# Essays in Hierarchical Time Series Forecasting and Forecast Combination

Christoph Weiss

Judge Business School
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

St. John's College

March 2018

Essays in Hierarchical Time Series Forecasting and Forecast Combination

Christoph Weiss

This dissertation comprises of three original contributions to empirical forecasting research. Chapter 1 introduces the dissertation.

Chapter 2 contributes to the literature on hierarchical time series (HTS) modelling by proposing a disaggregated forecasting system for both inflation rate *and* its volatility. Using monthly data that underlies the Retail Prices Index for the UK, we analyse the dynamics of the inflation process. We examine patterns in the time-varying covariation among product-level inflation rates that aggregate up to industry-level inflation rates that in turn aggregate up to the overall inflation rate. The aggregate inflation volatility closely tracks the time path of this covariation, which is seen to be driven primarily by the variances of common shocks shared by all products, and by the covariances between idiosyncratic product-level shocks. We formulate a forecasting system that comprises of models for mean inflation rate *and* its variance, and exploit the index structure of the aggregate inflation rate using the HTS framework. Using a dynamic model selection approach to forecasting, we obtain forecasts that are between 9 and 155 % more accurate than a SARIMA-GARCH(1,1) for the aggregate inflation volatility.

Chapter 3 is on improving forecasts using forecast combinations. The paper documents the software implementation of the open source R package for forecast combination that we coded and published on the official R package depository, CRAN. The GeomComb package is the only R package that covers a wide range of different popular forecast combination methods. We implement techniques from 3 broad categories: (a) simple non-parametric methods, (b) regression-based methods, and (c) geometric (eigenvector) methods, allowing for static or dynamic estimation of each approach. Using S3 classes/methods in R, the package provides a user-friendly environment for applied forecasting, implementing solutions for typical issues related to forecast combination (multicollinearity, missing values, etc.), criterion-based optimisation for several parametric methods, and post-fit functions to rationalise and visualise estimation results. The package has been listed in the official R Task Views for Time Series Analysis and for Official Statistics. The brief empirical application in the paper illustrates the package's functionality by estimating forecast combination techniques for monthly UK electricity supply.

Chapter 4 introduces HTS forecasting and forecast combination to a healthcare staffing context. A slowdown of healthcare budget growth in the UK that does not keep pace with growth of demand for hospital services made efficient cost planning increasingly crucial for hospitals, in particular for staff which accounts for more than half of hospitals' expenses. This is facilitated by accurate forecasts of patient census and churn. Using a dataset of more than 3 million observations from a large UK hospital, we show how HTS forecasting can improve forecast accuracy by using information at different levels of the hospital hierarchy (aggregate, emergency/electives, divisions, specialties), compared to the naïve benchmark: the seasonal random walk model applied to the aggregate. We show that forecast combination can improve accuracy even more in some cases, and leads to lower forecast error variance (decreasing forecasting risk). We propose a comprehensive parametric approach to use forecasts in a nurse staffing model that has the aim of minimising cost while satisfying that the care requirements (e.g. nurse hours per patient day thresholds) are met.

For Ruby

Declaration

I declare that this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text below. The contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification at the University of Cambridge, or any other university. Chapter 3 is, partly, joint work with Eran Raviv and Gernot Roetzer, though the majority of the work is my own.

In accordance with the regulations of the Judge Business School at the University of Cambridge, the dissertation contains fewer than 80,000 words of text.

Christoph Weiss
March 2018

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. Paul Kattuman. I am extremely thankful for the opportunity he provided me to pursue a PhD at the Cambridge Judge Business School (CJBS) and for his valuable advice and guidance – both on research and administrative matters – throughout the past 4 years. He has been an incredible source of encouragement, and a tremendous mentor.

I would also like to thank all faculty and other researchers for useful discussions and suggestions which contributed to my research, including but not limited to Vasco Carvalho, Andrew Harvey, Michael Freeman, Gernot Roetzer, Eran Raviv, as well as the participants of the 36th International Symposium on Forecasting, the 2016 International Conference of the Royal Statistical Society, and the Vienna Congress on Mathematical Finance 2016.

I am indebted to the Economic and Social Research Council (ESRC), the Cambridge Commonwealth, European & International Trust, the Qualcomm Fund, and St. John's College for providing financial assistance throughout my PhD.

Most importantly, I want to thank the special people in my life whom I can always count on: I am deeply grateful to my parents, Eveline and Erich, for their continuous support and encouragement; to my sister Olivia, for being the most intellectual person I have ever met and for having read every single word of this thesis; and to Roberta Spinicci, to whom this thesis is dedicated, for being an incredible source of inspiration and motivation to an extent far greater than she tends to give herself credit for, for always believing in me, for being the person I can rely on in my life when I need it the most.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

This dissertation consists of three self-contained chapters that contribute to the field of applied hierarchical forecasting and forecast combination. In the context of forecasting index-type variables, the two approaches are intertwined in an intriguing way: While hierarchical forecasting aims to maximise forecast accuracy by optimal reconciliation of forecasts at different levels of aggregation, forecast combination does so by combining forecasts at the same level of aggregation. An additional feature that conveniently links the two approaches is that hierarchical forecasting tends to result in a considerable number of candidate forecasts by combining different aggregation approaches and forecast approaches, while forecast combination relies on the existence of a sufficient number of candidate models that serve as input for the estimation of combination weights.

This introduction to the dissertation introduces the three substantive chapters, summarising the motivations, research questions, and main contributions. The second chapter, "Hierarchical Modelling and Forecasting System for Inflation Rate and Volatility" relates to the hierarchical time series forecasting literature. The third chapter "Forecast Combination in R Using the GeomComb Package" introduces a software contribution to applied forecast combination. The fourth chapter "Efficient Nurse Staffing: The Value of Hierarchical Time Series Forecasting and Forecast Combination" draws on the value of both approaches, applying them to a healthcare context and offering a systematic approach to hospital staffing.

1.1 Hierarchical Modelling and Forecasting System for Inflation Rate and Volatility

Adverse effects of inflation and inflation volatility on economic growth and welfare have been well documented in economic research (Friedman, 1977; Fischer, 1981; Holland, 1993). Yet, the discussion on the causal relationship between the two measures has caused considerable dichotomy in the research community: The "Friedman-Ball hypothesis" (e.g. Friedman, 1977; Ball and Cecchetti, 1990) documents a positive impact of inflation rate on inflation volatility. The "Cukierman-Melzer hypothesis" (e.g. Cukierman and Meltzer, 1986; Holland, 1995) suggests that causality is reversed, from inflation volatility to inflation rate. This ongoing argument has the effect that the data-generating process of inflation and its volatility is not very well understood.

Shedding some light on the dynamic drivers of inflation volatility in order to understand this important economic measure better is the aim of the first part of this chapter. Borrowing from the approaches in applied disaggregated research on the drivers of shocks to economic activity (e.g. Quah, 1994; Gabaix, 2011), we decompose aggregate inflation volatility into its disaggregated components (product-level variances and covariances), extending Comin and Mulani (2006)'s variance specification to a finer decomposition that splits up product-level price growth rates into common, industry, and product-level shocks. The detailed decomposition model contributes to our understanding of the inflation process and to literature on the potential impact of micro-level shocks on macroeconomic variables by documenting that aggregate volatility is driven by a combination of the variance of common shocks to the entirety of product prices, and product-level (co-)variances.

The second part of the chapter contributes to disaggregated forecasting literature by suggesting a modification of the hierarchical time series (HTS) forecasting framework by Athanasopoulos et al. (2009) and Hyndman et al. (2011) that allows the incorporation of index weights into the model without harming the required property of aggregation consistency. Building on the two-level decomposition from the first part, we formulate a joint hierarchical forecasting system for inflation rate and its volatility that takes note of the variables' hierarchical structure and – evaluating the resulting forecasts against a SARIMA-GARCH model – find overwhelming support for the use of hierarchical methods for inflation forecasting. Using a dynamic model selection model that switches between forecasting models conditional on in-sample estimates of time-varying inflation volatility (an extreme form of ensemble forecast),

the chapter also provides some evidence for the potential of revising hierarchical forecasts via forecast combination.

1.2   Forecast Combination in R Using the Geom-Comb Package

Ever since the seminal paper by Bates and Granger (1969), a large stock of literature has accumulated on different approaches to forecast combination – the integration of several forecasts for a single time series into one combined forecast. The proposed methods differ in their strategies to estimate combination weights.

A large number of empirical applications document the appealing features of combined forecasts that in many cases improve upon even the best component forecast's accuracy and can be used to reduce forecast risk, reducing forecast error variance and leading to higher consistency between in-sample and out-of-sample error distributions (Barrow and Kourentzes, 2016). Due to their data-driven nature, forecast combination has not been as widely employed in econometric applications as one should expect given the methods' appealing properties. However, econometric research has recently turned away from the assumption of the existence of one true data-generating process (Hansen, 2005), which led to a paradigm shift from model selection to model averaging and consequently an increasing interest in the potential of forecast combination techniques for econometric applications – recent empirical econometric research in the field includes Stock and Watson (2004) for output forecasting, Kapetanios et al. (2008) for inflation forecasting, Wright (2008) for exchange rate forecasting, and Rapach et al. (2010) for stock return forecasting.

Despite the methods' continuous and increasing popularity, the lack of comprehensive software implementations is astounding. We close this gap by introducing the R package **GeomComb**, a software contribution that we coded which is entirely focused on forecast combination: The package provides tools for data preparation that deal with two problems that are commonly related to forecast combination input data – missing values and multicollinearity. The package further implements static and dynamic estimation variants of 15 of the most widely employed forecast combination techniques, including statistics-based, regression-based, and eigenvector-based methods, and includes post-fit functions that facilitate the interpretation and visualisation of forecast combination results.

Our software package has made a valuable contribution to applied forecasting and forecasting research: Since its acceptance and release on the official R package depository, CRAN, in November 2016, the package has been added to the official R task views for 'Time Series Analysis' and 'Official Statistics & Survey Methodology', and has been downloaded 2605 times by forecasting practitioners and forecasters (as of 20 August 2017).

The dissertation chapter describes the package's functionality and illustrates its use in a brief application to monthly UK electricity supply data.

1.3 Efficient Nurse Staffing: The Value of Hierarchical Time Series Forecasting and Forecast Combination

A combination of neoliberal austerity (slowdown in healthcare budget growth) and demographic factors (population growth, ageing population) have led to a widening gap between healthcare expenditure and healthcare budget in the UK since 2010, causing record deficits for the National Health Service (Maguire et al., 2016). Operating within the financial constraints requires efforts to increase cost efficiency. With staff costs accounting for over 50 % of hospitals' expenditures, making nurse staffing more efficient can help hospitals to deliver better value care and do so in a cost-efficient manner. This is our focus, analysing a dataset with more than 3 million observations from a large UK hospital.

Exploiting the convenient link between hierarchical forecasting and forecast combination, (i.e. the former's tendency to produce a large number of candidate models and the latter's dependence on the existence of a sufficient number of candidate models) and applying our **GeomComb** package, this final chapter of this dissertation evaluates and extends extant hospital occupancy forecasting research. Taking note of the hospital structure, we show how these approaches can forecast patient census and patient churn more accurately than univariate forecast methods that have been proposed in the past for this purpose.

Subsequently, we propose a parametric nurse staffing model that systematically and efficiently uses the mean forecasts and the forecast error distributions for patient census and churn, and show that high quality of patient care (measured by the relative number of understaffed shifts) can only be assured if staff planning allows utilises (i) short-term adjustment of nurse deployment calling on temporary agency

nurses, (ii) accounting for patient churn in the estimation of workload, and (iii) adjusting the mean forecasts with a distributional factor to account for a target minimum of fully staffed shifts. Taking this best practice approach to nurse staffing forward, we show how it can be implemented in a cost-efficient manner through a constrained optimisation approach that is suited to save the hospital £3.5 million in staff costs annually.

The chapter further includes a discussion how the staffing model can be refined in future research by using non-linear forecasting techniques that specifically take into account the multimodal distribution of patient churn data and/or allowing for explanatory variables, such as average length-of-stay, in the churn forecasts. Keeping the model very flexible using a set of parameters allows a wide range of hospital service providers to directly and easily implement the proposed staffing model.

1.4 References

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman. 2009. Hierarchical Forecasts for Australian Domestic Tourism. *International Journal of Forecasting* 25:146–166.

Ball, L., and S. G. Cecchetti. 1990. Inflation and Uncertainty at Short and Long Horizons. *Brookings Papers on Economic Activity* 1/1990:215–254.

Barrow, D. K., and N. Kourentzes. 2016. Distributions of Forecasting Errors of Forecast Combinations: Implications for Inventory Management. *International Journal of Production Economics* pp. 24–33.

Bates, J. M., and C. W. J. Granger. 1969. The Combination of Forecasts. *Operations Research Quarterly* 20:451–468.

Comin, D., and S. Mulani. 2006. Diverging Trends in Aggregate and Firm Volatility. *Review of Economics and Statistics* 88:374–383.

Cukierman, A., and A. H. Meltzer. 1986. A Theory of Ambiguity, Credibility, and Inflation under Discretion and Asymmetric Information. *Econometrica* 54:1099–1128.

Fischer, S. 1981. Towards an Understanding of the Costs of Inflation: II. *Carnegie-Rochester Conference Series on Public Policy* 15:5 – 41.

Friedman, M. 1977. Nobel Lecture: Inflation and Unemployment. *Journal of Political Economy* 85:451–472.

Gabaix, X. 2011. The Granular Origins of Aggregate Fluctuations. *Econometrica* 79:733–772.

Hansen, B. 2005. Challenges for Econometric Model Selection. *Econometric Theory* 21:60–68.

Holland, A. S. 1993. Uncertain Effects of Money and the Link Between the Inflation Rate and Inflation Uncertainty. *Economic Inquiry* 31:39–51.

Holland, A. S. 1995. Inflation and Uncertainty: Tests for Temporal Ordering. *Journal of Money, Credit and Banking* 27:827–837.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. 2011. Optimal Combination Forecasts for Hierarchical Time Series. *Computational Statistics and Data Analysis* 55:2579–2589.

Kapetanios, G., V. Labhard, and S. Price. 2008. Forecasting Using Bayesian and Information-Theoretic Model Averaging: An Application to UK Inflation. *Journal of Business & Economic Statistics* 26:33–41.

Maguire, D., P. Dunn, and H. McKenna. 2016. How Hospital Activity in the NHS in England Has Changed Over Time. The King's Fund Report, December 2016.

Quah, D. 1994. One Business Cycle and One Trend From (Many,) Many Disaggregates. *European Economic Review* 38:605–614.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2010. Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Review of Financial Studies* 23:821–862.

Stock, J. H., and M. W. Watson. 2004. Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* 23:405–430.

Wright, J. H. 2008. Bayesian Model Averaging and Exchange Rate Forecasts. *Journal of Econometrics* 146:329–341.

Chapter 2

Hierarchical Modelling and Forecasting System for Inflation Rate and Volatility

2.1 Introduction

It is widely accepted that inflation and inflation volatility can distort saving, investment and resource allocation decisions (cf. Friedman, 1977; Fischer, 1981; Holland, 1993). Low and steady inflation rates are the avowed objective of most monetary authorities, many of which have adopted *inflation targeting*, acknowledging the negative consequences of inflation volatility for economic growth and welfare.[1] Notwithstanding their importance, the causal relationship between inflation and inflation volatility is not well understood. The "Friedman-Ball hypothesis" (e.g. Friedman, 1977; Cukierman and Wachtel, 1979; Ball and Cecchetti, 1990; Evans, 1991) suggests that average inflation rate impacts inflation volatility positively. The "Cukierman-Melzer hypothesis" (e.g. Cukierman and Meltzer, 1986; Holland, 1995) suggests that the causality runs the other way, from inflation volatility to inflation. Kim and Lin (2012) address the reverse causality question using a system of simultaneous equations and panel data for 105 countries, and find the relationship between inflation and its volatility to be bi-directional, consistent with both hypotheses. This

---

[1]Inflation targeting describes a central bank's medium-term goal to reach an explicitly announced target inflation rate through monetary policy actions. In the UK, inflation targeting was first adopted in October 1992 and currently a point inflation target of (annualised) 2 % is applied at all times. For a historical overview of inflation targeting in the UK, see Bean (2003).

points to the value of a modelling system that can forecast both jointly, which is the basic focus of this paper.[2]

In a comprehensive review of econometric models for inflation, Stock and Watson (2008) compare univariate time series models, backward-looking Phillips curve models, and models with other explanatory variables (e.g. term spread). They conclude that structural (Phillips curve-based) models do not improve upon the forecast accuracy of univariate models overall. "...for the last 15 years, economists have not produced a version of the Phillips curve that makes more accurate inflation forecasts than those from a naïve model" (Atkeson and Ohanian, 2001). Based on estimates from a large set of inflation forecasting models Faust and Wright (2013) agree, concluding that simple models that limit or avoid parameter estimation – e.g. a driftless random walk – are hard to beat.

Univariate time series approaches do not generally distinguish between models for aggregate variables such as the inflation rate, and models for constituent variables such as product-level inflation rates. In modelling aggregates, and particularly the volatility of aggregates, there are potential gains from taking note of *covariance patterns* among the constituents that are nested hierarchically in the aggregate. Indeed, the Great Moderation – the sharply lower aggregate economic volatility from mid-80s to 2007, which is held to be a reason for the poor forecasting performance of structural models (Stock and Watson, 2008) – could be explained in terms of changes in covariance patterns among firms in growth (Comin and Mulani, 2006). Further, in many contexts where the forecast of an aggregate variable (inflation rate) is of interest, aggregation-consistent forecasts of its constituents (product-level inflation rates) are also of interest. Finally, given the multiplicity of models, some of which work better than others in time phases that differ in terms of volatility, it would be useful to exploit the potential of switching between models in order to generate more accurate forecasts.

There is a gap in the literature in terms of modelling systems for inflation and inflation volatility that explicitly consider the way product-level inflation series combine in the aggregate inflation rate and address the above-mentioned issues. In this paper we apply aggregation consistent methods for hierarchical time series (HTS, Athanasopoulos et al., 2009; Hyndman et al., 2011) modelling to obtain forecasts of inflation and inflation volatility. A large number of alternative specifications of HTS

---

[2]The results of Granger causality tests for our sample support the view of a bi-directional causal relationship and are presented in Appendix 2.C.

models can be used in forecasting, and we offer a dynamic model selection approach
that improves the accuracy of forecasts.

2.2 Inflation Rate and Its Volatility as Aggregate Variables

2.2.1 Inflation Rate

The aggregate inflation rate (Y_t) is constituted as the weighted average of product-level inflation rates, $y_{i,t}$:

$$Y_t = \sum_{i=1}^{N_t} w_{i,t} y_{i,t}$$

where N_t is the number of products in the price index at time t (time-invariant N in this study) and $w_{i,t}$ are the products' weights (share) in the price index at time t.[3] A second level decomposition of product-level inflation rates, into common, industry, and idiosyncratic parts, can be helpful in understanding the inflation data generating process.

$$y_{i,t} = c_t + I_{i,t} + \epsilon_{i,t}$$

$$Y_t = \sum_{i=1}^{N} w_{i,t} y_{i,t} = \sum_{i=1}^{N} w_{i,t} \left(c_t + I_{i,t} + \epsilon_{i,t} \right)$$

where c_t is the part of a product's inflation rate that is shared with all products, attributable to common shocks – (weighted) average over all product price growth rates; $I_{i,t}$ is the part of ith product's inflation rate that it shares with products in the same industry, but not with products in other industries – (weighted) excess growth rate of the products in the industry that product i belongs to, relative to all

---

[3]In this study, the aggregate inflation rate is the month-to-month growth rate of the Retail Price Index (RPI). The product-level inflation rates, $y_{i,t}$, are the month-to-month growth rates of the respective product price indices, $p_{i,t}$, calculated as: $y_{i,t} = \frac{p_{i,t} - p_{i,t-1}}{(p_{i,t} + p_{i,t-1})/2}$. This preserves seasonality in the series that we take to our modelling exercises, rather than difference it out. This growth rate estimator is symmetric about zero, and bounded, allowing the treatment of entries, exits, and continuers on the same footing (Comin and Mulani, 2004; Davis et al., 2006). It allows for consistent aggregation, and is identical to log differences up to a second-order Taylor Series expansion. See Davis et al. (2006) and references therein for discussions of the appeal of this estimator.

products in the price index; and $\epsilon_{i,t}$ is the excess inflation rate for product i relative to the sum of the common and the industry parts relating to it – the residual.

Figure 2.1 illustrates this decomposition for 'Oil and Other Fuels'. It is obvious that the idiosyncratic part (green line) is the main driver of the product-level inflation rate (solid black line).
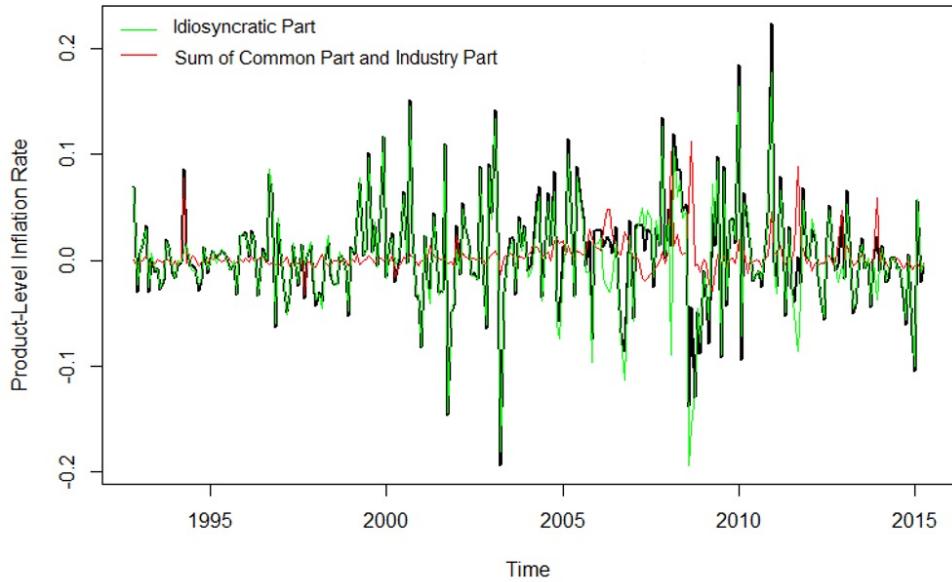
Fig. 2.1 Product-Level Inflation Rate Decomposition: Oil and Other Fuels.

2.2.2 Variance of Inflation Rate

While the inflation rate is observed, its volatility is an unobserved, latent variable – measures can be constructed in different ways.[4] Our focus is on *notional volatility* which corresponds to the ex-post sample-path of the inflation rate over a fixed time interval and can be measured without contingency on any specific model, non-parametrically; unlike *instantaneous volatility* which corresponds to the strength of the volatility process at a point in time and is model-specific. The simplest notional volatility measure is the simple moving average (SMA) variance. For a generic variable Y, over the chosen discrete time interval $m > 0$:

$$\hat{\sigma}^{2}_{Y_t}{}^{roll} = \frac{\sum_{\tau=t-m}^{t-1}(Y_\tau - \hat{Y}_t)^2}{m}$$

---

[4]See Andersen et al. (2010) for a comprehensive review of parametric and non-parametric volatility measurement.

where $\hat{Y}_t$ is the estimate of $E(Y_t)$, the time-varying expectation, the mean of Y over the interval (Comin and Mulani, 2006).[5] Conceptually, instantaneous volatility is the limiting value as $m \to 0$ in the $[t-m, t]$ interval, with m determining the bias-variance tradeoff of the estimator (larger values of m reducing variance, but increasing bias) in the interpretation of $\hat{\sigma}_{Y_t}^{2\ roll}$ as estimate of the 'current' variance of Y_t. The exponentially weighted moving average (EWMA) variance estimator which we use, is more appealing in that it places higher weights on more recent observations:

$$\hat{\sigma}_{Y_t}^{2\ roll} = \sum_{\tau=t-m}^{t-1} \alpha_\tau (Y_\tau - \hat{Y}_t)^2$$

where the weight scheme is given by: $\alpha_{\tau+1} = \lambda \alpha_\tau$, with $\lambda \in [0, 1]$ being the decay factor. A higher λ indicates slower decay, i.e. indicates strong persistence in volatility.[6]

The (rolling window EWMA) variance of the aggregate inflation rate, $\hat{\sigma}_{Y_t}^{2\ roll}$ can be written in terms of weighted (rolling window) estimators of product-level inflation rate variances and covariances. An exact decomposition in terms of a "variance component" (VC) and a "covariance component" (CC) is straightforward (Comin and Mulani, 2006):[7]

$$\hat{\sigma}_{Y_t}^{2\ roll} = \underbrace{\sum_i w_{i,t}^2 \hat{\sigma}_{y_{i,t}}^{2\ roll}}_{\text{VC}} + \underbrace{\sum_i \sum_{j \neq i} w_{i,t} w_{j,t} \hat{\sigma}_{y_{i,t}, y_{j,t}}^{roll}}_{\text{CC}}$$

Our focus in the volatility modelling part is on identifying the dynamic patterns in the way time-varying variances and covariances of the disaggregated components feed into the *volatility of the aggregate*, in order to forecast aggregate volatility better. Characterising time series dependencies in the variance component and the covariance component is a useful step in this. We work with estimates of product-level variances

---

[5]This volatility estimate is a modification of the measure used by Comin and Mulani (2006) – rather than symmetric about t, it is one-sided with respect to t. This makes it suitable for forecasting.

[6]EWMA models are related to GARCH-type models. In asset pricing, $E(Y_t)$, which is estimated as $\hat{Y}_t$, is usually assumed to be zero; this collapses the EWMA model to a zero-intercept IGARCH(1,1) – see Guo (2012, p. 193) for derivation:

$$\hat{\sigma}_{Y_t}^2 = \lambda \sigma_{t-1}^2 + (1 - \lambda) Y_{t-1}^2$$

Another appealing feature favouring the use of EWMA is the existence of variants that take note of distributional divergence from normality, which allows its use with great flexibility in our framework. See Lucas and Zhang (2016) for a review of robust EWMA, skewed EWMA, and fat-tailed skewed EWMA and their corresponding GARCH and GAS/DCS models.

[7]See Comin and Mulani (2004, p. 13) for a derivation of the variance identity.

$\hat{\sigma}^{2}_{y_{i,t}}{}^{roll}$, and covariances, $\hat{\sigma}^{roll}_{y_{it},y_{jt}}$ that are EWMA-smoothed.[8]

$$\hat{\sigma}^{2}_{y_{i,t}}{}^{roll} = \sum_{\tau=t-m}^{t-1} \alpha_{\tau}(y_{i,\tau} - \hat{y}_{i,t})^{2}$$

$$\hat{\sigma}^{roll}_{y_{i,t},y_{j,t}} = \sum_{\tau=t-m}^{t-1} \alpha_{\tau}(y_{i,\tau} - \hat{y}_{i,t})(y_{j,\tau} - \hat{y}_{j,t})$$

The time-varying volatility can be further decomposed into common, industry, and idiosyncratic parts. The Variance Component (VC) is:

$$\widehat{VC}_{t} = \underbrace{\sum_{i} w_{i,t}^{2}\hat{\sigma}^{2}_{c_{t}}{}^{roll}}_{\substack{VC \\ (\text{Var Common})}} + \underbrace{\sum_{i} w_{i,t}^{2}\hat{\sigma}^{2}_{I_{i,t}}{}^{roll}}_{\substack{VC \\ (\text{Var industry})}} + \underbrace{\sum_{i} w_{i,t}^{2}\hat{\sigma}^{2}_{\epsilon_{i,t}}{}^{roll}}_{\substack{VC \\ (\text{Var Idio.})}} +$$

$$+ \underbrace{2\sum_{i} w_{i,t}^{2}\hat{\sigma}^{roll}_{c_{t},I_{i,t}}}_{\substack{VC \\ (\text{Cov Common/Ind})}} + \underbrace{2\sum_{i} w_{i,t}^{2}\hat{\sigma}^{roll}_{c_{t},\epsilon_{i,t}}}_{\substack{VC \\ (\text{Cov Common/Idio.})}} + \underbrace{2\sum_{i} w_{i,t}^{2}\hat{\sigma}^{roll}_{I_{i,t},\epsilon_{i,t}}}_{\substack{VC \\ (\text{Cov Ind/Idio.})}}$$

The Covariance Component (CC) is:

$$\widehat{CC}_{t} = \underbrace{\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{2}_{c_{t}}{}^{roll}}_{\substack{CC \\ (\text{Var Common})}} + \underbrace{2\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{roll}_{c_{t},I_{j,t}}}_{\substack{CC \\ (\text{Cov Common/Ind})}} +$$

$$+ \underbrace{2\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{roll}_{c_{t},\epsilon_{j,t}}}_{\substack{CC \\ (\text{Cov Common/Idio.})}} + \underbrace{\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{roll}_{I_{i,t},I_{j,t}}}_{\substack{CC \\ (\text{Cov Ind/Ind})}} +$$

$$+ \underbrace{2\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{roll}_{I_{i,t},\epsilon_{j,t}}}_{\substack{CC \\ (\text{Cov Ind/Idio.})}} + \underbrace{\sum_{i}\sum_{j\neq i} w_{i,t}w_{j,t}\hat{\sigma}^{roll}_{\epsilon_{i,t},\epsilon_{j,t}}}_{\substack{CC \\ (\text{Cov Idio./Idio.})}}$$

---

[8]It is worth considering how inflation targeting can be incorporated into the model. One possibility is to assume that the expected inflation rate $E(Y_{t}) = \mu$ is the target inflation rate. The EWMA model under this assumption is:

$$\hat{\sigma}^{2}_{Y_{t}} = \lambda\sigma^{2}_{t-1} + (1-\lambda)Y^{2}_{t-1} - 2(1-\lambda)Y_{t-1}\mu + (1-\lambda)\mu^{2}$$

The volatility estimate is now a function of lagged inflation rate and target inflation rate as well.

We proceed to hierarchical time series modelling of the inflation system comprising mean inflation rate and inflation volatility, based on the above decompositions. We examine dynamic patterns in and between the variance and the covariance components, to distinguish high volatility / low volatility periods. We then proceed to forecast aggregate inflation rate and volatility, and compare results with extant volatility forecasting approaches.

2.3 Data and Descriptive Analysis

2.3.1 Data: UK Retail Price Index (RPI)

RPI is a long-standing measure of inflation in the UK, though it is no longer designated an official National Statistic (Office for National Statistics, 2013). RPI measures inflation with reference to the cost of a "representative basket of goods and services bought by consumers within the UK". It is calculated from the same basic price data as the CPI, and uses similar methodology in compiling and aggregating the constituent price indices. RPI covers 85 products.[9] Monthly data from January 1987 has been published by the Office for National Statistics. The weights are updated at the beginning of each year using the information on household spending. They are relatively unchanging, and may be considered exogenous.[10]

2.3.2 Descriptive Analysis

Inflation Rate. Figure 2.2 presents stylised facts of the monthly growth rate of the RPI and its volatility. The most notable features are the increased stability at a low inflation level after the introduction of inflation targeting (October 1992) and increased inflation volatility during recessions. The behaviour of inflation does not appear to be sensitive to changes in government or changes in the governance of the Bank of England. In the analysis that follows, we focus on the period after introduction of inflation targeting, which marks a clear structural break in the data generating process.

---

[9]For a full documentation of the representative products, see Annex C in Beeson (2016).

[10]The Herfindahl-Hirschman index (HHI) for the RPI lies in the range of 0.0210 to 0.0289 for the entire sample period from 1987 to 2015. For 85 items, a HHI of 0.0117 would indicate equal weighting, while a value of 1 would indicate concentration in a single product, suggesting that the RPI is relatively unconcentrated over the entire observation period.
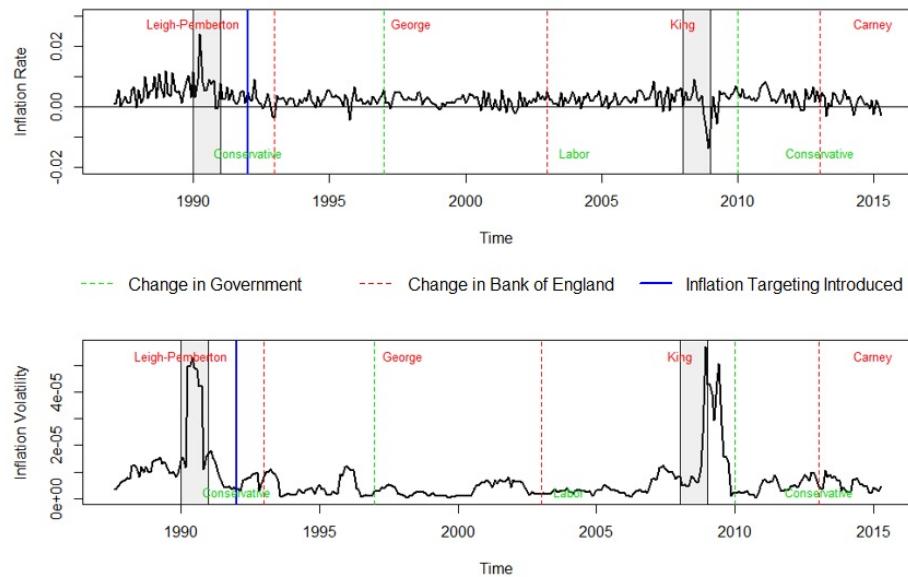
Fig. 2.2 Stylised Facts: UK Inflation – Impacts of Recessions and Monetary Policy.
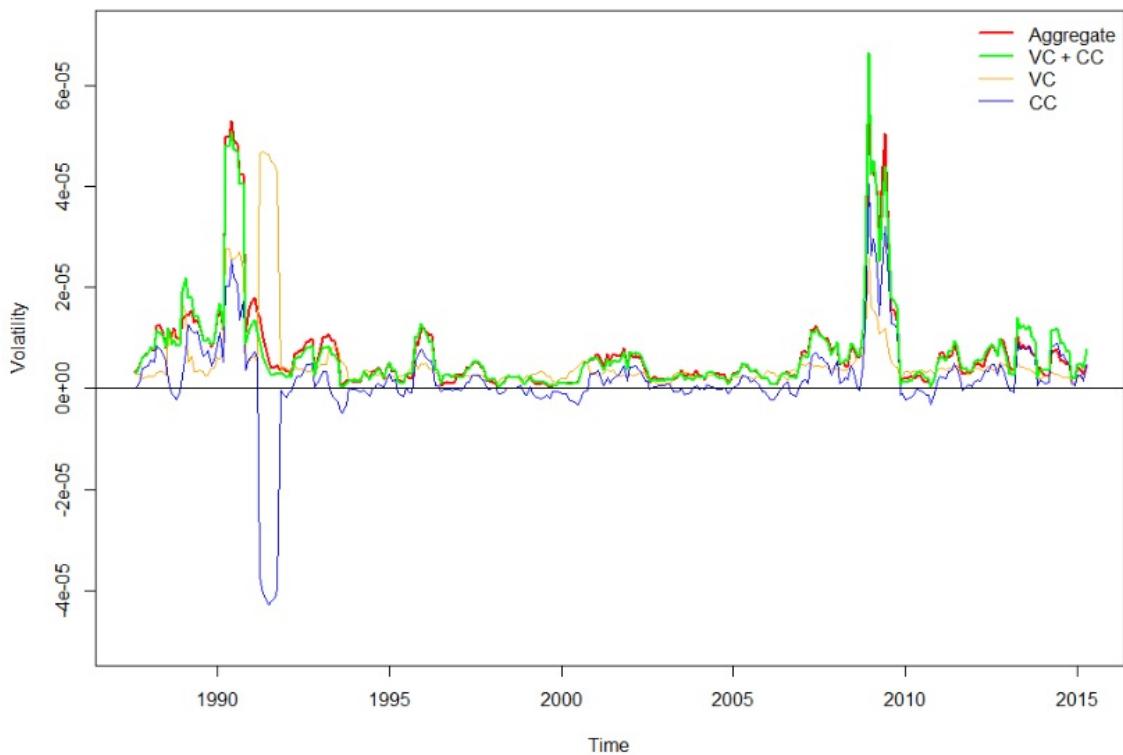
Fig. 2.3 Decomposition of Aggregate Inflation Variance into Variance (VC) and Covariance (CC) Components.

Inflation Volatility.   We begin with the decomposition of rolling sample variance of the aggregate inflation rate into its variance and covariance components (Figure 2.3). Two evident phases of high volatility appear to differ sharply from each other. In the early 90s, the recession saw high inflation and high volatility. The Great Recession saw negative inflation rate and high volatility. Both the variance and the covariance components differed in behaviour immediately following the first and the second high-volatility periods. A more detailed product-level analysis sheds some light on the underlying dynamics at play during these recession phases:

- The spikes in the Variance Component (the contribution of product-level variances to aggregate variance) are mostly due to a single product. The increased variance from April to October 1990 is caused by 'Council Tax and Rates' (responsible for 75 - 83 % of the Variance Component in this period); the increased value from April to October 1991 is caused by the same item (69 - 74 % of VC in this period), the spike in VC in December 2008 is due to the item 'Mortgage Interest Payments' that contributed 63 % that month.[11]

- Examining the Covariance Component, it is evident that product-level inflation rates co-move more strongly during recessions. The spikes in the CC are also linked to the items that cause increases in the VC: The main contributors to the positive spike in the CC from April to October 1990 are the positive covariances between 'Council Tax and Rates' and 'Rent', and between 'Council Tax and Rates' and 'Petrol and Oil' – all of which had positive product-level inflation rates. The main contributors to the CC from April to October 1991 were the negative covariances between 'Council Tax and Rates' (which showed a negative product-level inflation rate) with 'Beer on Sales', 'Rent' and 'Petrol and Oil' (which showed positive product-level inflation rates); as well as the positive covariance between 'Council Tax and Rates' and 'Mortgage Interest Payments' (which also showed a negative product-level inflation rate). The main contributors to the CC from December 2008 to July 2009 (note that the CC remains high for longer than the VC) are the positive covariances of 'Mortgage Interest Payments' with 'Gas' and with 'Petrol and Oil' (all of which showed negative product-level inflation rates).

---

[11]The time series of the monthly inflation rates corresponding to these items can readily be accessed on the ONS website: www.ons.gov.uk/economy/inflationandpriceindices/ timeseries/sgpr/mm23 (Council Tax and Rates), www.ons.gov.uk/economy/ inflationandpriceindices/timeseries/sgpn/mm23 (Mortgage Interest Payments). The specific drivers in play are likely to have been the short-lived poll tax in 1990 and 1991 (e.g. Ridge and Smith, 1991), and the sub-prime crisis in 2008 (e.g. Galati et al., 2011).

In recessions, we find that both the variance and covariance components of aggregate volatility tend to increase – with elevated variances of product-level inflation rates and elevated covariances across products. This is consistent with economic and financial contagion theory that builds on the idea that a shock to an individual item can cause a significant increase in co-movement between items. Dornbusch et al. (2000) find that while some degree of spill-over of micro-level shocks is expected even in tranquil times due to similarities between items (termed "fundamentals-based contagion" by Calvo and Reinhart, 1996), co-movement between items by far exceeds the amount that can be ascribed to macro-level shocks and fundamentals during periods of financial/economic turmoil. They link this excessive contagion to irrational phenomena – for instance financial panics, herd behaviour, loss of confidence, and extreme risk aversion.

In the immediate aftermath of the high-volatility phase we find that the covariance component is negative, while the variance component remains elevated. While the variance and covariance components are similar in magnitude, the dynamics of the covariance component drive the path of aggregate volatility, in particular in the aftermath of recessions. It is clearly important to understand covariation among product-level inflation rates.

Decomposition of the Variance and Covariance Components of Volatility.
We turn to the decomposition of the variance component, using the breakdown of the product-level inflation rate into common, industry, and idiosyncratic parts (Figure 2.4). Product-level inflation rate variances are mainly driven by the variances of their idiosyncratic part, which seems reasonable given our earlier finding that the idiosyncratic part is the dominant driver of product-level inflation rates and that the spikes in the VC are driven by high variances of individual products, rather than an increased variance of all products in the economy or an industry.

Figure 2.5 shows the decomposition of the covariance component (again using the breakdown into common, industry, and idiosyncratic at the product level). The variance of the common part appears to be the main driver of this component. This is intuitive, since the common part is designed to capture positive co-movement of the entirety of product-level inflation rates. Correspondingly, covariances between idiosyncratic parts capture a good part of the negative co-movement between product-level inflation rates. These patterns are consistent with the patterns found in the product-level analysis of recessions – during these phases, the CC is driven by a
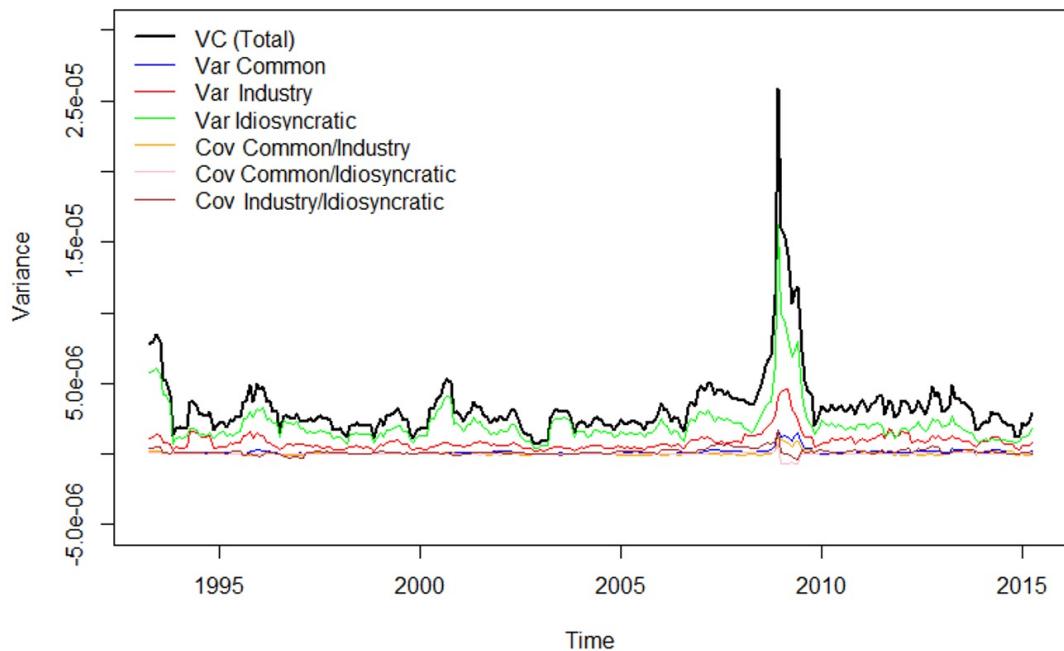
Fig. 2.4 Decomposition of Variance Component: Common, industry, and Idiosyncratic Part (Product-Level)
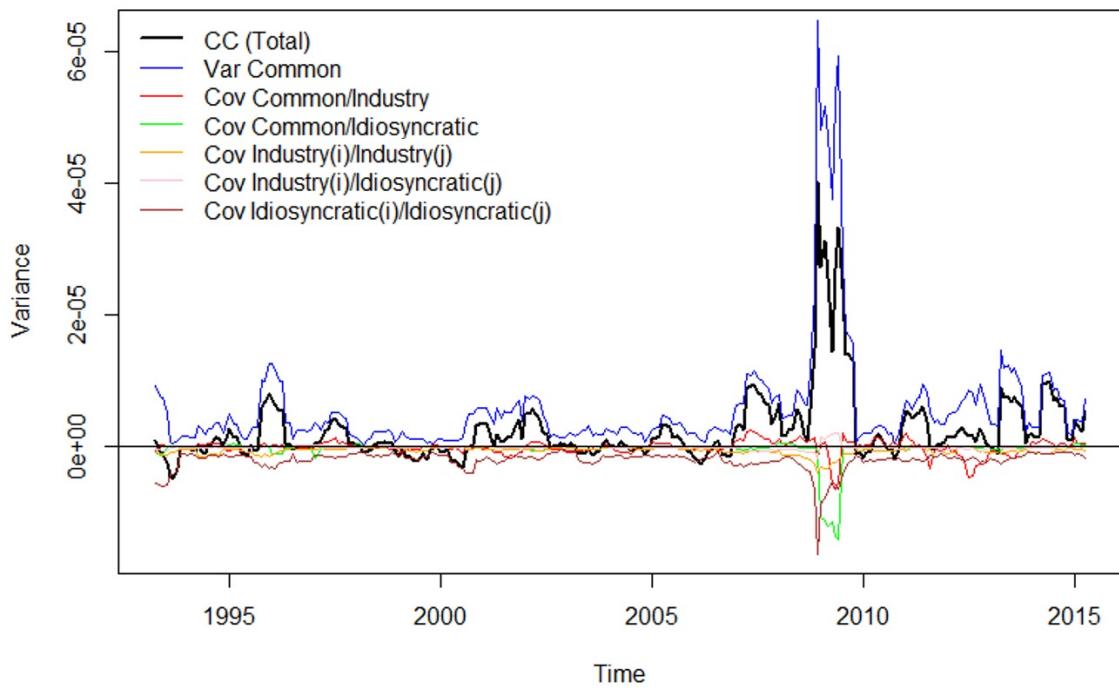
Fig. 2.5 Decomposition of Covariance Component: Common, industry, and Idiosyncratic Parts (Product-Level).

combination of stronger co-movement between all products, as well as covariances between individual products that are most responsible for the high aggregate volatility.

