Applied Methods for Forecasting Economic Aggregates and Their Components

Thesis

Submitted in partial fulfilment of the requirements
of the Degree of Doctor of Philosophy

by

Marcus Paul Andrew Cobb

School of Economics and Finance
Queen Mary University of London

September 2018

Statement of originality

I, Marcus Paul Andrew Cobb, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Marcus P. A. Cobb

Date: September 30th, 2018

Details of publications:

Preliminary versions of the chapters have been published as working papers at the University Library of Munich as: "Joint Forecast Combination of Macroeconomic Aggregates and Their Components," MPRA Paper 76556, "Improving Underlying Scenarios for Aggregate Forecasts: A Multi-level Combination Approach" MPRA Paper 88593, "Forecasting Economic Aggregates Using Dynamic Component Grouping," MPRA Paper 81585, and "Aggregate Density Forecasting from Disaggregate Components Using Large VARs," MPRA Paper 76849.

Abstract

This thesis focuses on improving the accuracy of forecasts for economic aggregates by developing applied methods that take advantage of the strengths from both the direct and bottom-up approaches. The starting point for developing each one of these is the idea that forecasting methods that may not in themselves provide an adequate answer in a particular setting can nevertheless contribute valuable information to the forecasting process. The challenge lies in identifying and appropriately incorporating the relevant information.

The first of the three methods focuses on increasing overall forecasting accuracy by jointly combining forecasts for an aggregate, any sub-aggregations and the components, from any number of models and measurement approaches. The framework seeks to benefit from each of the forecasting approaches, by accounting for their reliability in the combination process and exploiting the constraints that the aggregation structure imposes on the set of forecasts as a whole. The second method presented is one that forecasts economic aggregates using purpose-built groupings of components. The objective of developing such a method is to increase forecasting accuracy by transforming the data in a way that avoids the problems associated with disaggregate misspecification, while still allowing for distinct disaggregate dynamics to be picked up in the process. Finally, a method is developed to produce an aggregate density forecast from the density forecasts of its components in a way that considers the interaction between them. The motivation for doing this rests on the assumption that accounting for interdependencies should provide a more complete probabilistic assessment.

Overall, this research shows that there are benefits to be obtained from using alternative aggregation approaches. The gains from using these methods, however, depend critically both on how they are specified and on the particular dataset. In this context, combining many specifications appears as a way of obtaining consistently good results.

Contents

List of Tables

Chapter 2

Chapter 3

Chapter 4

List of Figures

Chapter 2

Chapter 3

Chapter 4

Chapter 1

Introduction

1.1 Motivation

The world can be seen as a highly complex and interconnected system, where millions of agents are constantly at work, pursuing objectives that are unknown to most of the rest of them. The economy is a part of this system, meaning that assessing its state and providing an outlook for its likely path involves interpreting large amounts of data in a coherent way. Economic aggregates are fundamental to this process, given that they synthesise the information from countless sources into relatively few figures. Bearing witness to this is the fact that with new releases, financial markets react, sentiments are affected and agents revise their assessment of the economy, adjusting their beliefs regarding the direction it is taking. In this context, considerable resources are devoted to predicting key economic variables, and accurate forecasts are valued greatly. This is especially the case at central banks and other policy-making institutions where decisions are taken to steer the economy in a certain direction and, therefore, have a direct impact on people's lives. The relevance of the decisions by central banks, for example, can be seen in the fact that their statements and minutes have a measurable effect on financial market trading and economic expectations, particularly when they call for changes in the views held by the committee members concerning the state of the economy (Reeves and Sawicki, 2007; Rosa, 2013).

In this context, it is not surprising that the literature on forecasting economic aggregates is vast and, with the ongoing increase in computational power and capacity, the research in this area is likely to continue to expand. Notwithstanding the many different methods that exist, there is consensus regarding the fact that it is very unlikely for a single one to outperform all of the others in every situation (Elliott and Timmermann, 2005). As pointed out by Alessi et al. (2014), the significant deterioration in forecasting power observed during the most recent financial crisis, in the case of many of the models that had previously performed well, only provides more evidence of this fact. In an in-depth evaluation of the effects of crisis periods on forecasting models' performance, Chauvet and Potter (2013) come to the conclusion that otherwise well-performing models often fail to deliver when faced with rapidly-changing conditions and abrupt increases in volatility. They also find that models perform very differently in expansions and recessions.[1] Because of the fact that empirical performance of any model is fundamentally determined by the data, forecasters usually resort to an array of different models to forecast economic variables. The challenge posed by doing so lies in being able to extract the relevant information from each one of them in a timely fashion (Hubrich and Skudelny, 2017).

As regards forecasting economic aggregates, there is a strand of research that has devoted itself to determining whether forecasting an aggregate directly is better in terms of aggregate accuracy than following a bottom-up approach: i.e. forecasting its components and summing them up. In theory, the bottom-up approach should outperform the direct method, but in practice it is the data and the aggregation structure that decide which of the methods is best (Lütkepohl, 1987). This is due to the fact that it is usually impossible to specify the underlying disaggregate processes perfectly. As Bermingham and D'Agostino (2014) point out, the bottom-up approach would be more sensitive than the direct method to structural breaks and abnormalities in the data, and macroeconomic time-series are known to be affected by these events relatively often (Stock and Watson, 1996).

There is a lot of evidence, however, that using the components to forecast an aggregate might improve its accuracy. The gains would be attributable to the fact that doing so allows underlying dynamics to be explored that are imperceptible or very hard to distinguish in the aggregate (Bache et al., 2010; Brüggemann and Lütkepohl, 2013). Motivated by this, the subject of this thesis is to improve the overall accuracy of forecasts for economic aggregates by developing applied methods that take advantage of the strengths of both direct and bottom-up approaches. This includes, but is not restricted to, improving the aggregate forecast in view of the fact that in some situations the accuracy of the underlying forecasting scenario is also of interest. The starting point for developing each of the methods is the idea that forecasting approaches that may not provide an adequate answer by themselves in a particular setting can nevertheless contribute valuable information to the forecasting process. As in the case of using different models, the challenge lies in identifying and appropriately incorporating the relevant information so that it serves its purpose. The first method presented is motivated by the fact that in certain situations forecasts for a number of sub-aggregations are required for analysis, in addition to one for the aggregate (Esteves, 2013; Espasa and Senra, 2017). In this context, the method involves developing a framework to im-

[1] This last point has been previously documented in Marcellino (2008) who finds that in recessions the more sophisticated models show a marked deterioration.

prove the forecasting accuracy of the whole underlying forecasting scenario. With this objective, a fully consistent forecasting scenario is produced by combining any number of forecasts for an aggregate, its components and any desired sub-aggregation. The framework seeks to benefit from the strengths of each of the forecasting approaches by accounting for their reliability in the combination process and exploiting the constraints that the aggregation structure imposes on the set of forecasts as a whole. The results for an empirical application of the method using Headline Inflation, a number of common sub-aggregations and the disaggregate components, show that there are gains in overall accuracy from including all of them together. In the following chapter, a method is proposed to increase overall accuracy by forecasting purpose-built groups of components. These groupings are selected on the basis of features of the data that would favour forecasting them together. The idea behind the approach is to form sets of groupings that allow the disaggregate dynamics to be modelled, while avoiding the problems of disaggregate misspecification. The method works in two stages. In the first, it uses a clustering technique to select a subset of groupings and, in the second, it selects a definitive grouping based on a criterion. In an exercise with CPI data a number of specifications are put to the test. The results show that some of them often outperform the best traditional alternatives. In the last section, a framework designed to produce an aggregate density forecast based on its components' forecasts is presented. The reason for doing this lies in the fact that accounting for the interdependencies between components could provide a more complete probabilistic assessment. The approach involves modelling the whole multivariate process using large Vector Autoregressions and empirical Bayesian methods. In an exercise using CPI and GDP data it is found that the framework performs well overall and, in particular, that it does better than the univariate alternatives in dealing with the shocks produced by the most recent global financial crisis. When combined with the direct approach, the combined forecasts are often significantly better than those of the direct approach on its own.

1.2 Outline

This section provides a general outline of the thesis and its contributions to previous research. A detailed outline with references to related literature can be found in the introduction of each chapter. The thesis is divided into five chapters. Chapter 2 presents a flexible framework for combining forecasts simultaneously for an aggregate, any number of sub-aggregations and its components, based on the merits of each one of them. The reasoning behind wanting to do this is that forecasting different valid representations of a series could provide a more complete picture of what to expect for its future path.[2] For economic aggregates, these valid forms of representation can come

[2] The underlying rationale is similar to the idea of different models helping to approximate the data generating process of a series.

from using different levels of aggregation and sub-aggregation. The research presented in this chapter extends the current literature on forecast combination in two ways. First, it incorporates multiple aggregation levels and representations into the same combination process and, second, it considers the reliability across aggregation levels of each individual forecast in the process. The method is formulated as the solution to a problem of reconciling all the different forecasts so that the outcome is a single set that is fully consistent. In practice, that means producing individual forecasts for the aggregate, all sub-aggregations and all the components and considering them as initial guesses. They are then updated on the basis of their relative reliability, so that they comply with the corresponding identities that define the aggregate. The problem is set as one of solving a sequence of analogous constrained quadratic programs. Under fairly mild assumptions a general analytical solution is derived, meaning that the algorithm that solves the entire problem is easy to implement and fast. An empirical application is performed using CPI data for France, Germany and the United Kingdom for the 1991-2015 period. In the exercise, the reliability weights are determined empirically using methods that are often used in the forecast combination literature. The results show that the method performs as well as equivalent traditional combination methods, in terms of aggregate accuracy, and equally well or better than the best single models in terms of disaggregate accuracy. This would suggest that by effectively imposing the aggregation structures on the individual forecasts, the framework translates the accuracy of the better-performing approaches into increases in disaggregate accuracy.

In Chapter 3 a method is developed to forecast economic aggregates using purpose-built groupings of components. The underlying idea is that summation can produce new series with properties that differ quite significantly from those originating them and, therefore, data could be transformed so that forecasting models perform better than with an exogenously determined disaggregation structure. The task involves finding groupings that avoid certain problems associated with disaggregate forecasting, while still allowing for distinct disaggregate dynamics to be picked up in the process. The work is novel in that it deals with the dimensionality problem associated with using disaggregate data by applying concepts from the statistical learning literature, while still retaining interpretability. In particular, a two-stage method that combines statistical learning techniques and traditional economic forecasting evaluation is developed. In the first stage, Agglomerative Hierarchical Clustering is used to reduce the size of the problem by choosing a subset of feasible groupings based on the commonality among the different components. In the second stage, selection procedures are used on the resulting hierarchies to produce the final aggregate forecasts. These selection procedures include choosing a single grouping based on some criterion and combining the whole subset of groups. In an empirical application using CPI data for France, Germany and the United Kingdom, different parameter choices are evaluated. The results show that some of the methods for selecting a unique grouping perform better than the best benchmark methods. They also show that the forecast combination methods perform well overall, suggesting that expanding the pool of forecasts by trying different combinations of components provides a way of benefiting from the strengths of forecast combination, without necessarily increasing the number of forecasting models.

In Chapter 4 the focus shifts to producing density forecasts for economic aggregates based on the density forecasts for their components. The motivation for developing such a method comes from the literature which argues that accounting for the dynamics of the underlying components can contribute to a better understanding of a wide range of uncertainties. The work extends the current literature in that a bottom-up framework that models the whole multivariate process is developed. This means that the interaction between components is taken into consideration. To do so, the methodology of large Bayesian VARs is used. The particular implementation allows for both fixed and time-varying parameter VARs and for stochastic volatility. Acknowledging that competing forecasts can serve a complementary role, the framework is extended so as to admit combining direct and bottom-up forecasts. In doing this, two commonly used approaches from the relevant literature are used: the linear and logarithmic opinion pools. The empirical application uses CPI and GDP data from France, Germany and the United Kingdom for the 1991-2015 time period. The results show that the multivariate methods are capable of producing bottom-up forecasts that are calibrated and perform equally well or better than comparable aggregate methods. Their advantage over the benchmark methods, however, is given mainly by their performance over the financial crisis. This is probably the reason why combining the direct and the multivariate bottom-up approaches is found to produce significant improvements in performance over the individual approaches.

Finally, Chapter 5 summarizes the main conclusions of this work.

Chapter 2

Improving Underlying Scenarios for Aggregate Forecasts: A Multi-level Combination Approach

2.1 Introduction

Macroeconomic aggregates play a fundamental role in assessing the state of the economy. Consequently, many different people and institutions devote considerable resources to predicting key economic variables. When it comes to policy-making institutions, however, the interest usually goes beyond that of the aggregate alone. As Esteves (2013) points out, these institutions often need to have detailed breakdowns of their aggregate forecasts. One obvious reason for this is that they may wish on occasions to provide the public with some additional information. It is likely, however, that the strongest reasons have to do with the analysis that remains within the institution. In the context of inflation forecasting, Espasa and Senra (2017) encourage looking beyond the aggregate alone, based on the argument that a similar Headline Inflation can correspond to very different inflation situations which, in turn, may require very different actions to be taken by the authorities.

Policy-making institutions find it relevant to look at a breakdown of components, because this can provide useful information concerning the components themselves, provide better understanding of the aggregate and increase aggregate forecasting accuracy (Espasa and Senra, 2017). In line with this view, it is not uncommon to find alternative disaggregation scenarios being presented within the same assessment, as a way of providing further insight for a particular topic. In this context, if forecasts for the aggregate, any sub-aggregations and the components are produced independently of one another, inconsistent and conflicting messages may appear. Because of this, practitioners typically rely on using a bottom-up approach that includes all the necessary components to produce a consistent underlying forecasting scenario (Esteves, 2013; Ravazzolo and Vahey, 2014).

However, using the bottom-up approach alone can mean that aggregate accuracy is negatively affected when compared with other methods. In particular, the relevant empirical literature points out that, depending on the scenario, the direct methods may produce more accurate forecasts than the bottom-up approach.[1] There are strong arguments in favour of using direct approaches instead of the bottom-up approach when the concern is aggregate accuracy alone. One of these is that, due to cancellation between components, aggregates can behave relatively smoothly even when the disaggregate data has a high degree of volatility (Hyndman et al., 2011). Another is that common factors that are relatively unimportant at an individual level may dominate the aggregate (Granger, 1987). A third is that bottom-up strategies that treat the disaggregate components independently would almost certainly be misspecified, because they cannot properly approximate the underlying multivariate process (Hendry and Hubrich, 2011). Practitioners who do not want to rely on the bottom-up approach have the alternative of simply using direct methods for the aggregate and reconciling the disaggregate forecasts when needed. Proceeding in this way, however, has the undesirable result that any useful information arising from the interactions between components is discarded (Hyndman et al., 2011). As pointed out previously, the direct method is better than the bottom-up approach in certain situations. In others, it is the latter that performs best. A forecaster who is concerned with overall accuracy will, therefore, want to benefit from both methods if possible.

If the concern were only for the aggregate, a popular way of dealing with competing forecasts would be simply to combine them. The idea of forecast combination was put forward quite a while ago in Bates and Granger (1969) and deals with the issue of exploiting the information contained in each individual forecast in the best possible way. The evidence in favour of using these methods as a way of increasing forecasting accuracy is substantial (Timmermann, 2006). As explained by Hoogerheide et al. (2010), a common justification for using forecast combination is that in many cases it is impossible to identify the true economic process and, therefore, different models play a complementary role in approximating it. Another is that, in the context of being unable to establish the single model that produces the smallest forecasting error

[1] This is supported by many empirical comparisons like: Espasa et al. (2002), Benalal et al. (2004), Hubrich (2005) and Giannone et al. (2014) for inflation in the Euro area; Marcellino et al. (2003), Hahn and Skudelny (2008), Burriel (2012) and Esteves (2013) for European GDP growth; and Zellner and Tobias (2000), Perevalov and Maier (2010) and Drechsel and Scheufele (2013) for GDP growth in specific industrialized countries.

in advance, combination appears as a way of hedging against choosing an exceedingly bad one (Hubrich and Skudelny, 2017). The second of these justifications seems particularly relevant for policy-makers, given their aversion to correcting their published assessments.[2]

Notwithstanding the extensive literature on combination methods, almost all of it deals with one variable at a time. In the context of forecasting economic aggregates and their components, this means disregarding the aggregation structure as a source of valuable information. A notable exception to this apparent omission in the combination literature is that of Hyndman et al. (2011). They propose a combination method to improve overall accuracy of an aggregate and its components, using the structure underlying the aggregate and any sub-aggregations. Specifically, they use individual forecasts for all levels of aggregation and combine them optimally. In their empirical application, they find that the method improves overall accuracy. Despite their good results, it is limited in at least two ways that could restrict its applicability to other problems. On the one hand, the combination weights are determined solely by the aggregation structure. On the other, the method can only handle a single hierarchical structure.

