S_4	Eucl_MnMx	102.49	106.93	109.83	111.87	113.36	114.66
	Pear_MnMx	***59.59***	***61.03***	***62.4***	63.33	64.25	65
	Eucl_Zsc	60.54	61.96	63.06	63.77	64.55	65.16
	Pear_Zsc	***59.59***	***61.03***	***62.4***	63.33	64.25	65
	Eucl_Pear_MnMx	60.33	61.99	63.41	64.29	65.12	65.9
	Pear_Pear_Zsc	59.92	61.31	62.49	***63.25***	**64.05**	**64.69**
	Eucl_MnMx_Zsc	**58.76**	**60.64**	**62.12**	**63.2**	***64.11***	***64.87***
	Pear_MnMx_Zsc	***59.59***	***61.03***	***62.4***	63.33	64.25	65
S_5	Eucl_MnMx	108.34	112.32	115.47	118.02	120.06	121.81
	Pear_MnMx	68.8	71.37	72.74	73.64	74.55	75.39
	Eucl_Zsc	68.54	71.2	72.68	73.76	74.72	75.52
	Pear_Zsc	68.8	71.37	72.74	73.64	74.55	75.39
	Eucl_Pear_MnMx	***67.95***	***70.65***	***72.32***	***73.38***	***74.37***	***75.27***
	Pear_Pear_Zsc	68.2	70.94	72.49	73.52	74.45	***75.27***
	Eucl_MnMx_Zsc	**66.73**	**69.61**	**71.45**	**72.72**	**73.85**	**74.71**
	Pear_MnMx_Zsc	68.8	71.37	72.74	73.64	74.55	75.39

Table 6.11: Summary of model performance on C-MSKF by *minimum* MASE using different clustering methods based on time series data. Results are obtained by taking the *minimum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	***0.45***	***0.57***	**0.61**	***0.67***	***0.72***	0.78
	Pear_MnMx	0.51	0.6	***0.62***	**0.66**	**0.69**	**0.74**
	Eucl_Zsc	0.51	0.6	***0.62***	**0.66**	**0.69**	**0.74**
	Pear_Zsc	0.51	0.6	***0.62***	**0.66**	**0.69**	**0.74**
	Eucl_Pear_MnMx	**0.41**	**0.54**	**0.61**	***0.67***	***0.72***	***0.77***
	Pear_Pear_Zsc	0.51	0.6	***0.62***	**0.66**	**0.69**	**0.74**
	Eucl_MnMx_Zsc	***0.45***	***0.57***	**0.61**	***0.67***	***0.72***	0.78
	Pear_MnMx_Zsc	0.51	0.6	***0.62***	**0.66**	**0.69**	**0.74**
S_2	Eucl_MnMx	0.92	1.11	1.27	1.41	1.47	1.52
	Pear_MnMx	0.78	0.93	1.13	1.28	***1.43***	1.54
	Eucl_Zsc	0.78	0.93	1.11	1.14	**1.19**	***1.25***
	Pear_Zsc	0.78	0.93	1.13	1.28	***1.43***	1.54
	Eucl_Pear_MnMx	***0.75***	***0.77***	***0.9***	***1.05***	**1.19**	**1.22**
	Pear_Pear_Zsc	0.78	0.93	1.02	1.14	**1.19**	1.25
	Eucl_MnMx_Zsc	**0.71**	**0.73**	**0.86**	**1.03**	**1.19**	**1.22**
	Pear_MnMx_Zsc	0.78	0.93	1.13	1.28	***1.43***	1.54
S_3	Eucl_MnMx	***1.24***	1.45	1.71	***1.9***	2.08	2.28
	Pear_MnMx	**0.81**	1.05	1.18	**1.31**	***1.51***	1.71
	Eucl_Zsc	**0.81**	1.05	***1.16***	**1.31**	***1.51***	***1.66***
	Pear_Zsc	**0.81**	1.05	1.18	**1.31**	***1.51***	1.71
	Eucl_Pear_MnMx	**0.81**	**0.94**	**1.15**	**1.31**	**1.48**	**1.58**
	Pear_Pear_Zsc	**0.81**	1.05	1.18	**1.31**	***1.51***	***1.66***
	Eucl_MnMx_Zsc	**0.81**	***0.98***	***1.16***	**1.31**	***1.51***	***1.66***
	Pear_MnMx_Zsc	**0.81**	1.05	1.18	**1.31**	***1.51***	1.71
S_4	Eucl_MnMx	1.07	1.27	1.51	1.69	1.86	2.08
	Pear_MnMx	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
	Eucl_Zsc	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
	Pear_Zsc	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
	Eucl_Pear_MnMx	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
	Pear_Pear_Zsc	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
	Eucl_MnMx_Zsc	**0.94**	***1.12***	***1.24***	***1.33***	***1.41***	***1.55***
	Pear_MnMx_Zsc	***0.95***	**1.05**	**1.14**	**1.2**	**1.34**	**1.52**
S_5	Eucl_MnMx	1.34	1.36	1.57	1.77	1.97	2.18
	Pear_MnMx	1.12	***1.24***	***1.42***	***1.56***	**1.67**	**1.77**
	Eucl_Zsc	***1.11***	**1.21**	**1.34**	**1.54**	1.75	***1.86***
	Pear_Zsc	1.12	***1.24***	***1.42***	***1.56***	**1.67**	**1.77**
	Eucl_Pear_MnMx	1.12	***1.24***	***1.42***	***1.56***	**1.67**	**1.77**
	Pear_Pear_Zsc	***1.11***	**1.21**	**1.34**	**1.54**	**1.67**	**1.77**
	Eucl_MnMx_Zsc	**1.06**	**1.21**	**1.34**	**1.54**	***1.7***	***1.86***
	Pear_MnMx_Zsc	1.12	***1.24***	***1.42***	***1.56***	**1.67**	**1.77**

Table 6.12: Summary of model performance on C-MSKF by *minimum* ME using different clustering methods based on time series data. Results are obtained by taking the *minimum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	**-0.1**	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	Pear_MnMx	***-0.09***	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
	Eucl_Zsc	***-0.09***	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	Pear_Zsc	***-0.09***	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
	Eucl_Pear_MnMx	***-0.09***	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	Pear_Pear_Zsc	***-0.09***	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	Eucl_MnMx_Zsc	***-0.09***	**-0.14**	**-0.18**	**-0.21**	**-0.24**	**-0.27**
	Pear_MnMx_Zsc	***-0.09***	**-0.14**	***-0.17***	**-0.21**	**-0.24**	**-0.27**
S_2	Eucl_MnMx	**-0.15**	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Pear_MnMx	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Eucl_Zsc	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Pear_Zsc	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Eucl_Pear_MnMx	***-0.14***	**-0.28**	***-0.34***	**-0.45**	**-0.52**	***-0.57***
	Pear_Pear_Zsc	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Eucl_MnMx_Zsc	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
	Pear_MnMx_Zsc	***-0.14***	**-0.28**	**-0.35**	**-0.45**	**-0.52**	**-0.58**
S_3	Eucl_MnMx	**-0.23**	**-0.33**	***-0.36***	***-0.42***	***-0.47***	***-0.54***
	Pear_MnMx	**-0.23**	**-0.33**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	Eucl_Zsc	**-0.23**	**-0.33**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	Pear_Zsc	**-0.23**	**-0.33**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	Eucl_Pear_MnMx	**-0.23**	**-0.33**	***-0.36***	***-0.42***	***-0.47***	***-0.54***
	Pear_Pear_Zsc	**-0.23**	**-0.33**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
	Eucl_MnMx_Zsc	**-0.23**	**-0.33**	***-0.36***	***-0.42***	***-0.47***	***-0.54***
	Pear_MnMx_Zsc	**-0.23**	**-0.33**	**-0.37**	**-0.43**	**-0.48**	**-0.56**
S_4	Eucl_MnMx	**-0.23**	**-0.26**	**-0.39**	**-0.52**	**-0.65**	**-0.78**
	Pear_MnMx	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Eucl_Zsc	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Pear_Zsc	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Eucl_Pear_MnMx	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Pear_Pear_Zsc	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Eucl_MnMx_Zsc	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
	Pear_MnMx_Zsc	***-0.22***	**-0.26**	**-0.39**	**-0.52**	**-0.65**	***-0.77***
S_5	Eucl_MnMx	**-0.16**	**-0.25**	**-0.26**	***-0.29***	**-0.36**	***-0.43***
	Pear_MnMx	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	Eucl_Zsc	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	Pear_Zsc	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	Eucl_Pear_MnMx	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	***-0.43***
	Pear_Pear_Zsc	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**
	Eucl_MnMx_Zsc	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	***-0.43***
	Pear_MnMx_Zsc	**-0.16**	**-0.25**	**-0.26**	**-0.3**	**-0.36**	**-0.44**