The above summary points to the potential for understanding phases of high aggregate volatility in terms of the dynamics in product-level inflation rates, paying attention to combining aggregate shocks (variance of common part) and industry/idiosyncratic shocks (covariances between industry and idiosyncratic parts).

Quah (1994) sounded an early warning of the potential fallacy of composition involved in modelling dynamics of macro aggregates ignoring the dynamic behaviour of disaggregates. The question whether aggregate volatility can be caused by shocks at microeconomic levels has been subject of much recent attention (e.g. Comin and Mulani, 2006; Gabaix, 2011; Carvalho and Gabaix, 2013). Abadir and Talmain (2002) have argued that common shocks are a more potent driver of aggregate fluctuations than idiosyncratic shocks. The disaggregated analysis of inflation suggests that aggregate volatility is driven by a combination of common shocks (through the CC) and idiosyncratic shocks (through the VC). We now turn to the question of using disaggregated information to forecast aggregate inflation and its volatility.

2.4   Forecasting Inflation Rate and Volatility: HTS

The issue of optimally forecasting contemporaneously aggregated variables – in particular the attempt of finding an optimal level of disaggregation for that purpose – is an ongoing debate in forecasting research (Hendry and Hubrich, 2006; Chen and Boylan, 2007, 2009). The approaches proposed in forecasting literature include: forecasting the aggregate using only aggregate information, forecasting the aggregate by aggregating forecasts of disaggregates, and forecasting the aggregate using information on disaggregates.

It is well established that if the data-generating process is known to the forecaster, then a procedure that aggregates the forecasts of disaggregates always outperforms direct forecasting of the aggregate, as the disaggregate information can be used optimally. But when the data-generating process is not known, as is common, the uncertainties in specification and estimation make the relative efficacies of the two approaches an empirical question. The Hierarchical Time Series (HTS) procedure of Athanasopoulos et al. (2009) is suitable for addressing this empirical question.

The starting point is to write the inflation rate as a hierarchical time series, noting that the index weights feature in the aggregation procedure. The decomposition involved is multi-stage, involving 1 aggregate series, 15 industry-level inflation

rates, 85 product-level inflation rates, and 255 inflation rates that correspond to the decomposition of product-level inflation rates into common, industry, and idiosyncratic parts). The total number of time series in a hierarchy with K levels is $n = 1 + n_1 + \cdots + n_K$, where n_i is the number of series at level i of the hierarchy (356 individual time series in our exercise).

Expressed in matrix form, the hierarchical time series ($\mathbf{Z}_t$) is obtained by applying a summation matrix ($\mathbf{S}$, of order $n \times n_K$) to aggregate the base-level series ($\mathbf{z}_{K,t}$). For the inflation rate, it is necessary to extend the simple aggregation in Athanasopoulos et al. (2009), to weighted aggregation.[12] To do this the simple summation matrix can be rewritten to contain the corresponding index weights of the bottom-level series.

As an illustration, consider 9 products (labelled 1 to 9) that aggregate into 4 industries (AA, AB, BA, BB), which then aggregate into core inflation (A) and non-core inflation (B), and finally into inflation (Y). The resulting hierarchical time series, can be expressed as follows:

$$
\underbrace{\begin{pmatrix}
P_{Y,t} \\
P^w_{y_A,t} \\
P^w_{y_B,t} \\
P^w_{y_{AA},t} \\
P^w_{y_{AB},t} \\
P^w_{y_{BA},t} \\
P^w_{y_{BB},t} \\
P^w_{y_1,t} \\
P^w_{y_2,t} \\
P^w_{y_3,t} \\
P^w_{y_4,t} \\
P^w_{y_5,t} \\
P^w_{y_6,t} \\
P^w_{y_7,t} \\
P^w_{y_8,t} \\
P^w_{y_9,t}
\end{pmatrix}}_{\mathbf{Z}_t}
=
\underbrace{\begin{pmatrix}
w_{1,t} & w_{2,t} & w_{3,t} & w_{4,t} & w_{5,t} & w_{6,t} & w_{7,t} & w_{8,t} & w_{9,t} \\
w_{1,t} & w_{2,t} & w_{3,t} & w_{4,t} & w_{5,t} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & w_{6,t} & w_{7,t} & w_{8,t} & w_{9,t} \\
w_{1,t} & w_{2,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & w_{3,t} & w_{4,t} & w_{5,t} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & w_{6,t} & w_{7,t} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{8,t} & w_{9,t} \\
w_{1,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & w_{2,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & w_{3,t} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & w_{4,t} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & w_{5,t} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & w_{6,t} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & w_{7,t} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{8,t} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{9,t}
\end{pmatrix}}_{\mathbf{S_t}}
\times
\underbrace{\begin{pmatrix}
P_{y_1,t} \\
P_{y_2,t} \\
P_{y_3,t} \\
P_{y_4,t} \\
P_{y_5,t} \\
P_{y_6,t} \\
P_{y_7,t} \\
P_{y_8,t} \\
P_{y_9,t}
\end{pmatrix}}_{\mathbf{z}_{K,t}}
$$

The incorporation of index weights into the summation matrix allows us to obtain forecasts of the unweighted series at each level of disaggregation as $\mathbf{z}_{K,t}$ contains the

---

[12]The original version of a hierarchical time series by Athanasopoulos et al. (2009) is a simple summation of base-level series. With weighted averages, such as the inflation rate, the index weights have to be incorporated into the summation matrix.

unweighted bottom-level series, and keeps the weights exogenous from the forecast estimation of the inflation rates at the different levels. The forecasts in $\mathbf{Z}_t$ contain the inflation rate at the aggregate level, and the *weighted* series at the disaggregated levels – indicated by the superscript w. It can easily be checked that $\mathbf{Z}_t$ contains the actual contribution of the respective series to the aggregate inflation rate, making the hierarchical time series aggregation consistent:[13]

$$\underbrace{\sum_{i=1}^{9} P_{y_i,t}^w}_{\text{Level 3}} = \underbrace{\sum_{i \in \{AA,AB,BA,BB\}} P_{y_i,t}^w}_{\text{Level 2}} = \underbrace{\sum_{i \in \{A,B\}} P_{y_i,t}^w}_{\text{Level 1}} = \underbrace{P_{Y,t}}_{\text{Level 0}}$$

Thus we can obtain valid forecasts of all series in the hierarchical time series, $\mathbf{Z}_t$, applying the different aggregation approaches for hierarchical time series (presented in the following section).

It is desirable to have forecasts of the unweighted series at the disaggregated levels – e.g. the forecast of industry level inflation, instead of the forecast of industry level inflation's contribution to aggregate inflation. This can be addressed easily: Noting that the h-step ahead forecasts of the quasi-exogenous index weights are $\hat{w}_{i,t+h|t} = w_{i,t}$, we can readily retrieve the forecasts for the unweighted series by dividing by the corresponding weights.

Aggregation Approaches. We consider the following aggregation approaches to forecasting the mean and variance of aggregate inflation, adapting the Hierarchical Time Series framework:

- *Top-down:* Forecast aggregate series and then disaggregate based on historical proportions (Gross and Sohl, 1990) or forecast proportions (Athanasopoulos et al., 2009). The methods can produce different forecasts for the disaggregated levels (Level 1 & Level 2), but all produce the same forecast for the aggregate level (Level 0). When the focus is on forecasting aggregate inflation rate/volatility, top-down methods are equivalent to an aggregate forecast.

- *Bottom-up:* Forecast disaggregate series at the lowest level and then aggregate up to forecast at higher levels. The argument in favour of the bottom-up approach is that the bottom-level series contain valuable information (e.g. different seasonal patterns in the bottom-level series). At the same time,

---

[13]This is a modification of the structure used in Capistrán et al. (2010), as their proposed approach does not satisfy aggregation consistency – a necessary property for HTS modelling.

smoothing noisy bottom-level series may lead to better forecasts of the aggregate. The empirical results are inconclusive. Top-down approaches do outperform bottom-up forecasts when the bottom-level data is noisy (e.g. Shlifer and Wolff, 1979; Fliedner, 1999; Hendry and Hubrich, 2006). There is also support the efficacy of bottom-up forecasts over top-down forecasts (e.g. Orcutt et al., 1968; Collins, 1976; Dunn et al., 1976; Dangerfield and Morris, 1992; Zellner and Tobias, 2000). While it seems intuitive that the bottom-up approach might provide higher bottom-level accuracy, and a top-down approach higher top-level accuracy, Hyndman et al. (2011) conclude that the bottom-up method performs significantly better than the conventional top-down method even for top-level forecasts.

Kahn (1998) suggests a hybrid approach, based on the argument that the efficacy of aggregation depends on the covariance structure of the constituent series (Tiao and Guttman, 1980; Kohn, 1982).

- *Middle-out*: A hybrid approach involving forecasting at intermediate levels, for aggregation to the higher levels (using a bottom-up approach), and disaggregation to the lower levels (using a top-down approach).

- *Optimal Combination*: Forecast each series in the hierarchy not heeding "aggregation consistency". Then optimally combine the forecasts to generate revised forecasts that are aggregation consistent and as close as possible to univariate forecasts. The reconciliation of forecasts is usually based on a Generalised Least Squares (GLS) estimator, but in practice reverts to OLS (Hyndman et al., 2011) or WLS (Hyndman et al., 2016a) due to the difficulty of estimating the covariance matrix of the reconciliation errors, which is non-identifiable as formally shown by Wickramasuriya et al. (2017). Several adjustments to the general approach have been proposed: Di Fonzo and Marini (2011) and Hyndman et al. (2016a) exploit the sparsity of the linear system, thereby making it possible to reduce computational complexity when a very large number of time series is involved. Van Erven and Cugliari (2015) propose the Game-Theoretically OPtimal (GTOP) method that guarantees that the total weighted quadratic loss of the reconciled forecasts will never be greater than the total weighted quadratic loss of the base forecasts. Wickramasuriya et al. (2017) recently proposed the Minimum Trace (MinT) reconciliation that minimises the sum of variances of the reconciled forecast errors under the assumption of unbiasedness.

Forecasting Approaches   We assess the forecast accuracy of each of these four aggregation approaches, using six different univariate forecasting methods. Three of the methods are commonly used in hierarchical time series forecasting and are included in the R package *hts* (Hyndman et al., 2016b) – ARIMA (Box and Jenkins, 1970), ETS (introduced/extended by: Pegels, 1969; Gardner, 1985; Hyndman et al., 2002; Taylor, 2003), and a naïve Random Walk forecast as benchmark. We use the selection algorithms based on minimising AICc proposed by Hyndman and Khandakar (2008) to fit the ARIMA and ETS models. In addition, the following methods, which have found favour in forecasting competitions as accurate, robust, and reliable, are also employed:

- *Damped Trend* (Gardner and McKenzie, 1985): These models deal with the problem that exponential smoothing methods with constant trend tend to over-forecast, by introducing a parameter that dampens the trend to a flat line some time in the future. The superior performance of the damped trend model compared to a range of other methods is documented (Fildes and Ord, 2002; Armstrong, 2006). We employ an additive damped trend model with additive errors and an additive seasonal component – this corresponds to an $ETS(A, A_d, A)$ model in the general notation of (Hyndman et al., 2008).[14]

- *Theta method* (Assimakopoulos and Nikolopoulos, 2000): The Theta method has been found to produce the most accurate forecasts for monthly data in the M3 forecasting competition (Makridakis and Hibon, 2000) and come to serve as a benchmark in more recent forecasting competitions (Athanasopoulos et al., 2011). It is also relatively simple and computationally fast (Nikolopoulos et al., 2012). The Theta method is applied to deseasonalised time series (usually based on the multiplicative classical decomposition).[15] The forecasts obtained with the Theta method are equivalent to Simple Exponential Smoothing with drift, where the drift is equal to half the slope of a linear regression fitted to the data (Hyndman and Billah, 2003).

---

[14]The damped trend model is part of the ETS framework. So the ETS algorithm will also select a damped trend specification for some of the series in the hierarchy. However, it is common to use it as a separate method – thereby forcing all series into a damped trend specification.

[15]Fiorucci et al. (2016b) argue that the seasonality test employed might not work well if the time series has one or more unit roots with a slow decay in the autocorrelation function. Since the inflation rate in our forecasting sample is stationary, this is not an issue – see 2.B for results of stationarity tests.

- *Dynamic Optimised Theta Model* (Fiorucci et al., 2016b) is a generalisation of the standard Theta method. DOTM produced more accurate forecasts than the standard Theta model for almost all combinations of type of data and frequency of the M3 time series (Fiorucci et al., 2016b) – the disadvantage is its higher computational intensity, which may be important when the number of base-level series is very large. With only 356 series (for the mean equation) and 15 series (for the variance equation), we have no difficulty in using DOTM.[16]

  Some special features of our data are to be noted. Thomakos and Nikolopoulos (2014) document that the Theta method performs especially well with trended series. Seasonality is a more dominant feature in our data than trend, and this might favour ARIMA and ETS models which incorporate seasonality in estimation. The Theta method only forecasts the deseasonalised series and then reseasonalises the data based on the multiplicative classical decomposition. The Theta method is inferior to other methods for forecasting monthly data with strong seasonality (Athanasopoulos et al., 2011).

The accuracy of the forecasts produced with these six forecasting methods is evaluated based on MAE[17]. Our forecast accuracy evaluation involves time series cross-validation based on training sets with a minimum of 180 observations, and constant length test sets. This involves separate analyses for 1 month, all the way up to 12 months horizons.[18]

Forecast Combination: Dynamic Model Switching. So far, we have been concerned with methods for determining the best individual forecast model, and aggregation procedure, for disaggregated inflation forecasting. In the next step, we take on board the fact that the models that produce most accurate forecasts often differ depending on different economic conditions. For example, Philips Curve-based models are more accurate during recessions, but do not consistently outperform the

---

[16]The variants of the Theta models were estimated using the *forecTheta* package in R (Fiorucci et al., 2016a). The provided seasonality test was used in order to choose between additive and multiplicative decomposition. Model parameters were optimised using the Nelder-Mead algorithm.

[17]Other standard measures such as RMSE, MPE, MAPE, MASE can be readily used. Since all the forecasts are computed for a single series – aggregate inflation rate, or aggregate inflation variance – we prefer MAE for its easy interpretation, because it is less sensitive to outliers than MSE or RMSE and it avoids common problems with MPE and MAPE in the presence of zero or very small values in some series. It should be noted that MAE is a scale-dependent measure and could not be used in a comparison between series, in which case MAPE or MASE (Hyndman and Koehler, 2006) should be used.

[18]The suitability of cross-validation for accuracy assessment with time series data is discussed in Bergmeir et al. (2015).

best univariate models at other times (Stock and Watson, 2008). We adopt a dynamic model selection (DMS) approach, which leverages the ability of different models to produce more accurate forecasts under different conditions. DMS is an extreme variant of forecast combination (i.e. weighted averaging of individual forecasts for a single time series) in which a combination weight of 1 is allocated to one of the candidate forecasting models, in a time-varying fashion depending on pre-defined criteria.

The literature on DMS is limited but promising: Belmonte and Koop (2013) use switching linear Gaussian state space models to forecast inflation; McMillan (2014) uses in-sample criteria in order to select between linear and nonlinear models for stock return forecasting; Buncic and Moretto (2015) use a dynamic model selection and averaging framework for copper price forecasts. Bagdatoglou et al. (2016) find that a dynamic model selection and averaging algorithm can improve forecasting accuracy for US inflation significantly compared to the UC-SV model of Stock and Watson (2008).

The forecast combination approach typically uses a linear combination of forecasts obtained from different models for the same time series, motivated by the view that all models of real world data generation processes are mis-specified, and combining forecasts across models can decrease model uncertainty and improve accuracy, by exploiting the different strengths of different models while compensating for their weaknesses.[19] One challenge of applying forecast combination to hierarchical time series is that the requirement of lack of high collinearity between the different forecast series is rarely satisfied. Similar methods – such as ETS and damped trend, or the Theta method – tend to produce collinear forecasts. As it relies on dynamic selection rather than averaging, the model switching approach that we employ does not break down under collinearity (McMillan, 2014). It can be seen as a boundary case of forecast combination.

---

[19]Different methods have been proposed and implemented: Clemen (1989) argues that simple averaging of all available forecasts is a successful and robust method; Armstrong (2001) suggests the use of trimmed means to avoid sensitivity to extreme values. Stock and Watson (2004) find that symmetric 5 % trimming performed about the same as simple averaging, while Jose and Winkler (2008) find that trimming of 10 - 30 % or Winsorising of 15 - 45 % can lead to improved accuracy compared to simple averages using data from the M3 forecasting competition; Granger and Ramanathan (1984) calculate the combination weights using an OLS regression; Aiolfi and Timmermann (2006) group forecasts into several clusters using a k-means algorithm, final forecasts are then obtained by averaging forecasts of the historically better performing cluster; and Hsiao and Wan (2014) introduce several ways of eigenvector-based forecast combination. For an in-depth discussion of different available methods, see Chapter 3 of this dissertation.

We have a number of candidate models – 26 of them, resulting from combining every aggregation approach with every forecasting approach. It is likely that different models forecast better for different parts of the sample. Choosing a single model out of many for the entire sample is likely to lead to poor out-of-sample forecasts. We noted earlier that the inflation rate series can be considered to have two volatility regimes, high and low. It is well-established that inflation rate and inflation volatility are strongly interconnected (Friedman-Ball hypothesis & Cukierman-Melzer hypothesis). The model switching approach allows us to use in-sample variance as criterion to select the best model for each point in time. The interconnection between inflation rate and volatility makes it more likely that changes in in-sample variance affect the underlying data generating process. This is the fundamental rationale for switching between models.[20] The model switching approach involves recursive estimation and in-sample forecast evaluation to guide the switching between forecasts produced using different models.

We adopt the following model switching rule (for both the mean and variance forecasts): Split up the validation set, $\mathbb{S}^{CV}$ into two subsets, a high-volatility set, $\mathbb{S}^{CV}_{high}(x)$ and a low-volatility set, $\mathbb{S}^{CV}_{low}(x)$. The elements of the high-volatility subset relate to those time points where the unconditional inflation volatility (as measured by the variance of the aggregate inflation rate over the past 6 months) exceeds the xth quantile of in-sample variance distribution, $\Phi(\sigma^2_{sample}, x)$:[21]

$$\mathbb{S}^{CV}_{high}(x) = \left\{ y_t \in \mathbb{S}^{CV} | \sigma^2_t \leq \Phi(\sigma^2_{sample}, x) \right\}$$

$$\mathbb{S}^{CV}_{low}(x) = \left\{ y_t \in \mathbb{S}^{CV} | \sigma^2_t > \Phi(\sigma^2_{sample}, x) \right\}$$

This leaves the choice of x – the threshold that optimally splits the support of the sample variance distribution, such that validation set is split into the two sub-samples in a way that allows us to minimise the MAE of the final forecast.

We adopt the following procedure:

1. Compute $\mathbb{S}^{CV}_{high}(x)$ and $\mathbb{S}^{CV}_{low}(x)$ for each possible x, i.e. for each percentile of the sample variance distribution,

2. For each value of x (100 cases), produce forecasts for $\mathbb{S}^{CV}_{high}(x)$ with all 26 candidate models and select the model with the highest accuracy (lowest MAE)

---

[20]McMillan (2014) used the best-fitting model of the previous period (according to AIC) as in-sample criterion to select a forecast model in a recursive framework.

[21]The switching rule is presented for the inflation rate for illustration, but applies to inflation volatility forecasts analogously.

for the given high-volatility subset – denote this optimal model for the high-volatility subset for a given x by $model_{high}^{opt}(x)$; for each x, produce forecasts for $\mathbb{S}_{low}^{CV}(x)$ with all 26 candidate models and select the model that shows the highest accuracy (lowest MAE) for the given low-volatility subset – denote this optimal model for the low-volatility subset for a given x by $model_{low}^{opt}(x)$

3. For each x, use $model_{high}^{opt}(x)$ to produce h-step ahead forecasts, $\hat{y}_{t+h|t}^{high}$, if $y_t \in \mathbb{S}_{high}^{CV}(x)$; and $model_{low}^{opt}(x)$ to produce h-step ahead forecasts, $\hat{y}_{t+h|t}^{low}$, if $y_t \in \mathbb{S}_{low}^{CV}(x)$. This yields forecasts for the entire cross-validation set, by combining the forecasts from the two subsets:

$$\hat{y}_{t+h|t}(x) = \begin{cases} \hat{y}_{t+h|t}^{high}, & \text{if } y_t \in \mathbb{S}_{high}^{CV}(x) \\ \hat{y}_{t+h|t}^{low}, & \text{if } y_t \in \mathbb{S}_{low}^{CV}(x) \end{cases}$$

4. Compute MAE for the forecasts of the cross-validation set for each x, and select the MAE-minimising x, x_{opt}, as a splitting point for our switching rule, which now becomes:

$$\hat{y}_{t+h|t}(x_{opt}) = \begin{cases} \hat{y}_{t+h|t}^{high}, & \text{if } y_t \in \mathbb{S}_{high}^{CV}(x_{opt}) \\ \hat{y}_{t+h|t}^{low}, & \text{if } y_t \in \mathbb{S}_{low}^{CV}(x_{opt}) \end{cases}$$

To summarise, the general approach is a data-driven procedure to select the forecast model depending on an in-sample criterion as it varies over time – in this study, the unconditional sample variance corresponding to the point in time at which the forecast is made. Due to the greater flexibility relative to a fixed forecast model, this approach has the potential to improve forecast accuracy.

Forecasting System: Inflation Mean & Variance. The empirical strategy is to estimate time-varying expectation of the inflation rate, fitting the model (e.g. ARIMA, ETS) to Y_t, and to use the fitted conditional mean $\hat{Y}_t$ as the expected inflation rate in the variance equation:[22]

---

[22]An alternative would be to think of the expected inflation rate as a linear combination of the mean model, $\hat{Y}_t$, and the target inflation rate, μ:

$$\gamma \hat{Y}_t + (1 - \gamma)\mu$$

The EWMA variance equation is then:

$$\hat{\sigma}^2_{Y_t} = \lambda\sigma^2_{t-1} + (1-\lambda)Y^2_{t-1} - 2(1-\lambda)Y_{t-1}\hat{Y}_t + (1-\lambda)\hat{Y}^2_t$$

The above equation illustrates how the fitted conditional mean enters the variance equation through the EWMA calculation. In order to keep the presentation simple we show this only for the aggregate inflation; however, this system that comprises of a mean model and a variance model can easily be applied at various levels of disaggregation (e.g. $\hat{y}_{it}$ entering into the EWMA calculation for $\hat{\sigma}^2_{y_{it}}$ at the product level).

To obtain the mean forecast, we use the subcomponents of the product-level inflation rates as input. These base-level series are combined with the index weights into a hierarchical time series. Forecasts are produced for the aggregate mean inflation rate using the aggregation approaches and forecast methods described above.

The mean model fit will be used for the calculation of EWMA-smoothed base-level (co)variances (and, consequently, to create the input series of the variance model). Before this, the optimal EWMA decay parameter, λ, must be determined. Our strategy, consistent with inflation volatility literature, is to use the conditional volatility as estimated by a GARCH(1,1) as proxy for the 'actual' values of volatility, and to estimate EWMA-smoothed aggregate volatility using a large range of potential decay parameter values. The RMSE-minimising λ can then be selected and – together with the fitted values of the base-level parts from the best mean model – used for the computation of the EWMA-smoothed base-level covariance matrix. By applying the index weights to the corresponding terms of the smoothed base-level covariance matrix, we directly compute the parts of variance component and the covariance component as inputs for the HTS volatility forecasts that are produced using the forecasting and aggregation methods described above.

2.5   Results

We now present an evaluation of forecast accuracy of the hierarchical time series based mean and volatility forecasts. With the exception of the Theta method

$$\begin{aligned}
\hat{\sigma}^2_{Y_t} = {} & \lambda\sigma^2_{t-1} + (1-\lambda)Y^2_{t-1} - 2(1-\lambda)\gamma Y_{t-1}\hat{Y}_t \\
& - 2(1-\lambda)(1-\gamma)Y_{t-1}\mu + (1-\lambda)\gamma\hat{Y}^2_t \\
& + (1-\lambda)(1-\gamma)\mu^2
\end{aligned}$$

variants, the models were estimated using data with no prior seasonal adjustment, extracting seasonality in the estimation.[23] For the Theta method models, the series were deseasonalised first, using the algorithm provided in Fiorucci et al. (2016a) which bases the choice between additive and multiplicative classical decomposition on a seasonality test, and were reseasonalised after the estimation. The tables in this section present the relative accuracy of the forecasts, i.e. $\frac{MAE_{model}}{MAE_{opt}}$ where MAE_{opt} is the accuracy of the best model.

Mean Model – Results.   The hierarchical time series data that is used to estimate the mean models has 4 levels of disaggregation (from lowest to highest): Level 0 – is the aggregate inflation rate, Level 1 – is the 15 industry-level inflation rates (e.g. Food, Housing, etc.), Level 2 – is the 85 product-level inflation rates (e.g. Bread, Furniture, Pet Care, etc.), and Level 3 – decomposes each product-level inflation rate into common, industry, and idiosyncratic parts. Consequently, the Bottom-Up approach works with the 255 base-level series and, using the summation matrix, aggregates the fitted values to obtain a fit for aggregate inflation rate; the Middle-Out (Level 2) approach fits the product-level series and aggregates to the aggregate inflation rate[24]; the Middle-Out (Level 1) approach fits the industry-level series and aggregates to the aggregate inflation rate; and the Top-Down approach works directly with the aggregate inflation rate series. The Optimal Combination approach fits all 356 constituent series of the hierarchical time series, at their different levels of disaggregation, and then applies the WLS method of Hyndman et al. (2016a) to reconcile the forecasts.[25] Table 2.1 presents relative forecast accuracy compared to the best model of the respective horizon for the full validation set – the absolute MAE value is presented in brackets for the best model for each horizon.

Accuracy assessment based on time series cross-validation provides clear support for the disaggregated approach. The best disaggregated model beats the best top-down univariate approach for all horizons, by a range between 8 % (for a one-step ahead forecast) and 21 % (for a twelve-step ahead forecast). For all horizons apart

---

[23]Alternatively, deseasonalised base-level series could be used. However, since the forecasting frameworks – ARIMA and Exponential Time Series Smoothing (including Damped Trend) – are able to incorporate seasonality in the model, we choose to use seasonal base-level series to avoid losing valuable base-level information.

[24]Middle-Out approaches apply a bottom-up approach to obtain estimates for levels of lower disaggregation and a top-down approach to obtain estimates for levels of higher disaggregation.

[25]It is worth noting that Wickramasuriya et al. (2017) document promising results for the recently proposed minimum trace (MinT) reconciliation method, making it likely that this algorithm could further improve the optimal combination forecasts – given the novelty of the method, its performance will have to be evaluated in future research.

Table 2.1 Cross-Validated Test Set Accuracy - Mean Model.

	1m	2m	3m	4m	5m	6m	7m	8m	9m	10m	11m	12m
Naïve												
Naïve	1.915	1.860	1.925	1.874	1.879	1.846	1.868	1.884	1.899	1.893	1.897	1.854
Top-Down												
ETS	1.080	1.116	1.170	1.178	1.202	1.229	1.246	1.262	1.263	1.257	1.254	1.257
ARIMA	1.172	1.160	1.163	1.167	1.174	1.186	1.197	1.206	1.207	1.208	1.209	1.212
Theta	1.093	1.117	1.151	1.165	1.190	1.210	1.224	1.234	1.232	1.229	1.227	1.231
Damped Trend	1.106	1.128	1.183	1.200	1.230	1.259	1.278	1.292	1.293	1.287	1.282	1.282
Dynamic Optimised Theta	1.093	1.115	1.147	1.160	1.184	1.203	1.217	1.226	1.224	1.221	1.220	1.223
Middle Out (Level 1)												
ETS	1.008	1.021	1.048	1.072	1.088	1.106	1.119	1.126	1.128	1.129	1.132	1.137
ARIMA	1.017	1.001	1.011	1.023	1.028	1.030	1.031	1.029	1.025	1.021	1.018	1.018
Theta	1.047	1.059	1.094	1.113	1.136	1.154	1.166	1.174	1.175	1.176	1.176	1.183
Damped Trend	1.014	1.019	1.060	1.089	1.115	1.140	1.157	1.169	1.172	1.174	1.178	1.183
Dynamic Optimised Theta	1.043	1.055	1.090	1.109	1.132	1.150	1.163	1.170	1.171	1.172	1.173	1.179
Middle Out (Level 2)												
ETS	1.018	1.003	1.027	1.039	1.048	1.062	1.073	1.079	1.080	1.082	1.084	1.087
ARIMA	1.021	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		(0.0023)	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0024)	(0.0025)	(0.0025)
Theta	1.085	1.069	1.084	1.092	1.099	1.116	1.125	1.132	1.132	1.132	1.139	1.149
Damped Trend	1.065	1.045	1.078	1.104	1.121	1.137	1.148	1.157	1.160	1.160	1.165	1.171
Dynamic Optimised Theta	1.110	1.084	1.090	1.096	1.104	1.121	1.131	1.138	1.137	1.137	1.144	1.154
Bottom Up												
ETS	1.088	1.114	1.169	1.201	1.230	1.250	1.258	1.269	1.265	1.261	1.263	1.262
ARIMA	1.092	1.076	1.079	1.083	1.088	1.088	1.089	1.085	1.080	1.077	1.072	1.069
Theta	1.154	1.166	1.204	1.218	1.238	1.259	1.272	1.282	1.284	1.284	1.284	1.288
Damped Trend	1.083	1.101	1.159	1.185	1.212	1.244	1.261	1.278	1.285	1.284	1.287	1.290
Dynamic Optimised Theta	1.144	1.157	1.193	1.207	1.227	1.247	1.260	1.268	1.269	1.270	1.270	1.274
Optimal Combination												
ETS	**1.000**	1.022	1.070	1.089	1.113	1.138	1.155	1.169	1.170	1.168	1.170	1.175
	(0.0022)											
ARIMA	1.061	1.053	1.060	1.068	1.076	1.086	1.094	1.098	1.098	1.099	1.099	1.101
Theta	1.087	1.109	1.144	1.158	1.182	1.202	1.216	1.226	1.225	1.222	1.220	1.224
Damped Trend	1.089	1.110	1.164	1.182	1.213	1.242	1.261	1.276	1.277	1.272	1.268	1.269
Dynamic Optimised Theta	1.087	1.108	1.140	1.153	1.176	1.196	1.209	1.219	1.217	1.215	1.213	1.217

from 1 month, the Middle Out (Level 2) ARIMA model produces the most accurate forecasts, i.e. the best approach to forecasting the conditional mean of aggregate inflation is to forecast product-level inflation rates (Level 2) using (seasonal) ARIMA models and then aggregate these forecasts.

These results are consistent with previous empirical findings on both aggregation approaches and forecasting methods: Our conclusions are similar to those of Athanasopoulos et al. (2011) about the dominance of ARIMA and ETS techniques compared to Naïve, Damped Trend, and Theta method for non-trended data that has strong seasonality. The results also support findings of a variety of studies evaluating the potential of disaggregated methods (e.g. Kahn, 1998) that conclude that bottom-up methods do not usually produce good aggregate forecasts: We find that the hybrid approaches (Middle-Out and Optimal Combination) outperform bottom-up; however, it should be noted that the best bottom-up approach outperforms top-down for all horizons except for 1-month forecasts – this suggests that the value of disaggregated information (in this case, mainly the different seasonal patterns of the bottom-level
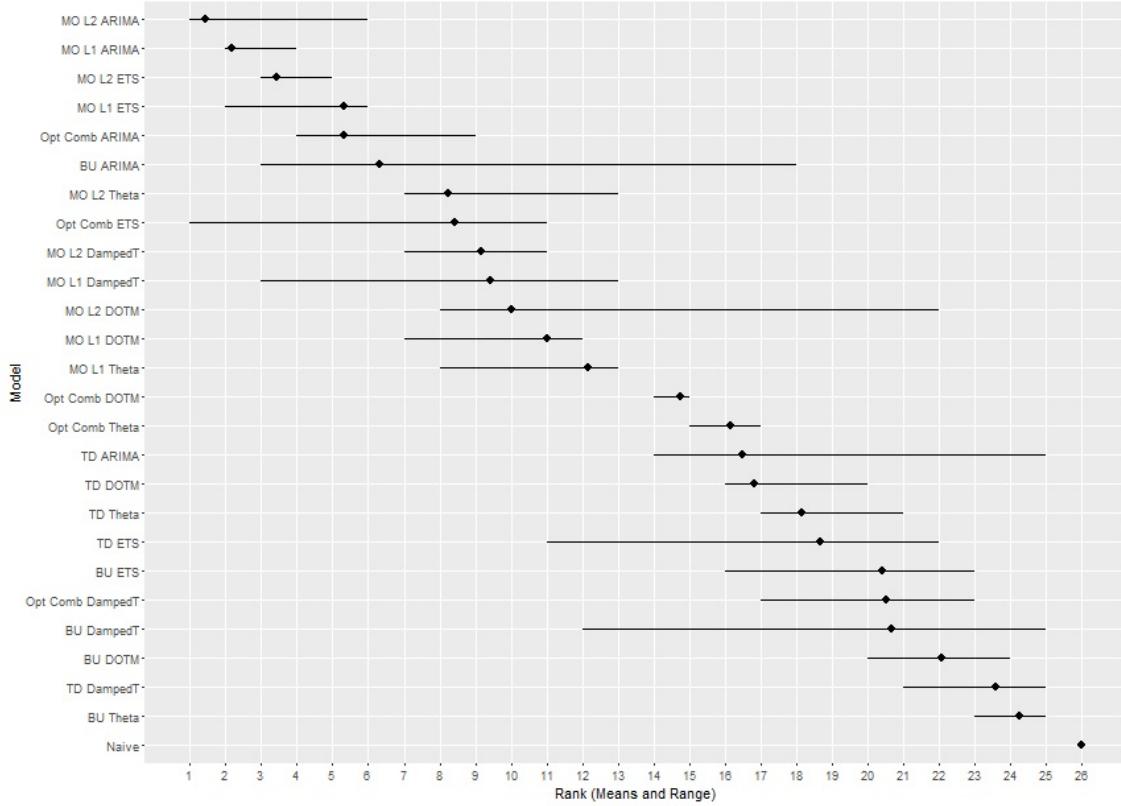
Fig. 2.6 Ranking (Mean and Range) of the Mean Model Approaches.

series) outweighs the cost of noisy bottom-level data. Finally, while both Theta method variants do not seem well-suited for forecasting seasonal, non-trended data, the dynamic optimised Theta method produces slightly better results than the original Theta method for 4 out of 5 aggregation approaches. All methods produce much better results than a naïve approach.

In order to present the information from the table in a concise way, Figure 2.6 plots the mean ranking, as well as the range of rankings, for each method out of the 26 models. This shows much better that (a) ARIMA and ETS methods are dominant for seasonal, non-trended monthly data, and (b) how valuable the disaggregated modelling approach is for inflation forecasting, given that the best aggregate model (Top Down ARIMA) has an average ranking of 16.5 out of 26 models.

Table 2.2 shows the test statistics and p-values of the Diebold-Mariano test (with a quadratic loss function) for predictive accuracy, comparing the performance of the best disaggregated model with the best aggregate model. For all horizons except for 1-month forecasts, the disaggregated model forecasts significantly better, at least at

Table 2.2 Diebold-Mariano Tests of Predictive Accuracy

	1m	2m	3m	4m	5m	6m
Best Aggregate	Opt Comb ETS	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA
Best Disagg.	TD ETS	TD DOTM	TD DOTM	TD DOTM	TD ARIMA	TD ARIMA
DM Statistic	-0.80	-1.62	-1.57	-1.41	-2.12	-2.19
p-value	0.210	0.054	0.059	0.080	0.018	0.015
	7m	8m	9m	10m	11m	12m
Best Aggregate	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA
Best Disagg.	TD ARIMA	TD ARIMA	TD ARIMA	TD ARIMA	TD ARIMA	TD ARIMA
DM Statistic	-2.28	-2.31	-2.20	-2.13	-2.04	-2.04
p-value	0.012	0.011	0.015	0.018	0.022	0.022

the 10 % level. The disaggregated models are seen to be relatively more valuable for multi-horizon forecasts.

Mean Model – Dynamic Model Switching. In order to assess the value of allowing for different models in low-volatility and high-volatility phases, we applied the dynamic model switching rule outlined in Section 2.4. Table 2.3 presents the optimal (MAE-minimising) quantiles ($\Phi(\sigma^2_{sample}, x_{opt})$) of the sample variance distribution that divide the validation set into the high-volatility and low-volatility subsets (denoted as 'Split Variance'), and the best models selected for the two subsets.

Table 2.3 Switching Rule Mean Model: Results for Different Horizons.

	1m	2m	3m	4m	5m	6m
Best High Vol	MO L1 ARIMA	MO L1 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA
Best Low Vol	BU ETS	BU ETS	BU ETS	BU ETS	BU DampedT	BU Theta
Split Variance	5.45e-06	5.42e-06	5.29e-06	5.29e-06	5.29e-06	5.29e-06
	7m	8m	9m	10m	11m	12m
Best High Vol	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L2 ARIMA	MO L1 ARIMA	MO L1 ARIMA
Best Low Vol	MO L2 Theta	MO L2 Theta	MO L2 ETS	MO L2 ETS	MO L2 ETS	MO L2 ETS
Split Variance	5.29e-06	5.29e-06	8.99e-06	8.99e-06	1.79e-05	2.29e-05

The application of the switching rule delivers some interesting insights in the previous findings: We find that the dominance of the Middle Out ARIMA approaches stems from their better performance at times of higher volatility. In phases of lower volatility, Bottom-Up ETS, Damped Trend, and Theta models (all of which belong to the exponential smoothing family) are best up to 6-months, and Middle Out (Level 2) Theta and ETS are best for longer horizons. It makes sense that bottom-up models are well-suited for forecasting at times of low volatility, as they contain valuable additional information and noise is often considered to be positively correlated with volatility (e.g. Bandi and Russell, 2006).
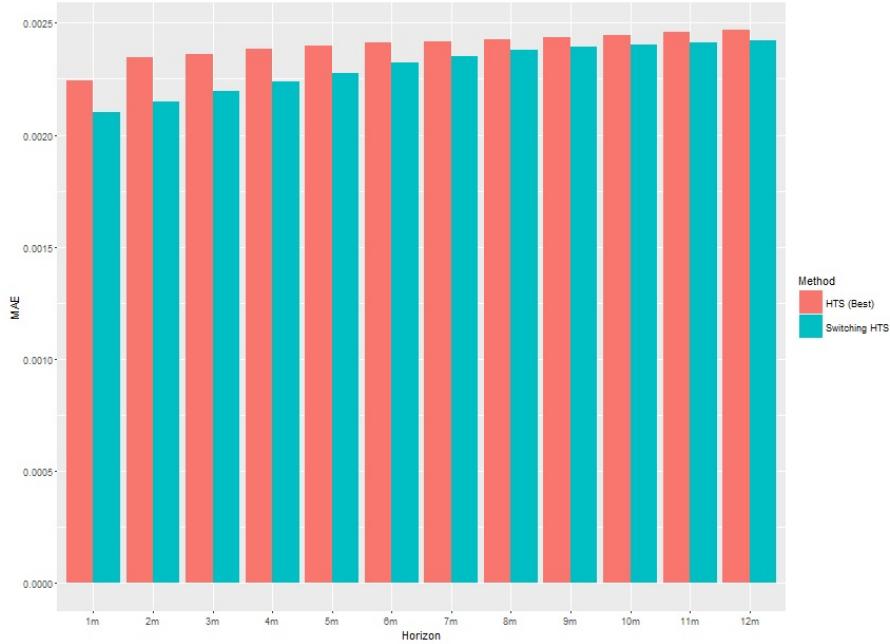
Fig. 2.7 Mean Model: MAEs of Switching-Rule vs. Best Single-Method HTS.

Figure 2.7 shows the MAEs of the switching-rule models in Table 2.3, comparing their accuracy to the best single-method HTS model. The switching-rule forecasts are between 1.8 % and 9.3 % more accurate compared to the best single-method HTS model, depending on horizon. For all horizons, the Diebold-Mariano test leads to the conclusion that the switching rule forecast significantly outperforms the best single-method HTS model, confirming the potential of the technique.

Variance Model − Results.   The fitted values of the mean model are used in the EWMA calculation of the base-level covariance matrix that is subsequently used to compute the parts of the VC and the CC – the inputs of the variance model. The decay parameter for EWMA, λ, was selected by the RMSE minimisation procedure described above, which returned a λ of 0.83 (Figure 2.8).

The hierarchical time series used as input for the variance models has 3 disaggregation levels (from lowest to highest): Level 0 – aggregate inflation variance, Level 1 – variance component (VC) and covariance component (CC), Level 2 – the 6 subparts each of VC and CC (Section 2.2.2).

Turning to the variance forecast results (Table 2.4), again based on a time series cross-validation, several aspects are noteworthy:
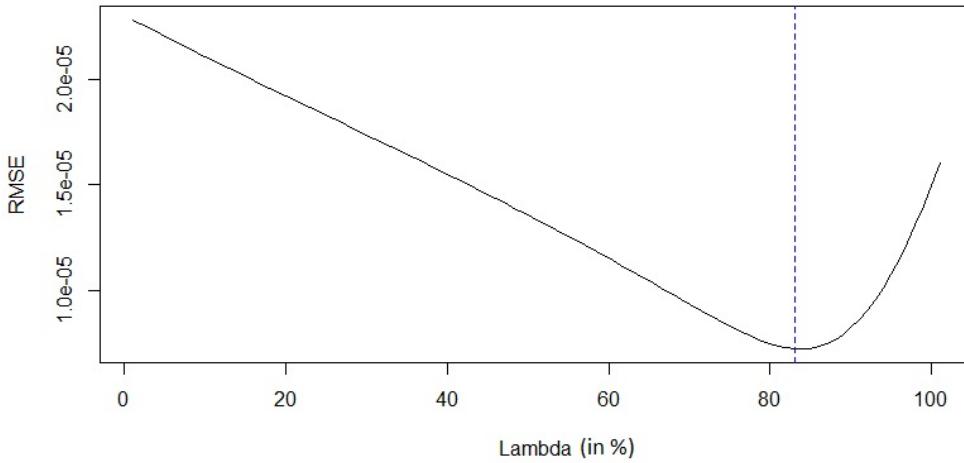
Fig. 2.8 RMSE-Optimisation of EWMA persistence parameter λ.

Table 2.4 Cross-Validated Test Set Accuracy - Variance Model.

|	1m	2m	3m	4m	5m	6m	7m	8m	9m	10m	11m	12m
Naïve												
Naïve	1.008	1.006	1.004	1.004	1.003	1.003	1.002	1.006	1.003	1.003	1.003	1.003
Top-Down												
ETS	1.350	1.365	1.388	1.519	1.621	1.694	1.783	1.872	1.993	2.206	2.504	2.899
ARIMA	1.306	1.386	1.417	1.427	1.426	1.422	1.407	1.402	1.396	1.410	1.434	1.459
Theta	1.075	1.065	1.054	1.048	1.041	1.036	1.032	1.034	1.030	1.031	1.033	1.034
Damped Trend	1.319	1.252	1.198	1.164	1.131	1.118	1.102	1.104	1.103	1.114	1.120	1.124
Dynamic Optimised Theta	1.071	1.062	1.052	1.046	1.039	1.034	1.030	1.031	1.027	1.028	1.029	1.029
Middle-Out												
ETS	1.244	1.241	1.248	1.265	1.272	1.278	1.283	1.299	1.310	1.333	1.358	1.385
ARIMA	1.108	1.038	1.029	1.020	1.019	1.004	1.000	**1.000**	1.001	1.012	1.025	1.036
								(4.21e-06)				
Theta	1.064	1.058	1.049	1.044	1.038	1.033	1.030	1.032	1.028	1.029	1.031	1.032
Damped Trend	1.231	1.195	1.174	1.166	1.157	1.158	1.157	1.161	1.163	1.179	1.193	1.205
Dynamic Optimised Theta	1.061	1.055	1.047	1.042	1.036	1.031	1.028	1.029	1.025	1.026	1.026	1.027
Bottom-Up												
ETS	1.273	1.209	1.116	1.071	1.029	1.004	1.002	1.017	1.025	1.033	1.033	1.031
ARIMA	1.141	1.254	1.257	1.240	1.235	1.212	1.194	1.187	1.182	1.192	1.206	1.225
Theta	1.002	1.002	1.001	1.001	1.001	1.001	1.001	1.006	1.002	1.003	1.003	1.004
Damped Trend	1.728	1.665	1.594	1.566	1.535	1.514	1.518	1.535	1.545	1.558	1.572	1.580
Dynamic Optimised Theta	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	1.004	**1.000**	**1.000**	**1.000**	**1.000**
	(1.84e-06)	(2.31e-06)	(2.77e-06)	(3.14e-06)	(3.46e-06)	(3.77e-06)	(4.03e-06)		(4.37e-06)	(4.43e-06)	(4.47e-06)	(4.50e-06)
Optimal Combination												
ETS	1.235	1.244	1.247	1.297	1.326	1.348	1.382	1.424	1.475	1.564	1.685	1.844
ARIMA	1.183	1.172	1.159	1.155	1.156	1.155	1.153	1.156	1.158	1.174	1.195	1.216
Theta	1.065	1.059	1.049	1.044	1.038	1.033	1.030	1.032	1.028	1.029	1.030	1.032
Damped Trend	1.194	1.154	1.118	1.107	1.095	1.097	1.095	1.103	1.106	1.122	1.133	1.140
Dynamic Optimised Theta	1.061	1.055	1.047	1.042	1.036	1.031	1.027	1.029	1.025	1.026	1.026	1.027

- The value of the disaggregated approach is understated in the variance forecasts. It should be noted that disaggregated models by far outperformed the aggregate models for the conditional mean of inflation and that these estimates were used as inflation expectation in the computation of the EWMA-smoothed bottom-level covariance matrix. Using inflation expectations of higher accuracy (compared to ones from aggregate models) improves the *input data* that the

variance forecasts are based on – this value is not incorporated in the forecast accuracy tables.