In terms of the combination weights, there are situations where information regarding the quality of the forecasts under consideration is available, in which case it could be desirable to be able to incorporate this information into the combination process. Such a situation could come up in a context where data is released asynchronously. Because of the lag in the publication of GDP, for example, current quarter growth is routinely estimated based on leading indicators (Antipa et al., 2012).[3] In many cases this involves estimating GDP components based on information that is usually not published at the same time, meaning that a new estimate can be produced with every new release (Bell et al., 2014; Higgins, 2014; Mogliani et al., 2017).[4] In this context, it should be expected that the relative reliability of the different forecasts would change significantly every time a model is run. Also, even if no prior information regarding the reliability of the forecasts is available, another reason why it may be desirable to have some control over the combination weights is that the combination literature highlights the gains that can be obtained from weighting different forecasts, based on their recent performance (Timmermann, 2006).

As regards allowing more than one hierarchy to be considered, the appeal lies in the fact that alternative measurement approaches and stratifications can provide valuable

[2] Goodhart (2004) argue that this aversion stems from the fact that perceived mistakes by the central banks could rapidly undermine the public's confidence in them.

[3] In Europe, for example, the first preliminary estimate of total GDP is released about 45 days after the end of the reference quarter and the first complete estimate about 65 days later.

[4] The Federal Reserve Bank of Atlanta's nowcasting tool, for example, is updated on average five or six times a month following nearly every major economic data release (Higgins, 2014).

information for the forecasting process. For example, based on the theoretical arguments given by Clark (2004), Peach et al. (2013) and Tallman and Zaman (2017) find significant improvements in aggregate accuracy from forecasting the prices of goods and services separately. Hargreaves et al. (2006) and Jacobs and Williams (2014) make a similar case for tradable and non-tradable inflation. Being able to consider both stratifications in the combination process may therefore be desirable. The same argument can be made for forecasts coming from different measurement approaches. Frale et al. (2011), for example, find gains in aggregate accuracy from combining forecasts from the production and expenditure approaches for measuring GDP, while Aruoba et al. (2013) do so for the income and expenditure perspectives.

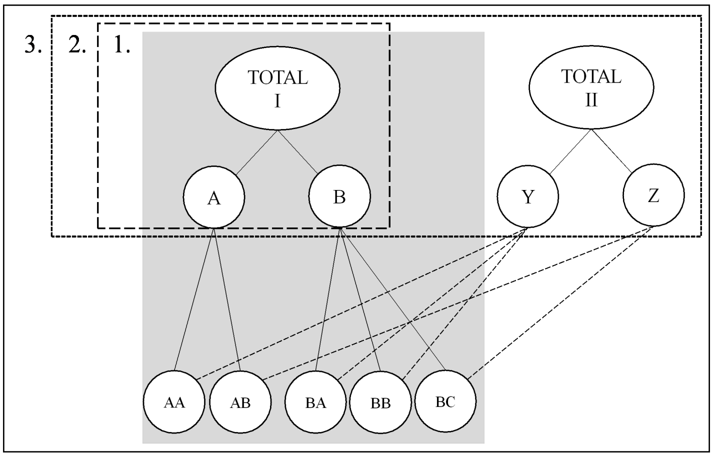
This chapter picks up on this point and, in order to improve overall accuracy of a disaggregate forecasting scenario, develops a framework that is flexible enough to incorporate both these aspects. For this purpose, it brings together the literature devoted to increasing forecasting accuracy through alternative disaggregation choices with that of forecast combination. The method consists of producing individual forecasts for all the series involved and considering them as initial guesses. They are then updated, based on their relative reliability, so that they comply with the identities that define the aggregate.

The rest of the chapter is organized as follows: Section 2.2 develops the framework that allows series from different levels of aggregation from any number of measurement approaches to be combined. Section 2.3 presents an empirical implementation using CPI data for France, Germany and the United Kingdom. Section 2.4 summarizes the conclusions.

2.2 A Framework for Combining Forecasts from Different Aggregation Levels and Alternative Measurement Approaches

The motivation for developing a multi-level combination method is that incorporating the information regarding the aggregation structure into the forecasting process of the components could improve their accuracy. Given that any set of component forecasts necessarily implies an aggregate forecast, it is also desirable that the multi-level method should exhibit the improvements that are expected from aggregate combination alone. For this reason, in what follows, the task of developing a multi-level combination framework is viewed as one of extending traditional single-variable combination methods so that they allow the bottom-up aggregate forecasts to be expressed in terms of the underlying component forecasts.

Figure 2.1: Different Aggregation Scenarios for an Aggregate



Note: Numbered squares highlight different aggregation scenarios: 1. A single one-level hierarchy, 2. Two different measurement approaches for the same aggregate, each based on a one-level hierarchy, and 3. A two-level hierarchy with two sets of non-nested sub-aggregations. The shaded rectangle highlights the type of hierarchical structure considered in Hyndman et al. (2011).

In this context, a property that is required in developing the multi-level method is that it should result in the same outcome as that of a comparable traditional method, if the circumstances are equivalent. An example for this is that, in a context where the direct aggregate and bottom-up forecasts are equally reliable, the final aggregate forecast that results from combining both aggregate forecasts should be the same as that of combining the direct aggregate forecasts with those of the components. In incorporating the components into the combination process, two additional properties are considered to be desirable. The first requires consistency between the reliability of the components and that of the resulting aggregate. Although it could be argued otherwise, it makes economic sense that if all components have the same reliability according to some measure, this should be equal to the reliability of the aggregate that results from adding up the components. The second additional property establishes that once the reliability of the different forecasts is taken into consideration, in line with considering the initial estimates as the best guesses, the combination procedure should result in each of the definitive forecasts deviating as little as possible from its initial estimate.

To have a notion of the forecasting setting under consideration, Figure 2.1 presents a simple picture of the general aggregation structure. It shows two different measurement approaches for the same aggregate, based on the same basic components. By considering non-nested sub-aggregations, the structure is not strictly hierarchical in the sense considered by Hyndman et al. (2011). Figure 2.1 also outlines the strategy for developing a method to solve such a combination problem. It consists in starting from a simple problem and progressively extending it to the more complex setting. The numbered squares illustrate this progression. The first two steps consider developing the necessary framework to solve the problem for a single one-level hierarchy and then extending it to admit multiple disjoint measurement approaches. The third and final step consists in using the results from both the previous settings to solve the combination for any number of levels and sub-aggregations. In practice, this is done by formulating the general problem as one of a succession of one-level combinations.[5]

2.2.1 One-Level Hierarchies

People working on the compilation of aggregate statistics regularly face the need to balance information from different sources in order to produce official statistics. In many of those applications, like the production of national accounts and social-accounting matrices, the reconciliation process involves a massive amount of data, with the result that procedures have been proposed over the years to iron out the differences (Dalgaard and Gysting, 2004). In a recent paper, Rodrigues (2014) casts the whole problem of balancing statistical economic data into a Bayesian framework. It suggests treating the data as stochastic processes, modelling the prior properties accordingly and finding the balanced posterior by means of relative entropy minimization.

The process proposed by Rodrigues (2014) equates to searching for a posterior distribution that is as close as possible to the prior while satisfying the required restrictions. Although the implementation is specific to balancing economic data, the principle behind the framework resembles the problem of any sort of forecast combination. The individual forecasts serve as best guesses, where different forecasts have different reliability and cross-sectional identities must be met. It establishes that a number of the conventional reconciliation methods are in fact particular cases of the general framework and shows that there is a one-to-one correspondence. Based on this correspondence, it is shown that it is possible to identify the conventional method's underlying assumptions and suggest using least squares approaches when uncertainty estimates are available.

2.2.1.1 Optimization Problem for a Single One-Level Hierarchy

The problem of combining direct aggregate forecasts with the components from a bottom-up approach is one of finding the set of forecasts that satisfies the required restrictions and is as close as possible to the preliminary figures. For this purpose, a least-squares formulation is used. This means letting the undefined criterion for "as close as possible" be governed by some quadratic loss function.

The problem for a one-level hierarchy is expressed as a general constrained quadratic

[5] The general framework is derived step-by-step in section 2.A of the Appendix.

program of the form:

$$\min_{\alpha,\beta} \sum_{i=1}^{A} f_{i,t}\left(y_{i,t}, \alpha_{i,t}, \varphi_{i,t}\right)^2 + \sum_{d=1}^{D}\sum_{j=1}^{N} g_{d,j,t}\left(q_{d,j,t}, \beta_{d,j,t}, \phi_{d,j,t}\right)^2 \tag{2.1}$$

subject to:

$$\left(1+\alpha_{1,t}\right) y_{1,t} - \sum_{j=1}^{N}\left(1+\beta_{1,j,t}\right) w_{1,j,t} q_{1,j,t} = 0$$

$$\left(1+\alpha_{1,t}\right) y_{1,t} - \left(1+\alpha_{i,t}\right) y_{i,t} = 0 \qquad \text{for } i = 2 \text{ to } A$$

$$\left(1+\beta_{1,n,t}\right) q_{1,n,t} - \left(1+\beta_{d,n,t}\right) q_{d,n,t} = 0 \qquad \text{for } d = 2 \text{ to } D,\ n = 1 \text{ to } N$$

where $y_{i,t}$ is the preliminary forecast for time t of the i-th aggregate model of a total of A, $\alpha_{i,t}$ is the percentage deviation of the definitive forecast from the preliminary, $\varphi_{i,t}$ is its exogenously chosen optimization weight and $f_{i,t}$ is some function of the three. Similarly, $q_{d,n,t}$ is the preliminary forecast for time t for component n of the d-th model of a total of D disaggregate models, $\beta_{d,n,t}$ is the percentage deviation of the definitive forecast from the preliminary, $\phi_{d,n,t}$ is its exogenously chosen optimization weight, $g_{d,n,t}$ is some function of the three and $w_{d,n,t}$ is the respective aggregation weight.[6]

2.2.1.2 An Analytic Solution for a Single Set of Forecasts

With the problem formulated in this way, in addition to the obvious influence of the reliability weights, it is the choice of loss function that ultimately determines the outcome. To facilitate finding an appropriate loss function, the problem is first restricted to that of combining one set of forecasts. That is, only one direct aggregate forecast and a single set of disaggregate forecasts. In this context, the following loss function is proposed:

$$\varphi_t \left(\alpha_t y_t\right)^2 + Q_t \sum_{j=1}^{N} \phi_{j,t} w_{j,t} q_{j,t} \beta_{j,t}^2 \tag{2.2}$$

with $Q_t = \sum_{j=1}^{N}\left(w_{j,t} q_{j,t}\right)$.

In deriving this particular loss function, the empirical success of the simple weighted averages is used as the foundation that is then extended to admit aggregates and components in the same problem. There is ample evidence suggesting that in practice simple methods often perform better than more involved procedures (Timmermann, 2006), with the equal-weighted average standing out as a benchmark that is hard to beat (Smith and Wallis, 2009; Elliott, 2017). To admit the components into the combination procedure, the proportional distribution approach proposed by Denton (1971) is

[6] All variables are in levels and for simplicity it is assumed that all components and aggregation weights are strictly positive.

used. As pointed out by Pavia-Miralles (2010), this is one of the most successful methods in the reconciliation literature, given its simplicity and overall good performance. The approach fits well within the framework as it involves minimizing the percentage deviation between the definitive series and the initial approximations.

Using this loss function and minimizing it subject to the restriction that the aggregate has to be equal to the sum of the components produces as the solution that the definitive aggregate forecast is:

$$\tilde{y}_t = \tilde{Q}_t = \frac{Q_t^2 + y_t \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)}{Q_t + \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \tag{2.3}$$

and the definitive forecast for any given component is:

$$\tilde{q}_{n,t} = \left(1 + \frac{\varphi_t}{\phi_{n,t}} \cdot \frac{y_t - Q_t}{Q_t + \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \right) q_{n,t} \tag{2.4}$$

From these results, the fulfilment of the desirable properties set out in the introduction to this section can be verified.

It is easy to see that the initial estimates of the components are modified by a factor that is the same for all components, except the first term, $\frac{\varphi_t}{\phi_{n,t}}$. If all components have equal reliability, that is $\phi_{n,t} = \phi_t$ for all n, the expression $\sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)$ is equal to $\frac{\varphi_t}{\phi_t} Q_t$ meaning that the property regarding the coherence between aggregate and disaggregate reliability weights is fulfilled. With this, equation (2.4) simplifies down to:

$$\tilde{q}_{n,t} = q_{n,t} + \frac{\varphi_t}{\phi_t + \varphi_t} \cdot (y_t - Q_t) \frac{q_{n,t}}{Q_t}$$

making obvious the proportional distribution of the difference between the preliminary aggregate forecasts between components. Likewise, equation (2.3) simplifies to the weighted average of the aggregate forecasts, meaning that the equivalence with the traditional combination methods under comparable circumstances is met.

The suggested loss function results in the desired outcome for one set of forecasts. If more than one set is considered for each variable, the outcome does not meet the aforementioned conditions. The problem can be avoided, however, simply by combining the multiple forecasts for the individual series before performing the combination of different levels and choosing the optimization weights so as to reflect the previous step.[7]

[7] This is shown in section 2.A of the Appendix.

2.2.1.3 Extension to multiple disjoint measurement approaches

On occasions, forecasts from more than one measurement approach may be available. An immediate example of this is the fact that there are three measurement perspectives for GDP. In this context, it could be beneficial to incorporate them into the same combination process. The one-level combination method developed in the previous section can easily be extended to do so.

For an aggregate that can be obtained as the sum of K alternative measurement approaches, where each approach is the result of the weighted sum of the respective strictly positive N_k components, let there be a direct aggregate forecast y and K distinct aggregate forecasts, each based on the corresponding N_k components' forecasts.

The minimization problem involving the aggregate reliability weight φ, the disaggregate reliability weights $\phi_{k,n}$ and the aggregation weights $w_{k,n}$, is:

$$\begin{aligned}\min_{\alpha,\beta} \varphi_t \left(\alpha_t y_t\right)^2 + \textstyle\sum_{k=1}^{K} \Big[Q_{k,t} \textstyle\sum_{j=1}^{N_k} \phi_{k,j,t} w_{k,j,t} q_{k,j,t} \left(\beta_{k,j,t}\right)^2 \\ +2\lambda_k \Big((1+\alpha_t) y_t - \textstyle\sum_{j=1}^{N_k} w_{k,j,t} (1+\beta_{k,j,t}) q_{k,j,t} \Big) \Big]\end{aligned}$$

Solving the problem subject to the corresponding constraints results in the definitive aggregate forecast being:

$$\tilde{y}_t = \frac{y_t + \displaystyle\sum_{k=1}^{K} \left(Q_{k,t} \cdot \frac{Q_{k,t}}{\chi_{k,t}} \right)}{1 + \displaystyle\sum_{k=1}^{K} \frac{Q_{k,t}}{\chi_{k,t}}} \tag{2.5}$$

where $\chi_{k,t} = \displaystyle\sum_{j=1}^{N_k} \frac{\varphi_t}{\phi_{k,j,t}} w_{k,j,t} q_{k,j,t}$. The definitive forecast for any given component is given by:

$$\tilde{q}_{k,n,t} = \left(1 + \frac{\varphi_t}{\phi_{k,n,t}} \cdot \frac{\tilde{y}_t - Q_{k,t}}{\chi_{k,t}} \right) q_{k,n,t} \tag{2.6}$$

As in the case of a single hierarchy, for more than one set of forecasts, the same combination process is followed, except that the multiple forecasts are combined in a prior step and optimization weights are chosen to reflect this.

2.2.1.4 Bounds for and Response to Reliability Weight Values

From the solution given by equation (2.3), it is clear that reliability weights have a defining impact on the final outcome. It is important, therefore, to establish the feasible region in which they guarantee a unique solution for the minimization problem. It is also useful to learn about the sensitivity of the final outcome to the choice of weights.[8]

From the solutions it is immediately clear that what matters is the relative reliability and therefore that the impact of a given value has to be examined in relation to the rest of the components. Considering as a starting point that all weights are set equal to some value, one extreme is to have no confidence in certain forecasts. If this were the case for the aggregate forecast only, this would mean making $\varphi_t = 0$ and therefore $\tilde{y}_t = Q_t$. On the other hand, if it were the case for a single component $n = 1$, making $\phi_{1,t} = 0$ means that this component absorbs all the deviation. This is clear from appreciating that $\frac{\varphi_t}{\phi_{1,t}} \to \infty$ and therefore that $\lim_{\phi_{1,t} \to 0} (1 + \alpha_t)\, y_t = y_t$. This means that the forecasts from all but this component are taken as given and that the definitive forecast $\tilde{q}_{1,t}$ is found residually. It also means that only one forecast can have a reliability weight equal to zero, otherwise the minimization problem has infinite solutions.