Table 6.13: Summary of model performance on C-MSKF by *minimum* MSE using different clustering methods based on time series data. Results are obtained by taking the *minimum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	***0.03***	**0.04**	**0.04**	**0.04**	***0.05***	**0.05**
	Pear_MnMx	***0.03***	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**
	Eucl_Zsc	***0.03***	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**
	Pear_Zsc	***0.03***	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**
	Eucl_Pear_MnMx	**0.02**	**0.04**	**0.04**	**0.04**	***0.05***	**0.05**
	Pear_Pear_Zsc	***0.03***	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**
	Eucl_MnMx_Zsc	***0.03***	**0.04**	**0.04**	**0.04**	***0.05***	**0.05**
	Pear_MnMx_Zsc	***0.03***	**0.04**	**0.04**	**0.04**	**0.04**	**0.05**
S_2	Eucl_MnMx	0.12	0.19	0.25	0.32	0.41	0.47
	Pear_MnMx	0.08	***0.12***	0.17	***0.22***	***0.26***	***0.29***
	Eucl_Zsc	0.09	***0.12***	***0.12***	**0.14**	**0.16**	**0.18**
	Pear_Zsc	0.08	***0.12***	0.17	***0.22***	***0.26***	***0.29***
	Eucl_Pear_MnMx	***0.07***	**0.09**	***0.12***	**0.14**	**0.16**	**0.18**
	Pear_Pear_Zsc	***0.07***	***0.12***	***0.12***	**0.14**	**0.16**	**0.18**
	Eucl_MnMx_Zsc	**0.06**	**0.09**	**0.11**	**0.14**	**0.16**	**0.18**
	Pear_MnMx_Zsc	0.08	***0.12***	0.17	***0.22***	***0.26***	***0.29***
S_3	Eucl_MnMx	0.34	0.46	0.59	0.69	0.83	0.98
	Pear_MnMx	0.15	0.22	0.34	0.39	0.49	0.58
	Eucl_Zsc	**0.13**	0.16	***0.21***	**0.26**	**0.32**	**0.38**
	Pear_Zsc	0.15	0.22	0.34	0.39	0.49	0.58
	Eucl_Pear_MnMx	***0.14***	**0.13**	**0.2**	***0.3***	***0.4***	***0.51***
	Pear_Pear_Zsc	**0.13**	0.16	***0.21***	**0.26**	**0.32**	**0.38**
	Eucl_MnMx_Zsc	**0.13**	***0.14***	***0.21***	**0.26**	**0.32**	**0.38**
	Pear_MnMx_Zsc	0.15	0.22	0.34	0.39	0.49	0.58
S_4	Eucl_MnMx	0.26	0.33	0.45	0.53	0.66	0.86
	Pear_MnMx	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
	Eucl_Zsc	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
	Pear_Zsc	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
	Eucl_Pear_MnMx	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
	Pear_Pear_Zsc	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
	Eucl_MnMx_Zsc	***0.19***	***0.28***	***0.36***	***0.43***	***0.48***	**0.55**
	Pear_MnMx_Zsc	**0.14**	**0.19**	**0.24**	**0.31**	**0.45**	***0.58***
S_5	Eucl_MnMx	0.37	0.48	0.63	0.74	0.84	1.04
	Pear_MnMx	0.29	0.49	0.67	0.83	1	1.17
	Eucl_Zsc	***0.25***	0.46	0.57	0.76	0.89	1.04
	Pear_Zsc	0.29	0.49	0.67	0.83	1	1.17
	Eucl_Pear_MnMx	0.26	***0.34***	***0.45***	***0.5***	***0.55***	***0.67***
	Pear_Pear_Zsc	***0.25***	0.49	0.67	0.83	0.98	1.12
	Eucl_MnMx_Zsc	**0.22**	**0.29**	**0.39**	**0.42**	**0.47**	**0.57**
	Pear_MnMx_Zsc	0.29	0.49	0.67	0.83	1	1.17

Table 6.14: Summary of model performance on C-MSKF by *minimum* sMAPE using different clustering methods based on time series data. Results are obtained by taking the *minimum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	***11.86***	***12.96***	***13.21***	***13.35***	***13.41***	***13.44***
	Pear_MnMx	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Eucl_Zsc	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Pear_Zsc	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Eucl_Pear_MnMx	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Pear_Pear_Zsc	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Eucl_MnMx_Zsc	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
	Pear_MnMx_Zsc	**10.3**	**11.37**	**11.54**	**11.62**	**11.77**	**12.53**
S_2	Eucl_MnMx	30.68	30.89	31.15	32.29	32.81	32.7
	Pear_MnMx	21.86	22.63	24.18	25.64	26.35	26.77
	Eucl_Zsc	21.86	22.63	24.18	25.64	26.35	26.77
	Pear_Zsc	21.86	22.63	24.18	25.64	26.35	26.77
	Eucl_Pear_MnMx	***19.72***	***21.7***	***20.73***	***20.52***	***21.34***	***21.82***
	Pear_Pear_Zsc	21.86	22.63	24.18	25.64	26.35	26.77
	Eucl_MnMx_Zsc	**18.28**	**21.11**	**20.25**	**20.2**	**21.18**	**21.77**
	Pear_MnMx_Zsc	21.86	22.63	24.18	25.64	26.35	26.77
S_3	Eucl_MnMx	***46.59***	47.64	50.7	53.18	53.96	53.73
	Pear_MnMx	**27.11**	***30.78***	***34.68***	***35.77***	36.37	37.75
	Eucl_Zsc	**27.11**	***30.78***	***34.68***	***35.77***	36.37	37.75
	Pear_Zsc	**27.11**	***30.78***	***34.68***	***35.77***	36.37	37.75
	Eucl_Pear_MnMx	**27.11**	***30.78***	***34.68***	***35.77***	***35.98***	***36.05***
	Pear_Pear_Zsc	**27.11**	***30.78***	***34.68***	***35.77***	36.37	37.75
	Eucl_MnMx_Zsc	**27.11**	**29.87**	**31.43**	**32.88**	**32.74**	**33.6**
	Pear_MnMx_Zsc	**27.11**	***30.78***	***34.68***	***35.77***	36.37	37.75
S_4	Eucl_MnMx	43.72	48.55	52.35	52.4	51.89	53.03
	Pear_MnMx	***34.88***	39.94	40.68	***38.6***	**39.66**	***40.92***
	Eucl_Zsc	**34.81**	**36.77**	**36.6**	**37.84**	**39.66**	***40.92***
	Pear_Zsc	***34.88***	39.94	40.68	***38.6***	**39.66**	***40.92***
	Eucl_Pear_MnMx	36.09	38.55	38.85	***38.6***	**39.66**	***40.92***
	Pear_Pear_Zsc	***34.88***	39.94	40.72	***38.6***	**39.66**	***40.92***
	Eucl_MnMx_Zsc	36.09	***38.47***	***38.78***	39.84	***39.84***	**39.74**
	Pear_MnMx_Zsc	***34.88***	39.94	40.68	***38.6***	**39.66**	***40.92***
S_5	Eucl_MnMx	45.6	**43**	***43.37***	44.81	46.11	47.23
	Pear_MnMx	45.81	45.67	44.26	44.86	***44.74***	***44.07***
	Eucl_Zsc	***42.12***	***43.44***	**41.52**	**42.73**	**44.06**	**44.04**
	Pear_Zsc	45.81	45.67	44.26	44.86	44.74	***44.07***
	Eucl_Pear_MnMx	45.81	46.03	44.01	***44.64***	44.98	***44.07***
	Pear_Pear_Zsc	43.55	***43.44***	**41.52**	**42.73**	**44.06**	**44.04**
	Eucl_MnMx_Zsc	**39.86**	***43.44***	**41.52**	**42.73**	**44.06**	44.18
	Pear_MnMx_Zsc	45.81	45.67	44.26	44.86	***44.74***	***44.07***