- For the variance forecast, the naïve model performs extremely well. This could be due to a number of different reasons:

  – Even though we used EWMA to compute the rolling-window covariances between series (rather than the equally-weighted SMA), the resulting smoothing of the input data removes some of the structure that might make more sophisticated models useful.

  – Over most of the observation period, the variance series is relatively stable at a low level – conditions that favour a naïve forecast, even more so given the lack of pronounced seasonality in the aggregate variance or its parts at the disaggregated levels. The only period that is characterised by high variance is the Great Recession, which caused a sudden spike in volatility that none of the models managed to capture very well;

  – The first part of this paper has established that since inflation targeting was introduced, VC (the part of aggregate variance due to bottom-level variances) and CC (the part of aggregate variance due to co-movement of the bottom-level units) are highly correlated (Correlation Coefficient = 84.94 %). This suggests that a middle-out approach cannot be expected to be of very high value for the variance forecast, as the 2 series do not add a lot of information compared to a forecast at the top level.

- The only models that can consistently beat the naïve forecast for all horizons are the Bottom-Up Theta model (except for 11 and 12-month forecasts) and the Bottom-Up Dynamic Optimised Theta model. This can be rationalised: First, the variance series are not dominated by seasonality as was the case for inflation rate itself. These are favourable conditions for Theta models. Second, while the Theta model – as special case of simple exponential smoothing with drift – is also a relatively simple forecasting model, it is designed to model the local curvature of the data (through the second Theta line). This can explain why these methods can actually beat the naïve forecast, which does not separately model long term and local trends. The Dynamic Optimised Theta model is designed to optimise the line for the local curvature more flexibly and it is unsurprising that the model outperforms (slightly) the original Theta model for all horizons. Third, the first part of this paper has shown that on the

bottom level (i.e. among the subparts of VC and CC), there is some variation –
while the VC is dominantly driven by the variance of the idiosyncratic part
and to some extent the variance of the industry part (two series again follow
similar patterns), the CC is driven by the variance of the common part and
the covariances of the idiosyncratic parts – two series that follow very different
patterns. This variation on the bottom-level, which was identified through the
two-stage decomposition, is information that improves forecast accuracy, even
if only slightly.[26]

- The good performance of the Middle Out ARIMA (which is among the best
  4 models for almost all horizons) seems like a curiosity at a first glance –
  ARIMA models do very poorly for all other aggregation approaches and the
  high correlation between the two series at the middle level also provides no
  explanation why this approach does relatively well. A more detailed analysis
  shows that this is due to Middle Out ARIMA being the best model in periods
  of high volatility (as presented in the switching model results). While we are
  able to identify its good performance at times of high volatility as reason for the
  good overall result, we have found no convincing explanation why the Middle
  Out ARIMA approach does so well in high-volatility phases – looking at the
  time points when volatility was above is 90 % quantile, we found that during
  these times, correlation between VC and CC were even higher with 92.69 %,
  which makes it very difficult to explain why this approach produces 30-40 %
  more accurate forecasts than the top-down ARIMA model.

Variance Model – Dynamic Model Switching. Table 2.5 presents the results
of the application of the switching rule.

Table 2.5 Switching Rule Variance Model: Results for Different Horizons.

	1m	2m	3m	4m	5m	6m
Best High Vol	Opt Comb DampedT	Opt Comb DampedT	Opt Comb DampedT	Opt Comb DampedT	Middle Out ARIMA	Middle Out ARIMA
Best Low Vol	BU DOTM	BU DOTM	Naive	Naive	Naive	Naive
Split Variance	1.64e-05	1.68e-05	2.78e-05	1.68e-05	2.78e-05	2.57e-05

	7m	8m	9m	10m	11m	12m
Best High Vol	Middle Out ARIMA	Middle Out ARIMA	Middle Out ARIMA	Middle Out ARIMA	Middle Out ARIMA	Middle Out ARIMA
Best Low Vol	Naive	BU ETS	Naive	Naive	Naive	Naive
Split Variance	2.78e-05	1.60e-05	2.78e-05	2.78e-05	2.78e-05	2.78e-05

---

[26]Diebold-Mariano tests for the best disaggregated model compared to a naïve forecast returned
p-values between 0.037 and 0.269 for the different horizons and therefore document that the forecast
performance of the disaggregated models is better than the naïve forecast, but only for two horizons
this improvement is significant at the 10 % level.
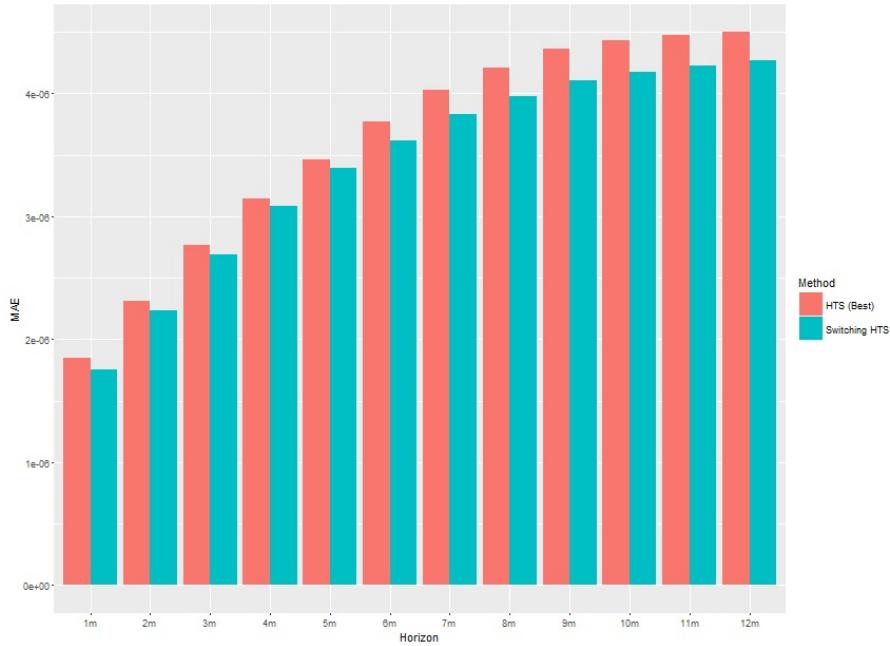
Fig. 2.9 Variance Model: Switching Rule vs. Best Single-Method HTS.

While overall the results coincide with the findings from the mean model section – simple models, such as naïve and Theta method tend to be best for low-volatility phases and more structured models, such as ARIMA are better suited for high-volatility phases – there are two notable features that are worth discussing: The Bottom-Up Dynamic Optimised Theta is only selected as the best model for short-term low-volatility forecasts – given that it is overall the best model for all horizons apart from one, this suggests that while not being much worse than a naïve model for longer-horizon forecasts in low-volatility phases, it produces more accurate forecasts than a naïve model in high-volatility phases; this can be ascribed to its ability to capture the local curvature of the series.

Another interesting aspect is that a model from the exponential smoothing family – Optimal Combination Damped Trend – is selected for short-horizon forecasts in high-volatility periods. However, this also shows the appeal of this switching-rule approach to some extent: Over the entire test set, the Optimal Combination Damped Trend model only ranks between 11 and 13 of the 21 candidate models even for 1- to 4-month forecasts due to its below-average forecasting performance in periods of low volatility – consequently, without the switching rule, we probably would not have identified this model's good ability to produce short-term forecasts in high-volatility phases. Figure 2.9 plots the improvement in MAE of the switching-rule models

compared to the best single-method HTS model for each horizon – the switching rule improves accuracy by between 1.8 % and 6.3 %, depending on horizon.

Validation of HTS Variance Models   For the purpose of validating the HTS variance models, we use GARCH-type models as benchmark. Estimating inflation uncertainty with GARCH-type models has a long history dating back to the seminal paper of Engle (1982) who first introduced the autoregressive conditional heteroskedasticity (ARCH) model, as well as Bollerslev (1986) who introduced the generalised ARCH model. Since then, GARCH-type models have become the most popular choice for characterising conditional inflation uncertainty of developed countries: e.g. Brunner and Hess (1993) for US CPI data; Joyce (1995), and Kontonikas (2004) for US inflation; and Grier and Perry (1998), and Fountas et al. (2000) for G7 countries.

Myriad extensions of the basic GARCH model have been employed to inflation modelling: A very popular model in the inflation literature is the Component GARCH (CGARCH) model by Engle and Lee (1993) that is equivalent to a restricted GARCH(2,2) model and decomposes conditional volatility into a permanent and a transitory component, thereby allowing for a time-varying long run volatility (e.g. used by Grier and Perry, 1998; Kontonikas, 2004). Common models in finance are the asymmetric volatility models Threshold GARCH (TGARCH), Exponential GARCH (EGARCH), and GJR-GARCH due to their ability to capture the "leverage effect" – the tendency of negative shocks to increase volatility by more than positive shocks of the same size. In the inflation context, previous empirical research has produced inconclusive evidence for asymmetric effects in conditional volatility: Grier and Perry (1998) employ a GJR-GARCH model and find no significant asymmetric effects for G7 countries. This coincides with e.g. Kontonikas (2004) who uses a GARCH-M model for monthly UK CPI data, and Shaikh and Salam (2014) using a GARCH model for inflation in Pakistan. In contrast, there are some studies that find supportive evidence for asymmetric conditional volatility in inflation rates: Moradi (2006) employs a TGARCH model for inflation in Iran, and Hossain (2014) uses an EGARCH model for Australia.

While the leverage effect is a widely accepted phenomenon in financial modelling, we do not find it surprising that there is a lack of evidence for it in inflation modelling. First, compared to financial markets, inflation targetting leads to a much more controlled environment in which inflation operates, leading to fewer extreme observations – so even if there was a leverage effect in inflation data, this exogenous

control of inflation levels and the low reporting frequency (leading to smaller sample sizes than with financial data) could make it difficult for models to pick up the effect. In particular large *negative* shocks seem rare in the inflation context for the UK. Second, and more importantly, there is no theoretical reason why a leverage effect should exist for inflation. In a stock context it is clear that a positive shock to a stock price is 'good news' and a negative shock is 'bad news'. Is this situation the same for inflation? Our view is that if the inflation level is close to the regulators' inflation target, only a stable inflation rate is 'good news', while large positive or negative shocks to the inflation rate are both 'bad news', thus making a significant asymmetric effect unlikely.

In order to identify the best model for the conditional mean for our RPI data, we use automated ARIMA selection and find that a $SARIMA(1,0,1)(1,0,1)_{12}$ model is the best fit. The ARCH-LM test rejected the null hypothesis of no ARCH effects (p-value: 0.25 %). ARCH/GARCH models are required.

Residual diagnostics revealed that the histogram of the residuals has a negative skew (Skewness: -0.896) and is leptokurtic (Kurtosis: 7.712). Previous research in the field documents that under these circumstances, the fit of GARCH-type models can be improved by using a leptokurtic error distribution – most commonly Student's t (proposed by Bollerslev, 1987) or GED (proposed by Nelson, 1991). Since the non-normality is primarily due to the largest outlier (a data point during the Great Recession), we also analysed normality without this outlier and find that without this data point, the histogram of the residuals approximates normality - the Jarque-Bera test statistic is 1.626 (p-value: 44.34 %). Given these results, we decided to estimate both GARCH-type models with Gaussian error distribution and ones with leptokurtic error distributions.

AIC-based selection returned a $SARIMA(1,0,1)(1,0,1)_{12}-GARCH(1,1)$ model with Gaussian error distribution as best fit for the inflation data (Table 2.6). We tried fitting asymmetric specifications (EGARCH, TGARCH), as well as Component GARCH due to its popularity in empirical inflation research, but none of the specifications were able to improve AIC compared to the GARCH(1,1). We also estimated the in-mean variants of all the mentioned models, but the in-mean effects were not significant. The AIC values of the estimated models are presented in Appendix 2.A.

Having found the optimal GARCH-type model for our data, we turned to the validation of our models comparing the best HTS variance model against the selected SARIMA-GARCH model. Thus, to test the quality of our disaggregated forecasting

Table 2.6 Best GARCH specification: $SARIMA(1,0,1)(1,0,1)_{12} - GARCH(1,1)$.

Variable	Coefficient	Std.Error	z-Statistic	p-value
AR(1)	0.657	0.182	3.615	0.0003
SAR(1)	0.983	0.008	125.442	0.0000
MA(1)	-0.420	0.226	-1.858	0.0632
SMA(1)	-0.887	0.027	-33.330	0.0000
Variance Equation				
C	1.60e-06	7.21e-07	2.221	0.0264
α (ARCH)	0.302	0.075	4.004	0.0001
β (GARCH)	0.395	0.175	2.261	0.0237

system for inflation mean and variance, we again apply time series cross-validation to compare the best disaggregated model's accuracy with the SARIMA-GARCH model. Given the previous optimisation of the decay parameter, we use the conditional variance from an EWMA model fitted to the full aggregate data as proxy for 'actual' inflation variance.

Table 2.7 Forecasting Performance: Single-Method HTS vs. Switching Rule vs. SARIMA-GARCH.

	Best Single HTS	Best Switching HTS	SARIMA-GARCH	Δ
1m	1.845e-06	1.754e-06	4.470e-06	2.55
3m	2.768e-06	2.687e-06	4.333e-06	1.61
6m	3.772e-06	3.615e-06	4.551e-06	1.26
9m	4.366e-06	4.104e-06	4.776e-06	1.16
12m	4.500e-06	4.271e-06	4.642e-06	1.09

Table 2.7 compares the MAE accuracy of the SARIMA-GARCH forecasts with those obtained from the best single-method HTS model and the best switching-rule model, respectively; Δ gives the relative advantage of the best switching-rule model compared to the SARIMA-GARCH. Accuracy is presented for selected horizons; the results are preserved in quality for the other horizons. Depending on horizon, the best switching-rule model forecasts 9 - 155 % more accurately than the SARIMA-GARCH model, while the biggest relative advantage of the HTS models is for short-horizon forecasts.

2.6   Discussion & Conclusions

We began by using a two-level decomposition of aggregate inflation volatility to shed light on the DGP of inflation volatility. Combining the variance/covariance

decomposition with the inflation rate decomposition into common, industry and idiosyncratic components adds to our understanding of the inflation process: The main drivers of the Variance Component (the part of aggregate volatility that is due to variances of product-level inflation rates) are the variances of idiosyncratic shocks. The main drivers of the Covariance Component (the part of aggregate volatility that is due to co-movement of product-level inflation rates) are the variances of the common shocks and the covariances of the product-level idiosyncratic shocks. The industry (product group) level plays only a subsidiary role.

Thus aggregate volatility is generated by a combination of common shocks (driving the covariation of product-level inflation rates) and idiosyncratic shocks (driving the product-level variances). Product-level analysis shows that episodes of high inflation volatility (generally in recessions) are often driven by a single or a few selected products, rather than a sweeping increase in product-level variances – the recession in the early 1990s was characterised by a shock to 'Council Tax & Rates', and covariation of other products with this item; the Great Recession was characterised by a shock to 'Mortgage Interest Payments', and covariation of other products with this item. Both high volatility phases also showed increased covariation of product-level inflation rates in general.

In the second part of the paper, we established the value of hierarchical time series modelling for forecasting the conditional mean and volatility of the aggregate inflation rate. For both the inflation rate and its volatility, the usefulness of explicitly considering the aggregation scheme underlying the RPI is evident from accuracy comparisons against conventional (aggregate) univariate modelling approaches. For the inflation rate (mean model), Middle Out Level 2 ARIMA produces the most accurate forecasts. For inflation volatility (variance model), the results respond to the call for an application of the recently introduced Dynamic Optimised Theta Model to a data set that is dominated by stationarity (by Fiorucci et al., 2016b). Bottom-Up DOTM produces the most accurate forecasts for inflation volatility. This can be attributed to the method's ability to capture local trends very well, and also to the heterogeneity of the Covariance Component's subparts – disaggregated level information that is ignored by aggregate models.

Finally, we presented an extreme variant of the forecast combination approach that appears to work well for inflation forecasting – this involves a dynamic switching-rule that, in a time-varying fashion, applies a combination weight of 1 to the best forecasting model based on an in-sample criterion (in our case the rolling sample variance), and a combination weight of 0 to all other models. While the predicted

inflation rate enters inflation volatility through the forecasting system in the standard way (the fitted mean model values are used to obtain the smoothed variance), the model switching approach lets inflation volatility affect the forecast of the inflation rate through the switching rule. This accommodates bi-directional relationship between inflation rate and its volatility. We find that the switching-rule approach can improve forecast accuracy compared to the best single-method HTS model – due to ARIMA models' superior performance in high-volatility episodes, and exponential smoothing-type models' superior performance in low-volatility episodes.

It is worth discussing some implications of our modelling approach in more detail: Given that our focus lies on forecasting aggregate inflation rate and volatility, some readers might ask why it is important to have aggregation consistent forecasts across the hierarchy. The reasons for aggregation consistency as a requirement of our modelling approach are both practical and statistical. First, in many cases applied inflation forecasting is interested in forecasts for the entire index, but simultaneously some of its components (for example, core inflation and non-core inflation), i.e. the hierarchical structure in itself is of interest in this context. Second, mean consistency implies that the variance of the aggregate series will be equal to the (weighted) sum of the variances and covariances of the bottom-level series. Since the mean forecasts directly feed into the variance forecasts through the variance identity, inconsistent forecasts (in the sense of aggregation consistency) would not be an accurate representation of the actual generating process underlying inflation volatility.

Another natural comment when talking about aggregation is the role played by the central limit theorem (CLT). While currently HTS forecasting is limited to point forecasts, probabilistic forecasting is a desirable extension of the approach in future research. Some first promising results for interval construction of HTS forecasts have been documented in a very recent paper by Taieb et al. (2017), who also take note of the challenges in creating probabilistic hierarchical forecasts: (a) the need to estimate the entire distribution of future observations, not only the mean (Kneib, 2013; Hothorn et al., 2014), (b) computing the distribution of hierarchical sums of random variables in high dimensions, and (c) accounting for the possible variety of distributions in the hierarchy; the distributions become more Gaussian with the aggregation level as a consequence of the CLT, while the series at lower levels often exhibit non-normality including multi-modality and high levels of skewness.

To summarise, this chapter has made several original contributions to research in the fields of hierarchical modelling and inflation forecasting:

- The additional detail of our variance decomposition compared to the original version of Comin and Mulani (2006) (a) helps with the identification of drivers of variance/covariance in different volatility regimes, and (b) can improve forecasts at the bottom level.[27]

- To the best of our knowledge this paper is the first that designs an aggregation consistent, weighted hierarchical time series, thereby providing a solution to the problem of combining index weighting with the HTS framework, which has the potential to turn HTS forecasting into a powerful forecasting tool in economics and finance.

- By applying HTS forecasting to an inflation context, we made a contribution in the ongoing debate about the optimal level of disaggregation (e.g. Hendry and Hubrich, 2006) – we show how (suboptimal) selection by the forecaster prior to estimation can be avoided and how this empirical question can be optimally resolved within the forecasting step itself instead.

- We proposed a dynamic switching model that accounts for the observed level of volatility in the selection of the optimal forecasting model and showed that this extreme case of forecast combination can lead to accuracy gains.

Overall, we find overwhelming support for the use of a hierarchical forecasting approach for UK inflation forecasting. The models can be extended and improved in several ways in future research. First, recent progress in Optimal Combination methods are designed to reconcile individual forecasts at the different levels of disaggregation to make them aggregate consistent – it is to be examined whether using more recently developed approaches (in particular Wickramasuriya et al., 2017) can further improve the disaggregated HTS forecasts and/or model switching based forecasts. Second, the UK has experienced major fluctuations in key macro variables over the last three decades. It should be useful to incorporate structural change in forecasting models – for example, by combining the HTS forecasting system with time-varying parameters inflation models – e.g. TVP-VAR models, TVP-FAVAR

---

[27]While it is true a hierarchical forecast of the *aggregate* does not use more information through our refined structure (industry and aggregate information is already part of the hierarchy), it is important to note that especially in an inflation context, forecasters are often interested in the inflation rate of components of the index. Assume for instance that an energy company wants to forecast the inflation rate of 'Oil and Other Fuels' – the original decomposition did not allow them to incorporate aggregate and industry information in the forecasting model in a straightforward way, while our refined structure does.

models, or regime-switching VAR models.[28] Our dynamic switching model is a first promising step on the way to the design of a hierarchical forecasting model that can incorporate both permanent and transitory shocks at different levels of aggregation. Third, the potential of incorporating explanatory variables in the hierarchical forecasting models should be explored – Barnett et al. (2012) find that models that include a large set of explanatory variables tend to do well for quarterly inflation forecasting. Aggregate explanatory variables can easily be incorporated in the HTS framework at least for a subset of forecasting methods. Fourth, despite the documented value of hierarchical forecasts and dynamic model switching, one challenge is that the estimation of the full set of hierarchical forecasts and of dynamic model switching is computationally expensive, making it difficult to apply in other practical contexts that require fast decision-making: For this purpose the results of this paper could be used to modify the recently proposed hierarchical selection algorithm by Nenova and May (2017). Finally, the issue of choosing a proxy for 'actual' inflation volatility remains open: one possible approach might be to use intra-month dispersion of product prices estimated using daily online prices as in the Billion Prices Project (Cavallo and Rigobon, 2016); another potential solution for this problem might be the extension of the recent probabilistic method proposed by Taieb et al. (2017) which could be modified to allow a joint estimation of mean and time-varying variance without the need to separately construct a proxy for volatility – similar to a GARCH approach, but incorporating the hierarchical information.

---

[28]Barnett et al. (2012) argue for the superiority of models with time-varying parameters primarily for the dataset that spans the inflation targeting, a clear structural break. See references therein for an overview of time-varying parameter methods.

2.7 References

Abadir, K., and G. Talmain. 2002. Aggregation, Persistence and Volatility in a Macro Model. *The Review of Economic Studies* 69:749–779.

Aiolfi, M., and A. Timmermann. 2006. Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics* 135:31–53.

Andersen, T. G., T. Bollerslev, and F. X. Diebold. 2010. Parametric and Non-parametric Volatility Measurement. In Y. Aït-Sahalia and L. P. Hansen (eds.), *Handbook of Financial Econometrics*, chap. 2, pp. 67–128. Amsterdam: Elsevier.

Arize, A. C. 2011. Are Inflation Rates Really Nonstationary? New Evidence from Non-Linear STAR Framework and African Data. *International Journal of Economics and Finance* 3:97–108.

Armstrong, J. S. 2001. Combining Forecasts. In J. S. Armstrong (ed.), *Principles of Forecasting*, pp. 417–439. Boston, MA: Springer.

Armstrong, J. S. 2006. Findings from Evidence-Based Forecasting: Methods for Reducing Forecast Error. *International Journal of Forecasting* 22:583–598.

Assimakopoulos, V., and K. Nikolopoulos. 2000. The Theta Model: A Decomposition Approach to Forecasting. *International Journal of Forecasting* 16:521–530.

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman. 2009. Hierarchical Forecasts for Australian Domestic Tourism. *International Journal of Forecasting* 25:146–166.

Athanasopoulos, G., R. J. Hyndman, H. Song, and D. C. Wu. 2011. The Tourism Forecasting Competition. *International Journal of Forecasting* 27:822–844.

Atkeson, A., and L. E. Ohanian. 2001. Are Phillips Curves Useful for Forecasting Inflation? *Federal Reserve Bank of Minneapolis Quarterly Review* 25:2–11.

Bagdatoglou, G., A. Kontonikas, and M. E. Wohar. 2016. Forecasting US Inflation Using Dynamic General-to-Specific Model Selection. *Bulletin of Economic Research* 68:151–167.

Ball, L., and S. G. Cecchetti. 1990. Inflation and Uncertainty at Short and Long Horizons. *Brookings Papers on Economic Activity* 1/1990:215–254.

Bandi, F. M., and J. R. Russell. 2006. Separating Microstructure Noise from Volatility. *Journal of Financial Economics* 79:655–692.

Barnett, A., H. Mumtaz, and K. Theodoridis. 2012. Forecasting UK GDP Growth, Inflation and Interest Rates Under Structural Change: A Comparison of Models with Time-Varying Parameters. Working Paper No. 450, Bank of England.

Bean, C. 2003. Inflation Targeting: The UK Experience. *Bank of England Quarterly Bulletin, Winter 2003* pp. 479–494.

Beeson, J. 2016. Consumer Price Inflation: 2016 Weights. Technical Report (22 March 2016), Office for National Statistics.

Belmonte, M., and G. Koop. 2013. Model Switching and Model Averaging in Time-Varying Parameter Regression Models. Working Paper No. 1302, University of Strathclyde Business School, Department of Economics.

Bergmeir, C., R. J. Hyndman, and B. Koo. 2015. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. Working Paper No. 10/15, Monash University, Department of Econometrics and Business Statistics.

Bollerslev, T. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31:307–327.

Bollerslev, T. 1987. A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics* 69:542–47.

Box, G. E. P., and G. M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control.* San Francisco, Calif.: Holden-Day Inc.

Brunner, A. D., and G. D. Hess. 1993. Are Higher Levels of Inflation Less Predictable? A State-Dependent Conditional Heteroscedasticity Approach. *Journal of Business and Economic Statistics* 11:187–197.

Buncic, D., and C. Moretto. 2015. Forecasting Copper Prices with Dynamic Averaging and Selection Models. *North American Journal of Economics and Finance* 33:1–38.

Calvo, S., and C. Reinhart. 1996. Capital Flows to Latin America : Is There Evidence of Contagion Effects? Policy Research Working Paper Series 1619, The World Bank.

Capistrán, C., C. Constandse, and M. Ramos-Francia. 2010. Multi-Horizon Inflation Forecasts Using Disaggregated Data. *Economic Modelling* 27:666–677.

Carvalho, V., and X. Gabaix. 2013. The Great Diversification and Its Undoing. *American Economic Review* 103:1697–1727.

Cavallo, A., and R. Rigobon. 2016. The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives* 30:151–178.

Chen, H., and J. E. Boylan. 2007. Use of Individual and Group Seasonal Indices in Subaggregate Demand Forecasting. *Journal of the Operational Research Society* 58:1660–1671.

Chen, H., and J. E. Boylan. 2009. The Effect of Correlation Between Demands on Hierarchical Forecasting. *Advances in Business and Management Forecasting* 6:173–188.

Clemen, R. T. 1989. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting* 5:559–583.

Collins, D. W. 1976. Predicting Earnings with Sub-Entity Data: Some Further Evidence. *Journal of Accounting Research* 14:163–177.

Comin, D., and S. Mulani. 2004. Diverging Trends in Macro and Micro Volatility: Facts. Working Paper No. 10922, National Bureau of Economic Research.

Comin, D., and S. Mulani. 2006. Diverging Trends in Aggregate and Firm Volatility. *Review of Economics and Statistics* 88:374–383.

Cukierman, A., and A. H. Meltzer. 1986. A Theory of Ambiguity, Credibility, and Inflation under Discretion and Asymmetric Information. *Econometrica* 54:1099–1128.

Cukierman, A., and P. Wachtel. 1979. Differential Inflationary Expectations and the Variability of the Rate of Inflation: Theory and Evidence. *The American Economic Review* 69:595–609.

Culver, S. E., and D. H. Papell. 1997. Is There a Unit Root in the Inflation Rate? Evidence from Sequential Break and Panel Data Models. *Journal of Applied Econometrics* 12:435–444.

Dangerfield, B. J., and J. S. Morris. 1992. Top-Down or Bottom-Up: Aggregate Versus Disaggregate Extrapolations. *International Journal of Forecasting* 8:233–241.

Davis, S. J., J. Haltiwanger, R. Jarmin, and J. Miranda. 2006. Volatility and Dispersion in Business Growth Rates: Publicly Traded Versus Privately Held Firms. *NBER Macroeconomics Annual* 21:107–156.

Di Fonzo, T., and M. Marini. 2011. Simultaneous and Two-Step Reconciliation of Systems of Time Series: Methodological and Practical Issues. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 60:143–164.

Dornbusch, R., Y. C. Park, and S. Claessens. 2000. Contagion: Understanding How It Spreads. *World Bank Research Observer* 15:177–97.

Dunn, D. M., W. H. Williams, and T. L. Dechaine. 1976. Aggregate Versus Sub-aggregate Models in Local Area Forecasting. *Journal of the American Statistical Association* 71:68–71.

Engle, R. F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50:987–1007.

Engle, R. F., and G. G. J. Lee. 1993. A Permanent and Transitory Component Model of Stock Return Volatility. Discussion Paper No. 92-44R, Department of Economics, UCSD.

Evans, M. 1991. Discovering the Link Between Inflation Rates and Inflation Uncertainty. *Journal of Money, Credit and Banking* 23:169–184.

Faust, J., and J. H. Wright. 2013. Forecasting inflation. *Handbook of Economic Forecasting* 2:1–56.

Fildes, R. A., and J. K. Ord. 2002. Forecasting Competitions – Their Role in Improving Forecasting Practice and Research. In M. P. Clements and D. F. Hendry (eds.), *Companion to Economic Forecasting*, pp. 322–353. Oxford: Blackwell.

Fiorucci, J. A., F. Louzada, and B. Yiqi. 2016a. *forecTheta: Forecasting Time Series by Theta Models*. URL http://CRAN.R-project.org/package=forecTheta. [Accessed on July 15, 2017], R package version 2.2.

Fiorucci, J. A., T. R. Pellegrini, F. Louzada, F. Petropoulos, and A. B. Koehler. 2016b. Models for Optimising the Theta Method and Their Relationship to State Space Models. *International Journal of Forecasting* 32:1151–1161.

Fischer, S. 1981. Towards an Understanding of the Costs of Inflation: II. *Carnegie-Rochester Conference Series on Public Policy* 15:5 – 41.

Fliedner, G. 1999. An Investigation of Aggregate Variable Time Series Forecast Strategies with Specific Subaggregate Time Series Statistical Correlation. *Computers and Operations Research* 26:1133–1149.

Fountas, S., M. Karanasos, and M. Karanassou. 2000. A GARCH Model of Inflation and Inflation Uncertainty with Simultaneous Feedback. Discussion Paper 2000/24, Department of Economics, University of York.

Friedman, M. 1977. Nobel Lecture: Inflation and Unemployment. *Journal of Political Economy* 85:451–472.

Gabaix, X. 2011. The Granular Origins of Aggregate Fluctuations. *Econometrica* 79:733–772.

Galati, G., S. Poelhekke, and C. Zhoua. 2011. Did the Crisis Affect Inflation Expectations? *International Journal of Central Banking* 7:167–207.

Gardner, E. S. 1985. Exponential Smoothing: The State of the Art. *Journal of Forecasting* 4:1–28.

Gardner, E. S., and E. McKenzie. 1985. Forecasting Trends in Time Series. *Management Science* 31:1237–1246.

Granger, C. W. J., and R. Ramanathan. 1984. Improved Methods of Combining Forecasts. *Journal of Forecasting* 3:197–204.

Grier, K. B., and M. J. Perry. 1998. On Inflation and Inflation Uncertainty in the G7 Countries. *Journal of International Money and Finance* 17:671–689.

Gross, C. W., and J. E. Sohl. 1990. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting* 9:233–254.

Guo, H. 2012. Estimating Volatilities by the GARCH and the EWMA Model of PetroChina and TCL in the Stock Exchange Market of China. In *Proceedings of the 6th International Scientific Conference Managing and Modelling of Financial Risks*, pp. 191–202. Ostrava.

Hendry, D. F., and K. Hubrich. 2006. Forecasting Economic Aggregates by Dis-aggregates. ECB Working Paper Series No. 589 (Feb 2006), European Central Bank.

Holland, A. S. 1993. Uncertain Effects of Money and the Link Between the Inflation Rate and Inflation Uncertainty. *Economic Inquiry* 31:39–51.

Holland, A. S. 1995. Inflation and Uncertainty: Tests for Temporal Ordering. *Journal of Money, Credit and Banking* 27:827–837.

Hossain, A. A. 2014. Inflation and Inflation Volatility in Australia. *Economic Papers: A Journal of Applied Economics and Policy* 33:163–185.

Hothorn, T., T. Kneib, and P. Bühlmann. 2014. Conditional Transformation Models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76:3–27.

Hsiao, C., and S. K. Wan. 2014. Is There an Optimal Forecast Combination? *Journal of Econometrics* 178:294–309.

Hyndman, R., and G. Athanasopoulos. 2014. *Forecasting: Principles and Practice.* OTexts.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. 2011. Optimal Combination Forecasts for Hierarchical Time Series. *Computational Statistics and Data Analysis* 55:2579–2589.

Hyndman, R. J., and B. Billah. 2003. Unmasking the Theta Method. *International Journal of Forecasting* 19:287–290.

Hyndman, R. J., and Y. Khandakar. 2008. Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software* 27:1–22.

Hyndman, R. J., and A. B. Koehler. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22:679–688.

Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder. 2008. *Forecasting with Exponential Smoothing: The State Space Approach.* Berlin, Heidelberg: Springer.

Hyndman, R. J., A. B. Koehler, R. D. Snyder, and S. Grose. 2002. A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods. *International Journal of Forecasting* 18:439–454.

Hyndman, R. J., A. J. Lee, and E. Wang. 2016a. Fast Computation of Reconciled Forecasts for Hierarchical and Grouped Time Series. *Computational Statistics and Data Analysis* 97:16–32.

Hyndman, R. J., E. Wang, A. Lee, and S. Wickramasuriya. 2016b. *hts: Hierarchical and Grouped Time Series.* URL http://CRAN.R-project.org/package=hts. [Accessed on July 15, 2017], R package version 5.0.

Jose, V. R. R., and R. L. Winkler. 2008. Simple Robust Averages of Forecasts: Some Empirical Results. *International Journal of Forecasting* 24:163–169.

Joyce, M. 1995. Modelling UK Inflation Uncertainty: The Impact of News and the Relationship with Inflation. Working Paper 30, Bank of England.

Kahn, K. B. 1998. Revisiting Top-Down Versus Bottom-Up Forecasting. *The Journal of Business Forecasting* 17:14–19.

Kim, D.-H., and S.-C. Lin. 2012. Inflation and Inflation Volatility Revisited. *International Finance* 15:327–345.

King, R. G., C. I. Plosser, J. H. Stock, and M. W. Watson. 1991. Stochastic Trends and Economic Fluctuations. *American Economic Review* 81:819–840.

Kneib, T. 2013. Beyond Mean Regression. *Statistical Modelling* 13:275–303.

Kohn, R. 1982. When is an Aggregate of a Time Series Efficiently Forecast by Its Past? *Journal of Econometrics* 18:337–349.

Kontonikas, A. 2004. Inflation and Inflation Uncertainty in the United Kingdom: Evidence from GARCH Modelling. *Economic Modelling* 21:525–543.

Lee, J. 2005. Estimating Memory Parameter in the US Inflation Rate. *Economics Letters* 87:207–210.

Lee, J., C. J. Huang, and Y. Shin. 1997. On Stationary Tests in the Presence of Structural Breaks. *Economics Letters* 55:165 – 172.

Lucas, A., and X. Zhang. 2016. Score-Driven Exponentially Weighted Moving Averages and Value-at-Risk Forecasting. *International Journal of Forecasting* 32:293–302.

Makridakis, S., and M. Hibon. 2000. The M3-Competition: Results, Conclusions and Implications. *International Journal of Forecasting* 16:451–476.

McMillan, D. G. 2014. Forecasting Stock Returns: Does Switching Between Models Help? In J. Ma and M. Wohar (eds.), *Recent Advances in Estimating Nonlinear Models: With Applications in Economics and Finance*, pp. 229–248. New York: Springer.

Moradi, M. A. 2006. A GARCH Model of Inflation and Inflation Uncertainty in Iran. *Quarterly Journal of the Economic Research* 6:122–145.

Nelson, C. R., and C. R. Plosser. 1982. Trends and Random Walks in Macroeconmic Time Series: Some Evidence and Implications. *Journal of Monetary Economics* 10:139 – 162.

Nelson, D. B. 1991. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* 59:347–370.

Nenova, Z. D., and J. H. May. 2017. Determining an Optimal Hierarchical Forecasting Model Based on the Characteristics of the Data Set: Technical Note. *Journal of Operations Management* 44:62–68.

Nikolopoulos, K., D. Thomakos, F. Petropoulos, A. Litsa, and V. Assimakopoulos. 2012. Forecasting S&P 500 with the Theta Model. *International Journal of Financial Economics and Econometrics* 4:73–78.

Noriega, A., C. Capistrán, and M. Ramos-Francia. 2013. On the Dynamics of Inflation Persistence Around the World. *Empirical Economics* 44:1243–1265.

Noriega, A. E., and E. de Alba. 2001. Stationarity and Structural Breaks — Evidence from Classical and Bayesian Approaches. *Economic Modelling* 18:503 – 524.

Office for National Statistics. 2013. Assessment of Compliance with the Code of Practice for Official Statistics. UK Statistics Authority Assessment Report 246 (March 2013).

Orcutt, G. H., H. W. Watts, and J. B. Edwards. 1968. Data Aggregation and Information Loss. *The American Economic Review* 58:773–787.

Pegels, C. C. 1969. Exponential Forecasting: Some New Variations. *Management Science* 15:311–315.

Quah, D. 1994. One Business Cycle and One Trend From (Many,) Many Disaggregates. *European Economic Review* 38:605–614.

Ridge, M., and S. Smith. 1991. Local Taxation: The Options and the Arguments. IFS Report Series No. 38, Institute for Fiscal Studies.

Russell, B., and A. Banerjee. 2008. The Long-Run Phillips Curve and Non-Stationary Inflation. *Journal of Macroeconomics* 30:1792–1815.

Shaikh, E. A., and R. B. M. Salam. 2014. Inflation Volatility Using GARCH-Family Models: Empirical Evidence from Pakistan. *Journal of Contemporary Issues in Business Research* 3:168–173.

Shlifer, E., and R. W. Wolff. 1979. Aggregation and Proration in Forecasting. *Management Science* 25:594–603.

Stock, J. H., and M. W. Watson. 2004. Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* 23:405–430.

Stock, J. H., and M. W. Watson. 2008. Phillips Curve Inflation Forecasts. NBER Working Paper No. 14322, National Bureau of Economic Research.

Taieb, S. B., J. W. Taylor, and R. J. Hyndman. 2017. Coherent Probabilistic Forecasts for Hierarchical Time Series. *Proceedings of the 34th International Conference on Machine Learning (ICML)* .

Taylor, J. W. 2003. Exponential Smoothing with a Damped Multiplicative Trend. *International Journal of Forecasting* 19:715–725.

Thomakos, D., and K. Nikolopoulos. 2014. Fathoming the Theta Method for a Unit Root Process. *IMA Journal of Management Mathematics* 25:105–124.

Tiao, G. C., and I. Guttman. 1980. Forecasting Contemporal Aggregates of Multiple Time Series. *Journal of Econometrics* 12:219–230.

Van Erven, T., and J. Cugliari. 2015. Game-Theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts. In A. Antoniadis, J.-M. Poggi, and X. Brossat (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics Vol. 217, pp. 297–317. Paris: Springer.

Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman. 2017. Optimal Forecast Reconciliation of Hierarchical and Grouped Time Series Through Trace Minimization. Working paper, Department of Econometrics and Business Statistics, Monash Business School.

Zellner, A., and J. Tobias. 2000. A Note on Aggregation, Disaggregation and Forecasting Performance. *Journal of Forecasting* 19:457–465.

Appendix

2.A   AIC Selection of Candidate GARCH-Type Models

Table 2.A1 shows AIC values of candidate GARCH-type models. In-mean variants were also estimated, but are not presented due to the insignificance of the in-mean effects in the output equations.

Table 2.A1 AIC Selection of GARCH-type Models.

Model	AIC
Assuming Constant Variance of Residuals	
$SARIMA(1,0,1)(1,0,1)_{12}$	-9.082
Adjusting for Non-Constant Variance of Residuals	
$SARIMA(1,0,1)(1,0,1)_{12}$-**GARCH(1,1)**	
Gaussian Error Distribution	-9.388
Student's t Error Distribution	-9.384
Generalised Error Distribution	-9.385
$SARIMA(1,0,1)(1,0,1)_{12}$-**EGARCH(1,1)**	
Gaussian Error Distribution	-9.378
Student's t Error Distribution	-9.374
Generalised Error Distribution	-9.375
$SARIMA(1,0,1)(1,0,1)_{12}$-**CGARCH(1,1)**	
Gaussian Error Distribution	-9.380
Student's t Error Distribution	-9.375
Generalised Error Distribution	-9.376

Note that the EGARCH specification has not led to improvements over the original GARCH specification for our sample. No evidence of a significant leverage effect.

2.B   Stationarity of Inflation Rates

It is worth discussing that none of the automated forecasting methods in the main chapter require prior transformation to a stationary time series – ARIMA takes the required number of differences within the estimation procedure, and exponential

smoothing-based methods do not assume stationarity in the first place. In fact, every ETS model is non-stationary (Hyndman and Athanasopoulos, 2014).

The use of such methods is appealing in general, but in particular in the inflation context, due to the large stock of literature on the stochastic nonstationarity of inflation rates that was sparked by the seminal paper of Nelson and Plosser (1982) who claim that most macroeconomic time series have a stochastic trend. Empirical studies that support this claim include: King et al. (1991), Lee (2005), Russell and Banerjee (2008), and Arize (2011). This stochastic nonstationarity implies that a shock to inflation has a permanent effect due to the presence of unit roots.

The nonstationary behaviour of inflation rates is not undisputed. Some authors find that inflation rates satisfy stationary – for instance Culver and Papell (1997), and Noriega et al. (2013) – which would mean that shocks to inflation are of transitory nature and die off over time.

Testing for stationarity using our sample led to interesting results. We applied both the KPSS test (a common test for stationarity) and the Augmented Dickey Fuller test (a common test for the presence of a unit root). The test statistics are presented in Table 2.B1 (p-values in brackets).

Table 2.B1 Results of Stationarity Test for UK RPI.

Test	Full Sample	Reduced Sample
ADF	-5.90 ($< 1\%$)	-6.34 ($< 1\%$)
KPSS	0.71 (1.26%)	0.07 ($> 10\%$)

While both tests indicated stationarity for the reduced sample that we used for forecasting (discarding data before inflation targeting was introduced), the tests give conflicting results when testing for the full sample: The KPSS test rejected stationarity at the 5% level, but the ADF test rejected the presence of a unit root even at the 1% level. This adds potential insight to the explanation why past research has not come to a consensus view on the stationarity of inflation rates – structural breaks can severely affect inference depending on what kind of test is used (e.g. Lee et al., 1997; Noriega and de Alba, 2001). This constitutes an interesting topic for future research, but exceeds the scope of our main chapter; for us, it is important that issues linked to stationarity do not jeopardise the validity of our forecasting results, which is satisfied as (a) none of our forecast methods require stationarity, and (b) we work with a reduced subsample that does not show any evidence of non-stationarity. Designing a hierarchical structural break model for the inflation rate is a topic for future research in the field; our switching model in the main chapter

is a promising first step in this direction that allows for temporary changes in the volatility regime and should be used as starting point for the design of a hierarchical inflation forecasting model that can accommodate both transitory and permanent shocks at different levels of aggregation.