The other extreme is to be completely confident about some forecasts. If this were the case for the aggregate forecast, this means making φ_t go to infinity. In such a case it is easy to see that $\lim_{\varphi_t \to \infty} (1 + \alpha_t)\, y_t = y_t$. On the other hand, for a single component $n = 1$, making $\phi_{1,t}$ go to infinity implies that $\frac{\varphi_t}{\phi_{1,t}} \to 0$. This means that the weight given to the direct forecast decreases but still remains positive. Taking it to the extreme and making all component weights go to infinity decreases to zero the weight given to the direct forecast. That is $\lim_{\phi_t \to \infty} (1 + \alpha_t)\, y_t = Q_t$ where $\phi_{n,t} = \phi_t$ for $n = 1$ to N.

For the purpose of allowing for some degree of combination it makes sense to restrict the aggregate forecasts by giving them finite reliability weights. For the components, on the other hand, one could have a weight that implies certainty, maybe due to the early release of relevant data. Following these guidelines, however, does not necessarily prevent nonsense results occurring. This might happen, for example, when some forecasts are considered to be as good as certain. Setting valid but contradictory reliability weights could result in unintended outcomes such as components measured in levels becoming negative, due to insufficient degrees of freedom in the combination procedure.

As regards the sensitivity of the outcome to different values of the reliability weights, it is possible to see how the solution in equation (2.3) is affected by varying φ_t and $\phi_{n,t}$ by looking at the effect of the reliability of one component when the rest are held constant. For this purpose, let $\phi_{i,t} = k\varphi_t$ and $\phi_{n,t} = \varphi_t$ for all other components. Using these weights results in the solution being:

$$\tilde{y}_t = \left(1 + \frac{k}{k - (k-1)s_i}\right)^{-1} \left(\frac{k}{k - (k-1)s_i} Q_t + y_t\right) \tag{2.7}$$

[8] For simplicity the analysis is performed for a single hierarchy and only one set of forecasts.

where $s_i = \frac{w_{i,t} q_{i,t}}{Q_t}$.

Not surprisingly, the additional weight that is given to the bottom-up forecast depends on the relative reliability of the component and its weight within the aggregate. If there were only one component -that is equivalent to having many components but giving them all the same reliability- $s_i = 1$ and Q_t would be given k times more weight than y_t. On the opposite side of the spectrum, as s_i tends to zero the extra weight given to Q_t converges to zero.

2.2.2 Multi-Level Hierarchies and Alternative Sub-aggregations

The approach for combining forecasts from two levels of aggregation may be useful in many settings. In others it may be too restrictive to be applicable. It can be used, however, as the basis for extending the approach to multiple levels of aggregation. The multi-level method involves deriving a set of consistent component forecasts for each forecast, for the aggregate and sub-aggregations, and then combining them to produce a definitive bottom-up forecast. The approach effectively breaks down the whole problem into a sequence of one-level combinations. In terms of the aggregate forecasts, it is shown that this is equivalent to combining the aggregate forecasts produced from different intermediate aggregation levels for the case of equal reliability weights. By construction, the result is a fully consistent forecasting scenario.[9]

2.2.2.1 An Aggregate Forecast Expressed as a Set of Reconciled Components

Let there be a single aggregate forecast y and a single set of disaggregate forecasts q_n for $n = 1$ to N, the aggregate reliability weight φ, the disaggregate reliability weights ϕ_n and the aggregation weights w_n. In this context, based on the one-level framework, the aggregate and component forecasts are given by equations (2.3) and (2.4). Then, to have a disaggregate scenario that is consistent with y taking q_n, for $n = 1$ to N, as the best guesses, it is enough to make the aggregate reliability arbitrarily large, $\varphi \to \infty$. With this, the y-consistent component forecasts are given by:

$$\hat{q}_{n,t}^{(y)} = \left(1 + \frac{y_t - Q_t}{\phi_{n,t} \cdot \sum_{j=1}^{N} \left(\frac{1}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \right) q_{n,t} \tag{2.8}$$

Having taken into consideration the relative reliability of the components in the process of producing the y-consistent components, the new set of forecasts can inherit the reliability of y. With this, definitive component forecasts can be produced by combining

[9] The derivation is shown in detail in section 2.B of the Appendix,

the original and y-consistent forecasts:

$$\begin{aligned}\tilde{q}^{alt}_{n,t} &= \tfrac{\phi_{n,t}q_{n,t}+\varphi_t\hat{q}^{(y)}_{n,t}}{\phi_{n,t}+\varphi_t} \\ &= \left(1+\tfrac{\varphi_t}{\phi_{n,t}}\cdot\frac{y_t-Q_t}{(\phi_{n,t}+\varphi_t)\sum_{j=1}^{N}\frac{1}{\phi_{j,t}}w_{j,t}q_{j,t}}\right)q_{n,t}\end{aligned}$$

For equal weights among components, that is $\phi_n = \phi$, the sum of them results in a definitive aggregate forecast that is the weighted average of both preliminary aggregate forecasts.

This result, which is valid for one level of disaggregation, is extendible to unlimited exhaustive groupings of components. Let there be S unique groupings of K_s sub-aggregations of components. The best guess of the decomposition of any sub-aggregation $y_{s,k}$ can be found using equation (2.8). That is:

$$\hat{q}^{(y_{s,k,t})}_{n,t} = \left(1+\tfrac{y_{s,k,t}-Q_{s,k,t}}{\phi_{n,t}\cdot\chi_{s,k,t}}\right)q_{n,t}$$

with $\chi_{s,k,t} = \sum\limits_{q_n\in y_{s,k}}\frac{1}{\phi_{n,t}}w_{n,t}q_{n,t}$ and $Q_{s,k,t} = \sum\limits_{q_n\in y_{s,k}} w_{n,t}q_{n,t}$.

Following the same process as for the one-level case, the definitive forecast for the components is given by:

$$\tilde{q}_{n,t} = \left[1+\frac{1}{\phi_{n,t}+\sum\limits_{s=1}^{S}\varphi_{s,k,t}}\cdot\sum_{s=1}^{S}\left(\tfrac{\varphi_{s,k,t}}{\phi_{n,t}}\cdot\tfrac{y_{s,k,t}-Q_{s,k,t}}{\chi_{s,k,t}}\right)\right]q_{n,t} \tag{2.9}$$

Summing up these forecasts for the case where all forecasts within the same grouping have the same reliability, results in the definitive aggregate being:

$$\tilde{y}_t = \frac{\phi_t Q_t+\sum\limits_{s=1}^{S}\varphi_{s,t}Y_{s,t}}{\phi_t+\sum\limits_{s=1}^{S}\varphi_{s,t}} \tag{2.10}$$

where $Y_{s,t} = \sum\limits_{k=1}^{K_s} y_{s,k,t}$.

It becomes clear that, under these circumstances, the definitive forecast is a weighted average of all the aggregate forecasts and, therefore, that for the case of equal weights, combining the aggregate forecasts produced from different aggregation levels is equivalent to the aggregate bottom-up forecast that results from imposing the different aggregate and intermediate forecasts on the component forecasts, and then combining all the resulting component forecasts.

2.2.2.2 Multi-level Combination Algorithm

The previous section shows that the process of combining many different aggregation levels and measurement approaches can be broken down into a series of one-level combinations involving each sub-aggregation and the components. With this, the procedure to generate the definitive aggregate forecast, and fully-consistent underlying scenario is described by the following algorithm:

1. Forecasting Step:
   (a) Produce individual forecasts for each of the models
   (b) Establish reliability weights for each of the forecasts
2. Single-variable Combination Step:
   (a) Combine all single variable forecasts
   (b) Establish reliability weights for each of the single variable forecasts
3. Multi-level Combination Step:
   (a) For each variable in all K sub-aggregations, perform a one-level combination with the bottom-level components assigning the variable an arbitrarily large reliability weight and using the components' own reliability weights.
   (b) For each of the N components, combine the original forecasts with the K sets of sub-aggregation consistent component forecasts, using for the latter the reliability weight of the corresponding sub-aggregation variable.
   (c) With the definitive component forecasts, use a bottom-up approach to produce the definitive forecasts for the aggregate and sub-aggregations.

As in the case of the one-level combination, caution should be taken in making sure that contradictory reliability weights are not used in the combination process. The possibility of clashes between reliability weights could increase, given that in each step of the multi-level combination the sub-aggregation is assigned an arbitrarily large weight.

2.3 Empirical Application

As an empirical application of the method, a forecasting exercise is performed using CPI data from France, Germany and the United Kingdom. Six different forecasting models and four different ways of establishing the combination weights are used within the framework. The evaluation is performed over the 2001-2015 period in a quarterly rolling scheme using a ten year window where in each period the models are re-estimated and a one-year-ahead quarterly forecast is generated. The aggregate forecasting accuracy is assessed by comparing the results with that of the single models and traditional forecast combinations. The forecasting accuracy of the components is evaluated against that of the single models.

2.3.1 Data and Sub-aggregations

For the exercise, CPI data for France, Germany and the United Kingdom is used. The data is quarterly and seasonally adjusted, spanning from 1991 to 2015 and available from the OECD statistics database. For all three countries the chosen lowest level of disaggregation are the twelve components presented in Table 2.1.

Table 2.1: Components Breakdown for Empirical Application

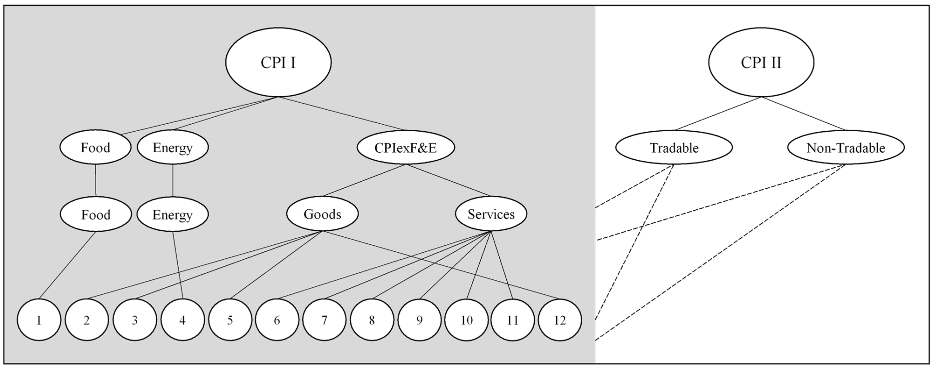
1. Food and non-Alcoholic beverages	6. Health
2. Alcoholic beverages, tobacco and narcotics	7. Transport
3. Clothing and footwear	8. Communication
4. Housing, water, electricity, gas and other fuels	9. Recreation and culture 10. Education
5. Furnishings, household equipment and maintenance	11. Restaurants and hotels 12. Miscellaneous goods and services

Regarding the sub-aggregations, three are chosen. The first two are in line with the extensive literature considering core measures for inflation. Aron and Muellbauer (2012) make a relatively extensive survey of studies that measure the benefits of removing certain components for forecasting, most of which find improvements from treating food and energy separately from the rest of CPI. The first sub-aggregation is therefore this breakdown. In line with Clark (2004), Peach et al. (2013) and Tallman and Zaman (2017), the second sub-aggregation separates the remaining CPI components from the first sub-aggregation in goods and services. The third follows Hargreaves et al. (2006) and Jacobs and Williams (2014), who similarly argue that the forces driving prices of tradables and non-tradables are very different in nature. They find significant improvements in aggregate accuracy from considering them separately.

The distribution of the different components among the sub-aggregation follows Johnson (2017) as closely as possible.[10] Taking all these factors into consideration, the aggregation structure for the empirical application is presented in Figure 2.2.

[10] The actual distribution is presented in section 2.D of the Appendix.

Figure 2.2: Aggregation Structure of the Empirical Application



2.3.2 Forecasting Models

The literature on inflation forecasting is vast meaning that choosing a subset of appropriate forecasting models for each component and sub-aggregation would not only be very time-consuming but necessarily an arbitrary process and therefore debatable. In this context, this empirical application uses a relatively reduced set of simple univariate and multivariate time-series models that are well proven and often serve as worthy benchmarks against which to compare more sophisticated approaches.[11]

Univariate models

Regardless of the numerous developments in econometric modelling, univariate methods continue to provide a strong benchmark against which to compare other models (Marcellino, 2008; Chauvet and Potter, 2013). They are also the methods used in many of the aggregate-disaggregate forecasting competitions and are therefore a reasonable starting point.

The first model is a random walk for the quarterly growth rate. The forecasts are produced using:

$$\hat{x}_{i,t+1|\mathrm{t}} = x_{i,t}$$

where $x_{i,t}$ is the first difference of the logarithm of the variable. The second is an autoregressive model of order one for the first differences of the variables, $x_{i,t} = a_i + \rho_i x_{i,t-1} + \epsilon_{i,t}$, where the forecasts are then produced using:

$$\hat{x}_{i,t+1|\mathrm{t}} = \hat{a}_i + \hat{\rho}_i x_{i,t}$$

It is worth mentioning that aggregating AR(1) processes will usually not result in an AR(1) process. In fact, the sum of an AR(p_1) and AR(p_2) process will be an ARMA(($p_1 + p_2$), max(p_1, p_2)). This means that assuming that both the components

[11] GARCH models and Dynamic Factor Models would be obvious candidates to include.

and all sub-aggregations actually follow AR(1) processes is an extremely strong assumption. However, as pointed out by Bermingham and D'Agostino (2014), in practical applications forecasts from overly parsimonious models may nevertheless outperform theoretically correct ones because of the associated low estimation error and parameter uncertainty. In this particular empirical application, allowing for higher lag orders does not alter its main findings and overall conclusion.[12]

Multivariate models

To account for the interdependence between components, Bayesian Vector Autoregressive models (BVARs) are also used. Following the implementation in Banbura et al. (2010), the estimated model is:

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}_1\mathbf{X}_{t-1} + \ldots + \mathbf{A}_5\mathbf{X}_{t-5} + \epsilon_t$$

and the forecasts are produced using:

$$\hat{\mathbf{X}}_{t+1|\mathrm{t}} = \hat{\mathbf{c}} + \hat{\mathbf{A}}_1\mathbf{X}_t + \ldots + \hat{\mathbf{A}}_5\mathbf{X}_{t-4}$$

In particular, four different approaches are used. In the first, separate VARs for each sub-aggregation are estimated for the variables in first differences. Specifically, than means estimating a first VAR including Gross Domestic Product (GDP) and Headline CPI alone, a second VAR including GDP, Food, Energy and CPI excluding Food and Energy, a third VAR including GDP, Food, Energy, Other Goods and Other Services, a fourth VAR including GDP, tradable items and non-tradable items and fifth VAR including GDP and the twelve components. The second approach follows the same structure as the first but the variables are differentiated according to a unit root test.[13] Following the notion in Hendry and Hubrich (2011), the third approach involves estimating a single large VAR that includes GDP and all CPI sub-aggregations and components with all variables in first differences. As before, the fourth approach follows the same structure as the third but the variables are differentiated according to a unit root test.

The smallest VARs, that is the two that include GDP and only Headline CPI, are estimated by OLS using two lags. All the others are estimated using five lags, and the choice of overall tightness, as in Banbura et al. (2010), is made so that the in-sample fit equals that of a two-variable VAR with five lags estimated by OLS over the first 10 years of the sample.

[12] Relative performance for AR models of up to order four and results of including them in the multilevel combination are presented in section 2.D of the Appendix.

[13] The differentiation is presented in section 2.D of the Appendix.

All this results in six sets of forecasts over the forecasting horizon for each one of the variables.

2.3.3 Empirical Reliability Weights

Even in the absence of relevant external knowledge, it may be desirable to determine reliability weights based on the properties of the preliminary estimates. Timmermann (2006) present an extensive survey on some of the suggestions from the combination literature for single variables and more become available from ongoing research (Hansen, 2008; Wei and Yang, 2012; Hsiao and Wan, 2014). Taking into consideration the ease with which each suggestion can be incorporated into the framework, four alternatives are suggested.

Scheme 1: Equal Weights

An obvious choice for the first set of weights is equal weights. This, because it serves as a natural benchmark against which to compare all the others and because in the traditional combination literature it has proved to perform remarkably well.