Table 6.15: Summary of model performance on C-MSKF by *maximum* MASE using different clustering methods based on time series data. Results are obtained by taking the *maximum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	3.33	3.85	4.35	4.77	5.22	5.72
	Pear_MnMx	1.12	1.25	1.37	1.53	1.69	***1.81***
	Eucl_Zsc	1.12	1.19	***1.3***	***1.46***	***1.61***	1.72
	Pear_Zsc	1.12	1.25	1.37	1.53	1.69	***1.81***
	Eucl_Pear_MnMx	***1.01***	***1.14***	**1.26**	**1.41**	**1.54**	**1.69**
	Pear_Pear_Zsc	1.12	1.25	1.37	1.53	1.69	***1.81***
	Eucl_MnMx_Zsc	**1**	**1.12**	**1.26**	**1.41**	**1.54**	**1.69**
	Pear_MnMx_Zsc	1.12	1.25	1.37	1.53	1.69	***1.81***
S_2	Eucl_MnMx	3.79	4.21	4.65	5.13	5.6	6.11
	Pear_MnMx	2	2.11	2.34	2.55	2.71	2.9
	Eucl_Zsc	***1.44***	***1.7***	***1.96***	***2.27***	***2.5***	***2.72***
	Pear_Zsc	2	2.11	2.34	2.55	2.71	2.9
	Eucl_Pear_MnMx	2	2.11	2.34	2.55	2.71	2.9
	Pear_Pear_Zsc	***1.44***	***1.7***	***1.96***	***2.27***	***2.5***	***2.72***
	Eucl_MnMx_Zsc	**1.4**	**1.61**	**1.89**	**2.2**	**2.44**	**2.66**
	Pear_MnMx_Zsc	2	2.11	2.34	2.55	2.71	2.9
S_3	Eucl_MnMx	3.56	4.02	4.36	4.77	5.2	5.76
	Pear_MnMx	2.24	2.67	3.14	3.56	3.97	4.38
	Eucl_Zsc	2.14	***2.41***	***2.67***	***2.92***	***3.22***	***3.53***
	Pear_Zsc	2.24	2.67	3.14	3.56	3.97	4.38
	Eucl_Pear_MnMx	***2.09***	2.52	3.01	3.44	3.85	4.24
	Pear_Pear_Zsc	2.14	***2.41***	***2.67***	***2.92***	***3.22***	***3.53***
	Eucl_MnMx_Zsc	**1.95**	**2.13**	**2.47**	**2.8**	**3.15**	**3.51**
	Pear_MnMx_Zsc	2.24	2.67	3.14	3.56	3.97	4.38
S_4	Eucl_MnMx	3.29	3.77	4.31	4.76	5.12	5.48
	Pear_MnMx	2.49	2.81	3.25	3.58	3.87	4.22
	Eucl_Zsc	***2.09***	**2.36**	***2.73***	**2.94**	**3.21**	**3.5**
	Pear_Zsc	2.49	2.81	3.25	3.58	3.87	4.22
	Eucl_Pear_MnMx	2.36	2.65	3.07	3.4	3.68	4.02
	Pear_Pear_Zsc	**2.07**	***2.43***	***2.73***	***2.99***	***3.27***	3.57
	Eucl_MnMx_Zsc	2.24	2.51	**2.69**	3.01	3.29	***3.54***
	Pear_MnMx_Zsc	2.49	2.81	3.25	3.58	3.87	4.22
S_5	Eucl_MnMx	3.28	3.86	4.25	4.62	4.96	5.29
	Pear_MnMx	2.62	3.09	3.49	3.87	4.26	4.7
	Eucl_Zsc	2.62	3.09	3.49	3.87	4.26	4.7
	Pear_Zsc	2.62	3.09	3.49	3.87	4.26	4.7
	Eucl_Pear_MnMx	***2.5***	***2.92***	**3.27**	**3.57**	**3.9**	**4.22**
	Pear_Pear_Zsc	2.62	3.09	3.49	3.87	4.26	4.7
	Eucl_MnMx_Zsc	**2.47**	**2.9**	***3.29***	***3.64***	***3.99***	***4.4***
	Pear_MnMx_Zsc	2.62	3.09	3.49	3.87	4.26	4.7

Table 6.16: Summary of model performance on C-MSKF by *maximum* ME using different clustering methods based on time series data. Results are obtained by taking the *maximum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	**0.11**	**0.14**	**0.17**	**0.21**	**0.23**	**0.26**
	Pear_MnMx	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Eucl_Zsc	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Pear_Zsc	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Eucl_Pear_MnMx	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Pear_Pear_Zsc	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Eucl_MnMx_Zsc	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
	Pear_MnMx_Zsc	**0.11**	**0.14**	**0.17**	**0.21**	***0.24***	***0.27***
S_2	Eucl_MnMx	**0.14**	**0.24**	**0.28**	**0.34**	**0.37**	**0.47**
	Pear_MnMx	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Eucl_Zsc	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Pear_Zsc	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Eucl_Pear_MnMx	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Pear_Pear_Zsc	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Eucl_MnMx_Zsc	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
	Pear_MnMx_Zsc	**0.14**	**0.24**	**0.28**	**0.34**	***0.38***	***0.48***
S_3	Eucl_MnMx	***0.2***	***0.28***	***0.31***	***0.35***	0.39	0.45
	Pear_MnMx	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
	Eucl_Zsc	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
	Pear_Zsc	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
	Eucl_Pear_MnMx	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
	Pear_Pear_Zsc	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
	Eucl_MnMx_Zsc	**0.19**	**0.27**	**0.29**	**0.33**	**0.36**	**0.41**
	Pear_MnMx_Zsc	***0.2***	**0.27**	**0.29**	**0.33**	***0.37***	***0.42***
S_4	Eucl_MnMx	**0.49**	**0.59**	**0.72**	**0.77**	**0.75**	**0.74**
	Pear_MnMx	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Eucl_Zsc	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Pear_Zsc	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Eucl_Pear_MnMx	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Pear_Pear_Zsc	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Eucl_MnMx_Zsc	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
	Pear_MnMx_Zsc	**0.49**	**0.59**	**0.72**	**0.77**	***0.76***	***0.75***
S_5	Eucl_MnMx	**0.27**	**0.37**	**0.39**	**0.38**	***0.55***	***0.7***
	Pear_MnMx	***0.28***	***0.38***	***0.41***	0.4	0.57	0.73
	Eucl_Zsc	***0.28***	***0.38***	0.42	0.41	0.57	0.73
	Pear_Zsc	***0.28***	***0.38***	***0.41***	0.4	0.57	0.73
	Eucl_Pear_MnMx	**0.27**	**0.37**	**0.39**	0.4	**0.52**	**0.67**
	Pear_Pear_Zsc	***0.28***	***0.38***	0.42	0.41	0.57	0.73
	Eucl_MnMx_Zsc	**0.27**	**0.37**	**0.39**	***0.39***	0.53	**0.67**
	Pear_MnMx_Zsc	***0.28***	***0.38***	***0.41***	0.4	0.57	0.73