2.C   Directionality of Causation

The main chapter motivated the design of a joint forecasting system of inflation rate and volatility with previous research in the field on the relationship between these two variables. While there is a heated debate on the directionality of causation between inflation rate and inflation volatility, it is accepted that both are strongly interlinked, which clearly makes separate models for these variables suboptimal.

The purpose of this chapter was the design of such a joint forecasting system, rather than solving the feud between the "Friedman-Ball" and "Cukierman-Meltzer" hypotheses. Nevertheless, it is worth discussing how our modelling approach and results are linked to this strand of research: Our first approach is by and large based on the "Friedman-Ball" hypothesis – we forecast inflation rates at the different aggregation levels and use these forecasts to inform the volatility model; this model does not take into consideration the possibility that inflation volatility can also have an impact on inflation rates. Hence, this first model is a step-wise model, rather than a joint forecasting system. Our second approach, the dynamic switching model, resolves this by allowing inflation volatility to affect the inflation rate forecast (through the selection of a forecasting model for the mean).

The finding that incorporating the mean forecasts into the variance model can improve accuracy suggests that inflation rate causes inflation volatility to some extent ("Friedman-Ball" hypothesis). The finding that the dynamic switching model does not lead to only one 'best' forecasting model irrespective of volatility suggests that volatility has a causal effect on inflation rate ("Cukierman-Meltzer" hypothesis). Our results therefore support the findings of Kim and Lin (2012) of a bi-directional causal relationship between these two variables, a feedback mechanism.

The hypothesis of a bi-directional causality between inflation and inflation volatility is further supported by a causality test for our sample: We first estimated a VAR(12) model[29], and subsequently performed Granger causality tests. The null hypothesis of this test is that Variable 1 does not Granger-cause Variable 2, so a

---

[29]A lag order of 12 was indicated by optimising AIC.

significant p-value indicates causality – for both directions, the p-value was significant at the 1 % level (Table 2.C1).

Table 2.C1 Results of Granger Causality Test for inflation rate and volatility.

Direction of Causation	Test Statistic	p-value
Infl. Rate → Infl. Vol.	4.97	8.23×10^{-8}
Infl. Vol. → Infl. Rate	2.35	0.62%

Chapter 3

Forecast Combination in R Using the GeomComb Package

3.1 Introduction

In the business of time series forecasting, accuracy is key. The advent of "black-box" statistical learning methods (such as *Artificial Neural Networks*) testifies to the fact that forecasting practitioners and researchers alike tend to favour accuracy over explainability of a forecast model. There is a strong consensus in the social sciences that the observed processes are too complex to be modelled perfectly. Excluding some natural sciences, it is generally undisputed that the underlying data-generating process (DGP) is unknown. Hence, statistical models are habitually too simple, mis-specified, and/or incomplete; a fact that is widely accepted in theory, but less widely applied in practice, in the field of econometrics, in which researchers often still hang on to the conceptual error of assuming one true DGP and putting too much focus on model selection to find the one true model. Hansen (2005) takes note of this and related misconceptions of econometric forecasting practice in his essay on the challenges for econometric model selection: "models should be viewed as approximations, and econometric theory should take this seriously".

Following the advice in Hansen (2005), we abandon the quest for the one *true* correctly specified model, so that we are free to include information from different models. In the context of forecasting, this means combining forecasts from different sources (models, experts). This emphasises the aforementioned shift in approaching practical econometric forecasting problems: model mis-specification cannot be fully rectified, but strategies can be found to mitigate its adverse effects on forecast quality.

Selecting a single "best" forecasting model bears the risk of ending up with a model which is only accurate when evaluated using some validation sample, yet might prove unreliable when applied to new data. In the past decades, ample empirical evidence on the merits of combining forecasts has piled up; it is generally accepted that the (mostly linear) combination of forecasts from different models is an appealing strategy to hedge against forecast risk. Even though reduction of forecast risk is the main argument for using combined forecasts, it should also be noted that there are cases when combined forecasts are more accurate than even its best component (e.g. Graefe et al., 2014). While this is not theoretically founded, it is somewhat intuitive that in a continuously changing environment different forecasting models deliver different results at different time periods. For example, it is reasonable for one method to perform well in a low-volatility regime, and for another method to perform poorly in that regime, but perform well in a high-volatility regime. Strong empirical support for this argument of change in relative performance of different methods over time is found among others by Elliott and Timmermann (2005).

"Hedging" against model risk by combining different forecasts is intuitive and appealing. Joined with accuracy gains, the idea is widely adopted in macroeconomics and finance, with application abound. Magnus and Wang (2014) explore growth determinants, Stock and Watson (2004) use forecast combination for output forecasting, Wright (2009) and Kapetanios et al. (2008) for inflation forecasting, Avramov (2002), Rapach et al. (2010) and Ravazzolo et al. (2007) for forecasting stock returns, Wright (2008) for exchange rate forecasts, Andrawis et al. (2011) for forecasting inbound tourism figures, Nowotarski et al. (2014) and Raviv et al. (2015) for electricity price forecasting, and Weiss (2017) for healthcare demand forecasting. More recently, forecast combinations have been applied not only in "first moment" applications, but for higher moments as well. For example in Christiansen et al. (2012) for volatility forecasting, and Opschoor et al. (2014) for Value-at-Risk forecasting.

The theoretical foundations of forecast combination date back five decades to the seminal papers of Crane and Crotty (1967) and Bates and Granger (1969). Yet even in the recent past, papers discussing new combination techniques are published in reputable journals and stimulate further research still (Hansen, 2007, 2008; Hansen and Racine, 2012; Elliott et al., 2013; Morana, 2015; Cheng and Yang, 2015). Two extensive reviews of the literature, techniques and applications of forecast combinations are Clemen (1989) and Timmermann (2006).

Given the topic's popularity in both theoretical and empirical research, it is somewhat surprising that very few combination techniques are readily available in R:

There are some packages that cover specialised specific topics related to combination methods, e.g. the **BMA** package by Raftery et al. (2017) for Bayesian model averaging, as well as the **opera** package by Gaillard and Goude (2016) and the **forecastHybrid** package by Shaub and Ellis (2017). However, as of now the only two R packages that are entirely devoted to forecast combination are the **ForecastCombinations** package by Raviv (2015), which focuses on regression-based combination methods, and the **GeomComb** package by Weiss and Roetzer (2016), which focuses on geometric, eigenvector-based combination methods. With the aim of improving user experience we have merged these two packages, providing one unified package for the widest range of forecast combination approaches available today in R. The package is flexible and provides enough guidance for users familiar or unfamiliar with the world of forecasting. We have made both regression-based and eigenvector-based combination methods available to users in a single standardised framework based on S3 classes and methods. The logic behind this choice is that comparing regression-based and eigenvector-based combination methods is often insightful – as pointed out by Hsiao and Wan (2014), the conditions under which these two approaches perform well differ from each other: regression-based methods tend to produce more accurate forecasts when one or a few of the individual forecasts are considerably better than the rest, while eigenvector-based methods perform better when the individual forecasts are in the same ballpark. This paper presents the functionalities made available in the package and demonstrates how to implement them in an empirical exercise.

The **GeomComb** package was created in version 3.2.5 of R, and imports the following packages: **forecast**, **ggplot2**, **Matrix**, **mtsdi**, **psych**, **quantreg**, **quadprog**, and **utils**. It is available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/web/packages/GeomComb/index.html.

The remainder of this paper is structured as follows. Section 3.2 reviews the forecast combination methods that are available in the **GeomComb** package, Section 3.3 provides a detailed implementation description using the package. Section 3.4 presents an empirical example – we combine univariate time series forecasts for UK energy supply. Section 3.5 concludes.

3.2   Forecast combinations

To fix notations, denote $F_{T \times P}$ as the matrix of forecast with dimension $T \times P$ where T is the number of rows and P is the number of columns (so we have P forecasts at each point in time). Denote f_i as the forecast obtained using model i, $i \in \{1 \dots P\}$.

When there is no danger of confusion, we omit the additional subscript t which denotes the time dimension of the forecast. Some combination methods require an ordering of the component forecasts. When this is the case, $f_{(i)}$ denotes the ith order statistic of the cross-section of component forecasts. Finally, the weight given to that forecast in the overall combined forecast is denoted as w_i, and the combined forecast as f^c.

3.2.1 Frequently used schemes for forecast combinations

Simple Combination Methods

1. *Simple Average.* The most intuitive approach to combine forecasts is using the average of all those forecasts. Over the years this innocent approach has established itself as an excellent benchmark, despite or perhaps because of its simplicity (e.g. Genre et al., 2013). The combined forecast is straightforwardly given by

$$f^c = \frac{1}{P} \sum_{i=1}^{P} f_i. \tag{3.1}$$

Clemen (1989) argues that this equal weighting of component forecasts is often the best strategy in this context. This is still true almost thirty years later and called the "forecast combination puzzle", a term coined by Stock and Watson (2004). A rigorous attempt to explain why simple average weights often outperform more sophisticated forecast combination techniques is provided in a simulation study by Smith and Wallis (2009), who ascribe this surprising empirical finding to the effect of finite-sample error in estimating the combination weights. Recently, Claeskens et al. (2016) provide a theoretical argumentation to these empirical findings. The authors make the fine case that lower estimation noise, when the weights are determined rather than estimated, goes a long way in explaining the puzzle. A more detailed overview of the empirical support for the "forecast combination puzzle" can be found in Graefe et al. (2014).

2. *Median.* Another fairly simple and appealing combination method is using the median of the component forecasts. The median is insensitive to outliers, which can be relevant for some applications. Palm and Zellner (1992) suggest

that simple averaging may not be a suitable combination method when some of the component forecasts are biased. This calls for the use of another location measures which is robust to outliers. The median method is an appealing, rank-based combination method that has been used in a wide range of empirical studies (e.g. Armstrong, 1989; McNees, 1992; Hendry and Clements, 2004; Stock and Watson, 2004; Timmermann, 2006).

For the median method, the combined forecast is given by:

- For odd P:

$$f^c = f_{(\frac{P}{2}+0.5)} \qquad (3.2)$$

- For even P:

$$f^c = \frac{1}{2}\left(f_{(\frac{P}{2})} + f_{(\frac{P}{2}+1)}\right) \qquad (3.3)$$

3. *Trimmed Mean.* Another outlier-robust location measure that is commonly used is the trimmed mean (e.g. Armstrong, 2001; Stock and Watson, 2004; Jose and Winkler, 2008).

   Using a trim factor λ (i.e. the top/bottom $100 \times \lambda\%$ are trimmed) the combined forecast is calculated as:

$$f^c = \frac{1}{P(1-2\lambda)} \sum_{i=\lambda P+1}^{(1-\lambda)P} f_{(i)} \qquad (3.4)$$

   Typically, we use $\lambda = 0.1$ indicating we trim the top and bottom 10% of the most extreme component forecasts, excluding those from the computation of the combined forecast. The trimmed mean is an interpolation between the simple average ($\lambda = 0$) and the median ($\lambda = 0.5$).

4. *Winsorised Mean.* Like the trimmed mean, the winsorised mean is a robust statistic that is less sensitive to outliers than the simple average. It takes a softer line when handling outliers: Instead of altogether removing them as in the trimmed mean approach, it caps outliers at a certain level. By capping outliers

rather than removing them, we allow for at least some degree of influence. For this reason, the measure is sometimes preferred, for example by Jose and Winkler (2008).

Let λ be the trim factor (i.e. the top/bottom $100 \times \lambda\%$ are winsorised) and $K = \lambda P$. The combined forecast is then calculated as:

$$f^c = \frac{1}{P} \left[K f_{(K+1)} + \sum_{i=K+1}^{P-K} K f_{(P-K)} \right] \tag{3.5}$$

5. *Bates/Granger (1969).* In their seminal paper, Bates and Granger (1969) introduced the idea of combining forecasts. Their approach builds on portfolio diversification theory and uses the diagonal elements of the estimated mean squared prediction error matrix in order to compute combination weights:

$$f^c = \sum_{i=1}^{P} f_i' \times \frac{\hat{\sigma}^{-2}(i)}{\sum_{j=1}^{P} \hat{\sigma}^{-2}(j)} \tag{3.6}$$

where $\hat{\sigma}^2(i)$ is the estimated mean squared prediction error of model i.

The approach ignores correlation between component forecasts due to difficulties in precisely estimating the covariance matrix.

6. *Newbold/Granger (1974).* Building on the earlier research by Bates and Granger (1969), the methodology of Newbold and Granger (1974) also extracts the combination weights from the estimated mean squared prediction error matrix.

Let Σ be the mean squared prediction error matrix of $F_{N \times P}$ and e be a $P \times 1$ vector of $(1, 1, ..., 1)'$. Newbold and Granger (1974)'s method is a constrained minimisation of the mean squared prediction error using the normalisation condition $e'w = 1$. This yields the following combined forecast:

$$f^c = F_{N \times P} \times \frac{\Sigma^{-1} e}{e' \Sigma^{-1} e} \tag{3.7}$$

While the method dates back to Newbold and Granger (1974), the variant of the method we use in the **GeomComb** package does not impose the prior

restriction that Σ is diagonal. This approach, used by Hsiao and Wan (2014), is a generalisation of the original method.

7. *Inverse Rank.* The inverse rank approach, suggested by Aiolfi and Timmermann (2006), ranks the forecast models based on their performance up to time N. The model with the lowest mean squared prediction error is assigned the rank 1, the model with the second lowest mean squared prediction error is assigned the rank 2, etc. The combined forecast is then calculated as follows:

$$f^c = \sum_{i=1}^{P} f_i' \times \frac{Rank_i^{-1}}{\sum_{j=1}^{P} Rank_j^{-1}} \tag{3.8}$$

Timmermann (2006) points out that this method, just like Bates and Granger (1969), also ignores correlations across forecast errors. However, the method is more robust to outliers, since total rankings are not likely to change dramatically by the presence of extreme forecasts.

Regression-based Combination Methods

8. *Ordinary Least Squares (OLS) regression.* The idea to use regression for combining forecasts was put forward by Crane and Crotty (1967) and successfully driven to the forefront by Granger and Ramanathan (1984). Using this approach, the combined forecast is a linear function of the individual forecasts where the weights are determined using a regression of the individual forecasts on the target itself:

$$y = \alpha + \sum_{i=1}^{P} w_i f_i + \varepsilon, \tag{3.9}$$

Using a portion of the forecasts to train the regression model, the OLS coefficients can be estimated by way of minimising the sum of squared errors in equation(8). The combined forecast is then given by:

$$f^c = \widehat{\alpha} + \sum_{i=1}^{P} \widehat{w}_i f_i, \tag{3.10}$$

An advantage of the OLS forecast combinations is that the combined forecast is unbiased due to the intercept in the equation, even if one of the individual

forecasts is biased. A disadvantage is that the method places no restriction on the combination weights (i.e. they do not add up to 1 and can be negative), which complicatesinterpretation, especially if the coefficients are non-convex.[1]

9. *Least Absolute Deviation (LAD) regression.* While the OLS regression minimises the coefficients in equation (3.10) by minimising the sum of squared errors, we may want to estimate those coefficients differently, minimising a different loss function[2], for example the absolute sum of squares. The reason is best explained using an example: Assume we have a model that performs well in general, yet every now and then misses the target by a very large margin. Such a model would be weighted more heavily under the LAD scheme than under the OLS scheme since those large but infrequent errors will be more heavily penalised using OLS. Whether this is beneficial depends on the user's preference and/or the cost of missing the target given the problem at hand. It should be noted that this lower sensitivity to outliers has another advantage: OLS weights can be unstable when predictors are highly correlated, which is the norm in forecast combination. Minor fluctuations in the sample can cause major shifts in the coefficient vector ('bouncing betas'), often leading to poor out-of-sample performance. This suggests that LAD combination should be favoured in the presence of highly correlated component forecasts (Nowotarski et al., 2014).

10. *Constrained Least Squares (CLS) regression.* Like the LAD approach, CLS addresses the issue of 'bounding betas'. It does so by minimising the sum of squared errors under some additional constraints. Specifically, we constrain the estimated coefficients $\{w_i\}$, allowing only for positive solutions: $w_i \geq 0 \; \forall i$, and to sum up to one: $\sum_{i=1}^{P} w_i = 1$. The solution requires numerical minimisation, but good optimisation algorithms are readily available: The **GeomComb** package relies on the function `solve.QP` available from the **quadprog** package (Turlach and Weingessel, 2013). To tackle problems with high (but imperfect) collinearity that can cause errors in the CLS estimation, we also implement

---

[1]if the combination of two individual forecasts is not convex, the resulting combined forecast will not necessarily lie between those individual forecasts. As an example, we can look at the first quarter (2014 or 2015) GDPplus series estimate, published by the Federal Reserve Bank of Philadelphia. The combined prediction of two individual estimates of the US GDP; one based on the expenditure-side and one based on the income-side, lied above the sum of those two estimate. Intuitively, this is hard to communicate (https://www.philadelphiafed.org/research-and-data/real-time-center/gdpplus).

[2]For a discussion of optimal forecast combinations under general loss functions, see Elliott and Timmermann (2004)

a revised Cholesky decomposition based on Ridge regression which has been proposed by Babaie-Kafaki and Roozbeh (2017) and can mitigate issues with multicollinearity.

Theoretically, the additional constraints set CLS sub-optimally compared to the OLS. It lacks the good asymptotic properties admitted by OLS. However, in practice it is often found to perform better, especially so when the individual forecasts are highly correlated. In addition, the CLS weights are more easily interpretable. It is hard to justify a non-convex linear combination of two forecasts, while CLS weights can be conveniently interpreted for example as percentages devoted to each of the individual forecasts.

11. *Complete subset regression.* The **GeomComb** package allows the relatively new idea of computing forecast combination weights using complete subset regression. The underlying idea is relatively straightforward: With P component forecasts that can serve as predictors in the regression model, we can form n regression models, each with a unique subset of predictors. n, the total number of combined forecasts from regression models is given by

$$n = \sum_{i=1}^{P} \binom{P}{i} = \sum_{i=1}^{P} \frac{P!}{i!(P-i)!} \tag{3.11}$$

In the most basic variant, the final combined forecast is obtained by taking the simple average over the cross-section of these n combined forecasts from complete subset regressions. The method is proposed by Elliott et al. (2013) who develop the theory behind this estimator and present favourable results from simulations and empirical application for US stock returns. Admittedly, the scheme is computationally expensive, and thus additional computational resources may be required if P is in the dozens (Elliott et al. (2013), Section 2.5.1 proposes a workaround based on random sampling from P). Additionally, since all n forecasts are returned, the user can freely refine the technique further – for example by choosing not to average over all n forecasts, but some partial subset. Using the median instead of the mean is another option that comes to mind.

Obtaining n combined forecasts from the complete subset regression also allows us to use a frequentist approach to forecast combination, also known as information-theoretic forecast combinations. In the **GeomComb** package, several information criteria are available in the complete subset regression

method, each with its own merits and weaknesses: By far the most common are the AIC (Akaike's information criterion) and the BIC (Bayesian information criterion, also known as the Schwarz information criterion). Both are supplied in addition to the corrected AIC (Hurvich and Tsai, 1989) and the Hannan Quinn information criterion (Hannan and Quinn, 1979).

Formally, the weight given to each forecast based on the information-theoretic forecast combinations is the following:

$$w_i = \frac{\exp(-1/2\varrho_i)}{\sum_{i=1}^{n} \exp(-1/2\varrho_i)}, \tag{3.12}$$

where ϱ_i is the information criterion for forecast i obtained using a regression with a specific combination of forecasts. The value of n is fixed as the number of possible combinations, and the combined forecast is given by:

$$f^c = \sum_{i=1}^{n} w_i \tilde{f}_i. \tag{3.13}$$

It is worth noting that this is a two-step combination method. The first step is the computation of n combined forecasts $\tilde{f}_i$ using the complete subset regression method with the original P forecasts as predictors; the second step is the combination of these combined forecasts using the weights based on information criteria.

One advantage of this frequentist approach to model averaging is that the amount of shrinkage enforced on each individual forecast is data driven. The specification of a shrinkage hyper-parameter, which is required in the corresponding Bayesian framework (e.g. Raftery et al., 2017) is spared from the user in this case.

Eigenvector-based Combination Methods

The eigenvector-based forecast combination methods, proposed by Hsiao and Wan (2014), are based on the idea of minimising the mean squared prediction error subject to a normalisation condition.

The most commonly used normalisation condition for this purpose is to require the combination weights to add up to one, i.e. $\sum_{i=1}^{P} w_i = 1$ (e.g. Newbold and Granger, 1974; Timmermann, 2006). Hsiao and Wan (2014) show that this normalisation condition leads to a constrained minimum of the mean squared prediction error

(MSPE), and propose a normalisation condition that leads to an unconstrained minimum of the MSPE:

$$\sum_{i=1}^{P} w_i^2 = 1 \tag{3.14}$$

This unconstrained minimum of the MSPE is the basis of the four eigenvector-based approaches in the **GeomComb** package.

12. *Standard Eigenvector Approach.* The standard eigenvector approach retrieves combination weights from the estimated MSPE matrix as follows: The P positive eigenvalues of the MSPE matrix are arranged in increasing order $(\Phi_1 = \Phi_{min}, \Phi_2, ..., \Phi_P)$, and κ_i denotes the eigenvector corresponding to Φ_i. Let $d_i = e'\kappa_i$ with e being a $P \times 1$ vector of $(1, 1, ...1)'$. The combination weights w are then chosen corresponding to the minimum of $(\frac{\Phi_1}{d_1^2}, \frac{\Phi_2}{d_2^2}, ..., \frac{\Phi_P}{d_P^2})$, denoted as κ_l, as:

$$w = \frac{1}{d_l}\kappa_l \tag{3.15}$$

The combined forecast is then obtained as usual:

$$f^c = \sum_{i=1}^{P} f_i' w_i \tag{3.16}$$

13. *Bias-Corrected Eigenvector Approach.* The bias-corrected eigenvector approach builds on the idea that if one or more of the component models yield biased forecasts, the accuracy of the standard eigenvector approach can be improved by eliminating the bias. It modifies the standard approach by decomposing forecast errors into three parts: model-specific bias, omitted common factors of all component models, as well as an idiosyncratic part that is uncorrelated across the component models.

The optimisation procedure to obtain combination weights coincides with the standard approach, except that we use as an input the centered MSPE matrix, i.e. after extracting the bias by subtracting the column means of the MSPE:

$$w = \frac{1}{\tilde{d}_l}\tilde{\kappa}_l \tag{3.17}$$

where $\tilde{d}_i$ and $\tilde{\kappa}_i$ are defined analogously to d_i and κ_i in the standard eigenvector approach with the only difference that they correspond to the spectral decomposition of the centered MSPE matrix rather than the original (uncentered) MSPE matrix.

The combined forecast is then obtained by:

$$f^c = \alpha + \sum_{i=1}^{P} f_i' w_i \qquad (3.18)$$

where the intercept α corrects for the potential bias.

14. *Trimmed Eigenvector Approach.*

   The standard approach is highly sensitive to the disparities in performance of different predictive models, i.e. the standard eigenvector approach's performance could be severely impaired by one or more component models that produce poor forecasts. This is due to treating uncertainties in the actual series, y, and the uncertainties of the component models, $F_{N \times P}$, symmetrically. For a detailed discussion of this so-called orthogonality principle, see Section 3 in Hsiao and Wan (2014). The trimmed eigenvector approach takes note of this issue.

   The idea of trimming the pool of input forecasts has been used by Aiolfi and Timmermann (2006) and is picked up by Hsiao and Wan (2014) using the eigenvector framework – the weights are computed exactly as in the standard eigenvector approach, but based on the MSPE matrix of the trimmed forecasts, after discarding particularly bad component models.

15. *Trimmed Bias-Corrected Eigenvector Approach.* The underlying methodology of the trimmed bias-corrected eigenvector approach is the same as the bias-corrected eigenvector approach: The weights are retrieved through the spectral decomposition of the centered MSPE matrix.

   The only difference to the bias-corrected eigenvector approach is that this method, like the trimmed eigenvector approach, pre-selects component models that serve as input for the forecast combination; only a subset of the available forecast models is retained, while the models with the worst performance are discarded, thereby combining the favourable modifications of the previous two methods.

3.3 Implementation

Table 3.1 Summary of the Main Functions Available in the **GeomComb** Package.

Function	Description
Data Preparation Functions	
foreccomb	Transform raw input data for forecast combination
cs_dispersion	Compute cross-sectional dispersion
Forecast Combination Functions	
Simple Methods	
comb_BG	Bates/Granger (1969) forecast combination
comb_InvW	Inverse Rank forecast combination
comb_MED	Median Forecast Combination
comb_NG	Newbold/Granger (1974) forecast combination
comb_SA	Simple Average forecast combination
comb_TA	Trimmed Mean forecast combination
comb_WA	Winsorised Mean forecast combination
Regression-Based Methods	
comb_CLS	Constrained Least Squares forecast combination
comb_CSR	Complete Subset Regression forecast combination
comb_LAD	Least Absolute Deviation forecast combination
comb_OLS	Ordinary Least Squares forecast combination
Eigenvector-Based Methods	
comb_EIG1	Standard Eigenvector forecast combination
comb_EIG2	Bias-Corrected Eigenvector forecast combination
comb_EIG3	Trimmed Eigenvector forecast combination
comb_EIG4	Trimmed Bias-Corrected Eigenvector forecast combination
Other Methods	
auto_combine	Automated grid-search forecast combination
rolling_combine	Rolling forecast combination (time-varying combination weights)
Post-Fit Functions	
plot.foreccomb_res	S3 method to plot results from a forecast combination model
summary.foreccomb_res	S3 method – summary of the forecast combination estimation

The main functions provided in the **GeomComb** package can be classified in 3 categories:

- Data Preparation

- Estimation of Forecast Combination

- Post-Fit Presentation of Results

In addition, some auxiliary functions are provided. Table 3.1 provides a list of the main functions.

3.3.1 Data Preparation

The **GeomComb** package considers as a starting point that the user has already obtained a set of component forecasts, either from survey data or using statistical techniques, and now seeks to improve accuracy by combining those component forecasts into one. If the user only has the actual time series data, other packages in R can be used to create a set of component models. For instance, the **forecastHybrid** package by Shaub and Ellis (2017) which creates several univariate forecasts using methods available in the mature and popular **forecast** package (Hyndman, 2017).

The method `foreccomb` is the workhorse in the data preparation step. It supports the user with transforming the raw input data to make sure that the estimation of the forecast combinations will run smoothly.

The call of `foreccomb` is:

```
foreccomb(observed_vector, prediction_matrix, newobs = NULL,
  newpreds = NULL, byrow = FALSE, na.impute = TRUE,
  criterion = "RMSE")
```

The function requires user input for the parameters `observed_vector` (a vector, the actual data) and `prediction_matrix` (a matrix, the set of component forecasts to be combined). The format of the input matrix is as follows: Each column contains the forecasts from one of the P component models. Each row corresponds to the cross-section of component forecasts at a specific point in time. A situation where the format of the data is reversed, meaning that rows correspond to forecast models and columns correspond to the time index, is handled by setting the argument `byrow` = `TRUE`.

The `foreccomb` function includes some convenient features that take note of the fact that in many cases combination methods are applied to survey forecasts and the challenges that come along with this:

- *Split into Training Set and Test Set.* The function allows the user to specify a training set (`observed_vector` and `prediction_matrix`)and a test set (`newobs` and `newpreds`) separately. This is useful since most combination functions have to estimate the weights (requiring part of the sample to be

dedicated to that task), while it is recommended that a test set is available to evaluate the model's performance on "new" data.

- *Missing Value Imputation.* Survey forecasts usually include missing values. This can be either because some of the survey participants did not respond or because the set of survey participants is changed. The `foreccomb` function provides two alternatives to deal with missing values: The default option (`na.impute = TRUE`) uses the missing value imputation algorithm from the **mtsdi** by Junger and de Leon (2012). It is a modified version of the EM algorithm for imputation that is specifically adjusted for multivariate time series data, accounting for correlation between the forecasting models and the time structure of the series itself. A smoothing spline is fitted to each of the time series at every iteration and the degrees of freedom of each spline are chosen by cross-validation. Alternatively, the argument can be set to `na.impute = FALSE`, which means the component forecast models that include any missing values are dropped prior to estimating forecast combinations, and the user is notified in the console if so relevant.

- *Handling Multicollinearity.* More often than not component forecasts that are used in the forecast combination are highly correlated. This can trouble the estimation process which does not handle well perfect collinearity. The `foreccomb` function has an inherent algorithm that checks the set of component forecasts for perfect multicollinearity, and if detected, drops one of the component models from the input data. The algorithm is designed to minimise the cost of dropping one or more models from the input data, in the sense that out of the models that cause perfect multicollinearity, it drops the least accurate forecast model. By default, Root Mean Squared Error (RMSE) is used as the accuracy metric, but alternatively the user may choose the Mean Absolute Error (MAE) or the Mean Absolute Percentage Error (MAPE) by changing the argument `criterion`.

The output of the `foreccomb` function is an object of `S3` class `foreccomb` that can be passed on to the estimation functions or the other auxiliary functions, for instance the function `cs_dispersion` which computes the cross-sectional dispersion of the set of component forecasts.

This is often helpful for selecting a suitable combination method: One of the main findings of Hsiao and Wan (2014) is that regression-based methods produce

more accurate forecasts when one or a few of the component forecasts are much better than the rest, while eigenvector-based methods perform better when there is low dispersion among the component forecasts. The `cs_dispersion` function can be used to compute and plot this cross-sectional dispersion using standard deviation (default), interquartile range, or range.

3.3.2   Estimation of Forecast Combination

The package provides the user with functions for the 15 estimation techniques for combined forecasts, which were described in Section 3.2. The estimation functions require an object of `S3` class `foreccomb` as input, which is obtained using the methods from the previous subsection.

Four of the methods include trimming, i.e. a pre-selected subset of the full set of component models that should be used in the estimation of combination weights. These are:

- Trimmed Mean (`comb_TA`)

- Winsorised Mean (`comb_WA`)

- Trimmed Eigenvector Approach (`comb_EIG3`)

- Trimmed Bias-Corrected Eigenvector Approach (`comb_EIG4`)

For these methods, the user has some flexibility. The package provides the option to set the trimming factor (or, for the eigenvector methods, the number of retained component models) manually. Otherwise, an inbuilt optimisation algorithm is used for choosing the trimming factor such that the combined forecast has the best possible fit. Again, this optimisation can be based on either RMSE, MAE, or MAPE, which are controlled by the argument `criterion`.

A simple simulation example:

```
R> actual <- rnorm(100)
R> forecasts <- matrix(rnorm(1000, 1), 100, 10)


R> input_data <- foreccomb(actual, forecasts)


R> # Manual Selection of Trimming Factor:
R> model1 <- comb_TA(input_data, trim_factor = 0.3)
```

```
R> # Assess accuracy of the combined forecast:
R> model1$AccuracyTrain
```

```
              ME RMSE  MAE    MPE   MAPE     ACF1 Theil's U
Training Set -1.07 1.52 1.27 146.41 486.35 -0.04805  1.698456
```

```
R> # Algorithm-Optimised Selection of Trimming Factor:
R> model2 <- comb_TA(input_data, criterion = "RMSE")
```

```
Optimisation algorithm chooses trim factor for trimmed mean approach.
Algorithm finished. Optimised trim factor: 0.1
```

```
R> # Assess accuracy of the combined forecast:
R> model2$AccuracyTrain
```

```
              ME RMSE  MAE    MPE   MAPE     ACF1 Theil's U
Training Set -1.06 1.51 1.27 134.72 489.43 -0.03678  1.692617
```

As can be seen, the automated selection of the trimming factor leads to an improved accuracy of the combined forecast.

The 15 methods included in the package all produce static combination weights, i.e. the models use the training set data to estimate combination weights, which will in turn be applied to all periods of the test set. The research community in the forecasting field is strongly divided in the assessment of the value of time-varying combination weights, since putting higher weights on more recent data tends to increase the parameter variance. Section 4.1 in Timmermann (2006) reviews the advantages and challenges of allowing for time-varying weights.

While the **GeomComb** mainly uses time-invariant combination weights, the user is provided with some flexibility. The `rolling_combine` function allows for the estimation of each of the methods with time-varying weights. The approach builds on the idea of time series cross-validation (Bergmeir et al., 2015), using the provided training set as a departure point to estimate starting weights, and then increasing the training set one step at a time and re-estimating the weights for the remaining

test set. However, this approach requires that the user provides a full test set, i.e. also providing observed values for the test set.

Finally, the package carefully considers not only experienced forecasting researchers or professionals. Exploring a wide range of combination techniques can be a daunting task for inexperienced forecasters. Therefore, the function `auto_combine` provides a quick and painless alternative. The function is based on a grid-search optimisation that returns the combined forecast with the best in-sample accuracy (using RMSE as accuracy metric in the default setting).

Table 3.2 Output Components of the Forecast Combination Estimation Methods.

Output	Description
For all methods	
`Method`	Returns the used forecast combination method
`Models`	Returns the individual input models that were used for the forecast combinations
`Weights`	Returns the combination weights obtained by applying the combination method to the training set
`Fitted`	Returns the fitted values of the combination method for the training set
`Accuracy_Train`	Returns a range of accuracy measures for the training set
`Forecasts_Test`	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set
`Accuracy_Test`	Returs a range of accuracy measures for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set
`Input_Data`	Returns the data forwarded to the method
Only for `comb_TA` and `comb_WA`	
`Trim Factor`	Returns the trim factor, λ
Only for `comb_OLS`, `comb_LAD`, `comb_EIG3` and `comb_EIG4`	
`Intercept`	Returns the intercept (bias correction)
Only for `comb_EIG3` and `comb_EIG4`	
`Top_Predictors`	Number of retained predictors
`Ranking`	Ranking of predictors that determines which models are removed in the trimming step

All of the estimation methods return an object of `S3` class `foreccomb_res` with the components presented in Table 3.2. The object can subsequently be passed on to post-fit functions.

3.3.3 Post-Fit Functions

Results from the estimation of combined forecasts can be passed on to two post-fit convenience functions: `summary` and `plot`, which are `S3` methods specific to the class `foreccomb_res`.

The summary function displays the output of the respective forecast combination in concise form, it displays the estimated combination weights (and the intercept, if the combination method includes one), as well as accuracy statistics for the training set and the test set.

The plot function will produce different plots based on the input data. If only a training set was provided, it plots the actual versus the fitted values; if a test set was also provided, it plots the combined forecasts as well. Another option for the user is a plot of the combination weights[3], obtained by setting `which = 2` in the function call. For the case of dynamic estimation, an additional weights plot is implemented: the evolution of the combination weights over time (`which = 3`).

3.4 UK Electricity Supply: An Empirical Example

The **GeomComb** package includes the dataset `electricity`, which is a multivariate time series of monthly UK electricity supply (in GWh) from January 2007 to March 2017, and 5 univariate time series forecasts for the same series and period. The observed data series is sourced from the International Energy Agency (IEA, 2017). The component models to be combined are the following cross-validated one-month univariate forecasts in the dataset:

- ARIMA (produced using the `auto.arima` function in the **forecast** package),

- ETS (produced using the `ets` function in the **forecast** package),

- Neural Network (produced using the `nnetar` function in the **forecast** package),

---

[3]In case the combined forecast is produced using time-varying combination weights, the weights plot displays only the average weight of the respective component model over the test set period.

- Damped Trend (produced using the `ets` function in the **forecast** package),

- Dynamic Optimised Theta Model (produced using the `dotm` function in the **forecTheta** package by Fiorucci et al. (2016)).

To illustrate the functionalities of the package, we apply 4 combination techniques: the simple average (`comb_SA`), the OLS regression (`comb_OLS`), the standard eigenvector approach (`comb_EIG1`) and the trimmed bias-corrected eigenvector approach (`comb_EIG4`). The selected methods span all three categories of combination techniques (statistics-based, regression-based, and eigenvector-based) and includes trimmed and bias-corrected methods and are therefore suitable to show the full functionality of the **GeomComb** package in this empirical context. using both the static and dynamic version of each. For the selected combination methods, we produce both static and dynamic forecasts, which gives us a total of 7 different time series of combined forecasts (not 8, since the static and dynamic versions of the simple average combination are identical).
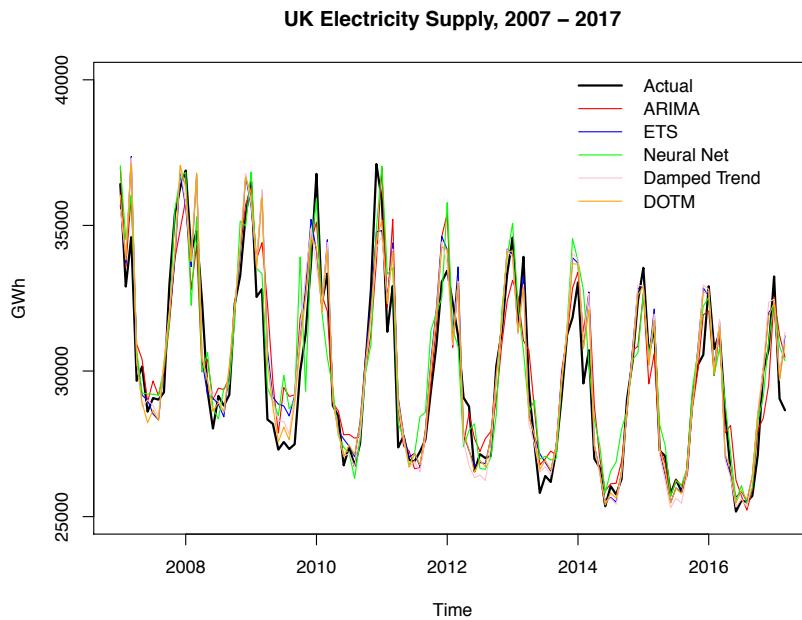

Fig. 3.1 UK Electricity Supply, 2007 - 2017: Actual Value and Forecasts.

For the purpose of this exercise, we use the first 84 months as training set, which leaves us with a test set size of 39. Figure 3.1 plots the actual series and the univariate forecasts. The forecasts (which are 1-month forecasts obtained via time

series cross-validation) track the actual series very well. None of them performs exceptionally poorly compared with the rest, which are conditions that tend to favour eigenvector approaches (Hsiao and Wan, 2014). As it is with most electricity data, the main difference between the individual models is in their ability to quickly recognise and adjust for turning points. For example, the neural nets model handles turning points well, but sometimes also overshoots, while the ARIMA model has a smoother behaviour around turning points.

First, we format the data correctly for the estimation of combination weights. This step ensures that all later operations would proceed without any hiccups. Since there are no missing values and no perfectly collinear columns in our dataset, this is relatively straightforward:

```
R> data(electricity)
R> train.obs <- electricity[1:84, 6]
R> train.pred <- electricity[1:84, 1:5]
R> test.obs <- electricity[85:123, 6]
R> test.pred <- electricity[85:123, 1:5]
R> input_data <- foreccomb(train.obs, train.pred,
                           test.obs, test.pred)
```

Once the object of `S3` class `foreccomb` is created, it can be fed into the estimation functions. We can look at the cross-sectional dispersion, to get a better idea of variability in the univariate forecasts.

```
R> cs_dispersion(input_data, measure = "SD", plot = TRUE)
```

Figure 3.2 shows that apart from a brief period of increased dispersion around the end of 2009, the cross-sectional standard deviation of the component forecasts is rather stable and low given the level of around 25,000 to 35,000 GWh. This begs the question whether conditions have fluctuated enough during the test set so that estimation of time-varying weights is beneficial, yet we proceed with it for this demonstration.

```
R> ###### ESTIMATION OF STATIC FORECAST COMBINATIONS ########
R> SA <- comb_SA(input_data)
R> OLS_static <- comb_OLS(input_data)
R> EIG1_static <- comb_EIG1(input_data)
R> EIG4_static <- comb_EIG4(input_data, criterion = "MAE")
```
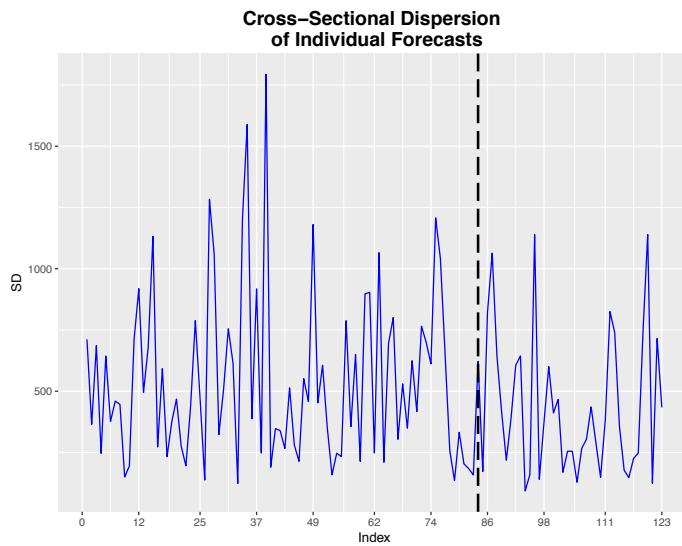
Fig. 3.2 Cross-Sectional Standard Deviation of the Component Forecasts.

```
R> ##### ESTIMATION OF DYNAMIC FORECAST COMBINATIONS
R> OLS_dyn <- rolling_combine(input_data, "comb_OLS")
R> EIG1_dyn <- rolling_combine(input_data, "comb_EIG1")
R> EIG4_dyn <- rolling_combine(input_data, "comb_EIG4",
              criterion = "MAE")
```

The 7 combined forecasts can be evaluated separately by looking at their summary measures, which we present here for the static Ordinary Least Squares approach:

```
R> summary(OLS_static)
```

```
Summary of Forecast Combination
-------------------------------


Method:  Ordinary Least Squares Regression


Individual Forecasts & Combination Weights:


        Combination Weight
arima           0.02152869
ets            -0.20646266
```

```
nnet             0.20992792
dampedt         -1.04349858
dotm             1.97991049


Intercept (Bias-Correction):  962.3229


Accuracy of Combined Forecast:


                        ME      RMSE      MAE         MPE      MAPE
Training Set -4.417560e-12 888.1433 697.8645 -0.08262472 2.254470
Test Set     -4.007742e+01 671.5214 536.0331 -0.24705122 1.841961


Additional information can be extracted from the combination object:
For fitted values (training set):  OLS_static$Fitted
For forecasts (test set):  OLS_static$Forecasts_Test
See  str(OLS_static)  for full list.
```

The output shows that the OLS combination puts an extremely high relative weight on the forecast from the Dynamic Optimised Theta model, which seems to be the best component forecast, which is rather surprising given that seasonality is an important feature in the analysed series and Theta models cannot incorporate seasonality into the estimation so far, relying on pre-estimation deseasonalising and post-estimation reseasonalising. Table 3.3 shows a comparison of the accuracies achieved by the combined forecasts. Since all forecasts are for the same series, it is reasonable to use MAE as accuracy metric.

Table 3.3 Mean Absolute Errors of Combined Forecasts.

Method	MAE Training Set	MAE Test Set
Simple Average	819.28	573.39
Ordinary Least Squares (static)	697.86	536.03
Ordinary Least Squares (dynamic)	697.86	533.47
Standard Eigenvector (static)	821.60	573.84
Standard Eigenvector (dynamic)	821.60	572.99
Trimmed Bias-Corrected Eigenvector (static)	785.30	540.18
Trimmed Bias-Corrected Eigenvector (dynamic)	785.30	541.98

The evaluation of accuracy delivers some interesting insight: All of the combination models perform better in the test set than in the training set, which is counter-intuitive, but is likely due to the increased dispersion among the component

forecasts in the early period. The results also clearly suggest that one or more of the component forecasts were biased. The two methods with an intercept (i.e. bias correction) perform best. Finally, allowing for time-varying combination weights does not seem to change test-set accuracy much compared with the models' static counterparts, suggesting that the estimated combination weights did not fluctuate a lot over time. There are some potential explanations why the OLS method performed extremely well in this case:

- With such stable conditions the risk of 'bouncing betas' (described in Section 3.2) is low,

- The OLS method produces unbiased forecasts even if one or more of the component forecasts are biased (which is why the trimmed bias-corrected eigenvector approach performed reasonably well too),

- One of the component forecasts is much better than the rest, a situation that is favourable for regression-based approaches, as pointed out by Hsiao and Wan (2014). These are also conditions under which sophisticated methods actually can largely improve upon a simple average combination.

Now that we learned that some of the combination models produced more accurate forecasts than the simple average, we address the next, very natural question: How well did the combination methods perform compared to the univariate component forecasts themselves? To shed more light on this question, Table 3.4 shows the MAE values for the univariate models during the test set period.

Table 3.4 Mean Absolute Errors of Univariate Component Forecasts.