Scheme 2: In-Sample Fit

Using in-sample fit to determine combination weights is not uncommon. Kapetanios et al. (2008) find promising results from using weights calculated using information criteria. Extending their particular approach to compare different series, however, is not straightforward. As an alternative, a normalization of the measure used by Banbura et al. (2010) to determine in-sample fit for their Bayesian VARs is implemented.

For this purpose, let the root mean square percentage error (RMSPE) at time u using information up to time p for the h-step ahead forecast of x_i be:

$$RMSPE_{i,u,p,h,v} = \sqrt{\frac{1}{v} \sum_{s=u-h-v}^{u-h} \left(\frac{x_{i,s+h}|p}{x_{i,s+h}} - 1 \right)^2} \tag{2.11}$$

where $x_{i,s+h}|p$ is the fitted value for x_i using the coefficients calculated at time p and v determines how much data is included in the measure. The latter is limited by the number of lags that are included in each model.

The weights based on in-sample fit are then defined as:

$$\omega^{ISP}_{i,t,h,v} = \frac{1}{RMSPE_{i,t,t,h,v}} \tag{2.12}$$

The reliability weights are calculated for every rolling window using the five most recent years of the window as evaluation sample.

Scheme 3: Out-of-Sample Past Performance

An obvious extension of the idea of weighting according to predictability is to weigh the different forecasts based on their recent out-of-sample performance. This approach goes as far back as Bates and Granger (1969). Empirical studies suggest that forecasts weighted by the inverse of their mean square forecasting error (MSFE) are found to work well in practice (Stock and Watson, 1999; Timmermann, 2006).

Following the same idea and arguments expressed for the in-sample fit weights, the weights based on out-of-sample past performance are defined as:

$$\omega^{OSP}_{i,t,h,v} = \frac{1}{RMSPE_{i,t,s,h,v}} \tag{2.13}$$

where in this case the s that goes into the formula as the time subscript is not a parameter, but the index in the sum embedded in equation (2.11). The reliability weights are calculated for every rolling window using the last two years as evaluation window.

Scheme 4: Optimal weights

In the context of single variable combinations Granger and Ramanathan (1984) address the problem of determining the optimal combination weights as a least-squares regression problem. Hyndman et al. (2011) extend the approach to a setting with variables from different aggregation levels. In their implementation, however, they only consider forecasts from one hierarchy. To enable combining forecasts from both sub-aggregations, an approximation is necessary.

The proposed approximation consists in treating all sub-aggregations as independent and calculating the weights following the procedure in Hyndman et al. (2011). A primary hierarchy is chosen and the weights for the other sub-aggregations are supplied from the other hierarchies, ensuring that they are consistent with those of the chosen primary hierarchy. As the weights from this method depend on the aggregation structure alone, they do not change from one period to the next.[14]

2.3.4 Forecasting Accuracy Evaluation

The forecasting accuracy is presented for different horizons by means of the model's MSFE relative to that of a benchmark model. That is, for variable i, horizon h and

[14] The derivation is presented in section 2.C of the Appendix

using model m, the relative MSFE is:

$$\text{RelMSFE}^{(i,h,m)} = \frac{\text{MSFE}_{T_0,T_1}^{(i,h,m)}}{\text{MSFE}_{T_0,T_1}^{(i,h,0)}}$$

with

$$\text{MSFE}_{T_0,T_1}^{(i,h,m)} = \frac{1}{T_1 - T_0 + 1} \sum_{t=T_0}^{T_1} \left(y_{i,t+h}^{(m)} | t - y_{i,t+h} \right)^2$$

where $y_{i,t+h}^{(m)}|t$ is the forecasted value for $t+h$ at time t and T_0 is the last period of actual data in the first sample used for the evaluation and T_1 is the last period of actual data in the last sample. As usual, a RelMSFE lower than one reflects an improvement over the benchmark model for which $m = 0$.

As regards measuring the overall forecasting accuracy of the components, this is done by comparing the cumulative absolute errors in the contribution to the aggregate level. For this purpose the cumulative absolute root mean square forecasting error for an aggregate with N components q_n, horizon h and using model m is defined as:

$$\text{CumRMSFE}_{T_0,T_1}^{(h,m)} = \sqrt{\frac{1}{T_1 - T_0 + 1} \sum_{t=T_0}^{T_1} \left(\sum_{n=1}^{N} w_{n,t+h} \cdot \text{abs} \left(q_{n,t+h}^{(m)} | t - q_{n,t+h} \right) \right)^2}$$

where $q_{n,t+h}^{(m)}|t$ is the forecasted value for $t+h$ at time t and T_0 is the last period of actual data in the first sample used for the evaluation and T_1 is the last period of actual data in the last sample. To evaluate the significance of the differences, for both the aggregations and components, the forecasts are compared using the Modified Diebold-Mariano test for equality of prediction accuracy proposed by Harvey et al. (1997).

2.3.5 Results

The forecasting application involves six different forecasting models and five different aggregation approaches. This means that for each country there are 30 alternative aggregate forecasts from which to choose. Table 2.2 presents the individual models' relative forecasting accuracy over the 2001-2015 sample for the three countries.

From inspecting the results, it becomes apparent that some of them occur in all three countries. One is that the dispersion in the performance of the different models is large, reaching 40% in the most extreme cases. Another is that the AR(1) models perform best and large BVARs that differentiate the variables according to the unit root test perform worst. Also, in all cases the best performing models show improvements of at least 15 to 20% over the aggregate random walk, depending on the horizon. Beyond that, however, differences appear. For France, for example, it would seem

Table 2.2: Single Model Aggregate Forecasting Errors by Sub-aggregation

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Headline CPI												
RW	1.00°°	1.00°°	1.00°°	1.00°°	1.00°	1.00°°	1.00	1.00	1.00	1.00	1.00	1.00
AR	0.91	0.82	0.73	0.67	0.88	0.80	**0.78**	**0.74**	0.96°°	0.91°°	0.93°°	0.94°°
SVDIF	**0.85**	0.80	0.77	0.72	0.87	0.83	0.88	0.87	0.92°°	0.88	0.90	0.90
SVDDIF	0.93	0.95	0.99°	1.00	0.92	0.92	0.99	1.00	0.94	0.92	0.95	0.95
LVDIF	1.13°°	1.03°°	0.96°°	0.91°°	0.87	0.90	0.96	0.94	1.03°°	1.04°	1.05°°	1.09°°
LVDDIF	1.17°°	1.11°°	1.07°°	1.02°°	1.02°°	1.06°	1.17	1.20	1.16°°	1.19	1.19	1.24
Sub-agg.1												
RW	1.02°°	1.01°°	1.01°°	1.02°°	1.00°	1.00°°	1.00	1.00	0.98	0.99	1.00	1.00
AR	0.91	0.84	0.76°°	0.71°°	0.87	0.80°	0.78	0.75	0.92°°	0.90°°	0.93°°	0.94°°
SVDIF	0.89	0.84	0.80	0.76	0.87	0.83	0.85	0.84	0.93°°	0.92	0.96	0.97
SVDDIF	0.89	0.87	0.88°°	0.87°	0.93°	0.93°	0.99	1.00	0.96°°	0.94	0.99	1.00
LVDIF	1.14°°	1.04°°	0.98°°	0.93°°	0.87	0.90	0.96	0.94	0.99°°	1.04°°	1.08°°	1.11°°
LVDDIF	1.19°°	1.12°°	1.06°°	1.02°°	1.02°°	1.06°	1.16	1.19	1.09°°	1.11°	1.10	1.14
Sub-agg.2												
RW	1.01°°	1.01°°	1.01°°	1.01°°	1.00°	1.00°°	1.00	1.00	1.17	1.11	1.10	1.10
AR	0.89	0.82	0.76°°	0.70°°	0.87	**0.79**	0.78	0.75	1.02	0.92°°	0.93°°	0.93°°
SVDIF	0.88	**0.80**	0.76	0.72	0.85	0.82	0.83	0.82	1.06°	0.98°°	1.01°	1.01
SVDDIF	0.90	0.88	0.88°°	0.88°	0.94°	0.93°	0.99	1.00	1.05	0.98°	1.00	1.02
LVDIF	1.14°°	1.05°°	0.99°°	0.94°°	0.87	0.90	0.96	0.94	1.08°°	1.05°°	1.06°°	1.10°°
LVDDIF	1.18°°	1.12°°	1.06°°	1.02°°	1.02°°	1.06°	1.16	1.19	1.18°°	1.14°°	1.11°	1.15°
Sub-agg.3												
RW	1.00°°	1.00°°	1.00°°	1.00°°	1.00°	1.00°°	1.00	1.00	0.79	0.87	0.89	0.87
AR	0.89	0.80	**0.73**	**0.66**	0.88	0.79	0.78	0.75	**0.75**	**0.77**	**0.79**	**0.80**
SVDIF	0.85	0.81	0.77	0.73	**0.85**	0.81	0.84	0.83	0.85°°	0.90	0.93	0.93
SVDDIF	0.95°	0.97	1.01°	1.01	0.91	0.91	0.98	1.00	0.82	0.90	0.93	0.92
LVDIF	1.13°°	1.04°°	0.97°°	0.92°°	0.87	0.90	0.96	0.94	1.02°°	1.05°	1.06°	1.06°
LVDDIF	1.17°°	1.12°°	1.08°°	1.03°°	1.01°°	1.05°	1.16	1.19	1.11°°	1.19°	1.20	1.22
Components												
RW	1.00°	1.00°°	1.01°°	1.01°°	1.00°	1.00°°	1.00	1.00	0.83	0.91	0.94	0.95
AR	0.89	0.82	0.77	0.71	0.88	0.80	0.78	0.76	0.79	0.81	0.84	0.84
SVDIF	0.95	0.85	0.79	0.73	0.87	0.84	0.87	0.84	0.87°°	0.92°°	0.95°°	0.96°°
SVDDIF	0.99°°	0.89	0.86°°	0.82°°	0.93°°	0.91	0.96	0.94	0.94°°	0.97°°	1.01°°	1.04°°
LVDIF	1.11°°	1.05°°	0.99°°	0.93°°	0.87	0.90	0.95	0.93	0.95°°	1.03°°	1.07°°	1.08°°
LVDDIF	1.13°°	1.06°°	1.00°°	0.94°°	1.01°°	1.05°	1.12	1.13	1.03°°	1.09°°	1.09°°	1.12°°

Note: Mean square forecasting error (MSFE) of each model relative to that of the direct approach using the random walk model for each horizon by sub-aggregation approach. The sub-aggregations are those of Figure 2.2. The models are a random walk with drift (RW), a first-differences autoregressive model of order one (AR), two small VARs including GDP and the series from each CPI sub-aggregation in first differences (SVDIF) and where each variable is differenced according to a unit root test (SVDDIF) and two large VARs including GDP and the series from all considered CPI sub-aggregations in first differences (LVDIF) and differenced according to a unit root test (LVDDIF). In bold the lowest MSFE for each horizon and country. ° and °° denote that the respective forecast is statistically worse than the best model for that country according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. Calculated for one to four steps ahead forecasts over the 2001-2015 period.

that forecasting the aggregate directly or using the separation between tradables and non-tradables, i.e. Sub-aggregation 3, results in the most accurate forecasts. Also, the improvements of the better models over the random walk and large BVARs are statistically significant for all sub-aggregations. For Germany, on the other hand, the choice of sub-aggregation does not seem to make much impact on the results with the differences between models not being statistically significant in most cases. Finally, for the United Kingdom using the AR(1) for Sub-aggregation 3, or with a bottom-up approach using the components, produces the best results. Using the AR(1) with other sub-aggregations produces forecasts that are significantly worse. As in the case of France, in this case too the large BVARs are significantly worse than the better models.

The relatively large differences between the performance of the single models support the concerns regarding choice of one model as being potentially risky in terms of forecasting accuracy. The appeal of forecast combination is that this is not necessary. Table 2.3 presents the MSFE for both traditional and multi-level combination. As a means of comparison, the results for the best and median single models for each country are included. The weighting schemes are equivalent for both combination approaches in the first three cases, that is equal weights, in-sample fit and out-of-sample performance. For optimal weights, however, the two methods are not equivalent. For the aggregate, the approach in Conflitti et al. (2015) is used. The method calculates the weights minimizing the MSFE and stipulating that weights should be non-negative and sum up to one. A result of the latter is that it trims off the worst-performing models. For the multi-level case, the approximation to the weighting scheme by Hyndman et al. (2011) is used.

It is immediately noticeable that the best-performing combinations show no improvements over the best single models. They do, however, tend to be half way between the minimum and median, therefore, supporting the view that, in a context where establishing the best model beforehand is not possible, combination will tend to reduce the possibility of choosing a very bad one. The differences between methods, however, are relatively small for all but the aggregate optimal weighting scheme. Its performance comes out as statistically worse than the best single models in most cases. There is hardly any difference between the aggregate accuracy of the multi-level combinations and their corresponding traditional counterparts. The only differences appear for France for the in-sample and out-of-sample weighting schemes where the multi-level versions are marginally worse and for the United Kingdom where they are marginally better for the out-of-sample weighting scheme. In terms of the comparative performance among weighting schemes for the multi-level combination, the in-sample comes out as worst of all. The differences between the other three schemes are marginal. Only for France at the longer horizons does the out-of-sample scheme look slightly better. Overall, the differences between the aggregate results from the traditional and

Table 2.3: Combination Aggregate Forecasting Errors

	Aggregate				Multi-level			
Horizon	1	2	3	4	1	2	3	4
France								
Single Models								
Minimum	0.85	0.80	0.73	0.66				
Median	0.99	0.98	0.97	0.92				
Combination								
Eq.W.	0.90	0.85	0.81°°	0.78°	0.90	0.85	0.81°°	0.78°
ISP	0.96°	0.90°	0.86°°	0.80°°	0.98°	0.91°	0.86°°	0.81°°
OSP	0.90	0.84	0.80°	0.75	0.91	0.85	0.80°°	0.76
OPT	1.19°°	1.06°°	0.97°°	0.90°°	0.91	0.85	0.82°°	0.78
Germany								
Single Models								
Minimum	0.85	0.79	0.78	0.74				
Median	0.90	0.91	0.96	0.94				
Combination								
Eq.W.	0.85	0.83	0.85	0.85	0.85	0.83	0.85	0.85
ISP	0.87	0.87	0.89	0.88	0.87	0.87	0.89	0.88
OSP	0.85	0.83	0.85	0.85	0.85	0.83	0.85	0.84
OPT	1.02°°	1.06°°	1.08	1.07	0.84	0.83	0.85	0.84
United Kingdom								
Single Models								
Minimum	0.75	0.77	0.79	0.80				
Median	0.97	0.98	1.00	1.00				
Combination								
Eq.W.	0.85	0.86	0.87	0.88	0.85	0.86	0.87	0.88
ISP	0.89°	0.92	0.93	0.95	0.89°	0.92	0.93	0.95
OSP	0.85	0.87	0.89	0.90	0.84	0.86	0.87	0.88
OPT	0.99°°	1.04°	1.06°	1.08°	0.85	0.86	0.87	0.88

Note: Mean square forecasting error of each combination method relative to that of the direct approach using the random walk model for each horizon. The combination weighting schemes are the simple average (EQ.W), in-sample fit (ISP), out-of-sample performance (OSP) and optimal weights (OPT). For the aggregate optimal weights we use the approach in Conflitti et al. (2015) that impose that weights should be non-negative and sum up to one. ° and °° denote that the respective forecast is statistically worse than the best single model within the sample according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. Calculated over the 2001-2015 period.

multi-level approaches seem negligible.

As regards disaggregate accuracy, Table 2.4 presents the cumulative MSFE of both traditional and multi-level combination for all sub-aggregations relative to that of the best single model within each approach for each horizon. For purposes of comparison, the median cumulative error of the single models is also presented.

The first thing to note from the distribution of the figures in bold, which denote improvements over the best single models, is that the positive impact of combination is significantly larger for the United Kingdom than for the other two countries. In this case, both the traditional and multi-level approaches show some improvement over the best methods for all sub-aggregations. The multi-level method, however, outperforms the traditional in all cases. The largest improvements are found for Sub-aggregation 2 for which the gains from using the out-of-sample weights go up to 16% with the differences being statistically significant for all horizons but the longest. The gains from the traditional method are quite moderate by comparison. As regards Sub-aggregations 1 and 3, the gains from the multi-level approach go up to 8 and 6% respectively while the traditional counterparts correspondingly achieve 5 and 3% at best. For France and Germany, on the other hand, there are also some improvements over the best models, but these are restricted to the one-step-ahead forecasts. In these cases the multi-level approach also performs equally well or better than the traditional combination in all cases, but the size of the improvements are smaller. In terms of overall performance, the combination methods tend to be well below the median cumulative error of the single models that can go as high as 17 to 42% over the best model depending on the horizon. In terms of relative performance of the different weighting schemes, one result from the aggregate outcome that is also present at the disaggregate level for all three countries, is that the in-sample weighting scheme comes out worst of all. For France, in fact, the differences with the best single model are statistically significant. As regards the other methods, however, differences appear at the disaggregate level. The equal and approximate optimal weights remain very similar, but the out-of-sample weighting scheme tends to outperform the others by a small margin, particularly for the longer horizons.