Table 6.17: Summary of model performance on C-MSKF by *maximum* MSE using different clustering methods based on time series data. Results are obtained by taking the *maximum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	***1.65***	2.25	2.89	3.66	**4.54**	**5.46**
	Pear_MnMx	**0.23**	0.36	0.47	***0.54***	**0.65**	**0.78**
	Eucl_Zsc	**0.23**	**0.31**	**0.4**	**0.53**	**0.65**	**0.78**
	Pear_Zsc	**0.23**	0.36	0.47	***0.54***	**0.65**	**0.78**
	Eucl_Pear_MnMx	**0.23**	***0.33***	***0.43***	**0.53**	**0.65**	**0.78**
	Pear_Pear_Zsc	**0.23**	**0.31**	**0.4**	**0.53**	**0.65**	**0.78**
	Eucl_MnMx_Zsc	**0.23**	***0.33***	***0.43***	**0.53**	**0.65**	**0.78**
	Pear_MnMx_Zsc	**0.23**	0.36	0.47	***0.54***	**0.65**	**0.78**
S_2	Eucl_MnMx	2.85	***4.24***	***6.04***	***9.23***	***12.5***	***14.76***
	Pear_MnMx	***0.89***	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Eucl_Zsc	***0.89***	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Pear_Zsc	***0.89***	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Eucl_Pear_MnMx	1	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Pear_Pear_Zsc	***0.89***	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Eucl_MnMx_Zsc	**0.78**	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
	Pear_MnMx_Zsc	***0.89***	**1.55**	**2.48**	**4.61**	**6.76**	**7.94**
S_3	Eucl_MnMx	3.28	4.29	6.12	8.19	10.55	13.66
	Pear_MnMx	1.64	2.63	4.93	7.49	9.99	13.23
	Eucl_Zsc	***1.3***	***2.19***	***4.24***	***6.53***	***8.94***	***12.25***
	Pear_Zsc	1.64	2.63	4.93	7.49	9.99	13.23
	Eucl_Pear_MnMx	1.64	2.63	4.93	7.49	9.99	13.23
	Pear_Pear_Zsc	***1.3***	***2.19***	***4.24***	***6.53***	***8.94***	***12.25***
	Eucl_MnMx_Zsc	**1.17**	**1.7**	**3.38**	**5.25**	**7.01**	**9.32**
	Pear_MnMx_Zsc	1.64	2.63	4.93	7.49	9.99	13.23
S_4	Eucl_MnMx	3.41	4.45	5.96	7.66	8.77	10.34
	Pear_MnMx	2.1	3.07	***4.03***	**5.19**	***6.27***	***7.14***
	Eucl_Zsc	***1.97***	***2.87***	**3.88**	***5.43***	6.56	7.63
	Pear_Zsc	2.1	3.07	***4.03***	**5.19**	***6.27***	***7.14***
	Eucl_Pear_MnMx	***1.97***	***2.87***	**3.88**	**5.19**	**5.98**	***7.14***
	Pear_Pear_Zsc	***1.97***	***2.87***	**3.88**	**5.19**	**5.98**	***7.14***
	Eucl_MnMx_Zsc	**1.4**	**2.25**	**3.88**	**5.19**	**5.98**	**7.02**
	Pear_MnMx_Zsc	2.1	3.07	***4.03***	**5.19**	***6.27***	***7.14***
S_5	Eucl_MnMx	3.83	5.1	7	9.15	10.97	***12.61***
	Pear_MnMx	2.64	3.18	4.77	7.35	11.13	15.76
	Eucl_Zsc	2.53	***3.1***	***4.7***	7.35	11.13	15.76
	Pear_Zsc	2.64	3.18	4.77	7.35	11.13	15.76
	Eucl_Pear_MnMx	***2.11***	3.18	4.77	***6.45***	***9.04***	**12.2**
	Pear_Pear_Zsc	2.53	***3.1***	***4.7***	7.35	11.13	15.76
	Eucl_MnMx_Zsc	**1.92**	**3.05**	**4.3**	**6.05**	**8.94**	12.71
	Pear_MnMx_Zsc	2.64	3.18	4.77	7.35	11.13	15.76

Table 6.18: Summary of model performance on C-MSKF by *maximum* sMAPE using different clustering methods based on time series data. Results are obtained by taking the *maximum* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	122.8	125.7	125.29	124.72	127.27	130.32
	Pear_MnMx	**43.24**	***36.69***	***32.03***	29.68	29.57	***29.22***
	Eucl_Zsc	**43.24**	***36.69***	***32.03***	29.96	29.58	30.09
	Pear_Zsc	**43.24**	***36.69***	***32.03***	29.68	29.57	***29.22***
	Eucl_Pear_MnMx	***43.46***	**35.9**	**30.97**	**28.6**	**27.68**	**27.98**
	Pear_Pear_Zsc	**43.24**	***36.69***	***32.03***	29.68	29.57	***29.22***
	Eucl_MnMx_Zsc	***43.46***	**35.9**	**30.97**	***29***	***29.44***	29.9
	Pear_MnMx_Zsc	**43.24**	***36.69***	***32.03***	29.68	29.57	***29.22***
S_2	Eucl_MnMx	145.09	147.67	152.03	154.63	157.51	159.88
	Pear_MnMx	**67.76**	***63.24***	***62.88***	***63.82***	***63.69***	**63.19**
	Eucl_Zsc	***67.95***	68.04	67.69	66.75	65.88	64.79
	Pear_Zsc	**67.76**	***63.24***	***62.88***	***63.82***	***63.69***	**63.19**
	Eucl_Pear_MnMx	**67.76**	***63.24***	***62.88***	***63.82***	***63.69***	**63.19**
	Pear_Pear_Zsc	***67.95***	68.04	67.69	66.75	65.88	64.79
	Eucl_MnMx_Zsc	**67.76**	**59.48**	**60.79**	**61.12**	**62.99**	***64.69***
	Pear_MnMx_Zsc	**67.76**	***63.24***	***62.88***	***63.82***	***63.69***	**63.19**
S_3	Eucl_MnMx	138.21	141.58	145.52	147.74	149.92	153.65
	Pear_MnMx	95.49	89.46	**85.99**	**88.2**	**89.14**	93.75
	Eucl_Zsc	95.49	***88.78***	***89.7***	***90.15***	***89.73***	**88.67**
	Pear_Zsc	95.49	89.46	**85.99**	**88.2**	**89.14**	93.75
	Eucl_Pear_MnMx	***95.27***	89.46	90.14	94.92	99.06	102.1
	Pear_Pear_Zsc	95.49	***88.78***	***89.7***	***90.15***	***89.73***	**88.67**
	Eucl_MnMx_Zsc	**83.59**	**85.15**	***89.7***	***90.15***	***89.73***	***88.8***
	Pear_MnMx_Zsc	95.49	89.46	**85.99**	**88.2**	**89.14**	93.75
S_4	Eucl_MnMx	167.89	166.79	166.99	165.83	167.6	168.82
	Pear_MnMx	101.25	***97.25***	**94.25**	***98.88***	102.57	106.96
	Eucl_Zsc	108.05	101.2	***94.43***	**95.44**	**98.77**	**102.96**
	Pear_Zsc	101.25	***97.25***	**94.25**	***98.88***	102.57	106.96
	Eucl_Pear_MnMx	98.6	106.64	109.48	108.5	110.82	113.44
	Pear_Pear_Zsc	**96.57**	**94.91**	**94.25**	***98.88***	102.57	106.96
	Eucl_MnMx_Zsc	***97.67***	102.04	100.28	100.13	***101.06***	***105.83***
	Pear_MnMx_Zsc	101.25	***97.25***	**94.25**	***98.88***	102.57	106.96
S_5	Eucl_MnMx	167.19	173.34	171.09	168.96	170.45	171.8
	Pear_MnMx	**119.09**	**118.14**	**116.66**	**116.57**	**117.6**	**122.79**
	Eucl_Zsc	**119.09**	**118.14**	**116.66**	119.16	125.86	***131.21***
	Pear_Zsc	**119.09**	**118.14**	**116.66**	**116.57**	**117.6**	**122.79**
	Eucl_Pear_MnMx	***128.11***	***131.76***	***131.72***	133.08	131.4	133.05
	Pear_Pear_Zsc	**119.09**	**118.14**	**116.66**	119.16	125.86	***131.21***
	Eucl_MnMx_Zsc	**119.09**	**118.14**	**116.66**	***117.74***	***125.84***	131.38
	Pear_MnMx_Zsc	**119.09**	**118.14**	**116.66**	**116.57**	**117.6**	**122.79**