Method	MAE Test Set
ARIMA	770.32
ETS	615.88
Neural Network	730.35
Damped Trend	660.08
DOTM	540.24

The results of the accuracy evaluation speak for themselves. Not only did two of the combined forecasts perform better or equally well as the best univariate forecast over the test set period, but also the forecast risk is considerably lower indeed. In the test set the range of MAEs of the different combined forecast methods is only 40, while the corresponding value for the univariate forecasting methods is 230. It is

worth noting that all of the combined forecasts perform considerably better than even the second-best univariate forecast, emphasising the appeal of forecast combination: In the ideal case, it is possible to end up with a forecast that is better than the best univariate forecast. Even if this is not the case, using forecast combination considerably decreases the risk of ending up with a poorly performing model.

Finally, we take a closer look at the results from the best combined forecast, which is the dynamic OLS combination method in this exercise.

```
R> ##### ACTUAL VS FITTED PLOT #####
R> plot(OLS_dyn)
```
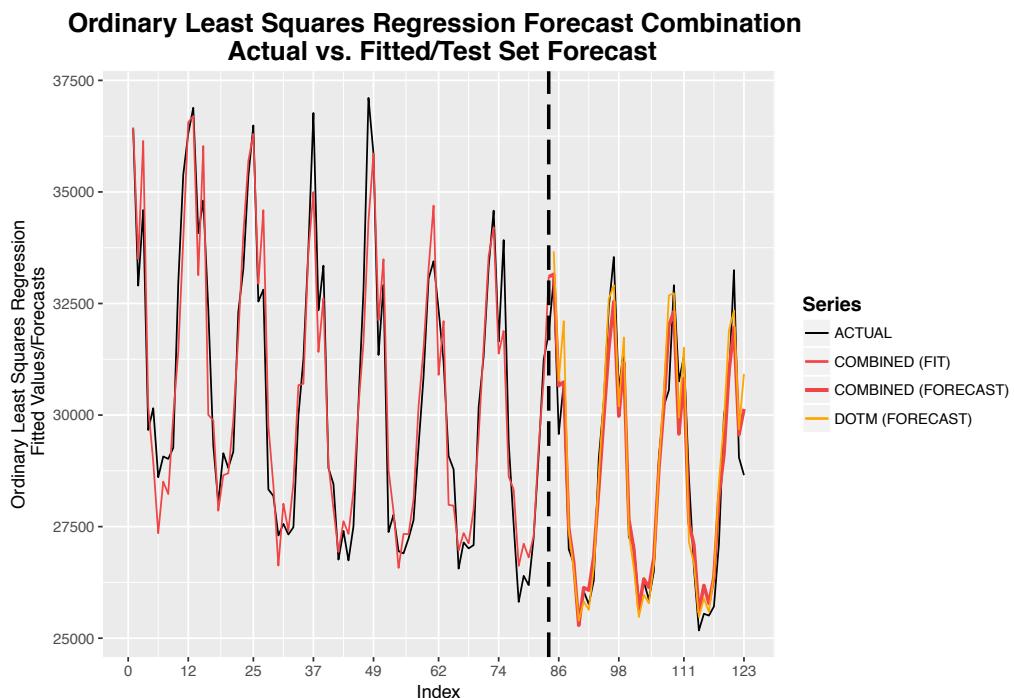

Fig. 3.3 Dynamic OLS Forecast Combination: Actual vs Fitted Plot.

```
R> ##### COMBINATION WEIGHTS #####
R> OLS_static$Weights
```

```
 fhatarima     fhatets    fhatnnet  fhatdampedt    fhatdotm
0.02152869 -0.20646266  0.20992792  -1.04349858  1.97991049
```

```
R> colMeans(OLS_dyn$Weights)
```

```
  fhatarima        fhatets      fhatnnet   fhatdampedt      fhatdotm
  0.002750948 -0.123811218   0.184846358  -1.098381593   2.001988985
```

```
R> par(mfrow = c(1,2))
R> plot(OLS_static, which = 2)
R> plot(OLS_dyn, which = 3)
```
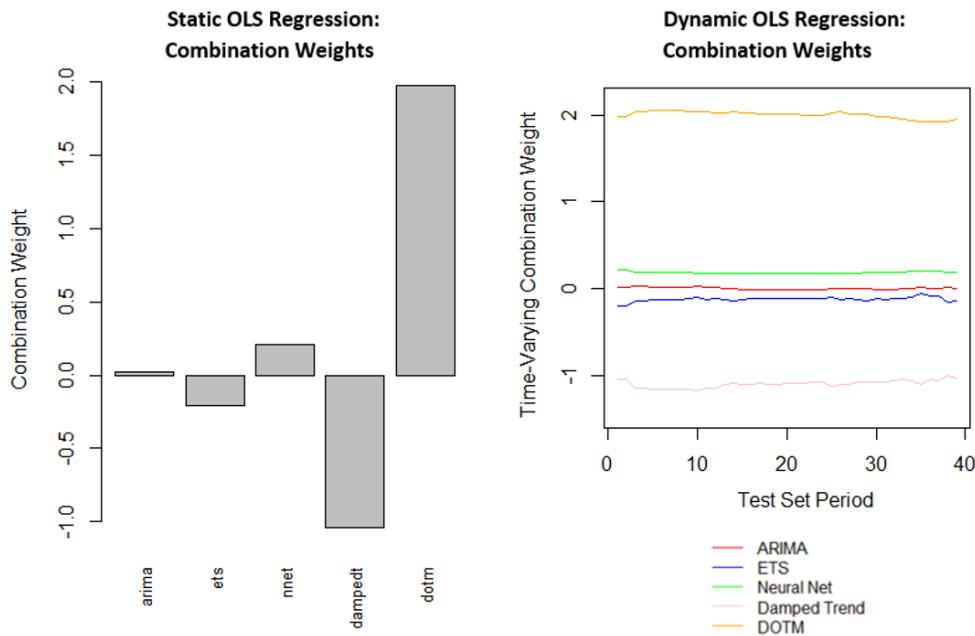

Fig. 3.4 (left) Combination Weights of the Static OLS Method; (right) Time-Varying Combination Weights of the Dynamic OLS Method.

Figures 3.3 shows how well the combined forecast obtained from the dynamic OLS method predicts the monthly electricity supply series; the best univariate component forecast (DOTM) is also plotted as benchmark. Figure 3.4 confirms the conjecture that the stable conditions (low cross-sectional dispersion and one very dominant univariate component forecast) do not cause a lot of fluctuation in the combination weights even when allowing for time-varying weights.

The weights graph also confirms another thing: That the weights obtained from OLS combination methods can be hard to interpret. It seems obvious that the method should put a high weight on the DOTM forecast, since it is the best univariate forecast by far. However, the reason why it assigns negative weights

to the ETS and Damped Trend forecasts (the second- and third-best univariate forecasts) is not very intuitive. A possible explanation might be that all three of these are exponential smoothing-type models, suggesting that the information obtained from the ETS and Damped Trend models is better captured by the DOTM model, while ARIMA and Neural Networks are not closely related modelling approaches to the Theta model, so that even though these models perform far worse on average, they capture information differently and might outperform the DOTM forecast in some periods for that reason, justifying their positive weights (however small) and explaining how the combined forecast can slightly outperform the best component forecast.

3.5 Discussion and Conclusions

Forecast combination is a useful strategy to hedge against model risk. Even if combined forecasts do not win over the most accurate component forecast, they generally avoid poor performance by circumventing the choice between individual methods (Timmermann, 2006). Instead of putting all eggs into one basket using model selection, these model averaging techniques are motivated by portfolio theory and diversify across component forecasts.

The best way to combine different forecasts has no theoretical underpinnings, a lot depends on the specifics of the data at hand. Since the seminal paper by Bates and Granger (1969), myriad combination strategies have been put forward in theoretical and empirical literature.

The **GeomComb** package categorises some of the most popular approaches into 3 groups: (a) simple statistics-based methods, (b) regression-based methods, and (c) eigenvector-based methods. Providing both regression-based combinations and eigenvector-based combinations to the users is considered useful, since the former tend to perform relatively better when one or a few component forecasts are much better than the rest, while the latter perform relatively better when all forecasts are in the same ballpark (Hsiao and Wan, 2014).

The package is designed to support users along the entire modelling process: data preparation, model estimation, and interpretation of results using summaries and plotting functionalities. It includes tools for data transformation that are designed to deal with two common issues in forecast combination prior to estimation – missing values and multicollinearity. The 15 combination methods are available in both static and dynamic variants, and users have the option to automate the selection algorithm

so that a good combination method is found based on training set fit. Finally, the package offers specialised functions for summarising and visualising the combination results.

While the current version of the package already provides a comprehensive toolset for forecast combination, there is scope for further extensions in future updates. First, additional combination methods that showed promising results recently can be added, for instance the factor-augmented regression approach by Cheng and Hansen (2015) and the AdaBoost algorithms reviewed by Barrow and Crone (2016). Second, additional algorithms for adaptive combination weights (cf. Timmermann, 2006) can be implemented to provide even more flexibility with dynamic estimation. Finally, we plan to design an adaptation for a forecast combination context of the mean absolute scaled error (Hyndman and Koehler, 2006) – the current gold standard for accuracy evaluation – using the in-sample MAE of the best component forecast as scaling factor.

3.6   References

Aiolfi, M., and A. Timmermann. 2006. Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics* 135:31–53.

Andrawis, R. R., A. F. Atiya, and H. El-Shishiny. 2011. Combination of Long Term and Short Term Forecasts, With Application to Tourism Demand Forecasting. *International Journal of Forecasting* 27:870–886.

Armstrong, J. S. 1989. Combining Forecasts: The End of the Beginning or the Beginning of the End? *International Journal of Forecasting* 5:585–588.

Armstrong, J. S. 2001. Combining Forecasts. In J. S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp. 417–439. Kluwer Academic Publisher.

Avramov, D. 2002. Stock Return Predictability and Model Uncertainty. *Journal of Financial Economics* 64:423–458.

Babaie-Kafaki, S., and M. Roozbeh. 2017. A Revised Cholesky Decomposition to Combat Multicollinearity in Multiple Regression Models. *Journal of Statistical Computation and Simulation* 87:2298–2308.

Barrow, D. K., and S. F. Crone. 2016. A comparison of AdaBoost algorithms for time series forecast combination. *International Journal of Forecasting* 32:1103–1119.

Bates, J. M., and C. W. J. Granger. 1969. The Combination of Forecasts. *Operations Research Quarterly* 20:451–468.

Bergmeir, C., R. J. Hyndman, and B. Koo. 2015. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. Working Paper No. 10/15, Monash University, Department of Econometrics and Business Statistics.

Cheng, G., and Y. Yang. 2015. Forecast Combination with Outlier Protection. *International Journal of Forecasting* 31:223–237.

Cheng, X., and B. E. Hansen. 2015. Forecasting with Factor-Augmented Regression: A Frequentist Model Averaging Approach. *Journal of Econometrics* 186:280–293.

Christiansen, C., M. Schmeling, and A. Schrimpf. 2012. A Comprehensive Look at Financial Volatility Prediction by Economic Variables. *Journal of Applied Econometrics* 27:956–977.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang. 2016. The Forecast Combination Puzzle: A Simple Theoretical Explanation. *International Journal of Forecasting* 32:754–762.

Clemen, R. T. 1989. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting* 5:559–583.

Crane, D. B., and J. R. Crotty. 1967. A Two-Stage Forecasting Model: Exponential Smoothing and Multiple Regression. *Management Science* 13:501–507.

Elliott, G., A. Gargano, and A. Timmermann. 2013. Complete Subset Regressions. *Journal of Econometrics* .

Elliott, G., and A. Timmermann. 2004. Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions. *Journal of Econometrics* 122:47–79.

Elliott, G., and A. Timmermann. 2005. Optimal Forecast Combinations Under Regime Switching. *International Economic Review* 46:1081–1102.

Fiorucci, J. A., F. Louzada, and B. Yiqi. 2016. *forecTheta: Forecasting Time Series by Theta Models.* URL https://CRAN.R-project.org/package=forecTheta. [Accessed on Sep 15, 2017], R package version 2.2.

Gaillard, P., and Y. Goude. 2016. *opera: Online Prediction by Expert Aggregation.* URL https://CRAN.R-project.org/package=opera. [Accessed on Sep 15, 2017], R package version 1.0.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann. 2013. Combining Expert Forecasts: Can Anything Beat the Simple Average? *International Journal of Forecasting* 29:108–121.

Graefe, A., J. S. Armstrong, R. J. Jones, and A. G. Cuzán. 2014. Combining Forecasts: An Application to Elections. *International Journal of Forecasting* 30:43–55.

Granger, C. W., and R. Ramanathan. 1984. Improved Methods of Combining Forecasts. *Journal of Forecasting* 3:197–204.

Hannan, E. J., and B. G. Quinn. 1979. The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 190–195.

Hansen, B. 2005. Challenges for Econometric Model Selection. *Econometric Theory* 21:60–68.

Hansen, B. E. 2007. Least Squares Model Averaging. *Econometrica* pp. 1175–1189.

Hansen, B. E. 2008. Least-Squares Forecast Averaging. *Journal of Econometrics* 146:342–350.

Hansen, B. E., and J. S. Racine. 2012. Jackknife Model Averaging. *Journal of Econometrics* 167:38–46.

Hendry, D. F., and M. P. Clements. 2004. Pooling of Forecasts. *The Econometrics Journal* 7:1–31.

Hsiao, C., and S. K. Wan. 2014. Is There An Optimal Forecast Combination? *Journal of Econometrics* 178:294–309.

Hurvich, C. M., and C.-L. Tsai. 1989. Regression and Time Series Model Selection in Small Samples. *Biometrika* 76:297–307.

Hyndman, R. J. 2017. *forecast: Forecasting functions for time series and linear models.* URL http://github.com/robjhyndman/forecast. [Accessed on Sep 15, 2017], R package version 8.1.

Hyndman, R. J., and A. B. Koehler. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22:679–688.

IEA. 2017. Monthly Electricity Statistics. URL http://www.iea.org/statistics/monthlystatistics/monthlyelectricitystatistics/. [Accessed on Sep 15, 2017].

Jose, V. R. R., and R. L. Winkler. 2008. Simple Robust Averages of Forecasts: Some Empirical Results. *International Journal of Forecasting* 24:163–169.

Junger, W., and A. P. de Leon. 2012. *mtsdi: Multivariate time series data imputation.* URL https://CRAN.R-project.org/package=mtsdi. [Accessed on Sep 15, 2017], R package version 0.3.3.

Kapetanios, G., V. Labhard, and S. Price. 2008. Forecasting Using Bayesian and Information-Theoretic Model Averaging: An Application to UK Inflation. *Journal of Business & Economic Statistics* 26:33–41.

Magnus, J. R., and W. Wang. 2014. Concept-Based Bayesian Model Averaging and Growth Empirics. *Oxford Bulletin of Economics and Statistics* 76:874–897.

McNees, S. K. 1992. The Uses and Abuses of 'Consensus' Forecasts. *Journal of Forecasting* 11:703–710.

Morana, C. 2015. Model Averaging by Stacking. Working Paper 310, University of Milan Bicocca, Department of Economics, Management and Statistics.

Newbold, P., and C. W. J. Granger. 1974. Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *Journal of the Royal Statistical Society, Series A* 137:131–165.

Nowotarski, J., E. Raviv, S. Trück, and R. Weron. 2014. An Empirical Comparison of Alternative Schemes for Combining Electricity Spot Price Forecasts. *Energy Economics* 46:395–412.

Opschoor, A., D. J. Van Dijk, and M. Van der Wel. 2014. Improving Density Forecasts and Value-at-Risk Estimates by Combining Densities. Discussion Paper 14-090/III, Tinbergen Institute.

Palm, F. C., and A. Zellner. 1992. To Combine or Not To Combine? Issues of Combining Forecasts. *Journal of Forecasting* 11:687–701.

Raftery, A., J. Hoeting, C. Volinsky, I. Painter, and K. Y. Yeung. 2017. *BMA: Bayesian Model Averaging.* URL https://CRAN.R-project.org/package=BMA. [Accessed on Sep 15, 2017], R package version 3.18.7.

Rapach, D. E., J. K. Strauss, and G. Zhou. 2010. Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *Review of Financial Studies* 23:821–862.

Ravazzolo, F., M. Verbeek, and H. K. Van Dijk. 2007. Predictive Gains from Forecast Combinations Using Time Varying Model Weights. Report 2007-26, Econometric Institute.

Raviv, E. 2015. *ForecastCombinations: Forecast combinations in R.* URL https://CRAN.R-project.org/package=ForecastCombinations. [Accessed on Sep 15, 2017], R package version 1.1.

Raviv, E., K. E. Bouwman, and D. van Dijk. 2015. Forecasting Day-Ahead Electricity Prices: Utilizing Hourly Prices. *Energy Economics* 50:227 – 239.

Shaub, D., and P. Ellis. 2017. *forecastHybrid: Convenient Functions for Ensemble Time Series Forecasts.* URL https://CRAN.R-project.org/package=forecastHybrid. [Accessed on Sep 15, 2017], R package version 0.4.1.

Smith, J., and K. F. Wallis. 2009. A Simple Explanation of the Forecast Combination Puzzle. *Oxford Bulletin of Economics and Statistics* 71:331–355.

Stock, J. H., and M. W. Watson. 2004. Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* 23:405–430.

Timmermann, A. 2006. Forecast Combinations. *Handbook of economic forecasting* 1:135–196.

Turlach, B. A., and A. Weingessel. 2013. *quadprog: Functions to solve Quadratic Programming Problems.* URL https://CRAN.R-project.org/package=quadprog. [Accessed on Sep 15, 2017], R package version 1.5-5.

Weiss, C. E. 2017. Efficient Healthcare Staffing: The Value of Hierarchical Time Series Fo. Working Paper, Judge Business School, University of Cambridge.

Weiss, C. E., and G. R. Roetzer. 2016. *GeomComb: (Geometric) Forecast Combination Methods.* URL https://CRAN.R-project.org/package=GeomComb. [Accessed on Sep 15, 2017], R package version 1.0.

Wright, J. H. 2008. Bayesian Model Averaging and Exchange Rate Forecasts. *Journal of Econometrics* 146:329–341.

Wright, J. H. 2009. Forecasting US Inflation by Bayesian Model Averaging. *Journal of Forecasting* 28:131–144.

Appendix

3.A Reference Manual for the GeomComb Package

Package 'GeomComb'

July 14, 2017

Type Package

Title Forecast Combination Methods

Version 1.1

Depends R (>= 3.0.2)

Imports forecast (>= 7.3),
  ForecastCombinations (>= 1.1),
  ggplot2 (>= 2.1.0),
  Matrix (>= 1.2-6),
  mtsdi (>= 0.3.3),
  psych (>= 1.6.9)

Suggests testthat (>= 1.0.2)

Author Christoph E. Weiss, Gernot R. Roetzer, and Eran Raviv

Maintainer Christoph E. Weiss <info@ceweiss.com>

Description Provides eigenvector-based (geometric) forecast
  combination methods; also includes simple approaches (simple average,
  median, trimmed and winsorized mean, inverse rank method) and regression-based
  combination. Tools for data pre-processing are available in order to deal with
  common problems in forecast combination (missingness, collinearity).

URL https://github.com/ceweiss/GeomComb

BugReports https://github.com/ceweiss/GeomComb/issues

License GPL (>= 2)

LazyData true

RoxygenNote 5.0.1

R topics documented:

---

auto_combine   *Automated Forecast Combination*

Description

Computes the fit for all the available forecast combination methods on the provided dataset with respect to the loss criterion. Returns the best fit method.

Usage

```
auto_combine(x, criterion = "RMSE", param_list = NULL)
```

Arguments

x	An object of class 'foreccomb'. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
criterion	Specifies loss criterion. Set criterion to either 'RMSE' (default), 'MAE', or 'MAPE'.
param_list	Can contain additional parameters for the different combination methods (see example below).

Details

The function `auto_combine` allows to quickly apply all the different forecast combination methods onto the provided time series data and selects the method with the best fit.

The user can choose from 3 different loss criteria for the best-fit evaluation: root mean square error (`criterion='RMSE'`), mean absolute error (`criterion='MAE'`), and mean absolute percentage error (`criterion='MAPE'`).

In case the user does not want to optimize over the parameters of some of the combination methods, `auto_combine` allows to specify the parameter values for these methods explicitly (see Examples).

The best-fit results are stored in an object of class 'foreccomb_res', for which separate plot and summary functions are provided.

Value

Returns an object of class `foreccomb_res` that represents the results for the best-fit forecast combination method:

Method	Returns the best-fit forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the best-fit combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)

# Evaluating all the forecast combination methods and returning the best.
# If necessary, it uses the built-in automated parameter optimisation methods
# for the different methods.
best_combination<-auto_combine(data, criterion = "MAPE")

# Same as above, but now we restrict the parameter ntop_pred for the method comb_EIG3 to be 3.
param_list<-list()
param_list$comb_EIG3$ntop_pred<-3
best_combination_restricted<-auto_combine(data, criterion = "MAPE", param_list = param_list)
```

| comb_BG | *Bates/Granger (1969) Forecast Combination Approach* |

Description

Computes forecast combination weights according to the approach by Bates and Granger (1969) and produces forecasts for the test set, if provided.

Usage

```
comb_BG(x)
```

Arguments

x   An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

In their seminal paper, Bates and Granger (1969) introduce the idea of combining forecasts. Their approach builds on portfolio diversification theory and uses the diagonal elements of the estimated mean squared prediction error matrix in order to compute combination weights:

$$w_i^{BG} = \frac{\hat{\sigma}^{-2}(i)}{\Sigma_{j=1}^N \hat{\sigma}^{-2}(j)}$$

where $\hat{\sigma}^2(i)$ is the estimated mean squared prediction error of the i-th model.

The combined forecast is then obtained by:

$$\hat{y}_t = \mathbf{f}_t{'}\mathbf{w}^{BG}$$

Their approach ignores correlation between forecast models due to difficulties in precisely estimating the covariance matrix.

Value

Returns an object of class foreccomb_res with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Bates, J. M., and Granger, C. W. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, **20(4)**, 451–468.

Timmermann, A. (2006). Forecast Combinations. In: Elliott, G., Granger, C. W. J., and Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, **1**, 135–196.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_BG(data)
```

---

comb_CLS   *Constrained Least Squares Forecast Combination*

Description

Computes forecast combination weights using constrained least squares (CLS) regression.

Usage

```
comb_CLS(x)
```

Arguments

x   An object of class 'foreccomb'. Contains training set (actual values + matrix of
                 model forecasts) and optionally a test set.

Details

The function is a wrapper around the constrained least squares (CLS) forecast combination implementation of the *ForecastCombinations* package.

Compared to the ordinary least squares forecast combination method, CLS forecast combination has the additional requirement that the weights, $\mathbf{w}^{CLS} = (w_1, \ldots, w_N)'$, sum up to 1 and that there is no intercept. That is, the combinations of comb_CLS are affine combinations.

This method was first introduced by Granger and Ramanathan (1984). The general appeal of the method is its ease of interpretation (the weights can be interpreted as percentages) and often produces better forecasts than the OLS method when the individual forecasts are highly correlated. A disadvantage is that if one or more individual forecasts are biased, this bias is not corrected through the forecast combination due to the lack of an intercept.

In addition to the version presented by Granger and Ramanathan (1984), this variant of the method adds the restriction that combination weights must be non-negative, which has been found to be almost always outperform unconstrained OLS by Aksu and Gunter (1992) and was combined with the condition of forcing the weights to sum up to one by Nowotarski et al. (2014), who conclude that even though the method provides a suboptimal solution in-sample, it almost always produces better forecasts than unrestricted OLS out-of-sample.

The results are stored in an object of class 'foreccomb_res', for which separate plot and summary functions are provided.

Value

Returns an object of class `foreccomb_res` with the following components:

`Method`	Returns the best-fit forecast combination method.
`Models`	Returns the individual input models that were used for the forecast combinations.
`Weights`	Returns the combination weights obtained by applying the combination method to the training set.
`Fitted`	Returns the fitted values of the combination method for the training set.
`Accuracy_Train`	Returns range of summary measures of the forecast accuracy for the training set.
`Forecasts_Test`	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
`Accuracy_Test`	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
`Input_Data`	Returns the data forwarded to the method.

References

Aksu, C., and Gunter, S. I. (1992). An Empirical Analysis of the Accuracy of SA, OLS, ERLS and NRLS Combination Forecasts. *International Journal of Forecasting*, **8(1)**, 27–43.

Granger, C., and Ramanathan, R. (1984). Improved Methods Of Combining Forecasts. *Journal of Forecasting*, **3(2)**, 197–204.

Nowotarski, J., Raviv, E., Trueck, S., and Weron, R. (2014). An Empirical Comparison of Alternative Schemes for Combining Electricity Spot Price Forecasts. *Energy Economics*, **46**, 395–412.

See Also

Forecast_comb, foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
```

```
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_CLS(data)
```

---

comb_CSR   *Complete Subset Regression Forecast Combination*

Description

Combine different forecasts using complete subset regressions. Apart from the simple averaging, weights based on information criteria (AIC, corrected AIC, Hannan Quinn and BIC) can be used.

Usage

```
comb_CSR(x)
```

Arguments

x   An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

OLS forecast combination is based on

$$obs_t = const + \sum_{i=1}^{p} w_i \widehat{obs}_{it} + e_t,$$

where obs is the observed values and $\widehat{obs}$ is the forecast, one out of the p forecasts available.

The function computes the complete subset regressions. So a matrix of forecasts based on all possible subsets of fhat is returned.

Those forecasts can later be cross-sectionally averaged (averaged over rows) to create a single combined forecast using weights which are based on the information criteria of the different individual regression, rather than a simple average.

Additional weight-vectors which are based on different information criteria are also returned. This is in case the user would like to perform the frequensit version of forecast averaging (see references for more details).

Value

Returns an object of class foreccomb_res with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights based on the different information criteria.
Fitted	Returns the fitted values for each information criterion.

Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Eran Raviv and Gernot R. Roetzer

References

Hansen, B. (2008). Least-squares forecast averaging *Journal of Econometrics*, **146(2)**, 342–350.

Kapetanios, G., Labhard V., Price, S. (2008). Forecasting Using Bayesian and Information-Theoretic Model Averaging. *Journal of Business & Economic Statistics*, **26(1)**.

Koenker R. (2005). *Quantile Regression. Cambridge University Press*.

Graham, E., Garganob, A., Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, **177(2)**, 357–373.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_NG, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_CSR(data)
```

---

comb_EIG1	*Standard Eigenvector Forecast Combination*

Description

Computes forecast combination weights according to the standard eigenvector approach by Hsiao and Wan (2014) and produces forecasts for the test set, if provided.

Usage

```
comb_EIG1(x)
```

Arguments

x   An object of class `foreccomb`. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

The standard eigenvector approach retrieves combination weights from the sample estimated mean squared prediction error matrix as follows: Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$, Σ is the (positive definite) mean squared prediction error matrix of $\mathbf{f}_t$ and $\mathbf{e}$ is an $N \times 1$ vector of $(1, \ldots, 1)'$. The N positive eigenvalues are then arranged in increasing order $(\Phi_1 = \Phi_{min}, \Phi_2, \ldots, \Phi_N)$, and $\mathbf{w^j}$ is defined as the eigenvector corresponding to Φ_j. The combination weights $\mathbf{w}^{EIG1} = (w_1, \ldots, w_N)'$ are then chosen corresponding to the minimum of $\left(\frac{\Phi_1}{d_1^2}, \frac{\Phi_2}{d_2^2}, \ldots, \frac{\Phi_N}{d_N^2}\right)$, denoted as $\mathbf{w}^l$, where $d_j = \mathbf{e}'\mathbf{w}^j$, as:

$$\mathbf{w}^{EIG1} = \frac{1}{d_l}\mathbf{w}^l$$

The combined forecast is then obtained by:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{EIG1}$$

The difference to extant methods that minimize the population mean squared prediction error (e.g., Newbold and Granger, 1974) is the normalization function. While previous approaches optimize MSPE under the constraint of $\mathbf{e}'\mathbf{w} = 1$, Hsiao and Wan (2014) show that this is dominated by using $\mathbf{w}'\mathbf{w} = 1$ as constraint in the optimization problem.

Value

Returns an object of class `foreccomb_res` with the following components:

Method   Returns the used forecast combination method.

Models   Returns the individual input models that were used for the forecast combinations.

Weights   Returns the combination weights obtained by applying the combination method to the training set.

Fitted   Returns the fitted values of the combination method for the training set.

Accuracy_Train   Returns range of summary measures of the forecast accuracy for the training set.

Forecasts_Test   Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.

Accuracy_Test   Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.

Input_Data   Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

Newbold, P., and Granger, C. W. J. (1974). Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *Journal of the Royal Statistical Society, Series A*, **137(2)**, 131–165.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_NG, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG1(data)
```

---

comb_EIG2 | *Bias-Corrected Eigenvector Forecast Combination*

Description

Computes forecast combination weights according to the bias-corrected eigenvector approach by Hsiao and Wan (2014) and produces forecasts for the test set, if provided.

Usage

```
comb_EIG2(x)
```

Arguments

x | An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

The bias-corrected eigenvector approach builds on the idea that if one or more of the predictive models yield biased predictions, the accuracy of the standard eigenvector approach can be improved by eliminating the bias. The optimization procedure to obtain combination weights coincides with the standard eigenvector approach, except that it is applied to the centered MSPE matrix after extracting the bias (by subtracting the column means of the MSPE).

The combination weights are calculated as:

$$\mathbf{w}^{EIG2} = \frac{1}{\tilde{d}_l} \tilde{\mathbf{w}}^l$$

where $\tilde{d}_j$ and $\tilde{\mathbf{w}}^j$ are defined analogously to d_j and $\mathbf{w}^j$ in the standard eigenvector approach, with the only difference that they correspond to the spectral decomposition of the centered MSPE matrix rather than the uncentered one.

The combined forecast is then obtained by:

$$\hat{y}_t = a + \mathbf{f}_t{}' \mathbf{w}^{EIG2}$$

where $a = E(y_t) - E(\mathbf{f}_t)' \mathbf{w}^{EIG2}$ is the intercept for bias correction. If the actual series and the forecasts are stationary, the expectations can be approximated by the time series means, i.e. the intercept is obtained by subtracting the weighted sum of column means of the MSPE matrix from the mean of the actual series. Forecast combination methods including intercepts therefore usually require stationarity.

Value

Returns an object of class `foreccomb_res` with the following components:

`Method`	Returns the used forecast combination method.
`Models`	Returns the individual input models that were used for the forecast combinations.
`Intercept`	Returns the intercept (bias correction).
`Weights`	Returns the combination weights obtained by applying the combination method to the training set.
`Fitted`	Returns the fitted values of the combination method for the training set.
`Accuracy_Train`	Returns range of summary measures of the forecast accuracy for the training set.
`Forecasts_Test`	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
`Accuracy_Test`	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
`Input_Data`	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

See Also

comb_EIG1, foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG2(data)
```

---

comb_EIG3   *Trimmed Eigenvector Forecast Combination*

Description

Computes forecast combination weights according to the trimmed eigenvector approach by Hsiao
and Wan (2014) and produces forecasts for the test set, if provided.

Usage

```
comb_EIG3(x, ntop_pred = NULL, criterion = "RMSE")
```

Arguments

x	An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
ntop_pred	Specifies the number of retained predictors. If NULL (default), the inbuilt optimization algorithm selects this number.
criterion	If ntop_pred is not specified, a selection criterion is required for the optimization algorithm: one of "MAE", "MAPE", or "RMSE". If ntop_pred is selected by the user, criterion should be set to NULL (default).

Details

The underlying methodology of the trimmed eigenvector approach by Hsiao and Wan (2014) is
the same as their standard eigenvector approach. The only difference is that the trimmed
eigenvector approach pre-selects the models that serve as input for the forecast combination, only
a subset of the available forecast models is retained, while the models with the worst performance
are discarded.

The number of retained forecast models is controlled via ntop_pred. The user can choose whether
to select this number, or leave the selection to the inbuilt optimization algorithm (in that case
ntop_pred = NULL). If the optimization algorithm should select the best number of retained
models, the user must select the optimization criterion: MAE, MAPE, or RMSE. After this
trimming step, the weights and the combined forecast are computed in the same way as in the
standard eigenvector approach.

The trimmed eigenvector approach takes note of the eigenvector approaches' property to treat y and
$\mathbf{f}$ symmetrically, which bears the risk that the (non-trimmed) eigenvector approaches' performance
could be severely impaired by one or a few models that produce forecasts much worse than the
average.

Value

Returns an object of class foreccomb_res with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Top_Predictors	Number of retained predictors.
Ranking	Ranking of the predictors that determines which models are removed in the trimming step.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

See Also

comb_EIG1 foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

## Number of retained models selected by the user:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG3(data, ntop_pred = 2, criterion = NULL)

## Number of retained models selected by algorithm:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG3(data, ntop_pred = NULL, criterion = "RMSE")
```

comb_EIG4 *Trimmed Bias-Corrected Eigenvector Forecast Combination*

Description

Computes forecast combination weights according to the trimmed bias-corrected eigenvector approach by Hsiao and Wan (2014) and produces forecasts for the test set, if provided.

Usage

```
comb_EIG4(x, ntop_pred = NULL, criterion = "RMSE")
```

Arguments

x	An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
ntop_pred	Specifies the number of retained predictors. If NULL (default), the inbuilt optimization algorithm selects this number.
criterion	If ntop_pred is not specified, a selection criterion is required for the optimization algorithm: one of "MAE", "MAPE", or "RMSE". If ntop_pred is selected by the user, criterion should be set to NULL (default).

Details

The underlying methodology of the trimmed bias-corrected eigenvector approach by Hsiao and Wan (2014) is the same as their bias-corrected eigenvector approach. The only difference is that the bias-corrected trimmed eigenvector approach pre-selects the models that serve as input for the forecast combination, only a subset of the available forecast models is retained, while the models with the worst performance are discarded.

The number of retained forecast models is controlled via ntop_pred. The user can choose whether to select this number, or leave the selection to the inbuilt optimization algorithm (in that case ntop_pred = NULL). If the optimization algorithm should select the best number of retained models, the user must select the optimization criterion: MAE, MAPE, or RMSE. After this trimming step, the weights, the intercept and the combined forecast are computed in the same way as in the bias-corrected eigenvector approach.

The bias-corrected trimmed eigenvector approach combines the strengths of the bias-corrected eigenvector approach and the trimmed eigenvector approach.

Value

Returns an object of class foreccomb_res with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Intercept	Returns the intercept (bias correction).
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Top_Predictors	Number of retained predictors.

Ranking	Ranking of the predictors that determines which models are removed in the trimming step.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

See Also

comb_EIG2 comb_EIG3 foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

## Number of retained models selected by the user:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG4(data, ntop_pred = 2, criterion = NULL)

## Number of retained models selected by algorithm:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_EIG4(data, ntop_pred = NULL, criterion = "RMSE")
```

---

comb_InvW	*Inverse Rank Forecast Combination*

Description

Computes forecast combination weights according to the inverse rank approach by Aiolfi and Timmermann (2006) and produces forecasts for the test set, if provided.

Usage

```
comb_InvW(x)
```

Arguments

x	An object of class `foreccomb`. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

In the inverse rank approach by Aiolfi and Timmermann (2006), the combination weights are inversely proportional to the forecast model's rank, $Rank_i$:

$$w_i^{InvW} = \frac{Rank_i^{-1}}{\Sigma_{j=1}^N Rank_j^{-1}}$$

The combined forecast is then obtained by:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{InvW}$$

This is a robust variant of the Bates/Granger (1969) approach and also ignores correlations across forecast errors.

Value

Returns an object of class `foreccomb_res` with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Aiolfi, M., and Timmermann, A. (2006). Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics*, **135(1)**, 31–53.

Bates, J. M., and Granger, C. W. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, **20(4)**, 451–468.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_BG, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_InvW(data)
```

---

comb_LAD	*Least Absolute Deviation Forecast Combination*

Description

Computes forecast combination weights using least absolute deviation (LAD) regression.

Usage

```
comb_LAD(x)
```

Arguments

x   An object of class 'foreccomb'. Contains training set (actual values + matrix of
                 model forecasts) and optionally a test set.

Details

The function is a wrapper around the least absolute deviation (LAD) forecast combination implementation of the *ForecastCombinations* package.

The defining property of comb_LAD is that it does not minimize the squared error loss like comb_OLS and comb_CLS, but the absolute values of the errors. This makes the method more robust to outliers – comb_LAD tends to penalize models, which have high errors for some observations, less harshly than the least squares methods would.

Optimal forecast combinations under general loss functions are discussed by Elliott and Timmermann (2004). The LAD method is described in more detail, and used in an empirical context, by Nowotarksi et al. (2014).

The results are stored in an object of class 'foreccomb_res', for which separate plot and summary functions are provided.

Value

Returns an object of class `foreccomb_res` with the following components:

Method	Returns the best-fit forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Intercept	Returns the intercept of the linear regression.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

References

Elliott, G., and Timmermann, A. (2004). Optimal Forecast Combinations Under General Loss Functions and Forecast Error Distributions. *Journal of Econometrics*, **122(1)**, 47–79.

Nowotarski, J., Raviv, E., Tr\"uck, S., and Weron, R. (2014). An Empirical Comparison of Alternative Schemes for Combining Electricity Spot Price Forecasts. *Energy Economics*, **46**, 395–412.

See Also

Forecast_comb, foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_LAD(data)
```

---

comb_MED   *Median Forecast Combination*

Description

Computes a 'combined forecast' from a pool of individual model forecasts using their median at each point in time.

Usage

```
comb_MED(x)
```

Arguments

x   An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \dots, f_{Nt})'$. For each point in time, the median method gives a weight of 1 to the median forecast and a weight of 0 to all other forecasts, the combined forecast is obtained by:

$$\hat{y}_t = median(\mathbf{f}_t)$$

The median method is an appealing simple, rank-based combination method that has been proposed by authors such as Armstrong (1989), McNees (1992), Hendry and Clements (2004), Stock and Watson (2004), and Timmermann (2006). It is more robust to outliers than the simple average approach.

Value

Returns an object of class foreccomb_res with the following components:

Method   Returns the used forecast combination method.

Models   Returns the individual input models that were used for the forecast combinations.

Weights   Returns the combination weights obtained by applying the combination method to the training set.

Fitted   Returns the fitted values of the combination method for the training set.

Accuracy_Train   Returns range of summary measures of the forecast accuracy for the training set.

Forecasts_Test   Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.

Accuracy_Test   Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.

Input_Data   Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Armstrong, J. S. (1989). Combining Forecasts: The End of the Beginning or the Beginning of the End?. *International Journal of Forecasting*, **5(4)**, 585–588.

Hendry, D. F., and Clements, M. P. (2004). Pooling of Forecasts. *The Econometrics Journal*, **7(1)**, 1–31.

McNees, S. K. (1992). The Uses and Abuses of 'Consensus' Forecasts. *Journal of Forecasting*, **11(8)**, 703–710.

Stock, J. H., and Watson, M. W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting*, **23(6)**, 405–430.

Timmermann, A. (2006). Forecast Combinations. In: Elliott, G., Granger, C. W. J., and Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, **1**, 135–196.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_SA, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_MED(data)
```

---

comb_NG   *Newbold/Granger (1974) Forecast Combination*

Description

Computes forecast combination weights according to the approach by Newbold and Granger (1974) and produces forecasts for the test set, if provided.

Usage

```
comb_NG(x)
```

Arguments

x   An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

Building on early research by Bates and Granger (1969), the methodology of Newbold and Granger (1974) also extracts the combination weights from the estimated mean squared prediction error matrix.

Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$, Σ is the (positive definite) mean squared prediction error matrix of $\mathbf{f}_t$ and $\mathbf{e}$ is an $N \times 1$ vector of $(1, \ldots, 1)'$.

Their approach is a constrained minimization of the mean squared prediction error using the normalization condition $\mathbf{e}'\mathbf{w} = 1$. This yields the following combination weights:

$$\mathbf{w}^{NG} = \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}$$

The combined forecast is then obtained by:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{NG}$$

While the method dates back to Newbold and Granger (1974), the variant of the method used here does not impose the prior restriction that Σ is diagonal. This approach, called VC in Hsiao and Wan (2014), is a generalization of the original method.

Value

Returns an object of class `foreccomb_res` with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Bates, J. M., and Granger, C. W. (1969). The Combination of Forecasts. *Journal of the Operational Research Society*, **20(4)**, 451–468.

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

Newbold, P., and Granger, C. W. J. (1974). Experience with Forecasting Univariate Time Series and the Combination of Forecasts. *Journal of the Royal Statistical Society, Series A*, **137(2)**, 131–165.

See Also

comb_BG, comb_EIG1, foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_NG(data)
```

---

comb_OLS   *Ordinary Least Squares Forecast Combination*

Description

Computes forecast combination weights using ordinary least squares (OLS) regression.

Usage

```
comb_OLS(x)
```

Arguments

x   An object of class 'foreccomb'. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

The function is a wrapper around the ordinary least squares (OLS) forecast combination implementation of the *ForecastCombinations* package.

The OLS combination method (Granger and Ramanathan (1984)) uses ordinary least squares to estimate the weights, $\mathbf{w}^{OLS} = (w_1, \dots, w_N)'$, as well as an intercept, b, for the combination of the forecasts.

Suppose that there are N not perfectly collinear predictors $\mathbf{f}_t = (f_{1t}, \dots, f_{Nt})'$, then the forecast combination for one data point can be represented as:

$$y_t = b + \sum_{i=1}^{N} w_i f_{it}$$

An appealing feature of the method is its bias correction through the intercept – even if one or more of the individual predictors are biased, the resulting combined forecast is unbiased. A disadvantage of the method is that it places no restriction on the combination weights (i.e., they do not add up to 1 and can be negative), which can make interpretation hard. Another issue, documented in Nowotarski et al. (2014), is the method's unstable behavior when predictors are highly correlated (which is the norm in forecast combination): Minor fluctuations in the sample can cause major

shifts of the coefficient vector ('bouncing betas') – often causing poor out-of-sample performance. This issue is addressed by the comb_LAD method that is more robust to outliers.

The results are stored in an object of class 'foreccomb_res', for which separate plot and summary functions are provided.

Value

Returns an object of class foreccomb_res with the following components:

Method	Returns the best-fit forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Intercept	Returns the intercept of the linear regression.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

References

Granger, C., and Ramanathan, R. (1984). Improved Methods Of Combining Forecasts. *Journal of Forecasting*, **3(2)**, 197–204.

Nowotarski, J., Raviv, E., Tr\"uck, S., and Weron, R. (2014). An Empirical Comparison of Alternative Schemes for Combining Electricity Spot Price Forecasts. *Energy Economics*, **46**, 395–412.

See Also

Forecast_comb, foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_OLS(data)
```

comb_SA	*Simple Average Forecast Combination*

Description

Computes forecast combination weights using simple average and produces forecasts for the test set, if provided.

Usage

```
comb_SA(x)
```

Arguments

x   An object of class `foreccomb`. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Details

Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$. The simple average gives equal weights to all predictors:

$$\mathbf{w}^{SA} = \frac{1}{N}$$

The combined forecast is then obtained by:

$$\hat{y}_t = \mathbf{f}_t'\mathbf{w}^{SA}$$

It is well-documented that simple average is a robust combination method that is hard to beat (e.g., Stock and Watson, 2004; Timmermann, 2006). This is often associated with the importance of parameter estimation error in sophisticated techniques – a problem that simple averaging avoids. However, simple averaging may not be a suitable combination method when some of the predictors are biased (Palm and Zellner, 1992).

Value

Returns an object of class `foreccomb_res` with the following components:

Method	Returns the used forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Palm, F. C., and Zellner, A. (1992). To Combine or not to Combine? Issues of Combining Forecasts. *Journal of Forecasting*, **11(8)**, 687–701.

Stock, J. H., and Watson, M. W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting*, **23(6)**, 405–430.

Timmermann, A. (2006). Forecast Combinations. In: Elliott, G., Granger, C. W. J., and Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, **1**, 135–196.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
comb_SA(data)
```

---

comb_TA   *Trimmed Mean Forecast Combination*

Description

Computes a 'combined forecast' from a pool of individual model forecasts using trimmed mean at each point in time.

Usage

```
comb_TA(x, trim_factor = NULL, criterion = "RMSE")
```

Arguments

x	An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
trim_factor	numeric. Must be between 0 (simple average) and 0.5 (median).
criterion	If trim_factor is not specified, an optimization criterion for automated trimming needs to be defined. One of "MAE", "MAPE", or "RMSE" (default).

Details

Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$. For each point in time, the order forecasts are computed:

$$\mathbf{f}_t^{ord} = (f_{(1)t}, \ldots, f_{(N)t})'$$

Using a trim factor λ (i.e., the top/bottom $\lambda\%$ are trimmed) the combined forecast is calculated as:

$$\hat{y}_t = \frac{1}{N(1-2\lambda)} \sum_{i=\lambda N+1}^{(1-\lambda)N} f_{(i)t}$$

The trimmed mean is an interpolation between the simple average and the median. It is an appealing simple, rank-based combination method that is less sensitive to outliers than the simple average approach, and has been proposed by authors such as Armstrong (2001), Stock and Watson (2004), and Jose and Winkler (2008).