These results suggest that using multi-level forecast combination can be beneficial in terms of disaggregate accuracy. The fact that the aggregate accuracy is practically the same as that of the equivalent traditional single-variable methods suggests that the benefits of achieving disaggregate consistency do not come at the cost of the aggregate accuracy. Furthermore, given that the multi-level combination method shows disaggregate forecasting accuracy that is similar to or better than those of both the best-performing single models and traditional combination, it would seem that the constraints it imposes on the disaggregate forecasts have the desired effect.

Table 2.4: Cumulative Disaggregate Forecasting Errors

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Sub-agg.1												
Single Model Median	1.12	1.17	1.23	1.30	1.07	1.16	1.28	1.34	1.09	1.11	1.14	1.18
Traditional Comb.												
Eq.W.	1.03	1.04	1.08	1.11	1.00	1.06	1.12	1.19	**0.97**	**0.96**	**0.95**	**0.98**
ISP	1.08°	1.10°	1.14°°	1.15°°	1.04	1.09	1.15	1.20	1.01	1.02	1.01	1.04
OSP	1.03	1.04	1.07	1.10	1.00	1.05	1.10	1.15	**0.98**	**0.97**	**0.96**	1.00
Multi-level Comb.												
Eq.W.	1.00	1.03	1.06	1.09	**0.99**	1.05	1.10	1.17	**0.93***	**0.95**	**0.94**	**0.97**
ISP	1.06	1.09	1.12°°	1.14°°	1.02	1.08	1.14	1.19	**0.96**	1.00	1.01	1.03
OSP	1.00	1.03	1.05	1.07	**0.99**	1.04	1.09	1.14	**0.92***	**0.94**	**0.94**	**0.96**
OPT	1.01	1.04	1.07	1.10	1.00	1.05	1.11	1.17	**0.94**	**0.96**	**0.95**	**0.96**
Sub-agg.2												
Single Model Median	1.12	1.23	1.27	1.34	1.10	1.17	1.24	1.29	1.06	1.13	1.15	1.16
Traditional Comb.												
Eq.W.	1.01	1.07	1.09	1.13°	1.03	1.06°	1.09	1.15	**0.95**	**0.97**	**0.96**	**0.98**
ISP	1.07	1.13°°	1.15°°	1.17°°	1.07°°	1.10°°	1.12	1.17	**0.99**	1.01	1.01	1.01
OSP	1.01	1.06	1.06	1.09	1.02	1.05	1.07	1.11	**0.96**	**0.96**	**0.96**	**0.98**
Multi-level Comb.												
Eq.W.	**0.98**	1.04	1.06	1.10	1.01	1.06	1.08	1.13	**0.84***	**0.89***	**0.90**	**0.91**
ISP	1.04	1.10°	1.12°	1.15°°	1.04	1.09°°	1.11	1.16	**0.87**	**0.93**	**0.94**	**0.95**
OSP	**0.98**	1.04	1.04	1.07	1.01	1.05	1.07	1.11	**0.84****	**0.89****	**0.89***	**0.90**
OPT	**0.99**	1.04	1.06	1.10	1.01	1.06°	1.08	1.14	**0.85***	**0.89***	**0.90**	**0.90**
Sub-agg.3												
Single Model Median	1.17	1.25	1.33	1.42	1.07	1.16	1.26	1.32	1.03	1.12	1.19	1.22
Traditional Comb.												
Eq.W.	1.06	1.06	1.11°°	1.16	1.00	1.06	1.11	1.16	**0.97**	**0.97**	1.00	1.03
ISP	1.14°°	1.12	1.15°°	1.19°°	1.03	1.08	1.14	1.19	1.04	1.04	1.08	1.10
OSP	1.07	1.05	1.08°°	1.12	1.00	1.04	1.09	1.13	**0.97**	**0.98**	1.00	1.02
Multi-level Comb.												
Eq.W.	1.05	1.04	1.08°°	1.13°	**0.99**	1.04	1.09	1.14	**0.94**	**0.96**	**0.98**	1.03
ISP	1.12°	1.11	1.13°°	1.17°°	1.02	1.07	1.12	1.17	**0.99**	1.01	1.05	1.09
OSP	1.06	1.04	1.06°	1.11	**0.99**	1.03	1.08	1.12	**0.94**	**0.95**	**0.98**	1.01
OPT	1.05	1.04	1.08°°	1.13	**0.99**	1.04	1.09	1.13	**0.94**	**0.95**	**0.97**	1.01
Components												
Single Model Median	1.12	1.18	1.22	1.24	1.05	1.10	1.16	1.19	1.09	1.14	1.16	1.18
Traditional Comb.												
Eq.W.	1.03	1.05	1.06	1.08	**0.99**	1.03	1.05	1.07	1.00	1.01	1.02	1.03
ISP	1.08°°	1.09°°	1.11°°	1.11°°	1.03	1.05	1.08	1.10	1.04	1.06°	1.07°	1.08
OSP	1.02	1.04	1.05	1.06	**0.99**	1.02	1.04	1.06	**0.99**	1.00	1.01	1.01
Multi-level Comb.												
Eq.W.	1.02	1.04	1.05	1.07	**0.99**	1.03	1.05	1.08	1.00	1.00	1.01	1.02
ISP	1.07°	1.08°°	1.10°°	1.11°	1.03	1.06	1.08	1.11	1.03	1.04	1.05	1.07
OSP	1.01	1.03	1.04	1.05	**0.99**	1.02	1.04	1.06	**0.98**	**0.98**	**0.99**	1.00
OPT	1.03	1.04	1.06	1.08	1.00	1.03	1.05	1.08	1.01	1.01	1.02	1.02

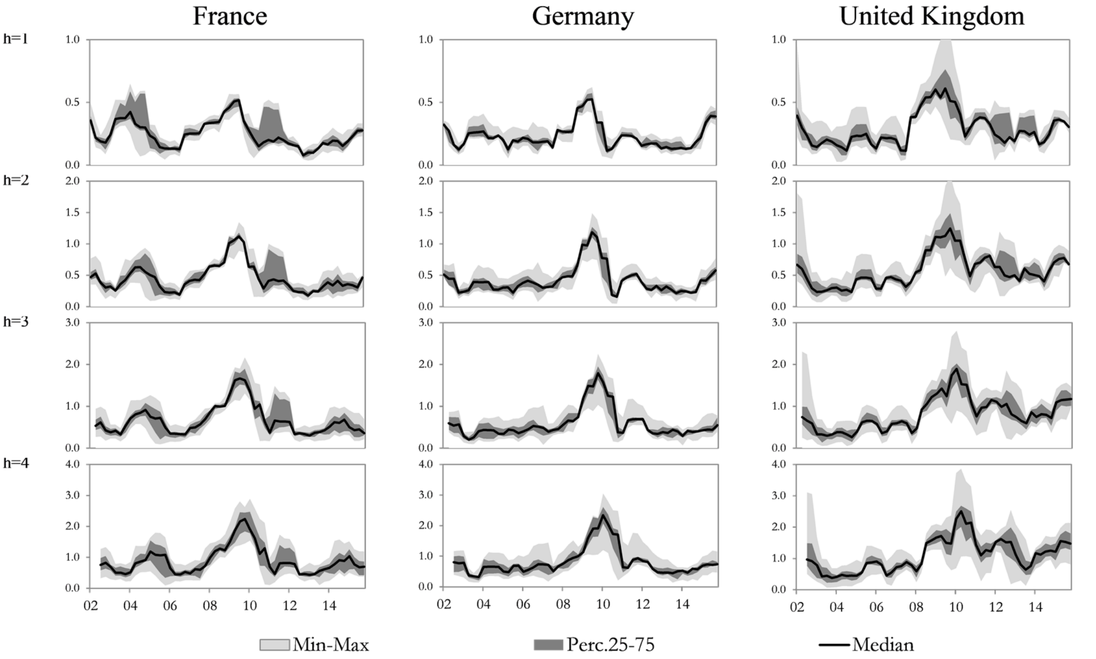
Note: Cumulative mean square forecasting error of the forecast that results from the combination approaches for each method relative to the minimum achievable from the single models for each horizon. The combination weighting schemes are the simple average (EQ.W), in-sample fit (ISP), out-of-sample performance (OSP) and optimal weights (OPT). ° and °° denote that the respective forecast is statistically worse than the best model for that country according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. * and ** denote that the respective forecast is statistically better than the best model for that country according to the same statistic and significance levels. Figures below one are highlighted in bold. Calculated over the 2001-2015 period.

As mentioned before, the impact of the combination method varies greatly between countries. The results for the United Kingdom seem very positive, while for the other two countries they are moderate at best. A possible explanation for these differences could come from the characteristics of the data or the features of the forecasting models. One of the arguments for using disaggregation is that modelling the aggregate can become very challenging if the components follow very different processes. On the contrary, if the disaggregate models are misspecified, forecasting the aggregate directly can lead to better results. There is a middle-ground, however, where forecasting the aggregate directly or through the bottom-up approach may give very similar results. This could be merely due to coincidence or the fact that the estimated processes for the aggregate end up being very similar. The results from the single models in Table 2.2 suggest that this may be the case for Germany and, to a lesser extent, for France. For the former, the results for each forecasting model are almost identical across sub-aggregations with the average difference between them being under 0.7 percentage points. On the opposite side of the spectrum, for the United Kingdom the differences appear comparatively large at 3.3 percentage points. France is between the two, with 1.8 percentage points.

This on its own, however, does not imply that there are no gains to be obtained from choosing different aggregation levels. Alternative sub-aggregations could perform well in different periods only to show similar results over the whole sample. Whether this is in fact the case for this particular empirical application can be examined to some extent in Figure 2.3. It presents the four-quarter rolling MSFE for the aggregate for all models and forecasting horizons. The figure shows the dispersion of the single models referred to as Min-max, the same measure trimming the best and worst-performing 25% and the median. The differences between countries are immediately obvious. For Germany, the dispersion of the forecasting errors is relatively low. For all horizons the middle 50% of the models are almost undistinguishable from the median and for the shorter horizons the whole distribution is very concentrated. All this suggests that across most models and sub-aggregations the difference in performance is relatively small. For the United Kingdom and France, on the other hand, the distribution of errors is fairly dispersed over most of the sample. Based on this analysis alone, it would appear that France should also show some positive results. As this is not the case, it would seem that there are other factors besides forecasting error dispersion that affect the performance of the multi-level combination. Nevertheless, the fact that the multi-level method performs equally well or better than the alternatives, both under apparently favourable and unfavourable circumstances, provides evidence of the robustness of the method and supports its use as a way of safeguarding against mistakenly picking an outstandingly bad model.

One non-trivial detail of the previous forecasting exercise is that the evaluation

Figure 2.3: Dispersion of the Rolling Forecasting Error



Note: Four-quarter rolling mean square forecasting error for each horizon. The Min-Max shaded area shows the span between the minimum and maximum MSFE from the 30 aggregate forecasts. The Perc.25-75 does the same but trims off the top and bottom 25%. Calculated as four-quarter moving windows over the 2001-2015 period.

period includes the end portion of what has been called the Great Moderation and the most recent financial crisis. A considerable body of literature has devoted itself to understanding the effects of these periods on forecasting models and Chauvet and Potter (2013) present a comprehensive review. Some of the conclusions state that many models that perform well in stable times fail completely with increases in volatility, and that models perform very differently in expansions and recessions. This could mean that the results from this empirical application could be overly influenced by the particular performance in the crisis years, simply because the forecasting errors could be massive. From Figure 2.3 the impact of the financial crisis is obvious for all three countries, for all horizons. Removing this period from the analysis, however, does not affect the overall results. These are, aggregate accuracy similar to that of comparable traditional single-variable combination methods and, in terms of disaggregate accuracy, cumulative forecasting errors that are low relative to the median of the single-models and similar to or better than the best-performing single models.[15]

Overall, in terms of the performance of the empirical weighting schemes, most of the gains of doing multi-level combination are picked up by the equal-weighted scheme. Some additional improvements are attainable, however, from using combination weights based on the recent out-of-sample performance of the models. Finding these additional

[15] The results of the exercise excluding the crisis years are provided in section 2.D of the Appendix.

gains supports the idea that being able to assign reliability weights subjectively to the forecasts from different levels can lead to an improvement in overall forecasting accuracy.

2.4 Conclusion

The framework developed in this chapter incorporates an aggregate, any number of sub-aggregations and its components into the same forecast combination process. The method performs the combination, relying on the merits of the individual forecasts, and acknowledges that for any realized outcome an aggregate is exactly the weighted sum of its components. This method makes use of disaggregate components and ensures that the accounting identities that underlie the aggregate are met, therefore delivering a completely consistent forecasting scenario. The method contributes to the existing literature in two aspects. First, it is flexible enough to incorporate forecasts from any number of models, measurement approaches and sub-aggregations. Second, it allows the use of weights that reflect the relative reliability of the preliminary forecasts themselves.

In the empirical application with CPI data from France, Germany and the United Kingdom, the multi-level combination framework provides similar aggregate forecasting accuracy to that of equivalent traditional forecast combination methods, and disaggregate accuracy equal to or better than those of the best-performing single models. In terms of the empirically determined weighting schemes, equal-weights attain most of the benefits from combination, but some additional gains are possible from using weights based on recent out-of-sample performance. All this suggests that this method could show an improvement over the traditional bottom-up approaches, in terms of disaggregate accuracy, when a fully consistent scenario is required. This is because some degree of interdependence is forced on the components' forecasts, no matter whether they are generated independently in the first place or not. Additionally, the possibility of establishing the weights could prove to be useful as a way of introducing external information or judgement into the forecasting process. This is something that Central Banks do regularly as a way of incorporating a broader assessment of relevant conditions that are not explicitly accounted for in their models (Alessi et al., 2014).

In terms of furthering research, one possibility is to explore its uses in settings where the asynchronous release of information means that at any given time some disaggregate data is known for the period of interest while other data has to be forecasted. Another possibility is to explore its use for density forecasting, in order to see how it affects the whole distribution. From an applied perspective, it would be interesting to enrich the set of models that are included in the combination process. Some obvious candidates would be to add factor models that may boost the performance of direct aggregate forecasts (Stock and Watson, 1998; Forni et al., 2005) and at the same time incorporate disaggregate methods that include interactions and common features between components within the process (Espasa and Mayo-Burgos, 2013; Esteves, 2013; Stock and Watson, 2015).

Appendix:

2.A Derivation for a One-level Combination

The premise upon which the derivation of the proposed method revolves around is that the solution to the multi-level combination should be equivalent to that of traditional single-variable methods if the conditions are comparable. This presentation explains, among other things, why common reconciliation procedures do not meet the requirements, but how, by working from them, the proposed loss function is found. Then the framework is developed, so that it works in a general setting.

Let there be a composite index that results from the simple sum of $N \geq 2$ strictly positive components and two forecasts for it. The first, y, comes from forecasting the aggregate directly, while the second one, Q, is the simple sum of the forecasts of its components q_n.

Result 1: Failure of the Equal Distribution of Differences

The additive deviation approach proposed by Denton (1971) finds the definitive values making the differences between them and the initial estimates equal in absolute terms.

The minimization problem for two aggregates can be written as:

$$\min_{\alpha,\beta} \left[(1+\alpha)\,y - y\right]^2 + \left[(1+\beta)\,Q - Q\right]^2 + 2\lambda\left[(1+\alpha)\,y - (1+\beta)\,Q\right] \tag{2.14}$$

The first order conditions imply that $\beta = -\alpha\frac{y}{Q}$ and $(1+\alpha)y = (1+\beta)Q$. Then replacing β in the latter gives:

$$(1+\alpha)y = Q - \alpha y \tag{2.15}$$

and solving for $(1+\alpha)y$ gives the simple average.

If the additive approach is used directly on the components' forecasts, however, the minimization problem is the following:

$$\min_{\alpha,\beta_n} (\alpha y)^2 + \textstyle\sum_{j=1}^{N} \left(\beta_j q_j\right)^2 + 2\lambda\left[(1+\alpha)y - \textstyle\sum_{j=1}^{N}(1+\beta_j)q_j\right] \tag{2.16}$$

This time the first order conditions imply that $\beta_n = -\alpha\frac{y}{q_n}$ for $n = 1$ to N and $(1+\alpha)y = \sum_{j=1}^{N}(1+\beta_j)q_j$. Solving for $(1+\alpha)\,y$, the aggregate forecast resulting from the combination is:

$$\tilde{y} = \frac{N \cdot y + \sum_{j=1}^{N} q_j}{N+1} = \frac{1}{N+1}\left(N \cdot y + Q\right) \tag{2.17}$$

that is different from the simple average, given that $N \geq 2$ and both aggregate forecasts are assumed to be distinct.