Table 6.19: Summary of model performance on C-MSKF by *median* MASE using different clustering methods based on time series data. Results are obtained by taking the *median* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	1.01	1.16	1.28	1.4	1.52	1.64
	Pear_MnMx	0.74	0.84	0.94	1.03	1.13	1.22
	Eucl_Zsc	***0.73***	0.84	0.93	***1.02***	1.12	***1.21***
	Pear_Zsc	0.74	0.84	0.94	1.03	1.13	1.22
	Eucl_Pear_MnMx	**0.7**	**0.8**	**0.89**	**0.97**	**1.06**	**1.15**
	Pear_Pear_Zsc	***0.73***	***0.83***	***0.92***	***1.02***	***1.11***	***1.21***
	Eucl_MnMx_Zsc	**0.7**	**0.8**	**0.89**	**0.97**	**1.06**	**1.15**
	Pear_MnMx_Zsc	0.74	0.84	0.94	1.03	1.13	1.22
S_2	Eucl_MnMx	2.67	3.03	3.31	3.67	3.99	4.29
	Pear_MnMx	1.14	1.28	1.43	1.62	1.8	1.98
	Eucl_Zsc	1.13	1.28	1.43	1.62	1.8	1.98
	Pear_Zsc	1.14	1.28	1.43	1.62	1.8	1.98
	Eucl_Pear_MnMx	***1.1***	***1.25***	***1.42***	***1.59***	***1.77***	***1.94***
	Pear_Pear_Zsc	1.12	1.26	***1.42***	1.6	1.79	1.96
	Eucl_MnMx_Zsc	**1.07**	**1.21**	**1.39**	**1.57**	**1.73**	**1.9**
	Pear_MnMx_Zsc	1.14	1.28	1.43	1.62	1.8	1.98
S_3	Eucl_MnMx	2.56	2.94	3.31	3.67	3.97	4.28
	Pear_MnMx	1.42	1.66	1.86	2.11	2.34	2.56
	Eucl_Zsc	1.41	1.63	1.83	2.05	2.27	2.51
	Pear_Zsc	1.42	1.66	1.86	2.11	2.34	2.56
	Eucl_Pear_MnMx	***1.37***	***1.58***	***1.8***	***2.04***	***2.26***	***2.48***
	Pear_Pear_Zsc	1.39	1.59	1.82	***2.04***	***2.26***	2.49
	Eucl_MnMx_Zsc	**1.34**	**1.54**	**1.76**	**2**	**2.21**	**2.42**
	Pear_MnMx_Zsc	1.42	1.66	1.86	2.11	2.34	2.56
S_4	Eucl_MnMx	2.54	2.84	3.16	3.49	3.82	4.12
	Pear_MnMx	1.49	1.71	1.95	2.2	2.41	2.64
	Eucl_Zsc	1.47	1.69	1.95	2.16	2.37	2.59
	Pear_Zsc	1.49	1.71	1.95	2.2	2.41	2.64
	Eucl_Pear_MnMx	1.47	***1.68***	1.95	2.2	2.38	2.64
	Pear_Pear_Zsc	***1.46***	***1.68***	***1.92***	***2.13***	***2.35***	***2.58***
	Eucl_MnMx_Zsc	**1.42**	**1.63**	**1.87**	**2.1**	**2.33**	**2.54**
	Pear_MnMx_Zsc	1.49	1.71	1.95	2.2	2.41	2.64
S_5	Eucl_MnMx	2.48	2.79	3.07	3.37	3.67	3.98
	Pear_MnMx	1.64	1.95	2.21	2.42	2.67	2.91
	Eucl_Zsc	***1.57***	***1.86***	***2.11***	***2.37***	***2.6***	***2.82***
	Pear_Zsc	1.64	1.95	2.21	2.42	2.67	2.91
	Eucl_Pear_MnMx	1.59	1.88	2.12	2.38	2.64	2.88
	Pear_Pear_Zsc	1.58	1.87	***2.11***	***2.37***	2.61	2.86
	Eucl_MnMx_Zsc	**1.53**	**1.81**	**2.07**	**2.31**	**2.56**	**2.81**
	Pear_MnMx_Zsc	1.64	1.95	2.21	2.42	2.67	2.91

Table 6.20: Summary of model performance on C-MSKF by *median* ME using different clustering methods based on time series data. Results are obtained by taking the *median* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
S_1	Eucl_MnMx	**0**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Pear_MnMx	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Eucl_Zsc	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Pear_Zsc	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Eucl_Pear_MnMx	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Pear_Pear_Zsc	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Eucl_MnMx_Zsc	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
	Pear_MnMx_Zsc	**0**	***0***	**-0.01**	**-0.01**	**-0.01**	**-0.01**
S_2	Eucl_MnMx	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Pear_MnMx	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Eucl_Zsc	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Pear_Zsc	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Eucl_Pear_MnMx	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Pear_Pear_Zsc	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Eucl_MnMx_Zsc	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
	Pear_MnMx_Zsc	**0.01**	**0.01**	**0.03**	**0.03**	**0.04**	**0.04**
S_3	Eucl_MnMx	**-0.02**	***-0.01***	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Pear_MnMx	**-0.02**	**-0.02**	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Eucl_Zsc	**-0.02**	***-0.01***	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Pear_Zsc	**-0.02**	**-0.02**	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Eucl_Pear_MnMx	**-0.02**	**-0.02**	**-0.03**	**-0.04**	**-0.05**	***-0.05***
	Pear_Pear_Zsc	**-0.02**	***-0.01***	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Eucl_MnMx_Zsc	**-0.02**	***-0.01***	**-0.03**	**-0.04**	**-0.05**	**-0.06**
	Pear_MnMx_Zsc	**-0.02**	**-0.02**	**-0.03**	**-0.04**	**-0.05**	**-0.06**
S_4	Eucl_MnMx	**0.02**	**0.05**	**0.08**	**0.09**	**0.11**	***0.12***
	Pear_MnMx	***0.03***	***0.06***	***0.09***	***0.1***	**0.11**	***0.12***
	Eucl_Zsc	***0.03***	***0.06***	**0.08**	**0.09**	**0.11**	***0.12***
	Pear_Zsc	***0.03***	***0.06***	***0.09***	***0.1***	**0.11**	***0.12***
	Eucl_Pear_MnMx	***0.03***	***0.06***	**0.08**	***0.1***	**0.11**	***0.12***
	Pear_Pear_Zsc	***0.03***	***0.06***	***0.09***	***0.1***	**0.11**	**0.11**
	Eucl_MnMx_Zsc	***0.03***	***0.06***	**0.08**	***0.1***	***0.12***	0.13
	Pear_MnMx_Zsc	***0.03***	***0.06***	***0.09***	***0.1***	**0.11**	***0.12***
S_5	Eucl_MnMx	**-0.01**	**0.03**	**0.03**	**0.05**	**0.06**	**0.06**
	Pear_MnMx	***0***	**0.03**	**0.03**	***0.06***	**0.06**	**0.06**
	Eucl_Zsc	***0***	**0.03**	***0.04***	***0.06***	**0.06**	**0.06**
	Pear_Zsc	***0***	**0.03**	**0.03**	***0.06***	**0.06**	**0.06**
	Eucl_Pear_MnMx	**-0.01**	**0.03**	**0.03**	***0.06***	***0.07***	**0.06**
	Pear_Pear_Zsc	***0***	**0.03**	**0.03**	***0.06***	**0.06**	**0.06**
	Eucl_MnMx_Zsc	***0***	**0.03**	***0.04***	***0.06***	**0.06**	**0.06**
	Pear_MnMx_Zsc	***0***	**0.03**	**0.03**	***0.06***	**0.06**	**0.06**

Table 6.21: Summary of model performance on C-MSKF by *median* MSE using different clustering methods based on time series data. Results are obtained by taking the *median* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	0.26	0.33	0.41	0.5	0.59	0.7
	Pear_MnMx	***0.09***	**0.11**	***0.14***	***0.17***	0.2	0.25
	Eucl_Zsc	***0.09***	**0.11**	***0.14***	***0.17***	0.2	0.25
	Pear_Zsc	***0.09***	**0.11**	***0.14***	***0.17***	0.2	0.25
	Eucl_Pear_MnMx	**0.08**	**0.11**	**0.13**	**0.16**	**0.18**	**0.22**
	Pear_Pear_Zsc	**0.08**	**0.11**	***0.14***	***0.17***	0.2	0.24
	Eucl_MnMx_Zsc	**0.08**	**0.11**	**0.13**	**0.16**	***0.19***	***0.23***
	Pear_MnMx_Zsc	***0.09***	**0.11**	***0.14***	***0.17***	0.2	0.25
S_2	Eucl_MnMx	1.17	1.54	1.94	2.38	2.83	3.23
	Pear_MnMx	0.26	0.34	0.46	0.61	0.8	0.97
	Eucl_Zsc	***0.24***	0.33	0.43	***0.58***	***0.74***	0.92
	Pear_Zsc	0.26	0.34	0.46	0.61	0.8	0.97
	Eucl_Pear_MnMx	***0.24***	***0.32***	0.43	0.59	0.75	0.92
	Pear_Pear_Zsc	***0.24***	***0.32***	***0.42***	***0.58***	***0.74***	***0.91***
	Eucl_MnMx_Zsc	**0.22**	**0.29**	**0.38**	**0.52**	**0.65**	**0.82**
	Pear_MnMx_Zsc	0.26	0.34	0.46	0.61	0.8	0.97
S_3	Eucl_MnMx	1.41	1.93	2.48	3	3.63	4.3
	Pear_MnMx	0.48	0.65	0.9	1.16	1.46	1.79
	Eucl_Zsc	0.44	0.63	0.82	1.11	1.4	1.71
	Pear_Zsc	0.48	0.65	0.9	1.16	1.46	1.79
	Eucl_Pear_MnMx	***0.41***	***0.6***	0.8	***1.05***	***1.34***	***1.64***
	Pear_Pear_Zsc	0.42	***0.6***	***0.79***	1.07	1.36	1.66
	Eucl_MnMx_Zsc	**0.4**	**0.56**	**0.75**	**0.97**	**1.22**	**1.52**
	Pear_MnMx_Zsc	0.48	0.65	0.9	1.16	1.46	1.79
S_4	Eucl_MnMx	1.56	1.96	2.51	3.08	3.76	4.49
	Pear_MnMx	0.68	0.98	1.35	1.76	2.23	2.75
	Eucl_Zsc	0.66	0.93	1.31	***1.74***	***2.18***	***2.66***
	Pear_Zsc	0.68	0.98	1.35	1.76	2.23	2.75
	Eucl_Pear_MnMx	***0.65***	0.95	1.3	1.75	2.21	2.79
	Pear_Pear_Zsc	***0.65***	***0.91***	***1.27***	***1.74***	2.22	2.73
	Eucl_MnMx_Zsc	**0.59**	**0.83**	**1.16**	**1.63**	**2.15**	**2.63**
	Pear_MnMx_Zsc	0.68	0.98	1.35	1.76	2.23	2.75
S_5	Eucl_MnMx	1.77	2.27	2.74	3.41	4.14	4.93
	Pear_MnMx	0.83	1.24	1.7	2.19	2.83	3.29
	Eucl_Zsc	0.78	***1.16***	**1.56**	**2.03**	***2.58***	***3.18***
	Pear_Zsc	0.83	1.24	1.7	2.19	2.83	3.29
	Eucl_Pear_MnMx	0.82	1.23	1.69	2.18	2.75	3.27
	Pear_Pear_Zsc	***0.76***	1.17	1.61	2.16	2.63	3.23
	Eucl_MnMx_Zsc	**0.73**	**1.13**	***1.57***	***2.04***	**2.57**	**3.16**
	Pear_MnMx_Zsc	0.83	1.24	1.7	2.19	2.83	3.29