This method allows the user to select λ (by specifying `trim_factor`), or to leave the selection to an optimization algorithm – in which case the optimization criterion has to be selected (one of "MAE", "MAPE", or "RMSE").

Value

Returns an object of class `foreccomb_res` with the following components:

`Method`	Returns the used forecast combination method.
`Models`	Returns the individual input models that were used for the forecast combinations.
`Weights`	Returns the combination weights obtained by applying the combination method to the training set.
`Trim Factor`	Returns the trim factor, λ.
`Fitted`	Returns the fitted values of the combination method for the training set.
`Accuracy_Train`	Returns range of summary measures of the forecast accuracy for the training set.
`Forecasts_Test`	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
`Accuracy_Test`	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
`Input_Data`	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Armstrong, J. S. (2001). Combining Forecasts. In: *Armstrong, J. S. (Ed.), Principles of Forecasting. Springer, Boston, MA*, 417–439.

Jose, V. R. R., and Winkler, R. L. (2008). Simple Robust Averages of Forecasts: Some Empirical Results. *International Journal of Forecasting*, **24(1)**, 163–169.

Stock, J. H., and Watson, M. W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting*, **23(6)**, 405–430.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_SA, comb_MED, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

## User-selected trim factor:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_TA(data, trim_factor=0.1)

## Algorithm-optimized trim factor:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_TA(data, criterion="RMSE")
```

---

comb_WA   *Winsorized Mean Forecast Combination*

Description

Computes a 'combined forecast' from a pool of individual model forecasts using winsorized mean at each point in time.

Usage

```
comb_WA(x, trim_factor = NULL, criterion = "RMSE")
```

Arguments

x	An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
trim_factor	numeric. Must be between 0 and 0.5.
criterion	If trim_factor is not specified, an optimization criterion for automated trimming needs to be defined. One of "MAE", "MAPE", or "RMSE" (default).

Details

Suppose y_t is the variable of interest, there are N not perfectly collinear predictors, $\mathbf{f}_t = (f_{1t}, \ldots, f_{Nt})'$. For each point in time, the order forecasts are computed:

$$\mathbf{f}_t^{ord} = (f_{(1)t}, \ldots, f_{(N)t})'$$

Using a trim factor λ (i.e., the top/bottom $\lambda\%$ are winsorized), and setting $K = N\lambda$, the combined forecast is calculated as (Jose and Winkler, 2008):

$$\hat{y}_t = \frac{1}{N}\left[K f_{(K+1)t} + \sum_{i=K+1}^{N-K} f_{(i)t} + K f_{(N-K)t} \right]$$

Like the trimmed mean, the winsorized mean is a robust statistic that is less sensitive to outliers than the simple average. It is less extreme about handling outliers than the trimmed mean and preferred by Jose and Winkler (2008) for this reason.

This method allows the user to select λ (by specifying `trim_factor`), or to leave the selection to an optimization algorithm – in which case the optimization criterion has to be selected (one of "MAE", "MAPE", or "RMSE").

Value

Returns an object of class `foreccomb_res` with the following components:

`Method`	Returns the used forecast combination method.
`Models`	Returns the individual input models that were used for the forecast combinations.
`Weights`	Returns the combination weights obtained by applying the combination method to the training set.
`Trim Factor`	Returns the trim factor, λ.
`Fitted`	Returns the fitted values of the combination method for the training set.
`Accuracy_Train`	Returns range of summary measures of the forecast accuracy for the training set.
`Forecasts_Test`	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
`Accuracy_Test`	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
`Input_Data`	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

References

Jose, V. R. R., and Winkler, R. L. (2008). Simple Robust Averages of Forecasts: Some Empirical Results. *International Journal of Forecasting*, **24(1)**, 163–169.

See Also

winsor.mean, foreccomb, plot.foreccomb_res, summary.foreccomb_res, comb_SA, comb_TA, accuracy

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

## User-selected trim factor:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_TA(data, trim_factor=0.1)

## Algorithm-optimized trim factor:
data<-foreccomb(train_o, train_p, test_o, test_p)
comb_TA(data, criterion="RMSE")
```

---

cs_dispersion	*Compute Cross-Sectional Dispersion*

Description

Computes (time-varying) dispersion measures for the cross section of individual model forecasts that are the input of forecast combination.

Usage

```
cs_dispersion(x, measure = "SD", plot = FALSE)
```

Arguments

x	An object of class foreccomb. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
measure	Cross-sectional dispersion measure, one of: "SD" = standard deviation (default); "IQR" = interquartile range; or "Range" = range.
plot	logical. If TRUE, evolution of cross-sectional forecast dispersion is plotted as ggplot.

Details

The available measures of scale are defined as in Davison (2003). Let $y_{(i)}$ denote the i-th order statistic of the sample, then:

$$Range_t = y_{(n),t} - y_{(1),t}$$

$$IQR_t = y_{(3n/4),t} - y_{(n/4),t}$$

$$SD_t = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^n \left(y_{i,t} - \bar{y}_t\right)}$$

Previous research in the forecast combination literature has documented that regression-based combination methods tend to have relative advantage when one or more individual model forecasts are better than the rest, while eigenvector-based methods tend to have relative advantage when individual model forecasts are in the same ball park.

Value

Returns a vector of the evolution of cross-sectional dispersion over the sample period (using the selected dispersion measure)

References

Davison, A. C. (2003). Statistical Models. *Cambridge University Press*.

Hsiao, C., and Wan, S. K. (2014). Is There An Optimal Forecast Combination? *Journal of Econometrics*, **178(2)**, 294–309.

See Also

foreccomb, sd, IQR, range

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
cs_dispersion(data, measure = "IQR")
```

---

electricity	*UK Electricity Supply 2007 - 2017*

Description

The `electricity` dataset contains monthly data on the total UK electricity supply in GWh from January 2007 to March 2017, as well as univariate time series forecasts for this series.

Usage

```
data(electricity)
```

Format

A multivariate time series of 123 observations; monthly, 2007-2017.

This data contains the following columns:

arima (ARIMA 1-month forecasts)

ets (ETS 1-month forecasts)

nnet (Neural Network 1-month forecasts)

dampedt (Damped Trend 1-month forecasts)

dotm (Dynamic Optimized Theta 1-month forecasts)

Actual (Observed values)

Source

International Energy Agency (2017). IEA Monthly Electricity Statistics. Available at http://www.iea.org/statistics/monthlystatistics/monthlyelectricitystatistics/

---

foreccomb   *Format Raw Data for Forecast Combination*

Description

Structures cross-sectional input data (individual model forecasts) for forecast combination. Stores data as S3 class foreccomb that serves as input to the forecast combination techniques. Handles missing value imputation (optional) and resolves problems due to perfect collinearity.

Usage

```
foreccomb(observed_vector, prediction_matrix, newobs = NULL,
  newpreds = NULL, byrow = FALSE, na.impute = TRUE, criterion = "RMSE")
```

Arguments

observed_vector

    A vector or univariate time series; contains 'actual values' for training set.

prediction_matrix

    A matrix or multivariate time series; contains individual model forecasts for training set.

newobs   A vector or univariate time series; contains 'actual values' if a test set is used (optional).

newpreds   A matrix or multivariate time series; contains individual model forecasts if a test set is used (optional). Does not require specification of newobs – in the case in which a forecaster only wants to train the forecast combination method with a training set and apply it to future individual model forecasts, only newpreds is required, not newobs.

byrow   logical. The default (FALSE) assumes that each column of the forecast matrices (prediction_matrix and – if specified – newpreds) contains forecasts from one forecast model; if each row of the matrices contains forecasts from one forecast model, set to TRUE.

na.impute   logical. The default (TRUE) behavior is to impute missing values via the cross-validated spline approach of the mtsdi package. If set to FALSE, forecasts with missing values will be removed. Missing values in the observed data are never imputed.

criterion   One of "RMSE" (default), "MAE", or "MAPE". Is only used if prediction_matrix is not full rank: The algorithm checks which models are causing perfect collinearity and the one with the worst individual accuracy (according to the chosen criterion) is removed.

Details

The function imports the column names of the prediction matrix (if byrow = FALSE, otherwise the row names) as model names; if no column names are specified, generic model names are created.

The missing value imputation algorithm is a modified version of the EM algorithm for imputation that is applicable to time series data - accounting for correlation between the forecasting models and time structure of the series itself. A smooth spline is fitted to each of the time series at each iteration. The degrees of freedom of each spline are chosen by cross-validation.

Forecast combination relies on the lack of perfect collinearity. The test for this condition checks if prediction_matrix is full rank. In the presence of perfect collinearity, the iterative algorithm identifies the subset of forecasting models that are causing linear dependence and removes the one among them that has the lowest accuracy (according to a selected criterion, default is RMSE). This procedure is repeated until the revised prediction matrix is full rank.

Value

Returns an object of class foreccomb.

Author(s)

Christoph E. Weiss, Gernot R. Roetzer

References

Junger, W. L., Ponce de Leon, A., and Santos, N. (2003). Missing Data Imputation in Multivariate Time Series via EM Algorithm. *Cadernos do IME*, **15**, 8–21.

Dempster, A., Laird, N., and Rubin D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39(1)**, 1–38.

See Also

mnimput, rankMatrix

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

## Example with a training set only:
foreccomb(train_o, train_p)

## Example with a training set and future individual forecasts:
foreccomb(train_o, train_p, newpreds=test_p)

## Example with a training set and a full test set:
foreccomb(train_o, train_p, test_o, test_p)

## Example with forecast models being stored in rows:
preds_row <- matrix(rnorm(1000, 1), 10, 100)
train_p_row <- preds_row[,1:80]
foreccomb(train_o, train_p_row, byrow = TRUE)
```

```
## Example with NA imputation:
train_p_na <- train_p
train_p_na[2,3] <- NA
foreccomb(train_o, train_p_na, na.impute = TRUE)

## Example with perfect collinearity:
train_p[,2] <- 0.8*train_p[,1] + 0.4*train_p[,8]
foreccomb(train_o, train_p, criterion="RMSE")
```

plot.foreccomb_res	*Plot results from forecast combination model*

Description

Produces plots for the results of a forecast combination method. Either an actual vs. fitted plot (which = 1) or a barplot of the combination weights (which = 2).

Usage

```
## S3 method for class 'foreccomb_res'
plot(x, which = 1, ...)
```

Arguments

x	An object of class 'foreccomb_res'.
which	Type of plot: 1 = actual vs. fitted, 2 = combination weights.
...	Other arguments passing to plot.default.

Value

A plot for the foreccomb_res class.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

See Also

foreccomb, summary.foreccomb_res

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
fit <- comb_EIG1(data)
plot(fit)
```

---

rolling_combine	*Dynamic Forecast Combination*

Description

Computes the dynamic version of the combined forecast for a method included in the GeomComb package.

Usage

```
rolling_combine(x, comb_method, criterion = NULL)
```

Arguments

x	An object of class 'foreccomb'. Must contrain full training set and test set.
comb_method	The combination method that should be used.
criterion	Specifies loss criterion. Set criterion to either 'RMSE', 'MAE', or 'MAPE' for the methods comb_TA, comb_WA, comb_EIG3, and comb_EIG4, or to 'NULL' (default) for all other methods.

Details

The function rolling_combine allows to estimate a dynamic version of the other combination methods of the package in a standardized way, i.e., it allows for time-varying weights. The function builds on the idea of time series cross-validation: Taking the provided training set as starting point, the models are re-estimated at each period of the test set using a revised (increased) training set.

Like univariate dynamic forecasting, the validation approach requires a full test set – including the observed values.

The results are stored in an object of class 'foreccomb_res', for which separate plot and summary functions are provided.

Value

Returns an object of class foreccomb_res that represents the results for the best-fit forecast combination method:

Method	Returns the best-fit forecast combination method.
Models	Returns the individual input models that were used for the forecast combinations.
Weights	Returns the combination weights obtained by applying the best-fit combination method to the training set.
Fitted	Returns the fitted values of the combination method for the training set.
Accuracy_Train	Returns range of summary measures of the forecast accuracy for the training set.
Forecasts_Test	Returns forecasts produced by the combination method for the test set. Only returned if input included a forecast matrix for the test set.
Accuracy_Test	Returns range of summary measures of the forecast accuracy for the test set. Only returned if input included a forecast matrix and a vector of actual values for the test set.
Input_Data	Returns the data forwarded to the method.

Author(s)

Christoph E. Weiss

References

Bergmeir, C., Hyndman, R.J., and Koo, B. (2015). A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. *Monash University, Deparment of Econometrics and Business Statistics*, Working Paper No. 10/15.

Timmermann, A. (2006). Forecast Combinations. *Handbook of Economic Forecasting*, **1**, 135–196.

See Also

foreccomb, plot.foreccomb_res, summary.foreccomb_res,

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)

#Static forecast combination (for example OLS):
static_OLS <- comb_OLS(data)

#Dynamic forecast combination:
dyn_OLS <- rolling_combine(data, "comb_OLS")
```

---

summary.foreccomb_res   *Summary of Forecast Combination*

Description

summary method for class 'foreccomb_res'. Includes information about combination method, combination weights assigned to the individual forecast models, as well as an accuracy evaluation of the combined forecast.

Usage

```
## S3 method for class 'foreccomb_res'
summary(object, ...)

## S3 method for class 'foreccomb_res_summary'
print(x, ...)
```

Arguments

`object`	An object of class 'foreccomb'. Contains training set (actual values + matrix of model forecasts) and optionally a test set.
`...`	potential further arguments (require by generic)
`x`	An object of class 'foreccomb'. Contains training set (actual values + matrix of model forecasts) and optionally a test set.

Author(s)

Christoph E. Weiss and Gernot R. Roetzer

See Also

foreccomb, plot.foreccomb_res,

Examples

```
obs <- rnorm(100)
preds <- matrix(rnorm(1000, 1), 100, 10)
train_o<-obs[1:80]
train_p<-preds[1:80,]
test_o<-obs[81:100]
test_p<-preds[81:100,]

data<-foreccomb(train_o, train_p, test_o, test_p)
fit<-comb_BG(data)
summary(fit)
```

Index

Chapter 4

Efficient Nurse Staffing: The Value of Hierarchical Time Series Forecasting and Forecast Combination

4.1   Introduction

The dominant cost driver in the majority of service settings (e.g. restaurants, call centres, hospitals) is labour. While service providers may perceive reductions in staffing as a way of containing their costs, matching labour supply with demand is non-trivial when customer arrivals (e.g. diners to a restaurant, callers to a helpline, patients to a hospital) are stochastic. At the same time, the costs of being understaffed may be high. For example, staffing shortages have been associated with reductions in clinical safety (Kuntz et al., 2015) and reduced hospital reimbursement (Powell et al., 2012), as well as worse restaurant sales performance (Tan and Netessine, 2014). Although such associations between staffing levels and service outcomes continue to be explored in the operations literature, less attention has been paid to the potential mitigators. Furthermore, those that have been discussed each carry costs, for example, pooling customers may result in mismatches between need and skill (e.g. senior doctors treating minor injuries), while flexible staffing (e.g. Freeman et al., 2016) demands that higher wages be paid to offset workers' inconvenience. An alternative to a change in process is to improve the accuracy of the forecast itself, thereby ensuring that only those staff necessary to achieve the desired service level

are scheduled for any given shift. While this can be viewed purely as a forecasting problem (see Section 4.4 for an overview in the context of nurse staffing), there are also service-specific characteristics and operational objectives that make this problem specialised. This paper explores this further in the context of hospital nurse staffing.

Hospitals are an especially compelling setting in which to investigate models of staff forecasting. First, staffing in hospitals is very costly, with staffing cost (from e.g. physicians, nurses, support staff) constituting over 50% of total hospital expenses (Guarin-Calvert, 2011; Hurst and Williams, 2012). Second, in some countries, mandatory inpatient nurse-to-patient ratios have been imposed to improve overall working conditions in hospitals and to achieve higher standards of care. Ensuring that sufficient staff are available is therefore critical from both a regulatory as well as a safety perspective (see Section 4.3 for more details). Third, hospitals face two very different types of demand: scheduled (elective) patients and unscheduled (emergency) patients. While admissions of scheduled patients are known to the hospital in advance, the arrivals of emergencies are unpredictable and volatile. Furthermore, variation in patient length of stay (which applies to both elective and emergency patients) adds another degree of demand stochasticity, compounded further by differences in the intensity of patient arrivals and patient care needs. Fourth, hospitals are complex organisations that typically operate as a "plant-within-a-plant" (Skinner, 1974). The hierarchical structure, with medical specialties (e.g. cardiology, neurosurgery, pediatrics) nested within divisions (e.g. general medicine, surgical medicine, women and children's), offers opportunities in forecasting future demand that has, as yet, been unexploited (see Section 4.4). In this paper, we bring together these features of the hospital and apply a recently introduced forecasting technique, hierarchical time series (HTS) forecasting, as well as a range of forecast combination techniques, to the hospital context. We show how this model can be used to reduce staff overage while ensuring that the required nurse-to-patient ratios are achieved to a specified service level (e.g. 95% of the time). The forecasting approach offers opportunities for reducing staffing costs while maintaining, or even reducing, the percentage of understaffed shifts.

Using seven years of detailed patient-level data from a large teaching and research hospital in the United Kingdom, we find that the most common approach to nurse staffing – based only on patient census (i.e. the number of patients residing in the hospital at a given time) while ignoring churn (admissions and discharges) – lead to chronic understaffing. In our constrained optimisation analysis (see 4.7), we conclude that there are 3 factors that can considerably reduce the risk of understaffing: (a)

the possibility of hiring temporary staff, (b) accounting for patient churn in the staffing model, (c) accounting for forecast uncertainty in the staffing model using the empirical forecast error distribution. We find that a staffing model based only on census numbers and only allowing for permanent staff implies a probability of understaffing of almost 80%, which is reduced to about 70% by allowing for hiring temporary staff. This can further be improved by accounting for patient churn in the staffing model, lowering the probability of understaffing to about 20% (which is a particularly impressive improvement considering that staff cost related to churn only contributes 10% to total staff cost). Finally, we show that a serious commitment to full staffing requires accounting for forecast uncertainty in the staffing model using the empirical forecast error distribution, which reduces the probability of understaffing to 1%. Further, we show that using hierarchical or combined forecasts gets relatively more beneficial with rising staffing model complexity, compared to the benchmark forecast (a seasonal random walk) – leading to savings in staff cost of roughly 5% in the staffing model that accounts for all of the mentioned factors.

4.2   Institutional Setting

Health services in the United Kingdom face unprecedented financial and operational pressures: In the financial year 2015/16, hospitals in England have run up a record deficit of £2.45bn, even while key performance measures have deteriorated. This is despite an annual increase in the Department of Health budget in real terms in the period 2009 to 2021.

The challenge that healthcare providers in the UK face is that the budget – despite increasing – has not kept pace with the significant increases in demand for healthcare services.

4.2.1   Healthcare Budget

The Department of Health budget has increased in real terms every single year in the past 15 years – however, the growth rate has slowed down considerably in recent years. The King's Fund (2017) report that the between 2009/10 to 2020/21, the budget will only grow by an annual 1.1 percent in real terms, which amounts to an unprecedented slowdown in funding growth, and is far below the long-term average of 4 percent per year in real terms since the National Health Service (NHS) was created. This healthcare austerity can be attributed to a combination of the

macroeconomic effects of the Great Recession, as well as a general tendency towards neoliberal austerity of Conservative governments since the early 1980s (Ruckert and Labonté, forthcoming) – its devastating effects on British healthcare are documented by Loopstra et al. (2016).

4.2.2 Healthcare Demand

UK Healthcare demand has increased considerably and steadily over the 2000s. Maguire et al. (2016) review the determinants of this trend:

- *Emergency Care.* The increased demand for emergency care is evident in two measures – first, attendances at major accident and emergency (A&E) departments have risen by 18% between 2003/04 and 2015/16, an average annual increase of 1.4%. Much more significantly, emergency admissions to hospitals via major A&E departments has increased by 65% over this period, an average of 4.3% per year. Maguire et al. (2016) argue that the majority of this increase can be ascribed to increased short-stay admissions driven by the introduction of a 4-hour waiting time standard for A&E departments. Overall, this means that over a fourth of the arrivals at major A&E departments are now admitted, while it was less than a fifth in 2003/04.

- *Elective Care.* Over the same period, elective care admissions have also increased by an annual 4.3% on average.

There are two main drivers of the observed sharp increase in healthcare demand. First, UK population has grown by 10% over the period from 2003 to 2015. Second and more importantly, the number of people over 85 has increased by 40% over this period. Jointly, these determinants account for a large part of the 50% increase in admissions over this period. It is eminently plausible that both factors will continue to rise in the near future.

4.2.3 Implications on the Micro Level

These macropolitical decisions and an ageing population have important implications at the micro level. Demand for healthcare services has kept increasing by approximately 4 % per year from 2003 and is expected to grow at that pace for the next few years, according to an NHS England projection (Maguire et al., 2016). Ever since the funding slowdown started in 2010, this has led to a widening funding gap in

the British healthcare system (Figure 4.1, based on data in Maguire et al., 2016), causing the NHS to run up sizeable deficits in the past few years.
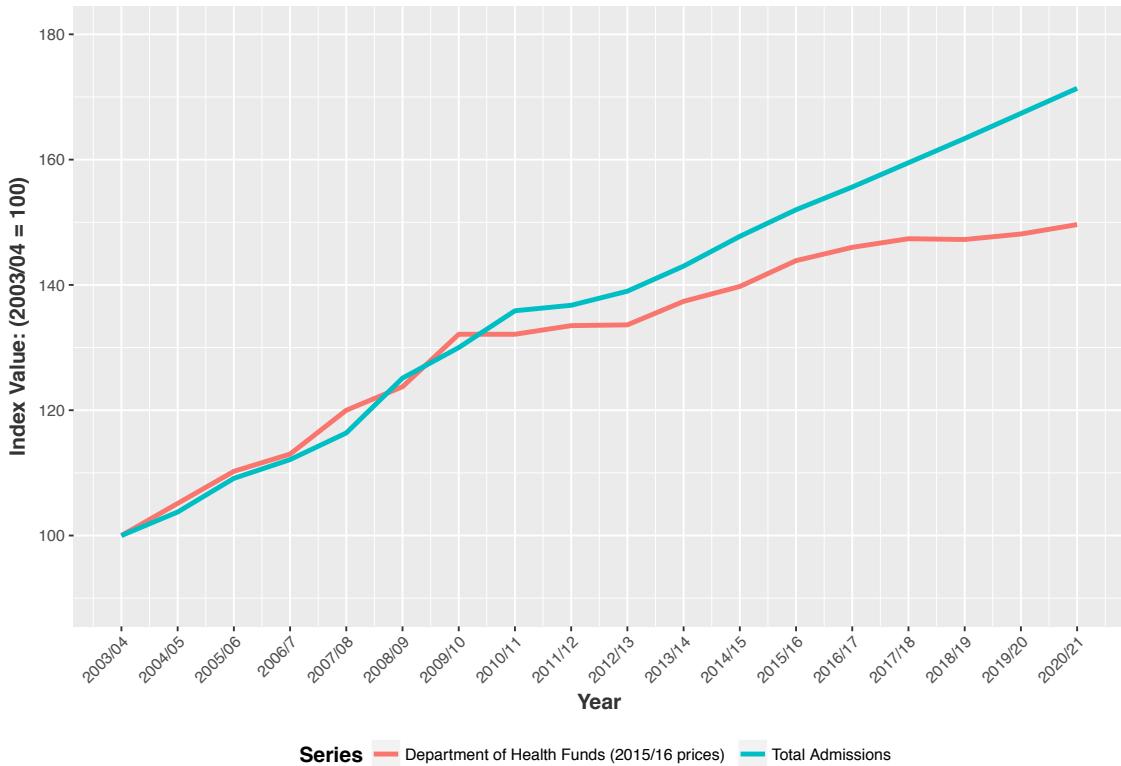

Fig. 4.1 Slowdown in Healthcare Budget Growth: Increasing Funding Gap.

Between 2017/18 and 2019/20, the situation cannot be expected to improve, since healthcare budget is only planned to increase by 0.6% annually (The King's Fund, 2017). The outlook for patient care is not a happy one – Robertson et al. (2017) argue that financial constraints have taken some time to adversely affect patient care, but their impact will only intensify, especially due to the cuts to staff and preventive services dictated by these constraints.

Since funding is unlikely to increase enough to close the funding gap any time soon, it is reasonable to expect that the pressure on healthcare providers to become more efficient will increase. The NHS was instructed to make savings of 4% per year over 2011-2015 (Hurst and Williams, 2012).

Robertson (2016) lists four possible responses of healthcare providers:

- running up deficits,

- improve productivity (delivering better value care),

- restrict access to services,

- dilute quality of services.

Running up sizeable deficits is not sustainable. The strategy to restrict access to services or dilute quality of services may bridge short-term financial constraints, but should be considered only as the last resort measure from the point of view of the needs of long-term healthcare management. Improving productivity is the unexceptionable approach to maintaining high service quality. Ham et al. (2016) argue that the NHS must focus on delivering better value care to the public. This requires a comprehensive and sustained commitment to quality improvement – including actions at both the micro level (tackling unwarranted variations in clinical care, reducing waste, becoming more patient-focused) and the macro level (ensuring that quality and safety are central topics of the health policy agenda).

4.3   Efficiency Gains Through Nurse Staffing

So what actions can be taken to deliver better value care? An immediate starting point is labour cost, which constitutes over 50% of hospital expenses (Guarin-Calvert, 2011; Hurst and Williams, 2012).

The growing funding gap has taken its toll and affected both short-term and long-term staffing structurally: cuts of front-line staff have created immense workload pressure, with the consequence that remaining staff are acting as shock absorbers, facing increased workload due to longer working hours to protect patient care. This led to higher perceived levels of stress and increased nurse absenteeism (Robertson et al., 2017). Empirical healthcare research on the relationship between workload, operational performance, and quality of care has established that workload is positively related with burnout in nurses (Greenglass et al., 2001; Vahey et al., 2004) and nurse absenteeism (Green et al., 2013), both of which are negatively related to nurses' professional efficacy. Furthermore, the adverse effect of increased workload on patient outcomes is well documented (e.g. Needleman et al., 2011; Kuntz et al., 2014).

This (admittedly, involuntary) development of hospitals towards an environment with even heavier workload than before is a major cause for concern. Obviously quality of care and finance are closely intertwined through opportunities to deliver better outcomes at lower cost (Alderwick et al., 2015). However, it should be noted that given the workforce's direct substantial impact on patient outcomes, this is

a highly complex cost-cutting exercise for the NHS, which has to take note of its effects on the delicate hospital ecosystem. Ham et al. (2016) argue that substantial productivity gains can only happen if the NHS manages to engage staff in this mission – making it crucial that the measures are perceived as a way to deliver better value rather than a simplistic focus on cost cutting.

A related challenge for efficient nurse staffing is the balance between permanent and temporary workforce. The NHS's struggle in recruitment and retention of permanent staff is widely acknowledged (National Audit Office, 2016). Since 55,000 of the NHS's 1.3 million workforce are citizens of other EU countries, the current Brexit negotiations – depending on outcome – could further exacerbate this problem (McKenna, 2016).

All told, the organisation of services to assure both high-quality patient care and cost efficiency is an important question for hospitals. The approach must deliver better value care without increasing workload to unsustainable levels (i.e. permanent understaffing). We take a constrained optimisation view: First, the modelling system aims at assuring a sustainable workload level in aid of productivity; second, the model minimises staff expenses under these circumstances.

4.3.1   Current Nurse Staffing Practices

Optimal nurse staffing is an enormous operational challenge, given the stochastic nature of both demand for hospital services and duration over which services are to be provided. A combination of both factors determines the key variables for nurse staffing: patient census (i.e. the number of patients residing in the hospital at a given time) and churn (i.e. the sum of admissions and discharges). The stochastic nature of these measures means that purely cost-efficient staffing (i.e. hospitals running close to full capacity) will occasionally cause unmanageably high levels of workload. This leads to a clear trade-off: Overstaffing is costly and unsustainable – a waste of financial resources that could be utilised to improve patient care by investment in new technology, funding of clinical trials, and research. Understaffing could jeopardise patient care and safety, as the research literature suggests: Several empirical studies document the impact of understaffing on a surge in deaths on weekends (e.g. Lamn, 1973; Rogot et al., 1976; Bell and Redelmeier, 2001; Aylin et al., 2010), an increase in medication errors (Frith et al., 2012), and higher probability of inpatient falls and hospital-acquired pressure ulcers (Staggs and He, 2013).

Nurse staffing is made more complex by the heterogeneous hierarchy of hospitals: Different divisions in the hospital will require different types of nurse deployment, and will differ in key measures for nurse staffing – for example due to the time-varying volatility of admissions to and discharges from the respective division. Chief nursing officers and other administrators predominantly optimise staffing on the division level. There are also advocates of centralised staffing who argue that only a centralised view of staffing needs can ensure an efficient allocation of staff throughout the entire organisation – i.e. an efficient matching of permanent and temporary nurses with patient care needs (Crist-Grundman and Mulrooney, 2011). This indicates the need for a combination of decentralised monitoring and centralised planning in order to achieve effective staffing and consequently a balance between quality, safety, labour costs, and staff satisfaction. Designing a staffing model that takes note of the hospital hierarchy is necessary.

4.4 Forecasting Demand for Hospital Services

4.4.1 Extant Forecasting Research for Nurse Staffing

The matching of health care demand with supply is the focus of a strand of literature that aims to forecast hospital admission numbers or occupancy using statistical models: for example, Lin (1989) forecasts hospital admission and discharge levels using ARIMA and Holt-Winters Exponential Smoothing models; Jones et al. (2008) forecast daily patient volumes at the emergency departments of three hospitals using several univariate time series models (SARIMA, Time Series Regression, Exponential Smoothing, and Artificial Neural Networks); Schweigler et al. (2009) uses SARIMA models to forecast short-term (4-hour and 12-hour) patient volume at three emergency departments; Jones et al. (2009) forecast emergency department volumes with a multivariate time series approach using series for different measures related to emergency department activity (arrivals, census, laboratory orders, etc.); Kam et al. (2010) use a multivariate SARIMA (a SARIMA model with meteorological explanatory variables) to forecast daily visits at an emergency department in Kore; Perry et al. (2010) use exponentially weighted moving average (EWMA) and fast orthogonal search models to forecast emergency department visits; and Koestler et al. (2013) use a Poisson Autoregressive model that is suitable for forecasting patient census, incorporating patient-level information into the model.

Most of the extant research on time series forecasting for healthcare demand documents promising results – however, previous empirical research focused forecasting on specific time series to forecast. To our best knowledge, there is no previous forecasting research that aims to create a comprehensive, integrated system that takes note of the hospital hierarchy in the sense that it incorporates information at all hierarchical levels (aggregate, divisions, specialties, etc.) into the forecast.

The above mentioned extant studies tend to focus on either census or churn as predicted time series, which makes them difficult to apply in this context – nurse allocation should be based on both census and churn (Tierney et al., 2013).

The purpose of predictions is to support chief nursing officers with efficiently allocating both permanent and temporary staff. Thus a good forecasting system for nurse staffing has to manage demand uncertainty in two stages, conceptually consistent with Bard and Purnomo (2005) and Kim and Mehrotra (2015): an early forecasting stage (planning stage) with the aim to find initial staffing levels and schedules and plan deployment of permanent staff accordingly, and a later forecasting stage (adjustment stage) to modify these schedules closer to the actual time of demand realisation, deploying temporary staff if required.

Nurse staffing should also take into account potential legal requirements such as mandatory nurse-to-patient ratios (Buerhaus, 1997; Ghosh and Cruz, 2005; Reiter et al., 2012). As of now, there is no mandatory nurse staffing legislation in the UK. Some professional organisations have produced guidelines for safe nurse staffing (e.g. Royal College of Nursing, 2010), focussed on specialised units, such as intensive care or neonatal services.

To summarise, these are the features of a comprehensive, integrated forecasting system for the purpose of nurse staffing:

- It applies to the entire hospital hierarchical system

- It takes account of the main determinants of nurse workload, patient census and churn;

- It allows for a planning stage (deploying permanent staff) and an adjustment stage (deploying temporary staff);

- It accounts for potential mandatory regulatory requirements and voluntary hospital targets;

- It allows for sensitivity analysis with respect to key parameters, e.g. the advised nurse hours per patient day (NHPPD) number for specific departments.

4.4.2 Proposed Forecasting Approach

We proceed to formulate a model comprising of a comprehensive forecasting system for nurse staffing by combining two powerful forecasting tools that suit the purpose very well: (i) Hierarchical Time Series (HTS) Forecasting and (ii) Forecast Combination.

Hierarchical Time Series (HTS) Forecasting

The hierarchical time series (HTS) framework by Athanasopoulos et al. (2009) and Hyndman et al. (2011) is a disaggregated forecasting framework that allows for the joint forecast of a variable that is observed on all levels of a hierarchy. In the hospital context, this means that using HTS it is possible to jointly forecast patient census for the entire hospital, each division, and each primary specialty. HTS forecasting is designed to generate aggregation-consistent forecasts, i.e. the sum of forecasts for all primary specialties feeding into a division will equal the forecast for the division, and the sum of forecasts for all divisions will equal the aggregate hospital forecast.

We illustrate the conceptual framework of a hierarchical time series using a simplified hospital structure: Assume a hospital only has 2 divisions, with 3 primary specialties feeding into the first division and 2 into the second. The simplified hospital structure is shown in Figure 4.2.
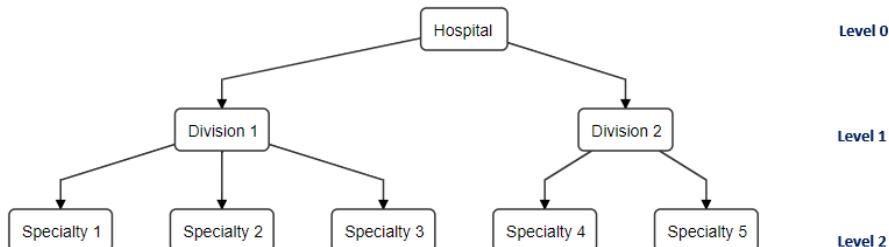

Fig. 4.2 Simplified Hospital Hierarchy.

Expressed in matrix form, this illustrative hospital structure (e.g. for patient census at all observed levels, denoted by the C in the matrix) constitutes a hierarchical time series ($\mathbf{Z}_t$), which is formally expressed in terms of the product of the summation matrix which takes note of the hierarchy ($\mathbf{S}$) with the bottom-level series ($\mathbf{z}_{K,t}$) – in our case the patient census values for the 5 primary specialties. The total number of time series in a hierarchy with K levels is given by $n = 1 + n_1 + \cdots + n_K$ where n_i is the number of series at level i of the hierarchy, so in this illustration $n = 1 + 2 + 5 = 8$. Of course the actual hospital structure with 8 divisions and 140

primary specialties feeding into them is a much more challenging application of HTS forecasting, constituting a hierarchy with a total of 149 time series.

$$
\underbrace{\begin{pmatrix} C_{\text{Hosp},t} \\ C_{\text{Div1},t} \\ C_{\text{Div2},t} \\ C_{\text{Spec1},t} \\ C_{\text{Spec2},t} \\ C_{\text{Spec3},t} \\ C_{\text{Spec4},t} \\ C_{\text{Spec5},t} \end{pmatrix}}_{\mathbf{Z}_t} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{S}} \times \underbrace{\begin{pmatrix} C_{\text{Spec1},t} \\ C_{\text{Spec2},t} \\ C_{\text{Spec3},t} \\ C_{\text{Spec4},t} \\ C_{\text{Spec5},t} \end{pmatrix}}_{\mathbf{z}_{K,t}}
$$

The current implementation of the HTS framework provides 4 different aggregation schemes for hierarchical forecasts:

- **Top-Down** methods forecast the aggregate time series and then disaggregate to produce forecasts for lower levels, using either historical proportions (Gross and Sohl, 1990) or forecast proportions (Athanasopoulos et al., 2009).

- **Bottom-Up** methods forecast the series at the bottom level and then aggregate up to forecasts at higher levels. Hyndman et al. (2011) find that the bottom-up aggregation scheme usually performs better than a top-down forecast, by virtue of utilising disaggregated information. However, noisy bottom-level series might introduce bias that diminish or even fully eliminate the positive value of disaggregated information.

- **Middle-Out** methods constitute a hybrid approach that forecast at intermediate levels of disaggregation, subsequently aggregating to form forecasts for more aggregated levels (using a bottom-up approach), and disaggregating to form forecasts for less aggregated levels (using a top-down approach).

- **Optimal Reconciliation**[1] approach starts by creating forecasts for each single series in the hierarchy without taking the hierarchical structure into account

---

[1]Note that the 'Optimal Reconciliation' forecast is often called 'Optimal Combination' forecast, like in Chapter 2 of this thesis. We use this terminology here to avoid confusing with forecast combination techniques that are also used in this study. The difference is worth explaining: While forecast combination techniques combine individual forecasts for the same time series, the 'Optimal Reconciliation' method forms a revised (aggregation-consistent) hierarchical forecast by optimally reconciling individual (aggregation-inconsistent) forecasts for the different time series of the underlying hierarchy.

at that point, i.e. the initial forecasts will not be aggregation-consistent. The forecasts are subsequently reconciled to satisfy the condition of aggregation consistency using one of the following methods: ordinary least squares (Hyndman et al., 2011), weighted least squares (Hyndman et al., 2016a), the game-theoretically optimal method (Van Erven and Cugliari, 2015), or the minimum trace method (Wickramasuriya et al., 2017). In this study, the 'Optimal Reconciliation' forecasts are computed with weighted least squares, using the R package **hts** by Hyndman et al. (2016b).

The individual forecasts in the HTS framework can be obtained from any time series forecasting model. In this study, we create forecasts using 4 different univariate forecasting approaches that have been popular in forecasting research:

- **Seasonal ARIMA** (Box and Jenkins, 1970)

- **ETS** (Pegels, 1969; Gardner, 1985; Hyndman et al., 2002; Taylor, 2003)

- **Damped Trend** (Gardner and McKenzie, 1985)

- **Theta Model** (Assimakopoulos and Nikolopoulos, 2000; Fiorucci et al., 2016b)[2]

For a detailed description of these forecasting methods, see Section 2.4 of this thesis. In addition, we also estimate a seasonal naïve random walk forecast that serves as benchmark. In our context this corresponds to simply using the values of census and churn for the current shift as the forecast for the same shift in the following week. In the selection of forecasting method, we required the methods to incorporate both non-stationarity and seasonality into the modelling procedure – while the Theta model in general does not take seasonality into account, the algorithm chosen for its estimation – using the R package **forecTheta** by Fiorucci et al. (2016a) – relies on deseasonalising the series prior to estimation using a seasonality test, and reseasonalising it after estimation.

Forecast Combination

Forecast combination describes the idea of combining (generally, linearly) different forecasts for the same time series, for example, from different models, or different

---

[2]For the Theta model forecast, we use the dynamically optimised Theta model by Fiorucci et al. (2016b), a generalisation of the standard Theta model that showed promising results in early empirical applications. Since the individual forecasts are used as input to forecast combination methods in this study, we refrain from estimating both versions of the Theta model in order to avoid excessively high levels of multicollinearity.

experts. First proposed almost five decades ago in the seminal paper by Bates and Granger (1969), myriad methods for combining forecasts have been proposed – each with its own merits and drawbacks. Overall, a large stock of empirical research has built up, documenting that forecast combination is an appealing strategy to decrease forecast risk (i.e. probability of obtaining a highly inaccurate forecast) and can in some cases improve accuracy, even compared to the best component forecast. Reviewing forecast combination techniques in detail is not the focus of this chapter and the reader is referred to excellent review studies, such as Clemen (1989), Armstrong (2001) or Timmermann (2006).

The idea of (linear) forecast combination methods is described in the following formula:

$$f^c{}_t = \sum_{i=1}^{N} w_{it} f_{it}$$

where N is the number of component forecasts that should be used in the forecast combination, f_{it} is the ith component forecast, w_{it} is the weight that the employed combination method assigned to the ith component forecast, and $f^c{}_t$ is the revised (combined) forecast.

In this study, we use the HTS forecasts as component forecasts and apply 14 different methods to compute combination weights in order to create revised (combined) forecasts. The combination methods we use are estimated in R using the **GeomComb** package by Weiss and Roetzer (2016) and can be categorised into 3 groups:

- **Simple Methods:** Simple Average (SA), Median (MED), Trimmed Mean (TM), Winsorised Mean (WM), Bates/Granger (BG), Newbold/Granger (NG), and Inverse Rank (InvR);

- **Regression-Based Methods:** Ordinary Least Squares (OLS), Constrained Least Squares (CLS), and Least Absolute Deviation (LAD);

- **Geometric (Eigenvector-Based) Methods:** Standard Eigenvector (EIG1), Bias-Corrected Eigenvector (EIG2), Trimmed Eigenvector (EIG3), and Trimmed Bias-Corrected Eigenvector (EIG4).

For a detailed description of the combination schemes, see Section 3.2.1 of this thesis and references therein.

4.5   The Data

The dataset we use for the empirical analysis reported in this chapter was provided by a large UK teaching hospital. It includes exact times of all inpatient admissions and discharges of both emergency and elective patients, including information on the primary specialty to which the cases were assigned, from 1st January 2007 to 20th December 2013. An advantage of having exact times of these churn events, instead of daily count data as in many previous forecasting studies, is the possibility to create counts for smaller intervals – for example, shift intervals.[3] This is valuable, as it is then possible to estimate the seasonal pattern shift by shift (i.e. the time series has a frequency of 14) instead of only weekdays (i.e. a frequency of 7).

The data includes a total of 3,400,053 churn events – 1,700,053 of which were admissions, the remaining 1,700,000, discharges. These were used to compute shift-level counts of churn and of census numbers for each primary specialty in the following way:

$$\text{Churn}_{i,t} = \text{Admissions}_{i,t} + \text{Discharges}_{i,t}$$
$$\text{Census}_{i,t} = \text{Census}_{i,t-1} + \text{Admissions}_{i,t} - \text{Discharges}_{i,t}$$

where i is the respective primary specialty and t is the time of the shift.

In this study, we limit our attention to two divisions of the hospital for illustration purposes – the medical and the surgical units; it should be noted that the method can readily be extended to forecast patient census and churn for the entire hospital hierarchy (i.e. aggregate, emergency/electives, divisions, primary specialties), so that the model can be used for staff planning at any desired level of disaggregation. Thus the model is suitable for hospitals with either centralised or decentralised staffing. Figures 4.3 to 4.6 display the time series behaviour of the census and churn series for the divisions analysed. While all the series show seasonality as a dominant feature (confirmed by the autocorrelation functions), the differences in the histograms are striking: the census series are very well-behaved and approximate normality; the churn series, however, very clearly follow two separate distributions, as indicated by the bimodal mixture distribution. This is expected in a hospital context: Given the strong demand for healthcare services, it is reasonable to expect hospitals to operate

---

[3]The hospital divides a day into 2 shifts: The day shift is from 7:30am to 7:29pm, the night shift is from 7:30pm to 7:29am.

close to full capacity at all times, keeping the census relatively stable; however, most admissions and in particular discharges will be handled during day shifts rather than at night time. This feature makes the churn series more challenging to forecast, since linear time series analysis usually assumes both Gaussian marginal and conditional distribution of the data (Wong and Li, 2000). At the same time, it emphasises the importance of using shift-level data – with daily data, all this information would be lost by aggregation of the two underlying distributions.[4]
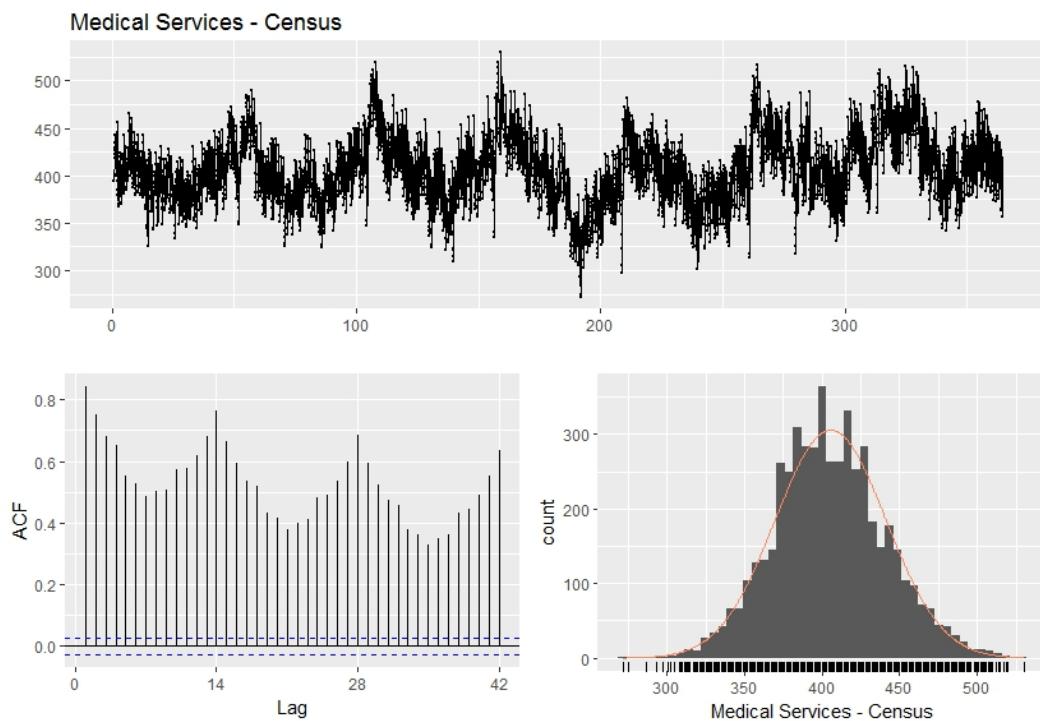

Fig. 4.3 Medical Services - Census: Time Series Behaviour.