Result 2: Failure of the Proportional Distribution of Differences

Following Denton (1971), the proportional deviation approach from the reconciliation literature finds the definitive values by making the differences between them and the initial estimates proportional. The minimization problem for two aggregates is therefore:

$$\min_{\alpha,\beta} \left[\frac{(1+\alpha)\,y - y}{y}\right]^2 + \left[\frac{(1+\beta)\,Q - Q}{Q}\right]^2 + 2\lambda\left[(1+\alpha)\,y - (1+\beta)\,Q\right] \tag{2.18}$$

where α and β are the percentage deviations of the definitive value from the initial estimates.

The first order conditions imply that $Q = -\frac{\beta}{\alpha}y$ and $(1+\alpha)\,y = Q + \beta Q$. The aggregate forecast resulting from solving the problem is then:

$$\tilde{y} = (1+\alpha)\,y = (1+\beta)\,Q = \left(\frac{y \cdot Q}{y^2 + Q^2}\right)(y + Q) \tag{2.19}$$

Using the inequality of arithmetic and geometric means shows that $0 \leq (y - Q)^2 = y^2 + Q^2 - 2yQ$. Then $2yQ \leq y^2 + Q^2$ and therefore:

$$\frac{y \cdot Q}{y^2 + Q^2} \leq \frac{1}{2}$$

meaning that the solution is strictly lower than an equal weighted average if both forecasts are distinct.

Result 3: A Loss Function for One Set of Forecasts

From comparing the two approaches it can be seen that the only difference between them is that the former eliminates the downward bias relative to the simple average present in the latter by penalizing deviations based on the relative size of each aggregate forecast. The same idea can be extended to find the appropriate penalty term for the components.

Including an unspecified weight η_n for the disaggregate components in equation (2.16) results in:

$$\min_{\alpha,\beta_n} (\alpha y)^2 + \sum_{j=1}^{N} (\beta_j \eta_j)^2 + 2\lambda \left[(1+\alpha)y - \sum_{j=1}^{N}(1+\beta_j)q_j\right] \tag{2.20}$$

This time the first order conditions imply that $\beta_n = -\frac{q_n}{\eta_n^2} \cdot \alpha y$ for $n = 1$ to N and $(1+\alpha)y = \sum_{j=1}^{N}(1+\beta_j)q_j$. Using this gives:

$$(1+\alpha)\, y = \sum_{j=1}^{N}(q_j) - \sum_{j=1}^{N}\left(\frac{q_j^2}{\eta_j^2} \cdot \alpha y\right)$$

Then matching with the intermediate step given by equation (2.15) results in:

$$Q - \sum_{j=1}^{N}\left(\frac{q_j^2}{\eta_j^2} \cdot \alpha y\right) = Q - \alpha y$$

Then solving for η_n the weight for the components is:

$$\eta_n = \sqrt{q_n \cdot Q}$$

With this, the loss function that produces the equal weighted result for the aggregate is:

$$(\alpha y)^2 + \sum_{n=1}^{N} q_j Q\,(\beta_j)^2 \tag{2.21}$$

Result 4: Incorporating Multiple Component Forecasts

If more than one set of forecasts for the same components are included in equation (2.21), a bias similar to that of equation (2.19) appears. This happens because not only the definitive aggregate forecasts have to coincide, but also those of the components.

This can be seen by extending the framework in equation (2.21) to a setting with D sets of disaggregate forecasts for the N components. The minimization problem may be written as:

$$\begin{aligned} \min_{\alpha,\beta_n} (\alpha y)^2 &+ \textstyle\sum_{d=1}^{D}\sum_{j=1}^{N} q_{d,j}Q_d\beta_{d,j}^2 \\ &+ 2\textstyle\sum_{d=1}^{D}\left(\lambda_d\left[(1+\alpha)y - \sum_{j=1}^{N}(1+\beta_{d,j})q_{d,j}\right]\right) \\ &+2\textstyle\sum_{d=2}^{D}\sum_{j=1}^{N}\left(\delta_{d,j}\left[(1+\beta_{1,j})q_{1,j} - (1+\beta_{d,j})q_{d,j}\right]\right) \end{aligned} \tag{2.22}$$

Simplifying the problem to the particular case with one aggregate, two disaggregate forecasts and $N = 2$ the first order conditions become:

1. $\frac{\partial}{\partial\alpha}: \quad \alpha y + \lambda_1 + \lambda_2 = 0$

2. $\frac{\partial}{\partial\beta_{1,n}}: \quad \beta_{1,n}Q_1 - \lambda_1 + \delta_n = 0 \qquad$ for $n = 1, 2$

3. $\frac{\partial}{\partial\beta_{2,n}}: \quad \beta_{2,n}Q_2 - \lambda_2 - \delta_n = 0 \qquad$ for $n = 1, 2$

4. $\frac{\partial}{\partial \lambda_d}$: $(1+\alpha)y - (1+\beta_{d,1})q_{d,1} - (1+\beta_{d,2})q_{d,2} = 0$ for $d = 1, 2$

5. $\frac{\partial}{\partial \delta_n}$: $(1+\beta_{1,n})q_{1,n} - (1+\beta_{2,n})q_{2,n} = 0$ for $n = 1, 2$

After some algebra using conditions 1, 2, 3 and 5, $(1+\beta_{1,n}) = q_{2,n}(Q_1 q_{2,n} + Q_2 q_{1,n})^{-1}(Q_1 + Q_2 - \alpha y)$ for $n = 1, 2$. Using this in the corresponding condition in 4. results in:

$$\begin{aligned} \tilde{y} &= \Phi(Q_1 + Q_2 - \alpha y) \\ &= \tfrac{\Phi}{1+\Phi}(y + Q_1 + Q_2) \end{aligned} \tag{2.23}$$

where

$$\Phi = \frac{Q_1^2 q_{2,1} q_{2,2} + Q_2^2 q_{1,1} q_{1,2}}{Q_1^2 q_{2,1} q_{2,2} + Q_1 Q_2 (q_{1,2} q_{2,1} + q_{1,1} q_{2,2}) + Q_2^2 q_{1,1} q_{1,2}}$$

For equation (2.23) to be the simple average it is necessary for $\frac{1+\Phi}{\Phi}$ to be equal to three. This is equivalent to saying that $\Phi^{-1} - 1$, that is given by:

$$\Phi^{-1} - 1 = \frac{Q_1 Q_2 (q_{1,2} q_{2,1} + q_{1,1} q_{2,2})}{Q_1^2 q_{2,1} q_{2,2} + Q_2^2 q_{1,1} q_{1,2}}$$

has to be equal to one.

To explore the circumstances under which this is in fact true, the second set of preliminary estimates is expressed as deviations from the first set, that is $q_{2,1} = \kappa_1 q_{1,1}$ and $q_{2,2} = \kappa_2 q_{1,2}$ where κ_1 and κ_2 can take any value. Assuming that $\Phi^{-1} - 1$ is in fact equal to one would result in:

$$\frac{Q_1(\kappa_1 q_{1,1} + \kappa_2 q_{1,2})(\kappa_1 q_{1,1} q_{1,2} + q_{1,1} \kappa_2 q_{1,2})}{Q_1^2 \kappa_1 \kappa_2 q_{1,1} q_{1,2} + (\kappa_1 q_{1,1} + \kappa_2 q_{1,2})^2 q_{1,1} q_{1,2}} = 1$$

Then:

$$(q_{1,1} + q_{1,2})(\kappa_1 q_{1,1} + \kappa_2 q_{1,2})(\kappa_1 + \kappa_2) = Q_1^2 \kappa_1 \kappa_2 + (\kappa_1 q_{1,1} + \kappa_2 q_{1,2})^2$$

$$(\kappa_1 q_{1,1} + \kappa_2 q_{1,2})(\kappa_1 q_{1,2} + \kappa_2 q_{1,1}) = Q_1^2 \kappa_1 \kappa_2$$

$$\kappa_1^2 q_{1,1} q_{1,2} + \kappa_2^2 q_{1,1} q_{1,2} = 2\kappa_1 \kappa_2 q_{1,1} q_{1,2}$$

$$\kappa_1^2 - 2\kappa_1 \kappa_2 + \kappa_2^2 = 0$$

that results in $(\kappa_1 - \kappa_2)^2 = 0$.

This condition only holds when $\kappa_1 = \kappa_2$, meaning that the outcome of equation (2.22) is a simple average only when the two sets of preliminary estimates are exactly the same or the second one is simply the first multiplied by a constant.

The problem that arises from trying to combine more than one set of forecasts directly in the multi-level combination framework can be avoided simply by combining the multiple forecasts for the individual series before performing the multi-level combination, and choosing the optimization weights so as to reflect the prior step.

Let the result for the prior step be:

$$y = \frac{1}{\Gamma}\sum_{i=1}^{A}\gamma_i y_i \qquad \text{and} \qquad q_n = \frac{1}{\Delta_n}\sum_{d=1}^{D}\delta_{d,n}q_{d,n} \tag{2.24}$$

with γ_i and $\delta_{d,n}$ being the reliability weights, $\Gamma = \sum_{i=1}^{A}\gamma_i$ and $\Delta_n = \sum_{d=1}^{D}\delta_{d,n}$.

The combination procedure remains unchanged except for the weights φ and ϕ_n, which are set to reflect the reliability of the combined forecasts y and q_n as opposed to the initial preliminary forecasts y_i and $q_{d,n}$.

In the case of equal reliability, for example, this means accounting for the fact that the problem as a whole involves A aggregate and D disaggregate forecasts. That is accomplished by setting $\varphi = A$ and $\phi_n = D$ making the solution for the aggregate forecast:

$$\tilde{y} = \frac{1}{A+D}\left(A \cdot y + D \cdot \sum_{j=1}^{N} w_j q_j\right) \tag{2.25}$$

By expanding the individual forecasts, given that γ_i and $\delta_{d,n}$ are equal to one, the definitive aggregate forecast is left in terms of the preliminary estimates:

$$\begin{aligned}\tilde{y} &= \tfrac{1}{A+D}\left(A \cdot \tfrac{1}{A}\textstyle\sum_{i=1}^{A} y_i + D \cdot \sum_{j=1}^{N}\tfrac{1}{D}w_j\sum_{d=1}^{D} q_{d,j}\right) \\ &= \tfrac{1}{A+D}\left(\textstyle\sum_{i=1}^{A} y_i + \sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right)\end{aligned} \tag{2.26}$$

that is the same as taking the simple average of all the available forecasts for the aggregate.

Result 5: One-level Combination for Multiple Measurement Approaches

For an aggregate that can be obtained as the sum of K alternative measurement approaches, where each approach is the result of the weighted sum of the respective strictly positive N_k components, let there be a direct aggregate forecast y and K distinct aggregate forecasts, each based on the corresponding N_k component's forecasts. The aggregation weights are assumed to be positive.

The minimization problem involving the aggregate reliability weight φ, the disaggregate reliability weights $\phi_{k,n}$ and the aggregation weights $w_{k,n}$, is:

$$\begin{aligned}\min_{\alpha,\beta}\varphi\left(\alpha y\right)^2+\textstyle\sum_{k=1}^{K}\Big[Q_k\sum_{j=1}^{N_k}\phi_{k,j}w_{k,j}q_{k,j}\left(\beta_{k,j}\right)^2\\ +2\lambda_k\left((1+\alpha)y-\textstyle\sum_{j=1}^{N_k}w_{k,j}(1+\beta_{k,j})q_{k,j}\right)\Big]\end{aligned}$$

The first order conditions are:

1. $\frac{\partial}{\partial\alpha}$: $\varphi\alpha y+\sum_{k=1}^{K}\lambda_k=0$
2. $\frac{\partial}{\partial\beta_{k,j}}$: $Q_k\phi_{k,j}\beta_{k,j}-\lambda_k=0$ for $j=1$ to N_k and $k=1$ to K
3. $\frac{\partial}{\partial\lambda_k}$: $(1+\alpha)y-\sum_{j=1}^{N_k}w_{k,j}(1+\beta_{k,j})q_{k,j}=0$

From 2., for any k, $\phi_{k,n}\beta_{k,n}=\frac{\lambda_k}{Q_k}$, and plugging in the corresponding restriction in 3. gives:

$$(1+\alpha)y=\textstyle\sum_{j=1}^{N_k}w_{k,j}q_{k,j}+\sum_{j=1}^{N_k}w_{k,j}\beta_{k,j}q_{k,j}$$

$$y+\alpha y=Q_k+\tfrac{\lambda_k}{Q_k}\textstyle\sum_{j=1}^{N_k}\left(\tfrac{1}{\phi_{k,j}}w_{k,j}q_{k,j}\right)$$

$$\lambda_k=\left[\textstyle\sum_{j=1}^{N_k}\tfrac{1}{\phi_{k,j}}w_{k,j}q_{k,j}\right]^{-1}Q_k\left(y+\alpha y-Q_k\right)$$

Then using 1. and dividing by φ:

$$\begin{aligned}\alpha y &= \sum_{k=1}^{K}\left(\left[\textstyle\sum_{j=1}^{N_k}\tfrac{\varphi}{\phi_{k,j}}w_{k,j}q_{k,j}\right]^{-1}Q_k\left(Q_k-y-\alpha y\right)\right)\\ &= \sum_{k=1}^{K}\frac{Q_k}{\chi_k}\left(Q_k-y-\alpha y\right)\end{aligned}$$

where $\chi_k=\sum_{j=1}^{N_k}\frac{\varphi}{\phi_{k,j}}w_{k,j}q_{k,j}$.

The previous equations can be manipulated as follows:

$$\alpha y=\textstyle\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}Q_k-\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}y-\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}\alpha y$$

$$\left(1+\textstyle\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}\right)\alpha y=\textstyle\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}Q_k-\left(1+\sum_{k=1}^{K}\tfrac{Q_k}{\chi_k}\right)y+y$$

Then the definitive aggregate forecast is seen to be a weighted average given by:

$$\tilde{y} = \frac{y + \sum_{k=1}^{K} \left(Q_k \cdot \frac{Q_k}{\chi_k} \right)}{1 + \sum_{k=1}^{K} \frac{Q_k}{\chi_k}} \tag{2.27}$$

The definitive component forecasts are obtained by combining 2. and λ_k:

$$Q_k \phi_{k,n} \beta_{k,n} - \varphi \tfrac{Q_k}{\chi_k} \left(y + \alpha y - Q_k \right) = 0$$

$$\beta_{k,n} = \tfrac{\varphi}{\phi_{k,n}} \tfrac{Q_k}{\chi_k} \left(y + \alpha y - Q_k \right) \tfrac{1}{Q_k}$$

with the final result being:

$$\tilde{q}_{k,n} = \left(1 + \frac{\varphi}{\phi_{k,n}} \cdot \frac{\tilde{y} - Q_k}{\chi_k} \right) q_{k,n} \tag{2.28}$$

For more than one set of forecasts, the same joint combination process is followed, only that the multiple forecasts are combined in a prior step and optimization weights are chosen to reflect this.

An example of choosing appropriate weights can be seen from the simple equal reliability scenario. Let there be a single aggregate forecast y and a single set of disaggregate forecasts q_n for $n = 1$ to N. The solution for the aggregate forecast is is given by equation (2.27) and is:

$$\tilde{y} = \frac{Q^2 + y \sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j q_j}{Q + \sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j q_j} \tag{2.29}$$

that involves the aggregate reliability weight φ, the disaggregate reliability weights ϕ_n and the aggregation weights w_n.