Table 6.22: Summary of model performance on C-MSKF by *median* sMAPE using different clustering methods based on time series data. Results are obtained by taking the *median* value across 30 replicates, 6 time series lengths and 6 forecasting horizons. Precisely, the best weight is determined based on $t = 17$ and $t \leq T$ are used for clustering stage. The best performing method is highlighted in bold face and the second best method is highlighted in italic bold face.

Scenarios	Combinations	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
S_1	Eucl_MnMx	35.74	33.77	32.87	31.5	30.98	30.38
	Pear_MnMx	23.39	22.72	21.83	21.66	21.35	21.13
	Eucl_Zsc	23.62	22.83	21.79	21.57	20.99	20.92
	Pear_Zsc	23.39	22.72	21.83	21.66	21.35	21.13
	Eucl_Pear_MnMx	***22.16***	***21.03***	**20.68**	**20.44**	***20.26***	***20.02***
	Pear_Pear_Zsc	22.88	22.6	21.71	***21.52***	21	21
	Eucl_MnMx_Zsc	**21.87**	**20.98**	***20.7***	**20.44**	**20.23**	**19.99**
	Pear_MnMx_Zsc	23.39	22.72	21.83	21.66	21.35	21.13
S_2	Eucl_MnMx	103.78	104.94	104.88	105.99	106.04	104.83
	Pear_MnMx	38.14	37.09	37.82	38.01	38.45	38.69
	Eucl_Zsc	37.84	37.6	38.02	38.49	38.65	39.18
	Pear_Zsc	38.14	37.09	37.82	38.01	38.45	38.69
	Eucl_Pear_MnMx	37.45	***36.36***	***36.52***	***37***	***37.29***	***37.55***
	Pear_Pear_Zsc	***37.25***	36.95	37.14	37.88	38.16	38.14
	Eucl_MnMx_Zsc	**36.15**	**35.73**	**35.82**	**35.76**	**36.18**	**36.73**
	Pear_MnMx_Zsc	38.14	37.09	37.82	38.01	38.45	38.69
S_3	Eucl_MnMx	102.66	106.27	110.77	113.99	115.89	117.3
	Pear_MnMx	56.32	57.2	59.14	58.45	57.9	59.28
	Eucl_Zsc	56.17	57.38	59.36	59.34	59.63	59.98
	Pear_Zsc	56.32	57.2	59.14	58.45	57.9	59.28
	Eucl_Pear_MnMx	**54.25**	***55.82***	***57.83***	***57.19***	57.42	57.61
	Pear_Pear_Zsc	55.3	56.85	58.19	58.21	***57.38***	***57.46***
	Eucl_MnMx_Zsc	***54.28***	**54.99**	**56.15**	**55.62**	**55.59**	**55.65**
	Pear_MnMx_Zsc	56.32	57.2	59.14	58.45	57.9	59.28
S_4	Eucl_MnMx	103.37	107.37	110.14	112.66	114.48	116.35
	Pear_MnMx	58.52	60.4	62.31	63.54	64.5	65.07
	Eucl_Zsc	59.06	61	62.49	63.8	64.81	64.92
	Pear_Zsc	58.52	60.4	62.31	63.54	64.5	65.07
	Eucl_Pear_MnMx	59.04	60.87	62.58	63.79	64.27	64.86
	Pear_Pear_Zsc	***58.27***	***60.31***	***61.82***	***63.33***	***64***	***64.36***
	Eucl_MnMx_Zsc	**56.57**	**60.03**	**61.62**	**62.79**	**63.39**	**64.14**
	Pear_MnMx_Zsc	58.52	60.4	62.31	63.54	64.5	65.07
S_5	Eucl_MnMx	111.33	113.98	116.42	119.85	121.13	123.17
	Pear_MnMx	66.37	69.97	72.3	71.94	73.03	73.87
	Eucl_Zsc	66.48	70.17	72.08	71.74	***71.81***	***72.93***
	Pear_Zsc	66.37	69.97	72.3	71.94	73.03	73.87
	Eucl_Pear_MnMx	***64.74***	***68.69***	***70.76***	***71.25***	72.04	73.17
	Pear_Pear_Zsc	65.07	69.89	72.08	71.79	72.25	73.17
	Eucl_MnMx_Zsc	**64.01**	**67.17**	**69.62**	**71**	**71.23**	**71.95**
	Pear_MnMx_Zsc	66.37	69.97	72.3	71.94	73.03	73.87

References

[1] J. Alon et al. “Discovering clusters in motion time-series data”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2003, pp. 375–381.

[2] F. M. Alvarez et al. “Energy time series forecasting based on pattern sequence similarity”. In: *IEEE Transactions on Knowledge and Data Engineering* 23.8 (2011), pp. 1230–1243.

[3] C. Bergmeir, R. J. Hyndman, and J. M. Benítez. “Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation”. In: *International Journal of Forecasting* 32.2 (2016), pp. 303–312.

[4] M. J. Brusco and J. D. Cradit. “A variable-selection heuristic for K-means clustering”. In: *Psychometrika* 66.2 (2001), pp. 249–270.

[5] M. Dash and H. Liu. “Feature selection for clustering”. In: *Pacific-Asia Conference on knowledge discovery and data mining*. Springer. 2000, pp. 110–121.

[6] R. Dubes and A. K. Jain. “Clustering methodologies in exploratory data analysis”. In: *Advances in Computers* 19 (1980), pp. 113–228.

[7] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[8] G. Duncan, W. Gorr, and J. Szczypula. *Comparative Study of Cross Sectional Methods for Time Series With Structural Changes*. Tech. rep. Carnegie Mellon University, 1994.

[9] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[10] J. Ernst, G. J. Nau, and Z. Bar-Joseph. “Clustering short time series gene expression data”. In: *Bioinformatics* 21.suppl_1 (2005), pp. i159–i168.

[11] P. Esling and C. Agon. “Time-series data mining”. In: *ACM Computing Surveys (CSUR)* 45.1 (2012), p. 12.

[12] D. Ö. Faruk. “A hybrid neural network and ARIMA model for water quality time series prediction”. In: *Engineering Applications of Artificial Intelligence* 23.4 (2010), pp. 586–594.

[13] P. H. Franses and D. Van Dijk. *Non-linear time series models in empirical finance*. Cambridge University Press, 2000.

[14] S. Frühwirth-Schnatter and S. Kaufmann. “Model-based clustering of multiple time series”. In: *Journal of Business & Economic Statistics* 26.1 (2008), pp. 78–89.

[15] T.-c. Fu. “A review on time series data mining”. In: *Engineering Applications of Artificial Intelligence* 24.1 (2011), pp. 164–181.

[16] R. Gnanadesikan, J. R. Kettenring, and S. L. Tsao. “Weighting and selection of variables for cluster analysis”. In: *Journal of Classification* 12.1 (1995), pp. 113–136.

[17] I. Guyon, U. Von Luxburg, and R. C. Williamson. “Clustering: Science or art”. In: *NIPS 2009 Workshop on Clustering Theory*. 2009, pp. 1–11.