---

[4]A case could be made for creating separate forecasts for (a) day-shift churn and (b) night-shift churn. While this is a valid approach, it should be seen as a measure of last resort, since it decreases the information that is used as input in either of the two univariate forecasting models – this approach is only advisable if the bimodal distribution carries over to the distribution of forecast errors of the full forecast model that uses both day-shift and night-shift churn, which would indicate that the model does not manage to capture the seasonal pattern well.
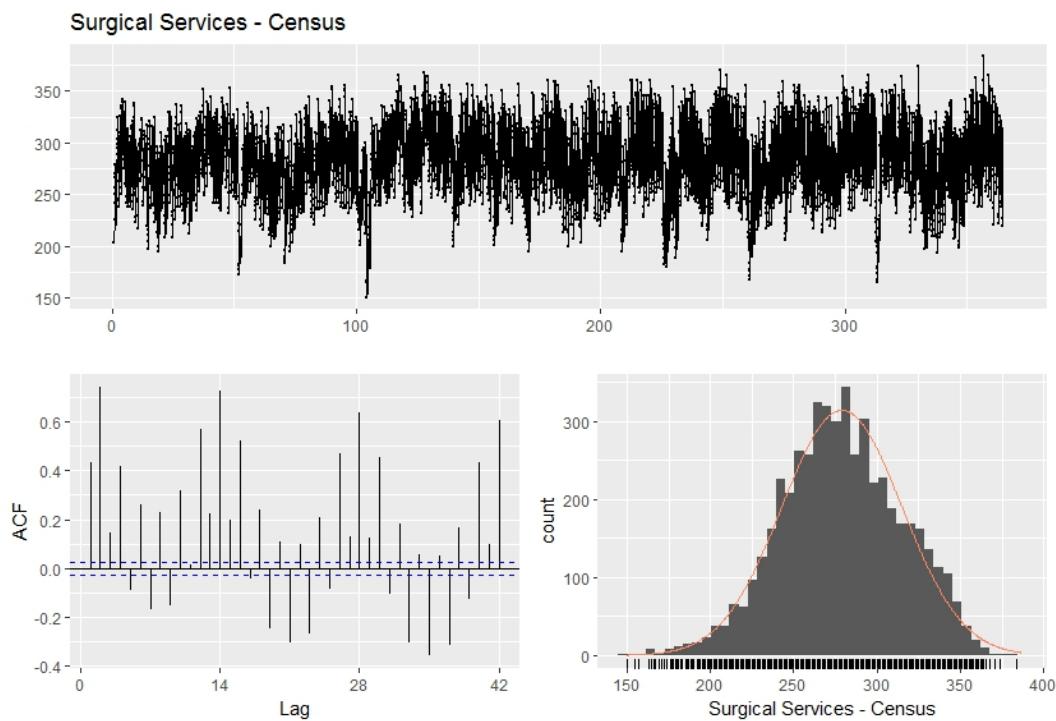
Fig. 4.4 Surgical Services - Census: Time Series Behaviour.
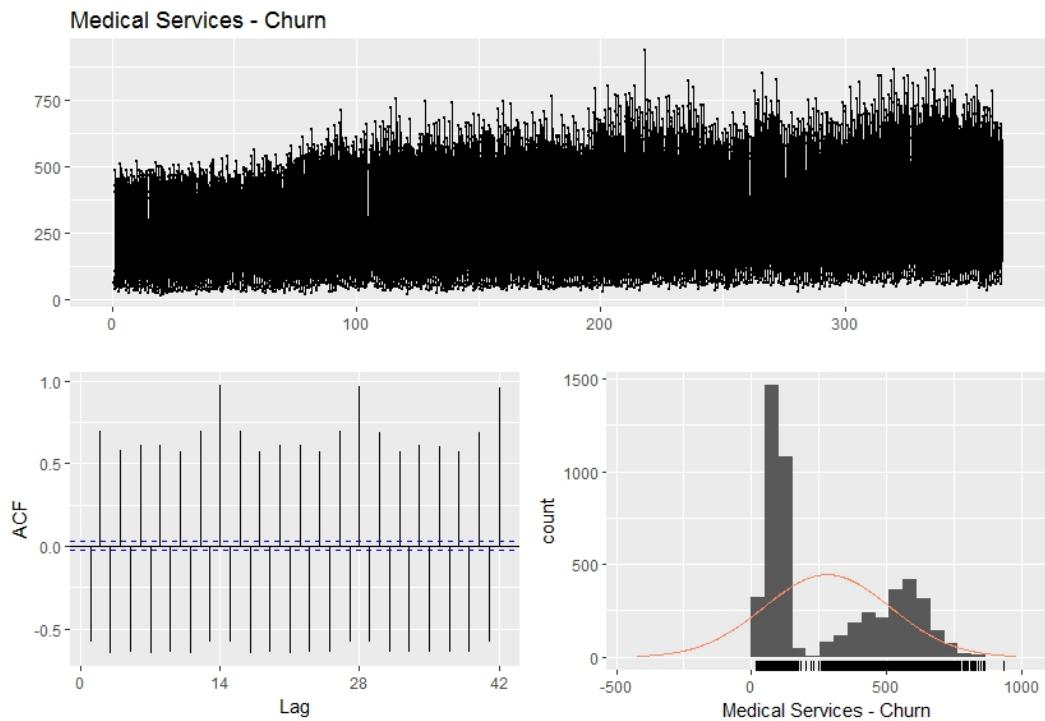
Fig. 4.5 Medical Services - Churn: Time Series Behaviour.
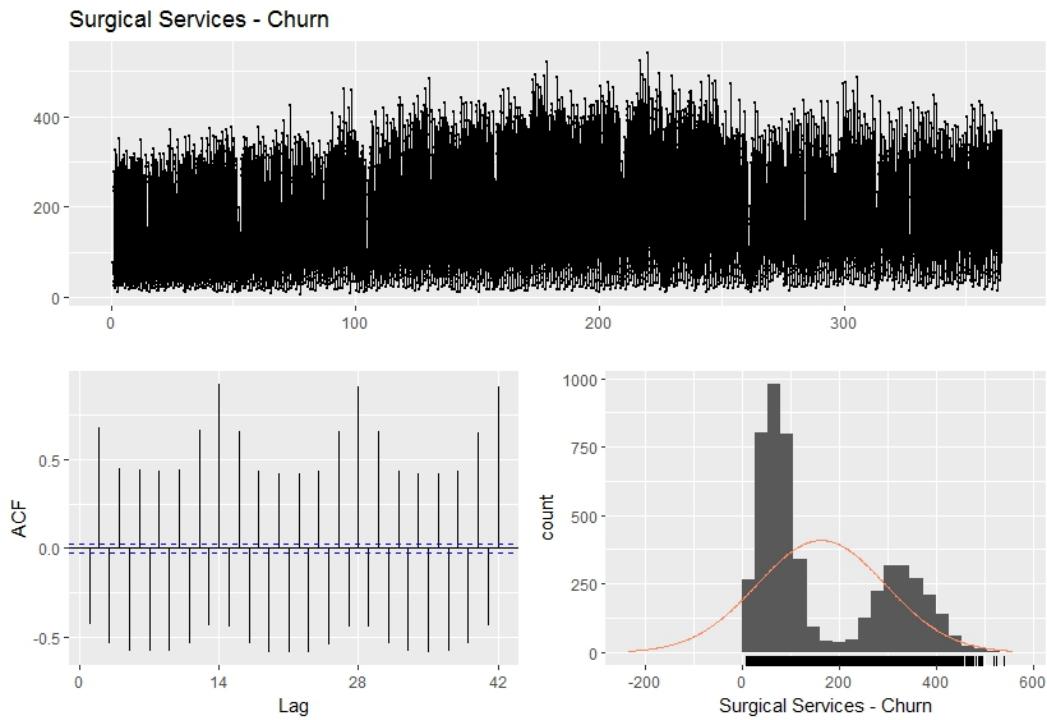
Fig. 4.6 Surgical Services - Churn: Time Series Behaviour.

There are 33 primary specialties feeding into the Medical Services division, and 42 feeding into the Surgical Services division. A full list is presented in Appendix 4.A. Figure 4.7 shows the (simplified) hierarchy for division-level census and churn forecast for only 2 specialties feeding into a division. This shows that for divisional census forecasts, the HTS framework provides 3 separate forecasts for each of the 4 forecasting approaches: a top-down forecast (forecasts divisional census directly), a bottom-up forecast (forecasts specialty-level census and subsequently aggregates), and the 'Optimal Reconciliation' approach – i.e. including the seasonal random walk forecast there are 13 separate HTS forecasts that can potentially be used as component forecasts for forecast combination methods. For divisional churn, each forecasting approach delivers 4 forecasts: top-down (forecasting divisional churn directly), middle-out (forecasting specialty-level churn and subsequent aggregation), bottom-up (forecasting specialty-level admissions and discharges and subsequent aggregation), and the 'Optimal Reconciliation' approach – giving us a potential of 17 separate forecasts (again, including the seasonal random walk) for combined churn forecasts.
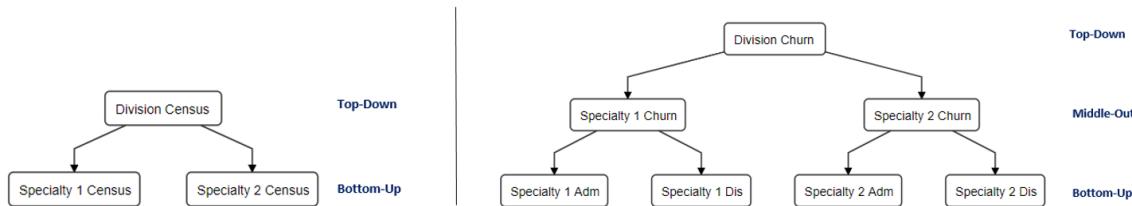
Fig. 4.7 Simplified Division-Level Hierarchy: (left) census, (right) churn.

4.6 Evaluation of Forecasting Results

Using the forecasting methods described above, candidate forecasts were produced for the following series:

- **Census – Medical Services:** 26 separate forecasting models (13 HTS; 13 Forecast Combinations – constrained least squares showed convergence issues and was discarded)

- **Census – Surgical Services:** 27 separate forecasting models (13 HTS; 14 Forecast Combinations)

- **Churn – Medical Services:** 27 separate forecasting models (13 HTS – SARIMA models were discarded due to estimation issues[5]; 14 Forecast Combinations)

- **Churn – Surgical Services:** 27 separate forecasting models (13 HTS – SARIMA models were discarded due to estimation issues; 14 Forecast Combinations)

To allow for a two-step staffing approach (planning step, adjustment step), we produced long-term (28-shift, i.e. 14-day) and short-term (2-shift, i.e. 1-day) forecasts. The forecasting procedure used a time series cross-validation approach (Bergmeir et al., 2015): Pseudo-out-of-sample HTS forecasts were estimated for the final 200 shifts (i.e. 100 days) of the sample by re-estimating the models with increasing training sets. Since most of the forecast combination techniques require a training set of component forecasts (in our case, of HTS forecasts) to establish stable combination

---

[5]Due to the very large number of series to forecast, HTS heavily relies on automated forecasting methods. While the criterion-based automated ARIMA estimation by Hyndman and Khandakar (2008) is popularly applied in forecasting research and practice, it is known to be prone to model misspecification when seasonal differencing is required, which is very likely part of the explanation for this method's bad performance. Another feature of this data that represents a major challenge for automated ARIMA modelling is its bimodal distribution explored in Section 4.5.

weights, we computed combined forecasts for the final 100 shifts, also using a time series cross-validation approach based on re-estimating the combination models with increasing training sets, allowing for time-varying combination weights.

For the accuracy evaluation of the forecasts, we use mean absolute percentage error (MAPE) – this is easy to interpret and consistent with previous research on forecasting hospital time series (e.g. Jones et al., 2008; Carvalho-Silva et al., forthcoming). Hyndman and Koehler (2006) document problematic performance of the measure in cases when a series includes values close to zero, but since the level of divisional patient census and churn is much higher than zero at all times, this is not concerning in this study.

Tables 4.1 to 4.4 present the MAPE values for the pseudo-out-of-sample forecasts for the final 100 shifts of the sample for the 4 series of interest. The results provide insights:

- Patient census can be forecast much more accurately than patient churn for both divisions.

- Forecasts for the medical division are slightly more accurate than the ones for the surgical division – this holds for both census and churn.

- Long-term forecasts that are used for scheduling permanent staff are not that much worse than short-term forecasts for census, suggesting that the series is mainly driven by the seasonal pattern; curiously, for churn the results show that the long-term forecasts are even slightly more accurate on average than short-term ones. The good overall performance of the long-term forecasts essentially means that the majority of staff needs can be filled with permanent staff, which may be positive for patient outcomes – Hockenberry and Becker (2016) find that a higher proportion of nursing hours provided by agency nurses significantly lowers both patient satisfaction and nurses' communication with patients.

- HTS forecasts and forecast combinations can improve upon the benchmark model – a seasonal random walk, that essentially coincides with the hospital's current staffing method – considerably: The long-term MAPE for the benchmark is 1.11 times the error from the best forecast for census in the medical division, 1.09 for census in the surgical division, 1.29 for churn in the medical division, and 1.11 for churn in the surgical division. This is even more severe for the short-term MAPE, for which the benchmark error is 1.54 times the

error from the best forecast for census in the medical division, 1.80 for census in the surgical division, 1.31 for churn in the medical division, and 1.26 for churn in the surgical division.

- For long-term forecasts, HTS models produce the best forecast for two of the analysed series, while forecast combinations perform best for the other two series. For short-term forecasts, forecast combination seems even more valuable, producing the best forecast for all but one of the series. For the HTS forecasts, the exponential smoothing-based methods ETS and Damped Trend tend to produce the most accurate census forecasts, which also holds for most of the churn forecasts, although a top-down Theta model produces the best long-term forecast for churn in the medical division.

- The results confirm that forecast combination can considerably reduce forecast risk, with the range of MAPE values that the different combination methods produce being much smaller than the correspondent ranges for individual forecast models.

The best forecast method identified can be used to produce the forecast of "patient days" per shift, which can then be used to compute the required number of nurses, using the hospital's target NHPPD value (see assumptions of the model in the following section). The hospital's quality assurance policy may require that the target NHPPD value is hit in 95 % of the shifts – i.e. only a maximum of 5 % of the shifts should be understaffed. Therefore, we also have to take into account the 95 % quantiles of the forecast error distributions in order to adjust the mean forecast, so that this requirement is satisfied. Figures 4.8 to 4.11 show the empirical distribution of percentage errors for the best model for each of the series of interest, comparing it to the corresponding distribution of the naïve forecast (the seasonal random walk) – the 95 % quantiles are indicated by the vertical lines. The plots show that in each case, the identified best forecast model also has a lower 95 % quantile of the percentage error distribution than the seasonal random walk forecast. This is in line with the results presented by Barrow and Kourentzes (2016): combined forecasts lead to reduced forecast error variance, and, – unlike for univariate forecasts – out-of-sample distributions are consistent with in-sample ones.

Table 4.1 MAPE Forecast Evaluation: Census – Medical Services.

Forecast Method	Long-Term (14d)	Short-Term (1d)
Seasonal Random Walk	3.16 %	3.47 %
HTS Forecasts		
Arima (Top Down)	4.48 %	2.62 %
Arima (Bottom Up)	4.42 %	2.76 %
Arima (Optimal Reconciliation)	4.45 %	2.62 %
ETS (Top Down)	2.91 %	2.40 %
ETS (Bottom Up)	2.91 %	2.28 %
ETS (Optimal Reconciliation)	2.91 %	2.40 %
Theta Model (Top Down)	3.26 %	2.70 %
Theta Model (Bottom Up)	3.46 %	3.10 %
Theta Model (Optimal Reconciliation)	3.26 %	2.70 %
Damped Trend (Top Down)	2.91 %	2.41 %
Damped Trend (Bottom Up)	**2.84 %**	2.29 %
Damped Trend (Optimal Reconciliation)	2.91 %	2.40 %
Combined Forecasts		
Simple Average	3.10 %	2.27 %
Median	2.89 %	**2.26 %**
Trimmed Mean	2.91 %	2.27 %
Winsorised Mean	2.94 %	2.29 %
Bates/Granger	3.03 %	2.29 %
Newbold/Granger	3.46 %	2.29 %
Inverse Rank	2.96 %	2.28 %
OLS	3.48 %	2.28 %
LAD	3.52 %	2.41 %
EIG1	3.12 %	2.28 %
EIG2	3.41 %	2.29 %
EIG3	2.95 %	2.28 %
EIG4	3.34 %	2.30 %

Table 4.2 MAPE Forecast Evaluation: Census – Surgical Services.

Forecast Method	Long-Term (14d)	Short-Term (1d)
Seasonal Random Walk	5.58 %	7.16 %
HTS Forecasts		
Arima (Top Down)	6.33 %	4.95 %
Arima (Bottom Up)	6.71 %	5.04 %
Arima (Optimal Reconciliation)	6.33 %	4.95 %
ETS (Top Down)	6.20 %	4.10 %
ETS (Bottom Up)	5.39 %	4.06 %
ETS (Optimal Reconciliation)	6.19 %	4.09 %
Theta Model (Top Down)	5.47 %	4.27 %
Theta Model (Bottom Up)	5.46 %	4.12 %
Theta Model (Optimal Reconciliation)	5.47 %	4.26 %
Damped Trend (Top Down)	5.50 %	4.05 %
Damped Trend (Bottom Up)	5.40 %	4.03 %
Damped Trend (Optimal Reconciliation)	5.50 %	4.05 %
Combined Forecasts		
Simple Average	5.62 %	4.12 %
Median	5.49 %	4.10 %
Trimmed Mean	5.49 %	4.13 %
Winsorised Mean	5.50 %	4.15 %
Bates/Granger	5.59 %	4.11 %
Newbold/Granger	5.57 %	3.99 %
Inverse Rank	5.49 %	4.04 %
OLS	5.19 %	4.09 %
CLS	5.59 %	**3.97 %**
LAD	**5.12 %**	4.26 %
EIG1	5.63 %	4.13 %
EIG2	5.59 %	4.19 %
EIG3	5.40 %	4.14 %
EIG4	5.39 %	4.16 %

Table 4.3 MAPE Forecast Evaluation: Churn – Medical Services.

Forecast Method	Long-Term (14d)	Short-Term (1d)
Seasonal Random Walk	16.19 %	16.87 %
HTS Forecasts		
ETS (Top Down)	13.70 %	13.55 %
ETS (Middle Out)	13.34 %	**12.88 %**
ETS (Bottom Up)	13.69 %	13.64 %
ETS (Optimal Reconciliation)	13.69 %	13.53 %
Theta Model (Top Down)	**12.52 %**	12.89 %
Theta (Middle Out)	14.58 %	14.79 %
Theta Model (Bottom Up)	14.62 %	15.40 %
Theta Model (Optimal Reconciliation)	12.56 %	12.93 %
Damped Trend (Top Down)	14.09 %	13.88 %
Damped Trend (Middle Out)	14.03 %	14.08 %
Damped Trend (Bottom Up)	13.90 %	14.00 %
Damped Trend (Optimal Reconciliation)	14.08 %	13.88 %
Combined Forecasts		
Simple Average	13.28 %	13.21 %
Median	13.50 %	13.01 %
Trimmed Mean	13.39 %	13.03 %
Winsorised Mean	13.52 %	13.01 %
Bates/Granger	13.35 %	13.20 %
Newbold/Granger	15.14 %	13.56 %
Inverse Rank	13.56 %	13.09 %
OLS	15.58 %	13.82 %
CLS	13.69 %	12.89 %
LAD	15.77 %	15.14 %
EIG1	13.26 %	13.22 %
EIG2	13.93 %	14.71 %
EIG3	13.61 %	12.89 %
EIG4	14.08 %	14.29 %

Table 4.4 MAPE Forecast Evaluation: Churn – Surgical Services.

| Forecast Method	Long-Term (14d)	Short-Term (1d)
Seasonal Random Walk	16.33 %	19.87 %
HTS Forecasts		
ETS (Top Down)	14.77 %	16.19 %
ETS (Middle Out)	15.85 %	16.72 %
ETS (Bottom Up)	14.78 %	16.17 %
ETS (Optimal Reconciliation)	14.78 %	16.17 %
Theta Model (Top Down)	17.42 %	17.80 %
Theta (Middle Out)	17.31 %	16.88 %
Theta Model (Bottom Up)	17.54 %	17.02 %
Theta Model (Optimal Reconciliation)	17.41 %	17.78 %
Damped Trend (Top Down)	16.03 %	17.32 %
Damped Trend (Middle Out)	15.83 %	16.74 %
Damped Trend (Bottom Up)	15.37 %	16.15 %
Damped Trend (Optimal Reconciliation)	16.02 %	17.30 %
Combined Forecasts		
Simple Average	14.71 %	15.93 %
Median	15.06 %	16.36 %
Trimmed Mean	14.78 %	16.23 %
Winsorised Mean	14.84 %	15.87 %
Bates/Granger	14.76 %	15.98 %
Newbold/Granger	16.24 %	17.32 %
Inverse Rank	14.78 %	15.97 %
OLS	17.61 %	19.03 %
CLS	14.81 %	**15.80 %**
LAD	16.33 %	16.10 %
EIG1	14.76 %	15.93 %
EIG2	**14.70 %**	16.07 %
EIG3	15.10 %	16.34 %
EIG4	16.07 %	14.74 %
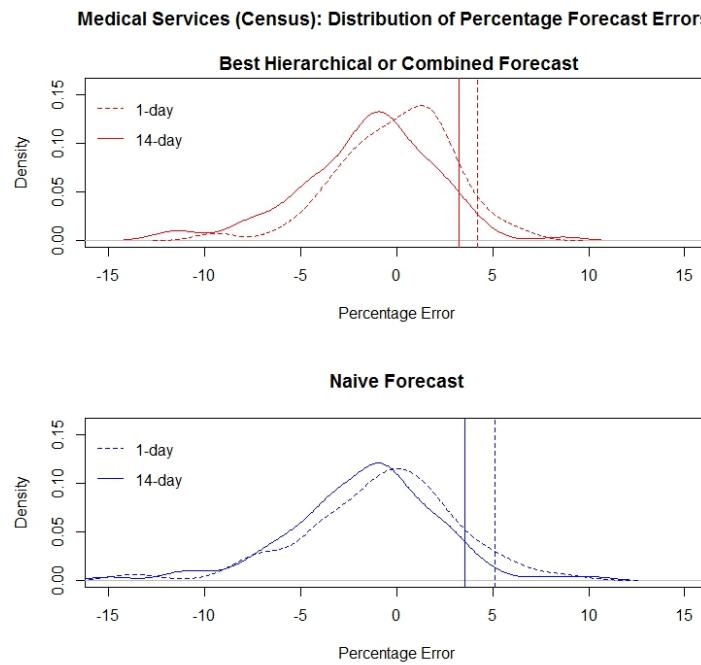
Fig. 4.8 Empirical Distribution of Percentage Forecast Errors: Census – Medical Services
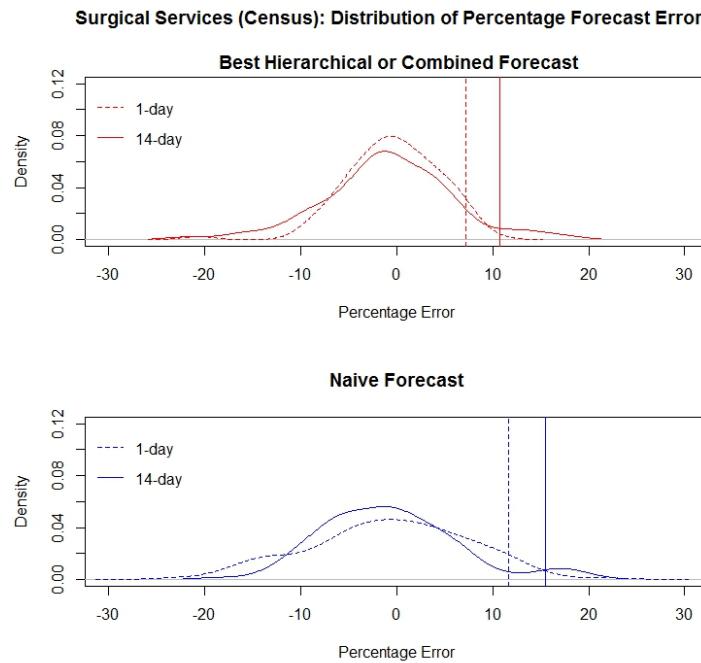

Fig. 4.9 Empirical Distribution of Percentage Forecast Errors: Census – Surgical Services
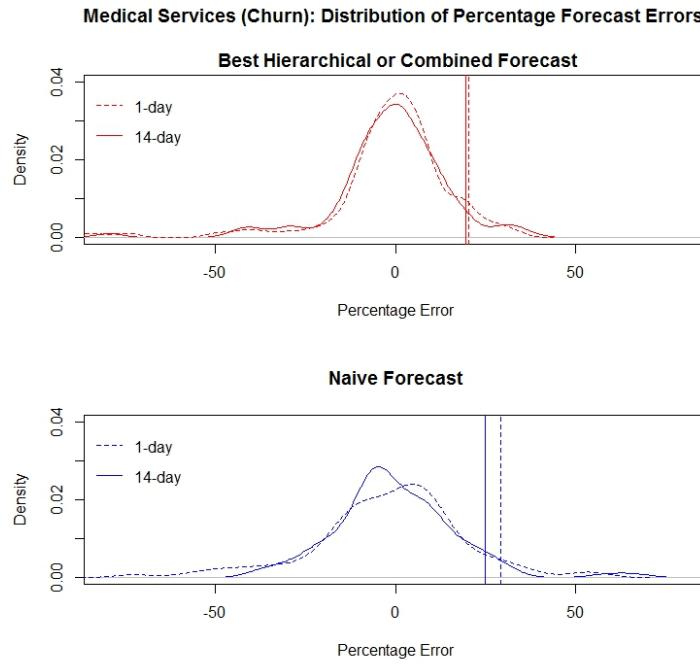
**Fig. 4.10** Empirical Distribution of Percentage Forecast Errors: Churn – Medical Services
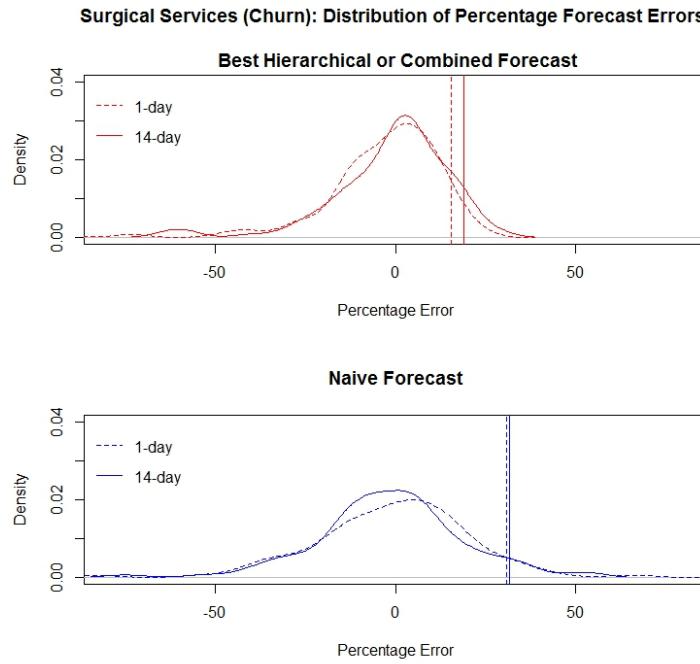
Fig. 4.11 Empirical Distribution of Percentage Forecast Errors: Churn – Surgical Services

4.7   Using the Forecasts in a Staffing Model

In this section, we show (i) how the forecasts can be used for staffing decisions, assuring that the target NHPPD value is satisfied in at least 95 % of the cases, and (ii)

how the best HTS forecasts and/or forecast combinations can improve cost-efficiency compared to the hospital's current staffing routine (essentially being described by the naïve seasonal random walk).

4.7.1   Parameters of the Staffing Model

The proposed cost model operates with the following parameters (assumptions):

- **Target nurse hours per patient day = 8 (α_1).** This may be reasonable in the medical and surgical divisions. Previous empirical studies (for the US) document that the median NHPPD in these divisions are 7.17 (Blegen et al., 2008) and that the mid range (the average between maximum and minimum values) is between 6.8 and 11.8 for US hospitals, where hospitals at the high end of this range tend to be teaching hospitals (Cavouras, 2002). Clarke and Donaldson (2008) find that staffing at the lower end of the continuum increases the risk of poor outcomes for patients and nurses, making it hazardous to staff at the lowest levels relative to peer units.

- **Target NHPPD value must be satisfied in at least 95 % of shifts (α_2).** While there is no mandatory nurse-to-patient ratio in the UK, this requirement has been set by internal hospital policy as quality assurance.

- **Total cost of permanent nursing staff per hour is £42.15 (α_3).** The hourly total cost of £41 for staff nurses and registered nurses was reported by Curtis (2012) for the year 2011/12. We adjusted the cost to 2013 prices (our forecasts are for 2013) using CPI inflation data from the Office of National Statistics (ONS), estimating the current total cost per hour at £42.15.

- **Temporary staff is 1.5 times as expensive as permanent staff (α_4).** While the average total cost per hour for permanent staff nurses and registered nurses is known to some accuracy, this is not the case for the cost of agency nurses. We rely on information adducted from the hospital, and assume a cost premium of 50% for temporary staff compared to permanent staff.

- **The time required to deal with a churn event (i.e. admission or discharge) is 0.5 nurse hours on average (α_5).** Again, information on the time requirement for churn events is scarce in the literature and we rely on the estimate by the hospital.

These parameters of the cost model for medical and surgical units can easily be varied to suit the other divisions – for instance, a staffing model for a critical care unit should obviously use a much higher target value for the NHPPD parameter, given the mid range between 14.6 and 25.5 for these departments (Cavouras, 2002).

4.7.2 Outline of the Staffing Model

The proposed staffing model utilises the patient census and churn forecasts, to compute the required number of both permanent nursing staff (based on the long-term forecasts) and agency nurses (based on the short-term forecasts). It is worth fixing notation: $\hat{x}_{i,t+h|t}$ denotes the h-step ahead forecast of model i for a time series x at time t, $\Phi(e_{i,j,h,t}; k)$ denotes the kth quantile of the empirical h-step percentage forecast error distribution of model i for series j at time t (note that the distribution is time-varying, as new forecasts are produced over time), C_t describes patient census for shift t (at the level the staff model is applied to, e.g. the 'Medical Services' division) and T_t denotes patient churn for shift t (based on the term 'turnover' to avoid notational confusion, rather than 'churn'). P_t denotes the number of permanent nursing staff scheduled for shift t, and A_t denotes the number of agency nurses scheduled for shift t. $\mathbb{1}_j$ denotes the indicator function, i.e. 1 if condition j is satisfied and 0 otherwise.

We specify 4 separate cases – each of which is a possible approach to nurse staffing:

1. **Case 1.** Permanent nursing staff is scheduled using the long-term census forecast (i.e. 28 shifts ahead). Churn is not accounted for in the nurse staffing decision. No adjustment stage using short-term forecasts, i.e. no hiring of agency nurses. No adjustment of the mean forecast to account for the target of hitting a NHPPD number of at least 8, in a minimum 95 % of the shifts.

   The number of required nurses for shift t in this scenario using forecasts from a model i (that can be specified by the user) is given by:

   $$P_t = \frac{1}{24} \times \alpha_1 \times \hat{C}_{i,t|t-28}$$
   $$A_t = 0$$

The predicted number of patients that is taken into account in staffing (in this case only the predicted census) is multiplied by $\frac{1}{24} \times alpha_1$ applying the NHPPD target. It is worth explaining the origin of this multiplicator, as it is also used in the other cases: The predicted number of patients is multiplied by 12 to give predicted number of patient hours in the shift and subsequently divided by 24 to predict the number of patient days in the shift, then we multiply by the target NHPPD value of α_1 to give required nurse hours in the shift and divide by 12 (the length of a shift) to obtain the required number of nurses, resulting in a factor of: $(12/24) * \alpha_1/12 = 1/24 * \alpha_1$.

2. **Case 2.** Same as Case 1, but allowing for an adjustment stage using short-term forecasts (i.e. 2-shift ahead) and hiring agency nurses accordingly.

   The number of required nurses for shift t in this scenario using forecasts from a model i is given by:

$$P_t = \frac{1}{24} \times \alpha_1 \times \hat{C}_{i,t|t-28}$$

$$A_t = \mathbb{1}_{\hat{C}_{i,t|t-2} > \hat{C}_{i,t|t-28}} \times \frac{1}{24} \times \alpha_1 \times \left(\hat{C}_{i,t|t-2} - \hat{C}_{i,t|t-28} \right)$$

   The number of permanent nurses hired, P_t, in this scenario is equal to Case 1. However, if the predicted patient census, 1 day ahead of the shift, is higher than the prediction 14 days ahead of the shift, additional agency nurses will be hired for the predicted excess census.

3. **Case 3.** Same as Case 2, but staffing decision takes into account patient churn.

   The number of required nurses for shift t in this scenario using forecasts from a model i is given by:

$$P_t = \frac{1}{24} \times \alpha_1 \times \hat{C}_{i,t|t-28} + \frac{1}{12} \times \alpha_5 \times \hat{T}_{i,t|t-28}$$

$$A_t = \mathbb{1}_{\hat{C}_{i,t|t-2} > \hat{C}_{i,t|t-28}} \times \frac{1}{24} \times \alpha_1 \times \left(\hat{C}_{i,t|t-2} - \hat{C}_{i,t|t-28} \right)$$

$$+ \mathbb{1}_{\hat{T}_{i,t|t-2} > \hat{T}_{i,t|t-28}} \times \frac{1}{12} \times \alpha_5 \times \left(\hat{T}_{i,t|t-2} - \hat{T}_{i,t|t-28} \right)$$

The factor of $1/12 * \alpha_5$ is applied to the predicted churn forecast, as a churn event takes a nurse α_5 hours, so that one nurse can cover $1/12 * \alpha_5$ churn events in a 12-hour shift.

4. **Case 4.** Same as Case 3, but in addition the forecasts are adjusted using the 95 % quantile of the relevant empirical percentage forecast error distribution:

$$P_t = \frac{1}{3} \times [1 + \Phi(e_{i,C,28,t}; \alpha_2)] \, \hat{C}_{i,t|t-28} + \frac{1}{24} \times [1 + \Phi(e_{i,T,28,t}; \alpha_2)] \, \hat{T}_{i,t|t-28}$$

$$A_t = \mathbb{1}_{\left[1 + \Phi(e_{i,C,2,t}; \alpha_2)\right]\hat{C}_{i,t|t-2} > \left[1 + \Phi(e_{i,C,28,t}; \alpha_2)\right]\hat{C}_{i,t|t-28}} \times \frac{1}{24} \times \alpha_1$$

$$\times \left\{ [1 + \Phi(e_{i,C,2,t}; \alpha_2)] \, \hat{C}_{i,t|t-2} - [1 + \Phi(e_{i,C,28,t}; \alpha_2)] \, \hat{C}_{i,t|t-28} \right\}$$

$$+ \mathbb{1}_{\left[1 + \Phi(e_{i,T,2,t}; \alpha_2)\right]\hat{T}_{i,t|t-2} > \left[1 + \Phi(e_{i,T,28,t}; \alpha_2)\right]\hat{T}_{i,t|t-28}} \times \frac{1}{12} \times \alpha_5$$

$$\times \left\{ [1 + \Phi(e_{i,T,2,t}; \alpha_2)] \, \hat{T}_{i,t|t-2} - [1 + \Phi(e_{i,T,28,t}; \alpha_2)] \, \hat{T}_{i,t|t-28} \right\}$$

As can be seen, the only change relative to Case 3 is that a factor of $[1 + \Phi(e_{i,j,h,t}, 95)]$ is assigned to the forecasts instead of the factor 1, thereby accounting for the condition that the target NHPPD value should be satisfied in at least 95 % of the shifts. The reasoning is as follows: Assume that the 95th quantile of the empirical percentage forecast error distribution is 11 %, i.e. based on the distribution it happens only 5 times in 100 that the observed value turns out to be more than 11 % higher than the forecast. If conditions are relatively stable, multiplying the forecast by a factor of 1.11 therefore should give us a 95 % probability of hitting the target, avoiding understaffed shifts.

Each of the staffing approaches above (Case 1 - 4) results in a different N_t, the number of total scheduled nurses for shift t:

$$N_t = P_t + A_t$$

which can be compared to $\tilde{N}_t$, the required number of nurses that is calculated using the actual (i.e. realised) numbers for patient census and churn:

$$\tilde{N}_t = \frac{1}{24} \times \alpha_1 \times C_{i,t} + \frac{1}{12} \times \alpha_5 \times T_{i,t}$$

A shift is understaffed if $N_t < \tilde{N}_t$.

The £ staff expenditure for a given shift, E_t, can be calculated in a straightforward way using the model parameters:

$$E_t = \alpha_3 \times (P_t + \alpha_4 \times A_t)$$

4.7.3   Results – Staffing Model

Table 4.5 shows the results of the staff model using (a) the best hierarchical forecast or forecast combination; versus (b) using a naïve (seasonal random walk) forecast.

Table 4.5 Results Staffing Model – Staff Cost and Risk of Understaffing.

Forecast Approach	Staff Cost (per Shift)	% Understaffed Shifts
Best Hierarchical or Combined Forecast		
Case 1	121,619.61 £	79.50 %
Case 2	124,009.52 £	72.50 %
Case 3	135,008.14 £	22.50 %
Case 4	144,193.46 £	0.50 %
Naïve Forecast		
Case 1	121,528.57 £	77.00 %
Case 2	125,481.39 £	68.50 %
Case 3	137,195.72 £	16.50 %
Case 4	150,895.31 £	1.00 %

The relationship shown in the table is intuitive: The more effort the hospital management undertakes to hit the target nurse-hours-per-patient-day level (by readjusting staff levels with agency staff, accounting for churn workload, and setting a target minimum of fully staffed shifts), the higher is the direct staff cost.

Using only patient census as input is the most common approach to nurse staffing (Baernholdt et al., 2010). However, it can clearly be seen that Cases 1 & 2 lead to more than two thirds of shifts being understaffed when taking into account actual workload (including churn events) in the evaluation – these staffing approaches do not come anywhere close to satisfying the hospitals commitment to quality care by setting the internal target of hitting a NHPPD value of 8 in 95 % of the shifts. A first very important step for the hospital to assure quality care is to account for churn in staff planning (Case 3), which reduces the number of understaffed shifts vastly for either forecasting approach – however, only adjusting staff levels for churn events and using the mean forecasts for census and churn still leads to around a fifth of the shifts being understaffed, due to the stochastic nature of demand. The best practice

approach for the hospital – if it takes the quality care target seriously – must be Case 4, adjusting the mean forecasts for census and churn by revising the forecasts using the distribution of percentage forecast errors. This modification to the staff planning procedure reduces the likelihood of understaffed shifts to extremely 1 % or 0.5 %, depending on model selection. It is interesting that the approach leads to the number of understaffed shifts being considerably lower than even the target maximum of 5 %. This could in general be due to a break in the series – if the series has become much more well-behaved and forecasts have become more accurate than for earlier periods as a consequence, using the empirical distribution of forecast errors could lead to overadjustment of the forecasts; however, this seems unlikely for 2 reasons: first, none of the forecast series display any obvious break in level or volatility, second, the forecast error distribution gets updated in each re-estimation step in the cross-validation, allowing for a time-varying forecast error distribution. The more plausible explanation why the number of understaffed shifts is even lower than the target maximum of 5 % is due to the idiosyncrasies of staffing: The model is designed such that permanent staff is scheduled 14 days in advance, and the 1-day ahead adjustment only concerns temporary staff, i.e. if the 1-day ahead forecast is higher than the 14-day ahead forecast, additional staff gets deployed, but if the opposite is the case, the permanent staff cannot be called off on such short notice.

A comparison of the best practice approach for the different forecast models shows that the best among the hierarchical and combined forecast approaches is considerably more cost-efficient than the simple staffing rule that the hospital currently applies, with a cost advantage of 4.44 % while also decreasing the risk of understaffed shifts further.[6] When employing the best practice staffing approach, the optimised hierarchical or combined forecast saves the hospital over £3.5 million annually. Comparing the best practice approach to the other staffing options, it is obvious that it increases the staff cost – however, these are merely direct staff costs; the decrease in understaffed shifts (from almost 80 % to less than 1 % when compared with Case 1) also decreases costs due to nurse burnout and absenteeism, and increases staff satisfaction as well as patient outcomes, quality of patient care and patient satisfaction, potentially reducing the hospital's legal costs.

---

[6]Note that the naïve approach slightly outperforms the 'best hierarchical or combined' approach for Case 1, while also generating lower risk of understaffing for that case. This might appear counter-intuitive, given that the best hierarchical or combined forecast approach had a better MAPE value – however, a more detailed look at the understaffed shifts is revealing: The naïve approach indeed leads to fewer understaffed shifts using this staffing approach, but when it does, it understaffs by 12.45 nurses on average, while the best hierarchical or combined approach understaffs by 11.89 nurses.

Table 4.6 shows the cost breakdown for an average shift applying the best practice staffing approach using the best hierarchical or combined forecast. It delivers some insight into the main cost factors: The vast majority (89.46 %) of the staff costs are accounted for by permanent staff hours related to the patient census (i.e. patient care excluding churn events). Most of the remaining staff costs are due to permanent staff dealing with churn events, accounting for another 8.81 % of total staff cost. Two conclusions follow immediately: (i) even though the staffing model results made it obvious that staff planning must take note of churn events to avoid the risk of understaffing, churn only accounts for a total of 9.28 % of staff costs, suggesting that it contributes a similar fraction of a nurse's workload; (ii) the best hierarchical or combined forecast approach is very good at forecasting demand 14 days in advance, keeping readjustments using temporary staff to a minimum: Agency nurses only account for 1.73 % of total staff cost.[7] These results emphasise the importance of the 14-day ahead census forecasts, as almost 90 % of total staff cost depend on their quality.

Table 4.6 Staff Costs per Average Shift: Best Hierarchical or Combined Forecast, Case 4.

		Staff Cost	...in % of Shift Cost
Total Cost		**144,193.46 £**	**100.00 %**
Medical Services		**83,300.20 £**	**57.77 %**
	Census (Permanent Staff)	73,730.47 £	51.13 %
	Census (Agency Staff)	978.72 £	0.68 %
	Churn (Permanent Staff)	8,158.55 £	5.66 %
	Churn (Agency Staff)	432.46 £	0.30 %
Surgical Services		**60,893.26 £**	**42.23 %**
	Census (Permanent Staff)	55,271.30 £	38.33 %
	Census (Agency Staff)	834.57 £	0.58 %
	Churn (Permanent Staff)	4,544.61 £	3.15 %
	Churn (Agency Staff)	242.78 £	0.17 %

Appendix 4.B presents the results of a sensitivity analysis with respect to staffing model parameters. The parametric approach allows healthcare providers to replicate our results with their own data and use the model for staffing by modifying the parameters (staff cost per hour, NHPPD values, etc.) according to their requirements.

---

[7]The corresponding value for the naïve forecast is 3.80 %. While this is not a vast difference, the lower dependence on agency nurses should be considered as an additional advantage of the hierarchical or combined forecast approach, as temporary staff are not only more expensive, but also decrease patient satisfaction and quality of care (Kane et al., 2007; Hughes et al., 2015).

4.8   Conclusion

An unprecedented slowdown in public healthcare funding growth, paired with rapid growth of demand for health services, has increased the pressure on NHS hospitals to delivering better value at lower cost.

In terms of this broad challenge, we evaluate the use of hierarchical time series forecasting and forecast combination approaches for hospital staffing, and present a comprehensive, integrated forecasting system for patient census and churn that takes into account the hospital structure. The results fill a gap in nurse staffing research, in going beyond the analysis of aggregated data that is common (Hughes et al., 2015).