If y and q_n for $n = 1$ to N are the result of a prior combination, that is $y = \frac{1}{\Gamma} \sum_{i=1}^{A} \gamma_i y_i$ and $q_n = \frac{1}{\Delta_n} \sum_{d=1}^{D} \delta_{d,n} q_{d,n}$ with γ_i and $\delta_{d,n}$ being the prior reliability weights, $\Gamma = \sum_{i=1}^{A} \gamma_i$ and $\Delta_n = \sum_{d=1}^{D} \delta_{d,n}$, the equivalence with the simple average of the initial forecasts can be shown by replacing them into the solution:

$$\tilde{y} = \frac{\left(\sum_{j=1}^{N} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)^2 + \left(\frac{1}{\Gamma} \sum_{i=1}^{A} \gamma_i y_i \right) \left(\sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)}{\left(\sum_{j=1}^{N} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right) + \left(\sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)} \tag{2.30}$$

Then incorporating the equal reliability of forecasts by setting $\gamma_i = \delta_{d,n,t} = 1$ and reflecting the number of forecasts that are involved in the first stage with $\varphi = A$ and $\phi_n = \phi = D$, the solution then simplifies as follows:

$$\begin{aligned}
\tilde{y} &= \frac{\left(\frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right)^2 + \left(\frac{1}{A}\sum_{i=1}^{A} y_i\right)\left(\frac{A}{D}\cdot\frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right)}{\left(\frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right) + \left(\frac{A}{D}\cdot\frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right)} \\
&= \frac{\left(\frac{1}{D}\sum_{d=1}^{D}\sum_{j=1}^{N} w_j q_{d,j}\right) + \left(\frac{A}{D}\sum_{i=1}^{I} y_i\right)}{1+\frac{A}{D}} \\
&= \frac{1}{A+D}\left(\sum_{i=1}^{A} y_i + \sum_{d=1}^{D} Q_d\right)
\end{aligned}$$

which is the simple average of all the aggregate forecasts.

2.B Foundation for the Multi-level Combination Method

The multi-level combination method consists on breaking down the whole problem into a sequence of one-level combinations. This is done by expressing any aggregate forecast in terms of a consistent set of component forecasts. It is shown that following this approach and combining the resulting components' forecasts is equivalent to combining the aggregate forecasts produced from different intermediate aggregation levels in the case of equal weights.

Let there be a single aggregate forecast y and a single set of disaggregate forecasts q_n for $n = 1$ to N, the aggregate reliability weight φ, the disaggregate reliability weights ϕ_n and the aggregation weights w_n. Based on the previous results, the definitive aggregate forecast for this one-level combination is given by:

$$\tilde{y} = \frac{Q^2 + y\sum_{j=1}^{N}\frac{\varphi}{\phi_j} w_j q_j}{Q + \sum_{j=1}^{N}\frac{\varphi}{\phi_j} w_j q_j} \tag{2.31}$$

where $Q = \sum_{j=1}^{N} w_j q_j$ and the components are obtained from:

$$\tilde{q}_n = \left(1 + \frac{\varphi}{\phi_n}\frac{y - Q}{Q + \sum_{j=1}^{N}\frac{\varphi}{\phi_j} w_j q_j}\right) q_n \tag{2.32}$$

With the objective of reconciling a set of components to an aggregate, equation

(2.32) can be rewritten as follows:

$$\begin{aligned}\tilde{q}_n &= \left(1+\frac{\varphi}{\phi_n}\cdot\frac{y-Q}{Q+\sum_{j=1}^{N}\frac{\varphi}{\phi_j}w_jq_j}\right)q_n\\ &= \left(1+\frac{\frac{\varphi}{\phi_n}\cdot(y-Q)}{Q+\sum_{j=1}^{N}\frac{\varphi}{\phi_j}w_jq_j}\right)q_n\\ &= \left(1+\frac{\frac{1}{\phi_n}\cdot(y-Q)}{\frac{1}{\varphi}Q+\sum_{j=1}^{N}\frac{1}{\phi_j}w_jq_j}\right)q_n\end{aligned}$$

Then, to have a disaggregate scenario that is consistent with the original forecast y, taking q_n, for $n = 1$ to N, as the best guesses, the aggregate reliability is made arbitrarily large so that:

$$\hat{q}_n^{(y)} = \left(1+\frac{y-Q}{\phi_n\cdot\sum_{j=1}^{N}\frac{1}{\phi_j}w_jq_j}\right)q_n \qquad (2.33)$$

Then, assigning the reliability of y to the y-consistent components and combining them with the original forecasts for the components results in:

$$\begin{aligned}\tilde{q}_n^{alt} &= \frac{\phi_nq_n+\varphi\hat{q}_n^{(y)}}{\phi_n+\varphi}\\ &= \frac{\phi_nq_n+\varphi q_n+\frac{\varphi(y-Q)}{\phi_n\cdot\sum_{j=1}^{N}\frac{1}{\phi_j}w_jq_j}\cdot q_n}{\phi_n+\varphi}\\ &= \left(1+\frac{\varphi}{\phi_n}\cdot\frac{y-Q}{(\phi_n+\varphi)\sum_{j=1}^{N}\frac{1}{\phi_j}w_jq_j}\right)q_n\end{aligned}$$

that is slightly different from $\tilde{q}_n$ in equation (2.32). For equal weights among components, however, that is $\phi_n = \phi$:

$$\begin{aligned}\tilde{q}_n^{alt} &= \left(1+\frac{\varphi}{\phi}\cdot\frac{y-Q}{(\phi+\varphi)\frac{1}{\phi}\sum_{j=1}^{N}w_jq_j}\right)q_n\\ &= \left(1+\frac{\varphi}{\phi+\varphi}\cdot\frac{y-Q}{Q}\right)q_n\\ &= \left(\frac{Q(\phi+\varphi)+\varphi(y-Q)}{\phi+\varphi}\right)\frac{q_n}{Q}\\ &= \left(\frac{\phi Q+\varphi y}{\phi+\varphi}\right)\frac{q_n}{Q}\end{aligned}$$

and by summing up the components the aggregate is:

$$\tilde{y} = \frac{\phi Q+\varphi y}{\phi+\varphi}$$

which is the same as is obtained from setting $\phi_n = \phi$ for the standard result in equation (2.32).

This is a useful result for a one-level disaggregation, but the process is in fact extendible to unlimited exhaustive groupings of components.

Let there be S unique groupings of K_s sub-aggregations of components. Then the best guess of the decomposition of any sub-aggregation $y_{s,k}$ can be found using equation (2.33). That is:

$$\hat{q}_n^{(y_{s,k})} = \left(1+\frac{y_{s,k}-Q_{s,k}}{\phi_n\cdot\chi_{s,k}}\right) q_n$$

with $\chi_{s,k} = \sum\limits_{q_n\in y_{s,k}} \frac{1}{\phi_n} w_n q_n$ and $Q_{s,k} = \sum\limits_{q_n\in y_{s,k}} w_n q_n$.

Then, by combining all the resulting component forecasts, the definitive one is obtained from:

$$\begin{aligned}
\tilde{q}_n &= \frac{\phi_n q_n+\sum\limits_{s=1}^{S}\varphi_{s,k}\hat{q}_n^{(y_{s,k})}}{\phi_n+\sum\limits_{s=1}^{S}\varphi_{s,k}} \\
&= \left[1+\frac{1}{\phi_n+\sum\limits_{s=1}^{S}\varphi_{s,k}}\cdot\sum_{s=1}^{S}\left(\frac{\varphi_{s,k}}{\phi_n}\cdot\frac{y_{s,k}-Q_{s,k}}{\chi_{s,k}}\right)\right] q_n
\end{aligned}$$

For the case where all forecasts within the same grouping have the same reliability, the aggregate is given by:

$$\begin{aligned}
\tilde{y} &= \sum_{n=1}^{N} w_n \left[1+\frac{1}{\phi+\sum\limits_{s=1}^{S}\varphi_s}\cdot\sum_{s=1}^{S}\left(\varphi_s\cdot\frac{y_{s,k}-Q_{s,k}}{Q_{s,k}}\right)\right] q_n \\
&= Q+\frac{1}{\phi+\sum\limits_{s=1}^{S}\varphi_s}\cdot\sum_{s=1}^{S}\varphi_s\cdot\sum_{n=1}^{N} w_n\left(\frac{y_{s,k}}{Q_{s,k}}\cdot q_n - q_n\right) \\
&= \frac{1}{\phi+\sum\limits_{s=1}^{S}\varphi_s}\cdot\left[Q\left(\phi+\sum_{s=1}^{S}\varphi_s\right)-\sum_{s=1}^{S}\varphi_s Q+\sum_{s=1}^{S}\varphi_s\sum_{k=1}^{K_s}\left(\frac{y_{s,k}}{Q_{s,k}}\cdot\sum_{q_n\in Q_{s,k}} w_n q_n\right)\right] \\
&= \frac{1}{\phi+\sum\limits_{s=1}^{S}\varphi_s}\cdot\left[\phi Q+\sum_{s=1}^{S}\varphi_s\sum_{k=1}^{K_s} y_{s,k}\right]
\end{aligned}$$

By making $Y_s = \sum\limits_{k=1}^{K_s} y_{s,k}$, it becomes clear that the definitive forecast is a weighted average of all the aggregate forecasts:

$$\tilde{y} = \frac{\phi Q+\sum\limits_{s=1}^{S}\varphi_s Y_s}{\phi+\sum\limits_{s=1}^{S}\varphi_s}$$

This shows that, for the case of equal weights, combining the aggregate forecasts produced from different aggregation levels is equivalent to the aggregate bottom-up forecast that results from imposing the different aggregate and intermediate forecasts on the component forecasts and then combining all the resulting component forecasts.

2.C Approximate Optimal Weights for Multiple Measurement Approaches

Hyndman et al. (2011) propose a method for obtaining consistent forecasts for a whole hierarchy of time series, using a regression approach. The data is described by

$$\mathbf{Y}_t = \mathbf{S}\mathbf{Y}_{K,t} \tag{2.34}$$

where $\mathbf{Y}_t$ is a vector containing the values for all the series in the hierarchy at time t, $\mathbf{S}$ is the aggregation matrix that defines the structure of the hierarchy and $\mathbf{Y}_{K,t}$ is a vector containing the values at time t for the series at the lowest level of the hierarchy (maximum disaggregation).

For a hierarchy composed of four components, two intermediate aggregations and the total, for example, the vector for lowest level would be $\mathbf{Y}_{2,t} = [\ y_{2,1,t} \quad y_{2,2,t} \quad y_{2,3,t}$ $y_{2,4,t}]'$, the vector for all observations would be $\mathbf{Y}_t = [\ y_{0,t} \quad y_{1,1,t} \quad y_{1,2,t} \quad \mathbf{Y}'_{2,t}\ ]'$ and the aggregation matrix would be:

$$\mathbf{S} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}'$$

Hyndman et al. (2011) propose using this same structure to find consistent definitive forecasts from a set of independent forecasts for all series. They set up the following problem for the forecasts at time h:

$$\widetilde{\mathbf{Y}}_h = \mathbf{S}\mathbf{P}\hat{\mathbf{Y}}_h \tag{2.35}$$

where $\hat{\mathbf{Y}}_h$ are the preliminary forecasts for all series, $\widetilde{\mathbf{Y}}_h$ are the consistent definitive forecasts for all series and $\mathbf{P}$ is a balancing matrix. They use the regression approach to find $\mathbf{P}$ and in particular assume that the forecast errors satisfy the same aggregation constraint as the data. Under these assumptions they find that the optimal balancing matrix is $\mathbf{P} = (\mathbf{S}'\mathbf{S})^{-1}\,\mathbf{S}'$ and that therefore

$$\widetilde{\mathbf{Y}}_h = \mathbf{S}\,(\mathbf{S}'\mathbf{S})^{-1}\,\mathbf{S}'\hat{\mathbf{Y}}_h \tag{2.36}$$

The optimal combination method, however, does not contemplate multiple alternative approaches. For the empirical application, an approximation is used. It consists of treating all sub-aggregations as independent and calculating the weights following the procedure in Hyndman et al. (2011). Then, a primary hierarchy is chosen and the weights for the other sub-aggregations are supplied from the other hierarchies, ensuring that they are consistent with those of the chosen primary hierarchy.

In the particular case of the empirical application depicted in Figure 2.2, the largest hierarchy is chosen as the primary one, in other words the three-level one on the left. Calculating the optimal weights using an appropriately built $\mathbf{S}$ matrix and the optimal weights formula, $\mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'$, provides the weights given to all series in the definitive aggregate forecast, except for those of the tradable-non-tradable sub-aggregation. The calculated weights are presented under the Hierarchy 1 heading in Table 2.5. The same procedure is followed for the second hierarchy and the resulting weights are shown in the same table under the Hierarchy 2 heading. As the weights given to the direct method in both hierarchies are not the same, the ratio between the direct approach and the sub-aggregations of the second hierarchy is preserved and the weights are adjusted proportionally so that the weights given to the direct approach in both cases match. The definitive approximate weights are presented in Table 2.5 under the Final heading.

Table 2.5: Approximate Optimal Weights for Empirical Application

	Hierarchy 1	Hierarchy 2	Final
	%	%	%
Headline CPI	56.3	63.1	56.3
1. Food and non-alcoholic beverages	14.6		14.6
2. Electricity, gas and other fuels	14.6		14.6
3. CPI excluding Food and Energy	27.2		27.2
1. Food and non-Alcoholic beverages	14.6		14.6
2. Electricity, gas and other fuels	14.6		14.6
3. Other goods	13.1		13.1
4. Other services	14.1		14.1
1. Tradable		30.8	27.5
2. Non-tradable		32.3	28.9
1. Food and non-Alcoholic beverages	14.6		14.6
2. Alcoholic beverages, tobacco and narcotics	3.3		3.3
3. Clothing and footwear	3.3		3.3
4. Housing, water, electricity, gas and other fuels	14.6		14.6
5. Furnishings, household equipment and maintenance	3.3		3.3
6. Health	2.3		2.3
7. Transport	2.3		2.3
8. Communication	2.3		2.3
9. Recreation and culture	2.3		2.3
10. Education	2.3		2.3
11. Restaurants and hotels	2.3		2.3
12. Miscellaneous goods and services	3.3		3.3

2.D Additional Information from the Empirical Application

Table 2.6: Differentiation for Empirical Application and Sub-aggregation Distribution

	France	Germany	United Kingdom	Good or Service	Tradable or Non-tradable
1. Food and non-alcoholic beverages	2	2	1	-	NT
2. Alcoholic beverages, tobacco and narcotics	2	2	1	Good	T
3. Clothing and footwear	1	1	1	Good	T
4. Housing, water, electricity, gas and other fuels	1	2	2	-	NT
5. Furnishings, household equipment and maintenance	2	2	1	Good	T
6. Health	1	1	1	Service	NT
7. Transport	1	1	1	Service	T
8. Communication	1	2	1	Service	NT
9. Recreation and culture	1	1	2	Service	T
10. Education	2	1	2	Service	NT
11. Restaurants and hotels	2	1	2	Service	NT
12. Miscellaneous goods and services	2	2	1	Good	NT

Note: Number of times the series is differentiated to make it stationary according to the parametric unit root test in Gomez and Maravall (1996). Sub-aggregation distribution based on the distribution in Johnson (2017).

Table 2.7: Aggregate Forecasting Errors of AR(p) Models

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Headline CPI												
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AR(1)	0.91	0.82	0.73	0.67	0.88	0.80	0.78	0.74	0.96	0.91	0.93	0.94
AR(2)	0.93	0.85	0.77	0.70	0.89	0.81	0.80	0.76	0.93	0.87	0.89	0.90
AR(3)	0.94	0.88	0.81	0.75	0.88	0.82	0.81	0.79	0.93	0.85	0.84	0.83
AR(4)	0.97	0.94	0.91	0.87	0.85	0.82	0.81	0.79	0.96	0.88	0.88	0.87
Sub-agg.1												
RW	1.02	1.01	1.01	1.02	1.00	1.00	1.00	1.00	0.98	0.99	1.00	1.00
AR(1)	0.91	0.84	0.76	0.71	0.87	0.80	0.78	0.75	0.92	0.90	0.93	0.94
AR(2)	0.93	0.87	0.80	0.75	0.90	0.83	0.81	0.77	0.87	0.83	0.86	0.87
AR(3)	0.95	0.89	0.83	0.77	0.89	0.83	0.80	0.77	0.85	0.82	0.83	0.83
AR(4)	0.96	0.90	0.85	0.81	0.88	0.83	0.81	0.78	0.93	0.91	0.92	0.92
Sub-agg.2												
RW	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.17	1.11	1.10	1.10
AR(1)	0.89	0.82	0.76	0.70	0.87	0.79	0.78	0.75	1.02	0.92	0.93	0.93
AR(2)	0.91	0.85	0.79	0.73	0.89	0.81	0.80	0.76	0.99	0.92	0.92	0.91
AR(3)	0.92	0.87	0.82	0.76	0.88	0.82	0.80	0.77	0.98	0.93	0.92	0.90
AR(4)	0.93	0.87	0.83	0.79	0.89	0.83	0.81	0.78	1.07	1.02	1.05	1.04
Sub-agg.3												
RW	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.79	0.87	0.89	0.87
AR(1)	0.89	0.80	0.73	0.66	0.88	0.79	0.78	0.75	0.75	0.77	0.79	0.80
AR(2)	0.91	0.83	0.76	0.69	0.90	0.84	0.81	0.77	0.78	0.83	0.84	0.84
AR(3)	0.92	0.85	0.79	0.73	0.89	0.83	0.80	0.77	0.81	0.86	0.88	0.88
AR(4)	0.91	0.87	0.83	0.79	0.87	0.83	0.81	0.79	0.85	0.92	0.97	0.99
Components												
RW	1.00	1.00	1.01	1.01	1.00	1.00	1.00	1.00	0.83	0.91	0.94	0.95
AR(1)	0.89	0.82	0.77	0.71	0.88	0.80	0.78	0.76	0.79	0.81	0.84	0.84
AR(2)	0.91	0.86	0.81	0.75	0.90	0.82	0.81	0.77	0.79	0.84	0.86	0.86
AR(3)	0.94	0.90	0.86	0.81	0.89	0.83	0.80	0.77	0.79	0.86	0.88	0.88
AR(4)	0.90	0.90	0.87	0.82	0.89	0.85	0.83	0.81	0.84	0.94	0.98	0.99

Note: Mean square forecasting error of each model relative to that of the direct approach using the random walk model for each horizon by sub-aggregation approach. The sub-aggregations are those of Figure 2.2. The models are a random walk with drift (RW), a first-differences autoregressive model of order one to four. Calculated for one to four steps ahead forecasts over the 2001-2015.