[18] J. Handl and J. Knowles. “An evolutionary approach to multiobjective clustering”. In: *IEEE transactions on Evolutionary Computation* 11.1 (2007), pp. 56–76.

[19] P. J. Harrison and C. F. Stevens. “A Bayesian approach to short-term forecasting”. In: *Operational Research Quarterly* (1971), pp. 341–362.

[20] A. C. Harvey and S. T.S. M. Forecasting. *the Kalman filter*. 1989.

[21] L. Hubert and P. Arabie. “Comparing partitions”. In: *Journal of Classification* 2.1 (1985), pp. 193–218.

[22] R. J. Hyndman. “Another look at forecast-accuracy metrics for intermittent demand”. In: *Foresight: The International Journal of Applied Forecasting* 4.4 (2006), pp. 43–46.

[23] R. E. Kass and D. Steffey. “Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)”. In: *Journal of the American Statistical Association* 84.407 (1989), pp. 717–726.

[24] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[25] E. Keogh and S. Kasetty. “On the need for time series data mining benchmarks: a survey and empirical demonstration”. In: *Data Mining and knowledge discovery* 7.4 (2003), pp. 349–371.

[26] G. Lance and W. Williams. “A generalized sorting strategy for computer classifications”. In: *Nature* 212.5058 (1966), pp. 218–218.

[27] M. Last, A. Kandel, and H. Bunke. *Data mining in time series databases*. Vol. 57. World scientific, 2004.

[28] M. H. Law, A. P. Topchy, and A. K. Jain. “Multiobjective data clustering”. In: *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. Vol. 2. IEEE. 2004, pp. II–II.

[29] T. W. Liao. “Clustering of time series data—a survey”. In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[30] M. Lorr. *Cluster analysis for social scientists*. Jossey-Bass Inc Pub, 1983.

[31] E. Lu and J. Handl. “Multicriterion Segmentation of Demand Markets to Increase Forecasting Accuracy of Analogous Time Series: A First Investigation”. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer. 2015, pp. 379–388.

[32] G. W. Milligan and M. C. Cooper. “A study of standardization of variables in cluster analysis”. In: *Journal of Classification* 5.2 (1988), pp. 181–204.

[33] G. W. Milligan and M. C. Cooper. “An examination of procedures for determining the number of clusters in a data set”. In: *Psychometrika* 50.2 (1985), pp. 159–179.

[34] S. Rani and G. Sikka. “Recent techniques of clustering of time series data: a survey”. In: *International Journal of Computer Applications* 52.15 (2012).

[35] D. Steinley. “Standardizing variables in K-means clustering”. In: *Classification, Clustering, and Data Mining Applications*. Springer, 2004, pp. 53–60.

[36] J. A. Stimson. “Regression in space and time: A statistical essay”. In: *American Journal of Political Science* (1985), pp. 914–947.

[37] A. M. Stoddard. “Standardization of measures prior to cluster analysis”. In: *Biometrics* (1979), pp. 765–773.

[38] G. P. Zhang. “Time series forecasting using a hybrid ARIMA and neural network model”. In: *Neurocomputing* 50 (2003), pp. 159–175.

Chapter 7

Conclusions and outlook

7.1 Journey

In this thesis, a number of research topics have been explored. The first research topic that was interrogated was the seat inventory control problem. Specifically, we were interested in modelling seat allocations in order to maximize the total revenue of train operators subject to certain conditions, *e.g.,* and the passenger demand on origin-to-destination journeys. Seat inventory control problems are interesting but pose a challenge regarding data collection as well as an intensive use of domain knowledge. Without research collaboration, open data was not sufficient for further analysis due to the nature of the problem, which requires the information associated with passenger demand on particular origin-to-destination journeys. Unfortunately, this piece of information is typically confidential. In addition to this, seat inventory control problems have been intensively studied over the last decades. It also presents larger challenges in defining the research gap. In light of this, we focused more on the aspects of data accessibility / methodology innovation.

In considering common challenges in the transport sector, time series forecasting (such as demand forecasting) was considered a competitive alternative to the seat inventory optimization problems. Additionally, taking into account my academic background in computer science, our work was further inspired by the idea of ensemble approaches. These approaches are commonly used in the machine learning field to boost the accuracy of an aggregated classifier / regressor via the aggregation of multiple models. Thus, we had started our exploration related to the idea of drawing power from multiple forecasts. Primarily, two approaches fall into this category, including combining forecasts and forecasting methods make use of analogies *e.g.,* the Bayesian Pooling method. Combining forecasts have received a wide range of applications in the forecasting field. This kind of method is regarded as post-hoc method, similar to ensemble methods in the machine learning area. These approaches have been reported to yield better forecasting results via the aggregation of multiple forecasting solutions. In essence, these methods do not impact on the mechanism of a forecasting algorithm. The main benefit of combining forecasts approaches is to reduce the risk of picking a wrong statistical model. Another group of methods that might be more promising and that have gained little attention are the Bayesian Pooling methods. Intrinsically, these forecasting methods update the estimated parameters over time by combining the estimates from a univariate time series and parameters estimated derived from pooled data. We expected that the forecasting accuracy of a statistical model might be significantly improved by correcting its parameters throughout the learning process, where analogies provide useful information additional to the past observations of a univariate time series. Given the potential of these methods, surprisingly limited work has been conducted in the forecasting field to explore suitable modeling approaches for the identification of analogies. As discussed in the paper (Stimson, 1985),the identification of suitable analogies is crucial for forecasting methods such as these that make use of analogies. This is despite the development of improved techniques being critical for the discernment of similarities between time series (Lee et al., 2007) and supporting the principled selection of analogies (Armstrong, 2001). Furthermore, following the paper *(Duncan, Gorr, and Szczypula, 1993)*, open data available in the public budgeting area can be applied to evaluate our ideas of analogy identification. As a result, we finally decided on the topic related to the improvement of forecasting accuracy using analogous time series.

7.2 Reflection

Although conventional forecasting methods typically make forecasts for a single series in isolation, almost all companies require methods that can simultaneously forecast a set of analogous time series. *e.g.,* the analysis of the sales of similar products fall into the same geographic area (Duncan, Gorr, and Szczypula, 2001). The use of analogies can be particularly useful when investigating problems for which there is little prior data available regarding the target series. This is often the case shortly after the launch of new products or where records are missing. The use of analogies can help to create opportunities for borrowing strengths from homogeneous time series to derive more reliable forecasts of the target series. Due to the significant role that analogies can play in forecasting contexts, therefore we have developed a methodological framework that enables the principled selection of analogies using multicriteria segmentation approaches.

For either judgmental forecasting or statistical forecasts, we aim to guide the selection of analogies with a consideration of multiple criteria. Our work develop data-driven approaches for the analogy identification using multiple criteria. Specifically, we explored the potential of using multiple information sources, distance metrics / standardization techniques as criteria for the segmentation stage. We demonstrated that by integrating individual criterion that carries uncorrelated noise is able to produce better results in the clustering of analogies, thus leads to an increased forecasting accuracy.

Moreover, we proposed multiple solutions for tackling the model selection problem in the multicriteria clustering context. Model selection in conventional clustering problems is an open question that has received widespread applications in the literature. Typically, statistical techniques such as the Silhouette Width, Gap statistics have been commonly employed in practice. However, their effectiveness was well studied in single-criterion clustering problems, but very limited applications can be found in scenarios where multiple Pareto-optimal clustering solutions may exist. As discussed in **Chapter 4**, a single solution is often required for further analysis in a forecasting context. It proposes a challenge to develop a technique that can determine the single best clustering solution in an objective manner. To our understanding, model selection plays an necessary role in the forecasting applications. Our work systematically analyzed promising model selection methods with or without consideration of an application context. It opens up opportunities for developing automated forecasting process using analogies that are identified by the multicriteria segmentation approach.

7.3 Limitations

In our current work, the major limitation of the work lies in the diversity of data. We used simulated data to conduct experiments in order to understand the sensitivity of different elements related to the forecasting process that makes use of analogies. We additionally used real data to evaluate the validity of these methods. In **Chapter 3**, we applied our methodological framework to both simulated and real data sets (US. personal income tax liability). However, manuscripts 2, 3 and 4 focused on using simulated data sets that were generated via the same set of equations as described in **Chapter 3**. Consequently, the limited diversity of the data in these manuscripts might reduce the reliability or weaken the generalizability of our proposed methods.