The proposed system provides guidance on managing staffing in two ways: First, it provides highlights measures to avoid risk of understaffing – (i) accounting for churn in nurse deployment, (ii) readjusting for predicted short-term demand changes using agency staff, and (iii) modifying mean forecasts using empirical forecast error distributions to incorporate a target minimum of fully staffed shifts. Second, using accurate forecasting techniques that incorporate disaggregated information in the model, staff cost is minimised under the 'quality of care' constraint.

This paper should be seen as an forerunner to a fully integrated, comprehensive time series forecast system for hospital staffing. Some improvements for future research are evident: One challenge is accurate modelling of shift-level churn time series, given their bimodal distribution. While churn only accounts for roughly 10 % of staff cost, it should be possible to further improve the forecast by applying non-linear time series forecasting techniques – which are better suited for multimodal data distributions – to the churn series. Candidate models to be tested in a hierarchical framework are the self-exciting threshold autoregressive model by Tong (1990), the multipredictor autoregressive time series model by Martin (1992), the mixture autoregressive model by Wong and Li (2000), the dynamic switching Markov chain model by Gouriéroux and Robert (2006), and the autoregressive conditional root model by Bec et al. (2008). There will be a trade-off between optimising fit and model usability – these non-linear methods are not widely known and are difficult to automate, raising the question whether their potentially improved fit can make up for the lost ease of implementation.

Another potential improvement for forecasting churn could be a forecasting model that includes regressors such as length of stay, which is negatively related to churn as documented by Unruh and Fottler (2006) and Hughes et al. (2015).

Finally, an avenue for future research could be the attempt to combine census and churn into one single hierarchy to make use of the hierarchical time series methods that take note of co-movement between series. However the correlation between these variables are low in our dataset (-0.039 for 'Surgical Services' and 0.192 for 'Medical Services'). It does appear that census and churn follow different data-generating processes.

The model presented in this chapter addresses a gap in health care research in enabling the combination of decentralised monitoring and centralised planning to achieve effective staffing: balancing quality, safety, labour cost reduction, and staff satisfaction.

4.9 References

Alderwick, H., R. Robertson, J. Appleby, P. Dunn, and D. Maguire. 2015. Better Value in the NHS: The Role of Changes in Clinical Practice. The King's Fund Report, July 2015.

Armstrong, J. S. 2001. Combining Forecasts. In J. S. Armstrong (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, pp. 417–439. Kluwer Academic Publisher.

Assimakopoulos, V., and K. Nikolopoulos. 2000. The Theta Model: A Decomposition Approach to Forecasting. *International Journal of Forecasting* 16:521–530.

Athanasopoulos, G., R. A. Ahmed, and R. J. Hyndman. 2009. Hierarchical Forecasts for Australian Domestic Tourism. *International Journal of Forecasting* 25:146–166.

Aylin, P., A. Yunus, A. Bottle, A. Majeed, and D. Bell. 2010. Weekend Mortality for Emergency Admissions: A Large, Multicentre Study. *Quality and Safety in Health Care* 19:213–217.

Baernholdt, M., K. Cox, and K. Scully. 2010. Using Clinical Data to Capture Nurse Workload: Implications for Staffing and Safety. *CIN: Computers, Informatics, Nursing* 28:229–234.

Bard, J. F., and H. W. Purnomo. 2005. Short-Term Nurse Scheduling in Response to Daily Fluctuations in Supply and Demand. *Health Care Management Science* 8:315–324.

Barrow, D. K., and N. Kourentzes. 2016. Distributions of Forecasting Errors of Forecast Combinations: Implications for Inventory Management. *International Journal of Production Economics* pp. 24–33.

Bates, J. M., and C. W. J. Granger. 1969. The Combination of Forecasts. *Operations Research Quarterly* 20:451–468.

Bec, F., A. Rahbek, and N. Shephard. 2008. The ACR Model: A Multivariate Dynamic Mixture Autoregression. *Oxford Bulletin of Economics and Statistics* 70:583–618.

Bell, C. M., and D. A. Redelmeier. 2001. Mortality Among Patients Admitted to Hospitals on Weekends as Compared With Weekdays. *The New England Journal of Medicine* 345:663–668.

Bergmeir, C., R. J. Hyndman, and B. Koo. 2015. A Note on the Validity of Cross-Validation for Evaluating Time Series Prediction. Working Paper No. 10/15, Monash University, Department of Econometrics and Business Statistics.

Blegen, M. A., T. Vaughn, and C. P. Vojir. 2008. Nurse Staffing Levels: Impact of Organizational Characteristics and Registered Nurse Supply. *Health Services Research* 43:154–173.

Box, G. E. P., and G. M. Jenkins. 1970. *Time Series Analysis: Forecasting and Control.* San Francisco, Calif.: Holden-Day Inc.

Buerhaus, P. I. 1997. What Is the Harm in Imposing Mandatory Hospital Nurse Staffing Regulations? *Nursing Economics* 15:66–72.

Carvalho-Silva, M., T. T. Monteiro, F. de Sá-Soares, and S. Dória-Nóbrega. forthcoming. Assessment of Forecasting Models for Patients Arrival at Emergency Department. *Operations Research for Health Care* .

Cavouras, C. A. 2002. Nurse Staffing Levels in American Hospitals: A 2001 Report. *Journal of Emergency Nursing* 28:40–43.

Clarke, S. P., and N. E. Donaldson. 2008. Nurse Staffing and Patient Care Quality and Safety. In R. G. Hughes (ed.), *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, chap. 25. Rockville, MD: Agency for Healthcare Research and Quality (US).

Clemen, R. T. 1989. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting* 5:559–583.

Crist-Grundman, D., and G. Mulrooney. 2011. Effective Workforce Management Starts with Leveraging Technology, While Staffing Optimization Requires True Collaboration. *Nursing Economics* 29:195–200.

Curtis, L. 2012. Unit Costs of Health & Social Care 2012. Technical report, Personal Social Services Research Unit, Canterbury.

Fiorucci, J. A., F. Louzada, and B. Yiqi. 2016a. *forecTheta: Forecasting Time Series by Theta Models*. URL http://CRAN.R-project.org/package=forecTheta. [Accessed on Sep 15, 2017], R package version 2.2.

Fiorucci, J. A., T. R. Pellegrini, F. Louzada, F. Petropoulos, and A. B. Koehler. 2016b. Models for Optimising the Theta Method and Their Relationship to State Space Models. *International Journal of Forecasting* 32:1151–1161.

Freeman, M., N. Savva, and S. Scholtes. 2016. Gatekeepers at Work: An Empirical Analysis of a Maternity Unit. *Management Science* .

Frith, K., E. Anderson, F. Tseng, and E. Fong. 2012. Nurse Staffing is an Important Strategy to Prevent Medication Errors in Community Hospitals. *Nursing Economics* 30:288–294.

Gardner, E. S. 1985. Exponential Smoothing: The State of the Art. *Journal of Forecasting* 4:1–28.

Gardner, E. S., and E. McKenzie. 1985. Forecasting Trends in Time Series. *Management Science* 31:1237–1246.

Ghosh, B., and G. Cruz. 2005. Nurse Requirement Planning: A Computer-Based Model. *Journal of Nursing Management* 13:363–371.

Gouriéroux, C., and C. Robert. 2006. Stochastic Unit Root Models. *Econometric Theory* 22:1052–1090.

Green, L., S. Savin, and N. Savva. 2013. "Nursevendor Problem": Personnel Staffing in the Presence of Endogenous Absenteeism. *Management Science* 59:2237–2256.

Greenglass, E. R., R. J. Burke, and L. Fiksenbaum. 2001. Workload and Burnout in Nurses. *Journal of Community & Applied Social Psychology* 11:211–215.

Gross, C. W., and J. E. Sohl. 1990. Disaggregation methods to expedite product line forecasting. *Journal of Forecasting* 9:233–254.

Guarin-Calvert, M. 2011. Assessment of Cost Trends and Price Differences for U.S. Hospitals. Working Paper 294, Federal Trade Commission.

Ham, C., D. Berwick, and J. Dixon. 2016. Improving Quality in the English NHS: A Strategy for Action. The King's Fund Report, February 2016.

Hockenberry, J. M., and E. R. Becker. 2016. How Do Hospital Nurse Staffing Strategies Affect Patient Satisfaction? *ILR Review* 69:890–911.

Hughes, R. G., K. Bobay, N. A. Jolly, and C. Suby. 2015. Comparison of Nurse Staffing Based on Changes in Unit-Level Workload Associated with Patient Churn. *Journal of Nursing Management* 23:390–400.

Hurst, J., and S. Williams. 2012. Can NHS Hospitals Do More With Less? Technical report, Nuffield Trust, London.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang. 2011. Optimal Combination Forecasts for Hierarchical Time Series. *Computational Statistics and Data Analysis* 55:2579–2589.

Hyndman, R. J., and Y. Khandakar. 2008. Automatic Time Series Forecasting: The Forecast Package for R. *Journal of Statistical Software* 27:1–22.

Hyndman, R. J., and A. B. Koehler. 2006. Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting* 22:679–688.

Hyndman, R. J., A. B. Koehler, R. D. Snyder, and S. Grose. 2002. A State Space Framework for Automatic Forecasting Using Exponential Smoothing Methods. *International Journal of Forecasting* 18:439–454.

Hyndman, R. J., A. J. Lee, and E. Wang. 2016a. Fast Computation of Reconciled Forecasts for Hierarchical and Grouped Time Series. *Computational Statistics and Data Analysis* 97:16–32.

Hyndman, R. J., E. Wang, A. Lee, and S. Wickramasuriya. 2016b. *hts: Hierarchical and Grouped Time Series*. URL http://CRAN.R-project.org/package=hts. [Accessed on Sep 15, 2017], R package version 5.0.

Jones, S. S., R. S. Evans, T. L. Allen, A. Thomas, P. J. Haug, S. J. Welch, and G. L. Snow. 2009. Multivariate Time Series Approach to Modeling and Forecasting Demand in the Emergency Department. *Journal of Biomedical Informatics* 42:123–139.

Jones, S. S., A. Thomas, R. S. Evans, S. J. Welch, P. J. Haug, and G. L. Snow. 2008. Forecasting Daily Patient Volumes in the Emergency Department. *Academic Emergency Medicine* 15:159–170.

Kam, H. J., J. O. Sung, and R. W. Park. 2010. Prediction of Daily Patient Numbers for a Regional Emergency Medical Center using Time Series Analysis. *Healthcare Informatics Research* 16:158–165.

Kane, R. L., T. Shamliyan, C. Mueller, S. Duval, and T. J. Wilt. 2007. Nurse Staffing and Quality of Patient Care: Evidence Report/Technology Assessment No. 151. AHRQ Publication 07-E005, Agency for Healthcare Research and Quality.

Kim, K., and S. Mehrotra. 2015. A Two-Stage Stochastic Integer Programming Approach to Integrated Staffing and Scheduling with Application to Nurse Management. *Operations Research* 63:1431–1451.

Koestler, D. C., H. Ombao, and J. Bender. 2013. Ensemble-Based Methods for Forecasting Census in Hospital Units. *BMC Medical Research Methodology* 13:67.

Kuntz, L., R. Mennicken, and S. Scholtes. 2014. Stress on the Ward: Evidence of Safety Tipping Points in Hospitals. *Management Science* 61:754–771.

Kuntz, L., S. Scholtes, and S. Sülz. 2015. Separate and Concentrate: Accounting for Process Uncertainty in the Design of General Hospitals. Working paper, Cambridge Judge Business School.

Lamn, H. 1973. The Lost Weekend in Hospitals. *The New England Journal of Medicine* 289:923.

Lin, W. T. 1989. Modeling and Forecasting Hospital Patient Movements: Univariate and Multiple Time Series Approaches. *International Journal of Forecasting* 5:195–208.

Loopstra, R., M. McKee, S. Katikireddi, D. Taylor-Robinson, B. Barr, and D. Stuckler. 2016. Austerity and Old-Age Mortality in England: A Longitudinal Cross-Local Area Analysis, 2007-2013. *Journal of the Royal Society of Medicine* 109:109–116.

Maguire, D., P. Dunn, and H. McKenna. 2016. How Hospital Activity in the NHS in England Has Changed Over Time. The King's Fund Report, December 2016.

Martin, V. L. 1992. Threshold Time Series Models as Multimodal Distribution Jum Processes. *Journal of Time Series Analysis* 13:79–94.

McKenna, H. 2016. Five Big Issues for Health and Social Care After the Brexit Vote. The King's Fund Report, June 2016.

National Audit Office. 2016. Managing the Supply of NHS Clinical Staff in England. Technical report, National Audit Office, London. URL http://www.nao.org.uk/report/managing-the-supply-of-nhs-clinical-staff-in-england/. [Accessed on Sep 15, 2017].

Needleman, J., P. Buerhaus, and V. Pankratz. 2011. Nurse Staffing and Inpatient Hospital Mortality. *The New England Journal of Medicine* 364:1037–1045.

Pegels, C. C. 1969. Exponential Forecasting: Some New Variations. *Management Science* 15:311–315.

Perry, A. G., K. M. Moore, L. E. Levesque, C. W. Pickett, and M. J. Korenberg. 2010. Emergency Department Visits for Respiratory Illness Using Telehealth Ontario Calls. *Canadian Journal of Public Health* 101:464–469.

Powell, A., S. Savin, and N. Savva. 2012. Physician Workload and Hospital Reimbursement: Overworked Physicians Generate Less Revenue Per Patient. *Manufacturing & Service Operations Management* 14:512–528.

Reiter, K. L., D. W. Harless, G. H. Pink, and B. A. Mark. 2012. Minimum Nurse Staffing Legislation and the Financial Performance of California Hospitals. *Health Services Research* 47:1030–1050.

Robertson, R. 2016. Six Ways in which NHS Financial Pressures Can Affect Patient Care. The King's Fund Report, March 2016.

Robertson, R., L. Wenzel, J. Thompson, and A. Charles. 2017. Understanding NHS Financial Pressures: How Are They Affecting Patient Care? The King's Fund Report, March 2017.

Rogot, E., R. Fabsitz, and M. Feinleib. 1976. Daily Variation in USA Mortality. *American Journal of Epidemiology* 103:198–211.

Royal College of Nursing. 2010. Guidance on Safe Nurse Staffing Levels in the UK. Tech. rep., Royal College of Nursing, London.

Ruckert, A., and R. Labonté. forthcoming. Health Inequities in the Age of Austerity: The Need for Social Protection Policies. *Social Science & Medicine* .

Schweigler, L. M., J. S. Desmond, M. L. McCarthy, K. J. Bukowski, E. L. Ionides, and J. G. Younger. 2009. Forecasting Models of Emergency Department Crowding. *Academic Emergency Medicine* 16:301–308.

Skinner, W. 1974. The Focused Factory. *Harvard Business Review* May - June:113–123.

Staggs, V., and J. He. 2013. Recent Trends in Hospital Nurse Staffing in the United States. *The Journal of Nursing Administration* 43:388–393.

Tan, T., and S. Netessine. 2014. When Does the Devil Make Work? An Empirical Study of the Impact of Workload on Worker Productivity. *Management Science* 60:1574–1593.

Taylor, J. W. 2003. Exponential Smoothing with a Damped Multiplicative Trend. *International Journal of Forecasting* 19:715–725.

The King's Fund. 2017. The NHS Budget and How it Has Changed. URL https://www.kingsfund.org.uk/projects/nhs-in-a-nutshell/nhs-budget. [Accessed on Sep 15, 2017].

Tierney, S., P. Seymour-Route, and S. Crawford. 2013. Weighted Staffing Plans for Better Prediction of Staffing Needs. *The Journal of Nursing Administration* 43:461–467.

Timmermann, A. 2006. Forecast Combinations. *Handbook of economic forecasting* 1:135–196.

Tong, H. 1990. *Non-Linear Time Series*. Oxford University Press.

Unruh, L. Y., and M. D. Fottler. 2006. Patient Turnover and Nursing Staff Adequacy. *Health Services Research* 41:599–612.

Vahey, D. C., L. H. Aiken, D. M. Sloane, S. P. Clarke, and D. Vargas. 2004. Nurse Burnout and Patient Satisfaction. *Medical Care* 42:II57–II66.

Van Erven, T., and J. Cugliari. 2015. Game-Theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts. In A. Antoniadis, J.-M. Poggi, and X. Brossat (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics Vol. 217, pp. 297–317. Paris: Springer.

Weiss, C. E., and G. R. Roetzer. 2016. *GeomComb: (Geometric) Forecast Combination Methods*. URL https://CRAN.R-project.org/package=GeomComb. [Accessed on Sep 15, 2017], R package version 1.0.

Wickramasuriya, S. L., G. Athanasopoulos, and R. J. Hyndman. 2017. Optimal Forecast Reconciliation of Hierarchical and Grouped Time Series Through Trace

Minimization. Working paper, Department of Econometrics and Business Statistics, Monash Business School.

Wong, C. S., and W. K. Li. 2000. On a Mixture Autoregressive Model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62:95–115.

Appendix

4.A   List of Primary Specialties

The following lists describe the primary specialties that feed into the divisions 'Medical Services' and 'Surgical Services':

- **Medical Services:** Acute Medicine, Bone Density Scanning, Cardiology, Clinical Pharmacology, Dermatology, Gastro Physiology, Gastro-Enterology, Gauchers, General Medicine, Genito-Urinary Medicine, Geriatric Medicine, Hepatology, Hepatology Biopsy, Hepatology Chronic, Hepatology ERCP, Hepatology Other, Infectious Diseases, Medical Assessment, Metabolic Bone Disease, Nephrology, Paedatric Allergy, Paediatric Dermatology, Paediatric Rheumatology, Rehabilitation, Respiratory Physiology, Rheumatology, Special Vasculitis, Stroke Medicine, Thoracic Medicine, Thoracic Surgery, Vasculitis, Vasculitis Treatment

- **Surgical Services:** Bladder Cancer WL, Breast Surgery, Colorectal Surgery, ENT, EPS Assessment, EPS Follow-up, EPS Surgery (Primary), General Surgery, HPB & Pancreas Surgery, Liver Transplant Assessment, Liver Transplant Follow Up, Liver Transplantation, Maxillo-Facial Surgery, Medical Ophthalmology, Ophthalmology, Oral Surgery, Orthodontics, Orthopaedic, Paediatric ENT, Paediatric Gastro-intestinal Surgery, Paediatric Maxillo-Facial Surgery, Paediatric Ophthalmology, Paediatric Orthopaedic, Paediatric Plastic Surgery, Paediatric Transplant NOS, Paediatric Trauma, Pancreas Transplant, Pancreas Transplant Follow-Up, Plastic Surgery, Renal Donor, Renal Transplant, Renal Transplant Follow Up, Restorative Dentistry, Small Bowel Assessment, Small Bowel Follow Up, Small Bowel Transplant, Transplant Surgery NOS, Trauma, Upper Gastro-intestinal Surgery, Urology, Vascular Access, Vascular Surgery

4.B   Visualised Results of Sensitivity Analysis

The following figures show the effect of changing the staffing model parameters on staff costs and risk of understaffing. The parameters were allowed to vary as follows:

- For the target nurse-hours-per-patient-day parameter values between 6.8 and 11.8 were considered, since this was the reported mid range for medical and surgical divisions in Cavouras (2002).

- For the target minimum percentage of fully staffed shifts values between 60 % and 100 % were tested.

- The hourly cost for permanent staff was set to values between 30 and 60.

- The cost factor for agency staff was allowed to vary between 1 and 5 times the cost for permanent staff.

- The average nurse time per churn event was set to values between 6 minutes and 1 hour for the sake of the sensitivity analysis.

The dotted black vertical lines in the graphs indicate the values that were used in the actual staffing model in this paper.
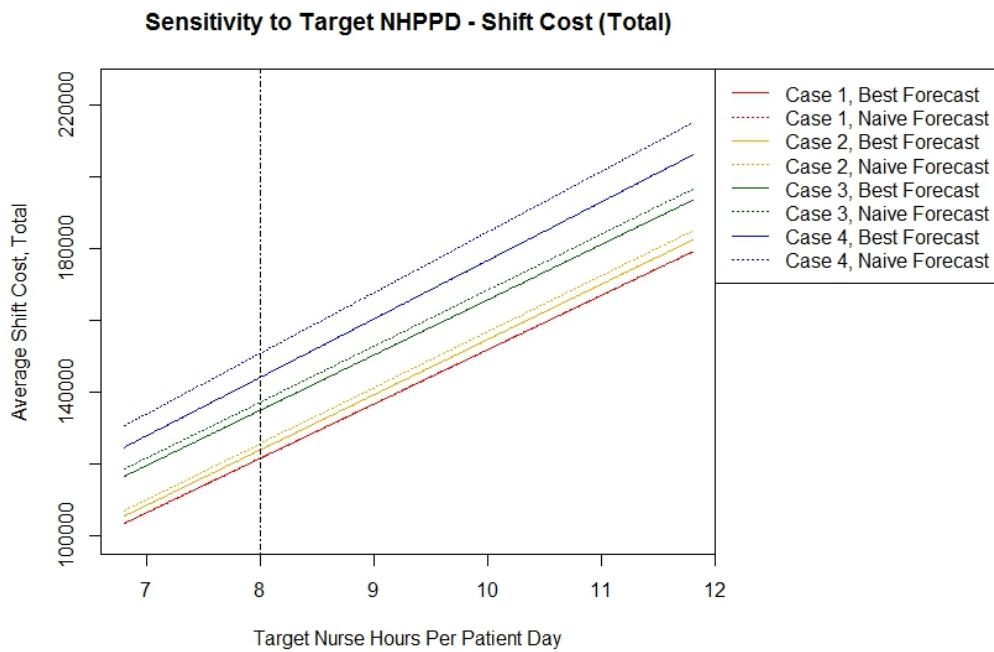
Fig. 4.B1 Cost Sensitivity to Target Nurse-Hours-Per-Patient-Day Values.
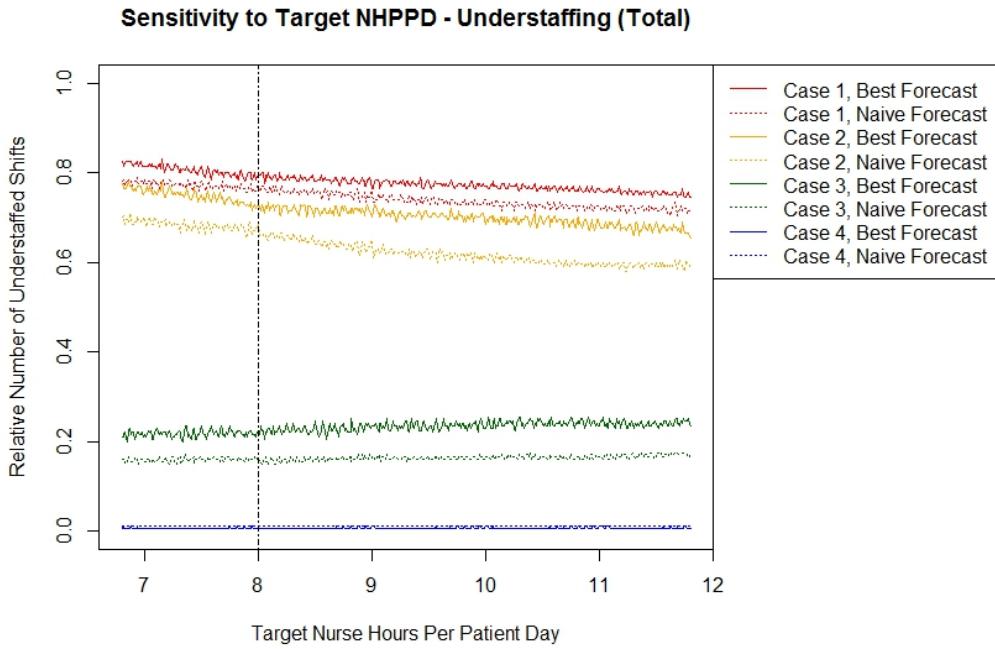
Fig. 4.B2 Understaffing Risk Sensitivity to Target Nurse-Hours-Per-Patient-Day Values.
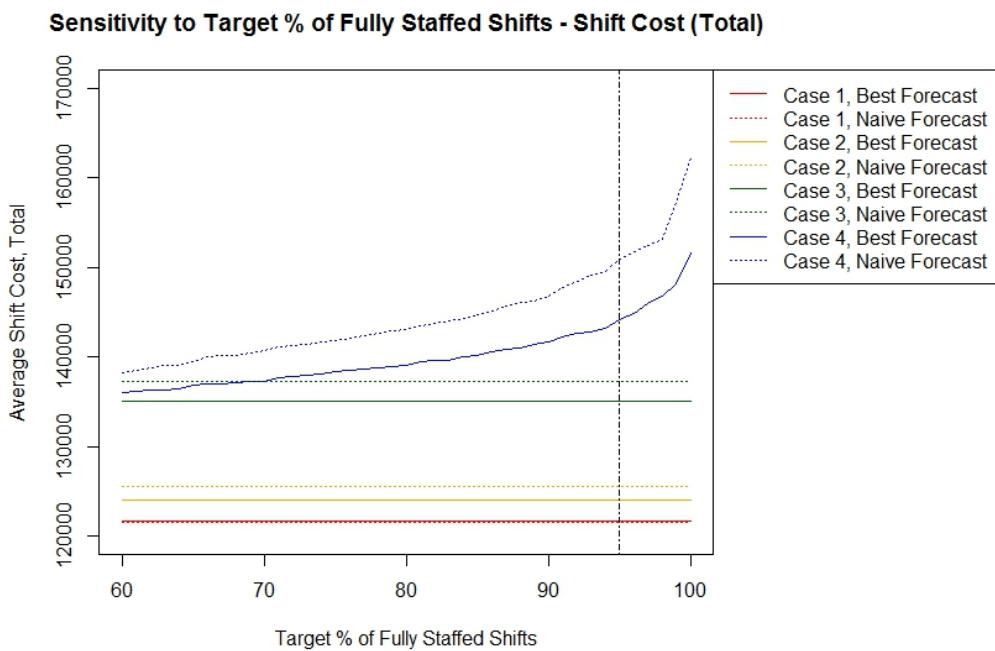
Fig. 4.B3 Cost Sensitivity to % of Fully Staffed Shifts Target.
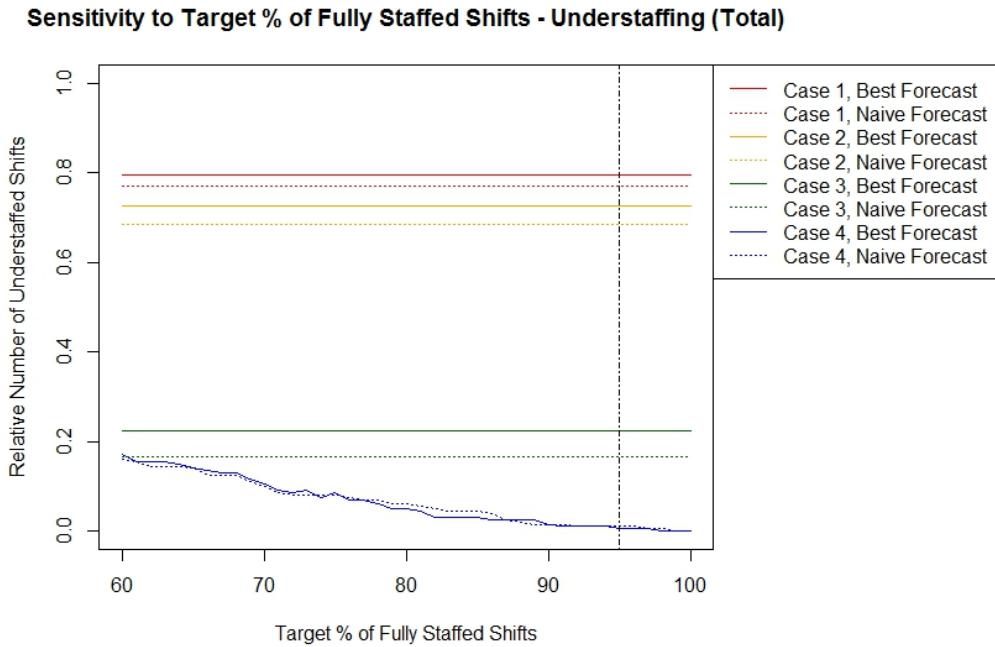
**Fig. 4.B4** Understaffing Risk Sensitivity to % of Fully Staffed Shifts Target.
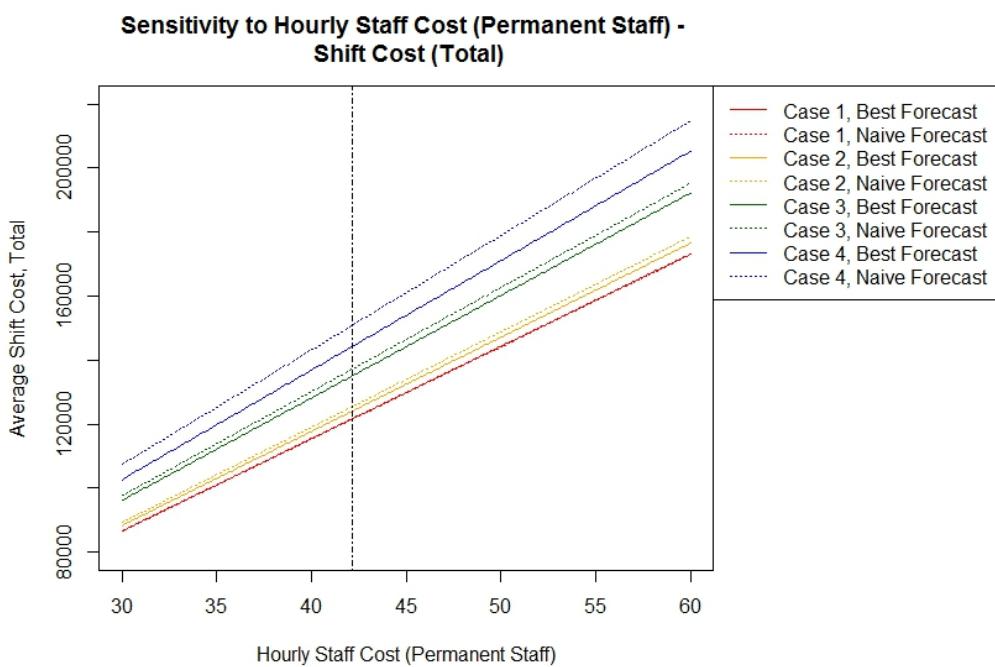
**Fig. 4.B5** Cost Sensitivity to Hourly Cost of Permanent Staff.
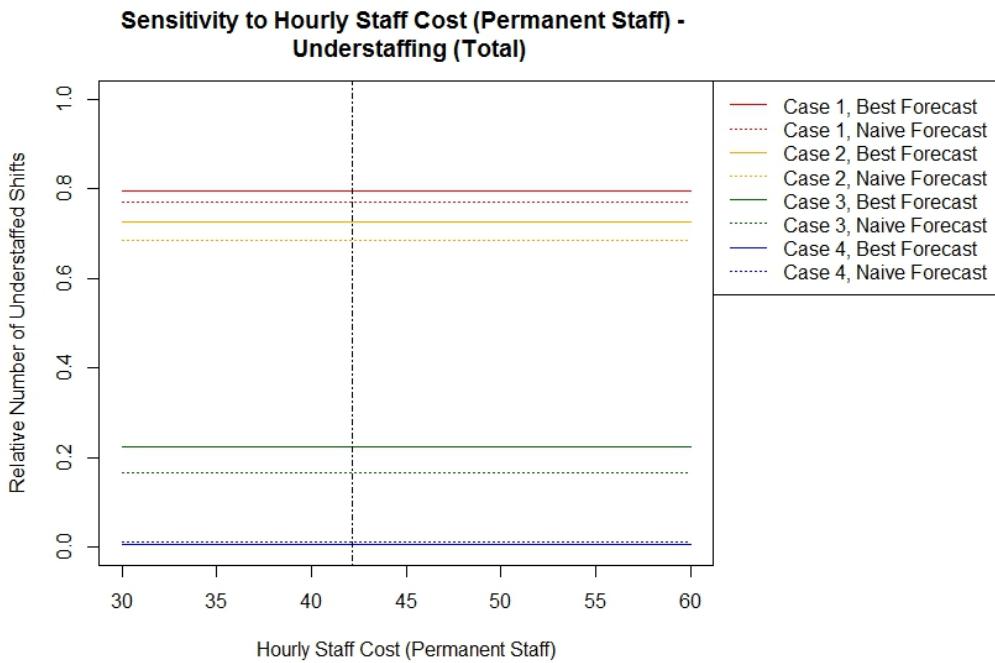
Fig. 4.B6 Understaffing Risk Sensitivity to Hourly Cost of Permanent Staff.
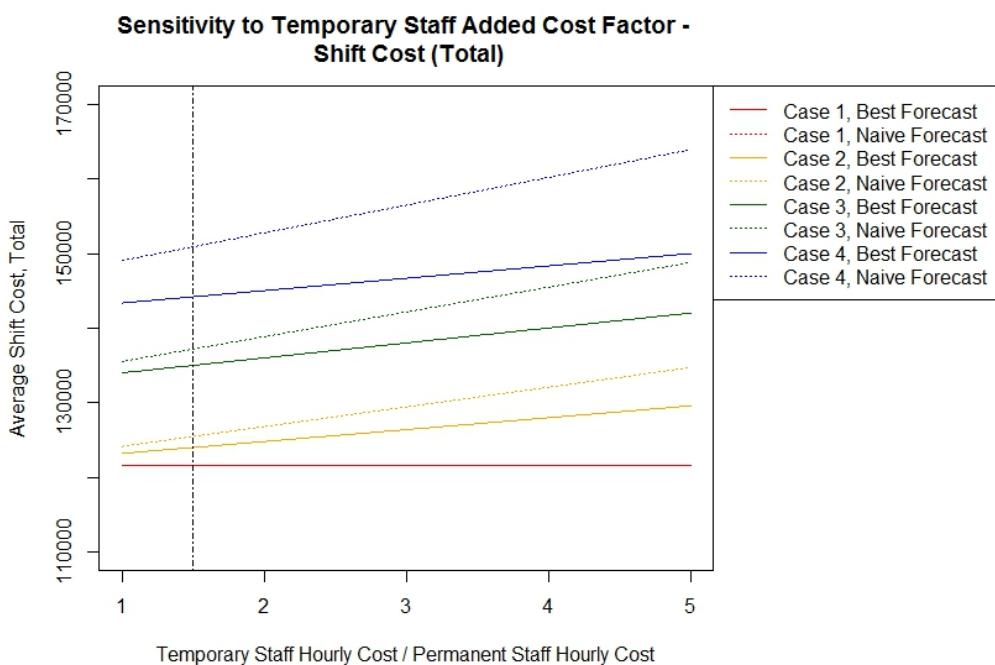

**Fig. 4.B7** Cost Sensitivity to Added Cost Factor of Temporary Staff.
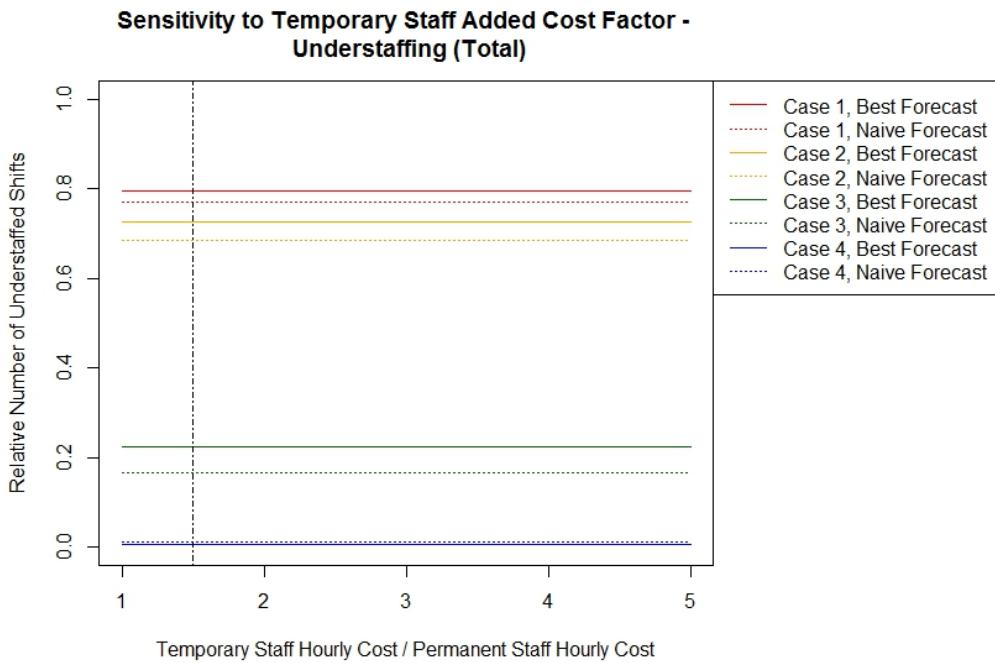
Fig. 4.B8 Understaffing Sensitivity to Added Cost Factor of Temporary Staff.
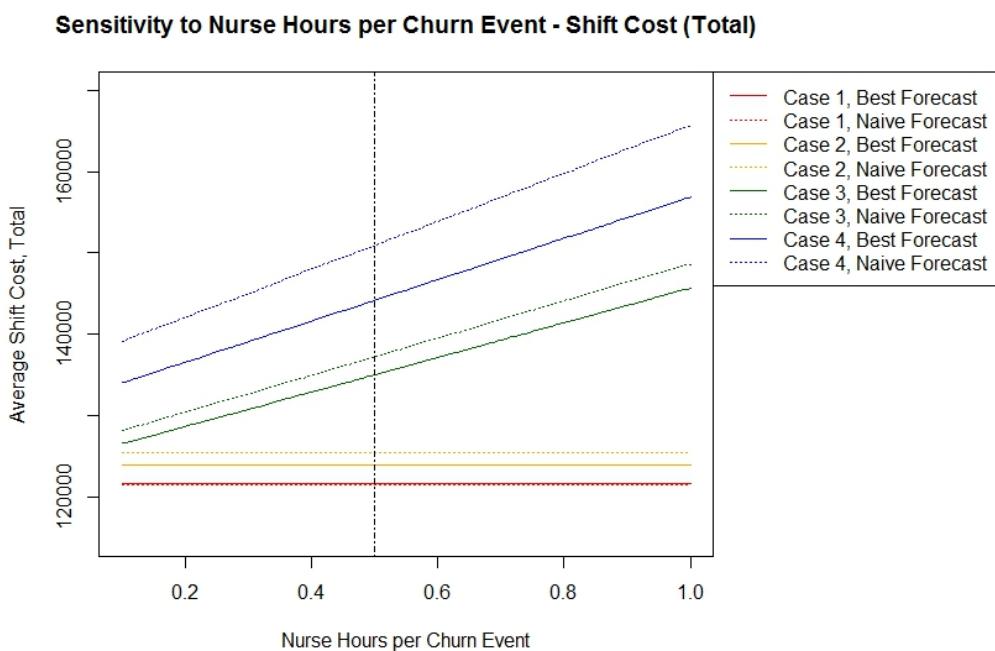

**Fig. 4.B9** Cost Sensitivity to Nurse Hours Required per Churn Event.
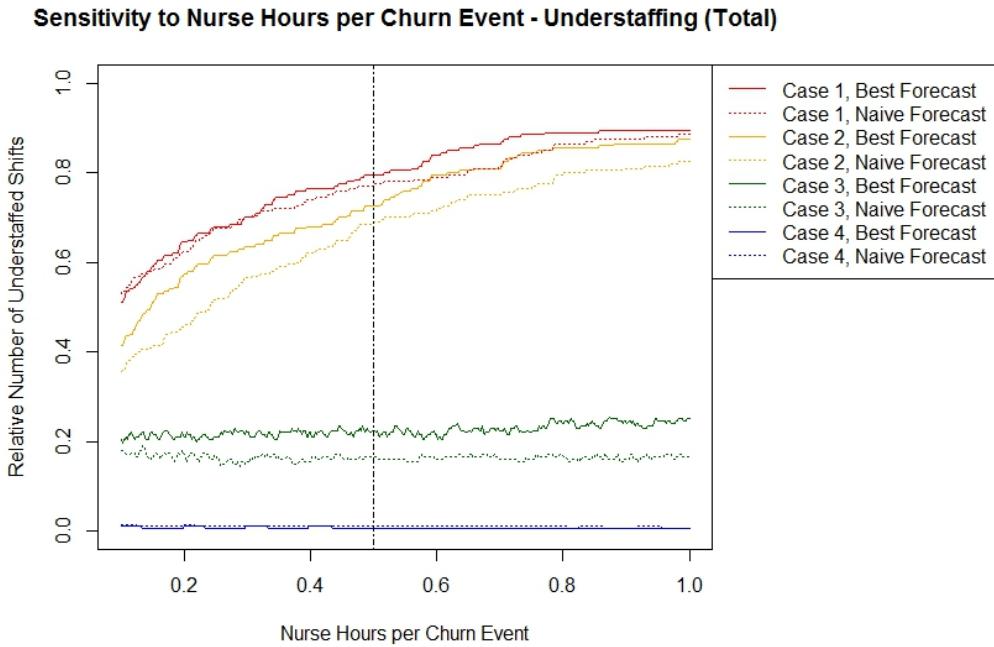
Fig. 4.B10 Understaffing Risk Sensitivity to Nurse Hours Required per Churn Event.

Chapter 5

Conclusions

This thesis has made contributions to a growing body of research that is concerned with optimally incorporating disaggregated information into aggregate forecasts. The linking theme between the chapters is the integration of a multitude of time series forecasts into a single one, either vertically (hierarchical forecasting) or horizontally (forecast combination).

The first study in this thesis, "Hierarchical Modelling and Forecasting System for Inflation Rate and Volatility" contributes to research on the impact of microeconomic shocks on macroeconomic volatility. A two-stage variance decomposition furthers our understanding of the data generating process of inflation rate and its volatility. We show that aggregate volatility is generated by a combination of macro shocks (driving the covariation of product-level inflation rates) and micro shocks (driving the product-level variances). Through a product-level analysis, we find that episodes of high inflation volatility are often driven by a single or a few selected products, rather than a general increase in product-level variances. The second part of the paper formulates a forecasting system for inflation rate and volatility, building on the variance decomposition from the first part. We explore the value of hierarchical time series forecasting in the inflation context and find that bringing in the scheme of aggregating product-level inflation rates, and of distinguishing between their common, industry, and idiosyncratic parts, leads to (statistically) significant improvements in forecast accuracy for both inflation rate and its volatility. Finally, we suggest a dynamic switching model based on in-sample inflation volatility, which is an extreme case of forecast combination, and find that this can further improve forecast accuracy, indicating that there is no single best forecasting model, i.e. different models perform better in different volatility regimes.

The second study in this thesis, "Forecast Combination in R Using the GeomComb Package" documents a software contribution. The GeomComb package in R was developed by us with the aim to provide a comprehensive toolset for forecast combination, a statistical approach to forecasting that is based onn the (weighted) averaging of several individual forecasts for the same time series, which can often improve upon the best individual forecast and is a suitable strategy to reduce model risk. The package provides functions that can be used for all steps of the forecasting procedure: (a) *data processing* – functions dealing with common forecast combination issues such as missing data and multicollinearity; (b) *forecast estimation* – 15 different static and dynamic forecast combination methods ranging from simple statistics-based methods to more sophisticated regression-based and eigenvector-based methods; (c) *resutls interpretation* – summary and plotting functions allowing the users to rationalise forecast combination results. The functionalities of the package are demonstrated using UK electricity supply data.

The third study in this thesis, "Efficient Nurse Staffing: The Value of Hierarchical Time Series Forecasting and Forecast Combination", contributes to the healthcare operations literature. Combining the value of HTS forecasting and forecast combination techniques, we show how the approaches can be used to minimise permanent and temporary nurse staff cost in a large UK teaching hospital. We formulate a parametric straffing model, a constrained optimisation, and find that the current norm in nurse staffing (based only on patient census and ignoring workload related to patient churn) lead to chronic understaffing issues, jeopardising patient safety and staff satisfaction. We show how understaffing can be avoided by incorporating three factors into the staffing model: (a) the possibility to hire temporary staff, (b) taking into account workload related to patient churn, and (c) accounting for forecast uncertainty using the empirical forecast error distribution. By adding these factors to the staffing model, the probability of understaffing can be decreased from almost 80% to 1%. The value of hierarchical forecasting and forecast combination is evident when using the staff model that aims to avoid understaffing, reducing staff cost by 5% compared to the benchmark forecast (a seasonal random walk). Due to its parametric nature, the model is very flexible regarding the cost and quality inputs, so that it can be readily employed by health care providers worldwide for cost-efficient nurse staffing that aims to avoid understaffing.