In terms of experimental settings, we varied the values of the following factors such as the noise level of time series data, time series lengths, forecasting horizons. The main concern is to evaluate the sensitivity of different forecasting methods by varying the values of different elements that might impact the accuracy of forecasting methods. Different choices of forecasting origin may also cause the distinctions in forecasting accuracy among methods. Thus, our findings show limitations in interpreting the impact of forecasting origin on different segmentation approaches (CF, TS, MC) and thus producing the forecasting results.

Here, the weighted-sum method can also been extended to accommodate more than two criteria at a distance function level. Nevertheless, the difficulty underlying the process might stem from the model selection step. For example, angle-based methods have been applied to identify the best "knee" point in a two-feature space (see **Chapter ??**). These methods might not be well extended to cater for more than three criteria. Further steps will be involved by comparing the solutions across every 2-dimensional spaces.

7.4 Generalizations

Overall, the methodological framework proposed in **Chapter 3** is expected to generalize to real-world settings. The framework here is also expected to accommodate other options of forecasting algorithms, which can exploit information from analogies, in the forecasting stage. As the C-MSKF method is used for illustration purpose, we consider the forecasting algorithm in the forecasting stage can be replaced by statistical forecasting methods that draw information from analogies. For instance, the Cross-sectional Exponential Smoothing method (Duncan, Gorr, and Szczypula, 1994) can also be an alternative.

Among the proposed model selection approaches, the MC$_{SilHist}$ method proposed in **Chapter ??** and **3** showed the most promising results in both simulated and real data. We would expect our concepts of model selection can be generalized to the multicriteria clustering problems, where multiple criteria are present, *e.g.,* data information sources, distance metrics / standardization techniques.

The last main focus of our work lies in the employment of bagging techniques. Our findings show that bagging methods show strengths in producing more reliable forecasting results via the aggregation of multiple forecasts. Our proposed bagging strategies are expected to generalize to forecasting framework that exploit information from analogies.

7.5 Implication

As discussed in **Chapter 3**, by varying the range of cluster numbers, the multicriteria clustering approach shows consistently superior forecasting results to single-criterion clustering approaches. This suggests that multicriteria clustering approach continues to benefit from the use of complementary information sources, even in a scenario where the correct number of clusters is overestimated.

By investigating the relationship before, we provided new insight into the relationship between the accuracy of the segmentation stage and the performance of a forecasting algorithm that makes use of analogies. Throughout the experiments, our results showed that the improved clustering quality of analogies demonstrated a positive impact on the forecasting accuracy performance. This suggests the forecasting framework might benefit from improved quality of analogies.

Further, we conducted a systematic study that compares the performance of single-criterion and multicriteria segmentation approaches related to forecasting. Our findings imply that by integrating multiple criteria, with uncorrelated noise associated with individual criterion, the multicriteria segmentation approach shows superior capability in boosting the homogeneity of analogies and lead further boost in the forecasting accuracy.

To address model selection problem in the context of multicriterion segmentation problems, our findings confirmed that the conclusion of **Chapter 3** that model selection (clustering quality) is best evaluated in a problem-specific context. We believe the insights here could be meaningful for later studies that focuses on addressing model selection problem in clustering applications.

7.6 Further research

As discussed earlier, various sets of simulated data should be used to test the property of the methods proposed in the thesis. Currently, the same set of equations has been applied to generate simulated data across manuscripts 1,2, 3 and 4, although the data produced can be slightly different. To increase the diversity of the data, we consider expanding our methodologies in different application contexts not limited to the personal income tax liability data such as crime data. This is because crime rate might also be associated with the fluctuation in macro economy. This meets the basic assumptions underlying the C-MSKF algorithms.

At the methodological level, all multicriteria clustering approaches proposed in this thesis are limited to the combination of two criteria. Our future work could be extended to account for more criteria where desirable, but this might raise issues concerning time-complexity. This is because, as the number of criteria increases, the number of possible trade-off clustering solutions may grow exponentially due to the larger number of combinations between weights. Given this, further work should be done to tackle the issue of time-complexity.

To understand the strengths and weakness of different forecasting methods, our future work should take into account the influence of forecasting origin that might impact on the performance of statistical forecasting methods. The determination of forecasting origin would probably influence the number of historical observations after the structural change. Due to the differences in responsiveness of a forecasting method, the choice of forecasting origin determines the latest observations that after the structural change. As the recent history increases, C-MSKF methods might lose advantages in drawing power from analogies as historical observations might well represent the history of the model.

As suggested in **Chapter 3**, the use of two information sources is superfluous in the absence of noise in the individual information sources, and can only be beneficial in the presence of uncorrelated noise. Further work might take into account the correlated noise in the experiments in order to understand the impact of correlation on the following forecasting stage.

References

[1] J. S. Armstrong. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Vol. 30. Springer Science & Business Media, 2001.

[2] G. Duncan, W. Gorr, and J. Szczypula. “Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting”. In: *Management Science* 39.3 (1993), pp. 275–293.

[3] G. Duncan, W. Gorr, and J. Szczypula. *Comparative Study of Cross Sectional Methods for Time Series With Structural Changes*. Tech. rep. Carnegie Mellon University, 1994.

[4] G. T. Duncan, W. L. Gorr, and J. Szczypula. “Forecasting analogous time series”. In: *Principles of forecasting*. Springer, 2001, pp. 195–213.

[5] W. Y. Lee et al. “Providing support for the use of analogies in demand forecasting tasks”. In: *International Journal of Forecasting* 23.3 (2007), pp. 377–390.

[6] J. A. Stimson. “Regression in space and time: A statistical essay”. In: *American Journal of Political Science* (1985), pp. 914–947.

Appendix A

Implemented 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.

$$F_{t+h} = X_t \tag{A.1}$$

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.

$$F_{t+h} = X_t + \frac{h}{t-1}(X_t - X_1) \tag{A.2}$$

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.

$$F_{t+1} = \alpha X_t + (1-\alpha)F_t \tag{A.3}$$

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:

$$\begin{aligned} L_t &= \alpha X_t + (1-\alpha)(L_{t-1} + T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \\ F_{t+h} &= L_t + hT_t \end{aligned} \tag{A.4}$$

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, Everette, and McKenzie, 1985). Forecasts for Damped method can be calculated as:

$$\begin{aligned} L_t &= \alpha X_t + (1-\alpha)(L_{t-1} + \phi T_{t-1}) \\ T_t &= \beta(L_t - L_{t-1}) + (1-\beta)\phi T_{t-1} \\ F_{t+h} &= L_t + \sum_{i=1}^{h} \phi^h T_t s \end{aligned} \tag{A.5}$$

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:

$$\begin{aligned} X_t &= T_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, V_\varepsilon) \\ T_t &= T_{t-1} + S_t + \gamma_t, \quad \gamma_t \sim N(0, V_\gamma) \\ S_t &= S_{t-1} + \rho_t, \quad \rho_t \sim N(0, V_\rho) \end{aligned} \tag{A.6}$$

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:

$$\begin{pmatrix} T_t \\ S_t \end{pmatrix} \sim N\left[\begin{pmatrix} m_t \\ b_t \end{pmatrix}, C_t = \begin{pmatrix} V_{11,t} & V_{12,t} \\ V_{12,t} & V_{22,t} \end{pmatrix}\right] \tag{A.7}$$

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 ($j \in \{\text{no change}, \text{step change}, \text{slope change}, \text{tr}$
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 West and Harrison

(1998) 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:

$$m_t^{(j,k)} = m_{t-1}^{(j)} + b_{t-1}^{(j)} + A_{1,t}^{(j,k)} e_t^{(j)}$$

$$b_t^{(j,k)} = b_{t-1}^{(j)} + A_{2,t}^{(j,k)} e_t^{(j)}$$

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

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

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

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

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

$$e_t^{(j)} = X_t - (m_{t-1}^{(j)} + b_{t-1}^{(j)})$$

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

$$A_{2,t}^{(j,k)} = r_{12,t}^{(j,k)} / V_e^{(k)} t$$

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

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

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

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

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 $q_t^{(k)} = \sum\limits_j p_t^{(j,k)}$ 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, Gorr, and Szczypula, 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:

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

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

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

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.

$$\begin{pmatrix} T_{it} \\ S_{it} \end{pmatrix} \sim N \left[\begin{pmatrix} m_{it} \\ b_{it} \end{pmatrix}, C_t = \begin{pmatrix} V_{11,it}^{(j)} & V_{12,it}^{(j)} \\ V_{12,it}^{(j)} & V_{22,it}^{(j)} \end{pmatrix} \right] \quad \text{(A.8)}$$

$m_{it,}^{(j)}, b_{it}^{(j)}$ 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 West and Harrison (1998) 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.

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

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

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

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

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

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.