Table 2.8: Cumulative Disaggregate Forecasting Errors for Exercise Including AR(p)

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Sub-agg.1												
Single Model Median	1.10	1.10	1.12	1.16	1.04	1.05	1.07	1.11	1.07	1.04	1.06	1.09
Traditional Comb.												
Eq.W.	1.03	1.03	1.05	1.07	**0.97**	1.00	1.03	1.07	**0.97**	**0.96**	**0.95**	**0.97**
ISP	1.05	1.07	1.09°°	1.11°°	**0.99**	1.02	1.05	1.09	**0.99**	**0.99**	**0.98**	1.01
OSP	1.03	1.02	1.04	1.06	**0.97**	**0.99**	1.02	1.05	**0.98**	**0.96**	**0.96**	**0.99**
Multi-level Comb.												
Eq.W.	**0.98**	1.00	1.02	1.05	**0.97**	**0.99**	1.02	1.06	__0.93**__	**0.94**	**0.93**	**0.95**
ISP	1.02	1.05	1.06°	1.08°	**0.99**	1.02	1.05	1.09	**0.94**	**0.97**	**0.97**	**0.99**
OSP	**0.98**	1.00	1.01	1.02	**0.97**	**0.99**	1.01	1.05	__0.92**__	**0.93**	**0.92**	**0.95**
OPT	**0.99**	1.01	1.03	1.05	**0.97**	**0.99**	1.02	1.07	**0.95**	**0.94**	**0.93**	**0.95**
Sub-agg.2												
Single Model Median	1.07	1.10	1.11	1.12	1.05	1.06	1.04	1.05	1.04	1.07	1.12	1.12
Traditional Comb.												
Eq.W.	1.00	1.03	1.04	1.06	1.00	1.00	1.00	1.05	**0.96**	**0.95**	**0.94**	**0.95**
ISP	1.02	1.08°	1.08	1.09°°	1.02	1.03	1.03	1.06	**0.98**	**0.98**	**0.97**	**0.98**
OSP	1.00	1.02	1.02	1.03	**0.99**	**0.99**	**0.99**	1.02	**0.96**	**0.95**	**0.95**	**0.97**
Multi-level Comb.												
Eq.W.	**0.97**	1.01	1.02	1.05	**0.98**	1.00	1.00	1.04	__0.84**__	__0.87**__	__0.88**__	__0.90*__
ISP	1.00	1.06	1.06	1.08°	1.00	1.02	1.03	1.06	__0.86*__	__0.90**__	__0.91*__	**0.92**
OSP	**0.97**	1.01	1.00	1.01	**0.98**	**0.99**	**0.99**	1.02	__0.84**__	__0.87**__	__0.88**__	__0.89*__
OPT	**0.98**	1.01	1.02	1.05	**0.98**	1.00	1.00	1.04	__0.86**__	__0.87**__	__0.88**__	__0.89*__
Sub-agg.3												
Single Model Median	1.12	1.09	1.12	1.17	1.07	1.06	1.06	1.09	1.04	1.09	1.15	1.17
Traditional Comb.												
Eq.W.	1.05	1.01	1.03	1.07	**0.97**	1.00	1.02	1.06	**0.97**	**0.97**	**0.98**	**0.99**
ISP	1.10°	1.06	1.07°	1.09°	**0.99**	1.02	1.04	1.07	1.02	1.00	1.03	1.05
OSP	1.06	1.01	1.01	1.04	**0.97**	**0.99**	1.01	1.03	**0.97**	**0.97**	**0.99**	1.00
Multi-level Comb.												
Eq.W.	1.03	1.00	1.03	1.07°	**0.97**	**0.99**	1.01	1.04	**0.94**	**0.94**	**0.96**	1.00
ISP	1.09	1.05	1.07°	1.10°°	**0.98**	1.01	1.03	1.07	**0.97**	**0.97**	1.01	1.04
OSP	1.04	1.00	1.01	1.05	**0.97**	**0.98**	1.00	1.03	**0.94**	**0.94**	**0.96**	**0.99**
OPT	1.04	1.00	1.02	1.07	**0.96**	**0.98**	1.01	1.04	**0.95**	**0.94**	**0.95**	**0.98**
Components												
Single Model Median	1.05	1.07	1.10	1.13	1.03	1.04	1.06	1.06	1.05	1.08	1.12	1.13
Traditional Comb.												
Eq.W.	1.00	1.02	1.03	1.04	**0.96**	**0.98**	**0.99**	1.01	**0.97**	**0.98**	1.00	1.01
ISP	1.04	1.05	1.06°°	1.06°	**0.99**	1.01	1.01	1.03	1.01	1.02	1.04	1.04
OSP	1.00	1.01	1.02	1.02	**0.96**	**0.98**	**0.98**	1.00	**0.97**	**0.98**	**0.99**	1.00
Multi-level Comb.												
Eq.W.	1.00	1.01	1.02	1.03	**0.96**	**0.98**	**0.99**	1.01	**0.99**	**0.97**	**0.99**	1.00
ISP	1.03	1.04	1.04	1.05	**0.99**	1.01	1.01	1.03	1.00	1.00	1.01	1.03
OSP	**0.99**	1.00	1.00	1.00	**0.96**	**0.98**	**0.98**	1.00	**0.97**	__0.96*__	**0.97**	**0.98**
OPT	1.01	1.01	1.02	1.03	**0.97**	**0.99**	**0.99**	1.02	1.00	**0.98**	**0.99**	1.00

Note: Cumulative mean square forecasting error of the forecast that results from the combination approaches for each method relative to the minimum achievable from the single models for each horizon. The combination weighting schemes are the simple average (EQ.W), in-sample fit (ISP), out-of-sample performance (OSP) and optimal weights (OPT). ° and °° denote that the respective forecast is statistically worse than the best model for that country according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. * and ** denote that the respective forecast is statistically better than the best model for that country according to the same statistic and significance levels. Calculated over the 2001-2015 period excluding 2008 and 2009.

Table 2.9: Single Model Aggregate Forecasting Errors Excluding Crisis

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Headline CPI												
RW	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00	1.00	1.00	1.00
AR	0.80	0.71	0.67	0.73	0.84	0.77	0.78	0.80	1.00$^{\circ\circ}$	0.97$^{\circ\circ}$	0.99$^{\circ\circ}$	1.01
SVDIF	0.79	0.74	0.72	0.78	0.85	0.84$^{\circ\circ}$	0.95$^{\circ\circ}$	1.00°	0.92	0.85	0.85	0.85
SVDDIF	0.86	0.82	0.82	0.85	0.88	0.83$^{\circ\circ}$	0.85$^{\circ\circ}$	0.86°	0.93°	0.90	0.91	0.90
LVDIF	1.13$^{\circ\circ}$	1.10$^{\circ\circ}$	1.10$^{\circ\circ}$	1.19$^{\circ\circ}$	0.81	0.74	0.77°	0.81°	0.96°	0.95	0.99	1.04$^{\circ\circ}$
LVDDIF	1.21$^{\circ\circ}$	1.14$^{\circ\circ}$	1.13$^{\circ\circ}$	1.24$^{\circ\circ}$	0.95$^{\circ\circ}$	0.91$^{\circ\circ}$	1.04$^{\circ\circ}$	1.16$^{\circ\circ}$	0.98°	1.01	1.03	1.07$^{\circ\circ}$
Sub-agg.1												
RW	1.03$^{\circ\circ}$	1.03$^{\circ\circ}$	1.03$^{\circ\circ}$	1.04	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00	1.00	1.00	1.01
AR	0.85$^{\circ\circ}$	0.78$^{\circ\circ}$	0.75$^{\circ\circ}$	0.82$^{\circ\circ}$	0.84	0.76	0.76	0.79	0.94	0.91$^{\circ\circ}$	0.95$^{\circ\circ}$	0.98
SVDIF	0.84	0.78	0.76	0.82	0.80	**0.70**	**0.67**	**0.71**	0.87	0.82	0.85	0.86
SVDDIF	0.88$^{\circ\circ}$	0.82	0.81	0.87	0.88	0.82$^{\circ\circ}$	0.85$^{\circ\circ}$	0.87°	0.92°	0.89	0.93	0.94
LVDIF	1.16$^{\circ\circ}$	1.13$^{\circ\circ}$	1.13$^{\circ\circ}$	1.22$^{\circ\circ}$	0.80	0.74	0.76	0.81°	0.95°	0.96	1.00	1.07$^{\circ\circ}$
LVDDIF	1.25$^{\circ\circ}$	1.19$^{\circ\circ}$	1.17$^{\circ\circ}$	1.28$^{\circ\circ}$	0.95$^{\circ\circ}$	0.91$^{\circ\circ}$	1.04$^{\circ\circ}$	1.16$^{\circ\circ}$	1.00$^{\circ\circ}$	1.03	1.06	1.11°
Sub-agg.2												
RW	1.02$^{\circ\circ}$	1.02$^{\circ\circ}$	1.02$^{\circ\circ}$	1.03	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.27	1.18	1.16	1.16
AR	0.81$^{\circ\circ}$	0.74$^{\circ\circ}$	0.73°	0.80	0.83	0.75	0.76	0.79	1.09	0.96$^{\circ\circ}$	0.98$^{\circ\circ}$	1.00
SVDIF	0.82	0.76	0.75	0.83	0.80	0.71	0.68	0.71	1.06	0.90	0.90	0.90
SVDDIF	0.89$^{\circ\circ}$	0.83	0.82	0.88	0.90	0.83$^{\circ\circ}$	0.85$^{\circ\circ}$	0.87°	1.10	0.95	0.95	0.94
LVDIF	1.18$^{\circ\circ}$	1.13$^{\circ\circ}$	1.14$^{\circ\circ}$	1.23$^{\circ\circ}$	0.80	0.74	0.77	0.81°	1.11°	1.02	1.01	1.08$^{\circ\circ}$
LVDDIF	1.25$^{\circ\circ}$	1.19$^{\circ\circ}$	1.18$^{\circ\circ}$	1.30$^{\circ\circ}$	0.95$^{\circ\circ}$	0.91$^{\circ\circ}$	1.05$^{\circ\circ}$	1.16$^{\circ\circ}$	1.17$^{\circ\circ}$	1.09°	1.07	1.10$^{\circ\circ}$
Sub-agg.3												
RW	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	1.00$^{\circ\circ}$	**0.71**	0.83	0.84	**0.81**
AR	**0.78**	**0.69**	**0.67**	**0.73**	0.84	0.75	0.75	0.78	0.71	**0.79**	**0.83**	0.86
SVDIF	0.79	0.74	0.73	0.78	0.79	0.71	0.71	0.75	0.76	0.83	0.85	0.84
SVDDIF	0.88°	0.84°	0.84	0.87	0.86	0.81°	0.83$^{\circ\circ}$	0.86$^{\circ\circ}$	0.73	0.83	0.85	0.81
LVDIF	1.13$^{\circ\circ}$	1.10$^{\circ\circ}$	1.10$^{\circ\circ}$	1.19$^{\circ\circ}$	0.80	0.74	0.76	0.81°	0.91°	0.94	0.97	1.00°
LVDDIF	1.20$^{\circ\circ}$	1.14$^{\circ\circ}$	1.14$^{\circ\circ}$	1.24$^{\circ\circ}$	0.95$^{\circ\circ}$	0.90$^{\circ\circ}$	1.04$^{\circ\circ}$	1.16$^{\circ\circ}$	0.96$^{\circ\circ}$	1.06	1.08	1.09$^{\circ\circ}$
Components												
RW	1.01$^{\circ\circ}$	0.99°	0.98°	0.99	1.00$^{\circ\circ}$	1.01$^{\circ\circ}$	1.01$^{\circ\circ}$	1.01$^{\circ\circ}$	0.81	0.90	0.91	0.91$^{\circ\circ}$
AR	0.79	0.73	0.74	0.82	0.84	0.76	0.76	0.80	0.78	0.84	0.90	0.93
SVDIF	0.88	0.79	0.77	0.80	**0.79**	0.72	0.73	0.76	0.83	0.89°	0.95	0.99
SVDDIF	0.96$^{\circ\circ}$	0.85	0.84	0.91	0.86	0.81$^{\circ\circ}$	0.82$^{\circ\circ}$	0.84	0.87	0.91	0.97	1.02
LVDIF	1.12$^{\circ\circ}$	1.11$^{\circ\circ}$	1.12$^{\circ\circ}$	1.21$^{\circ\circ}$	0.81	0.75	0.77	0.80°	0.87	0.93	0.99	1.04
LVDDIF	1.18$^{\circ\circ}$	1.13$^{\circ\circ}$	1.12$^{\circ\circ}$	1.22$^{\circ\circ}$	0.96$^{\circ\circ}$	0.90$^{\circ\circ}$	1.00$^{\circ\circ}$	1.10$^{\circ\circ}$	0.91°	1.00	1.04	1.10°

Note: Mean square forecasting error of each model relative to that of the direct approach using the random walk model for each horizon by sub-aggregation approach. The sub-aggregations are those of Figure 2.2. The models are a random walk with drift (RW), a first-differences autoregressive model of order one (AR), two small VARs including GDP and the series from each CPI sub-aggregation in first differences (SVDIF) and where each variable is differenced according to a unit root test (SVDDIF) and two large VARs including GDP and the series from all considered CPI sub-aggregations in first differences (LVDIF) and differenced according to a unit root test (LVDDIF). $^{\circ}$ and $^{\circ\circ}$ denote that the respective forecast is statistically worse than the best model for that country according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. Calculated for one to four steps ahead forecasts over the 2001-2015 period excluding 2008 and 2009.

Table 2.10: Combination Aggregate Forecasting Errors Excluding Crisis

	Aggregate				Multi-level			
Horizon	1	2	3	4	1	2	3	4
France								
Single Models								
Minimum	0.78	0.69	0.67	0.73				
Median	0.98	0.92	0.91	0.95				
Combination								
Eq.W.	0.88°	0.81	0.78	0.83	0.88°	0.81	0.78	0.83
ISP	0.95$^{\circ\circ}$	0.89°	0.87	0.93	0.97$^{\circ\circ}$	0.90$^{\circ\circ}$	0.88	0.94
OSP	0.87°	0.79	0.75	0.80	0.89°	0.81	0.77	0.82
OPT	1.25$^{\circ\circ}$	1.12$^{\circ\circ}$	1.11$^{\circ\circ}$	1.16$^{\circ\circ}$	0.88°	0.81	0.78	0.82
Germany								
Single Models								
Minimum	0.79	0.70	0.67	0.71				
Median	0.86	0.81	0.82	0.85				
Combination								
Eq.W.	0.80	0.74	0.73	0.76	0.80	0.74	0.73	0.76
ISP	0.81	0.75	0.76	0.80	0.81	0.75	0.76	0.80
OSP	0.80	0.74	0.73	0.75	0.80	0.73	0.72	0.75
OPT	0.96$^{\circ\circ}$	0.92$^{\circ\circ}$	0.93°	0.97°	0.80	0.74	0.74	0.76
United Kingdom								
Single Models								
Minimum	0.71	0.79	0.83	0.81				
Median	0.94	0.94	0.97	1.00				
Combination								
Eq.W.	0.79	0.79	0.81	0.83	0.79	0.79	0.81	0.83
ISP	0.81	0.83	0.85	0.89	0.80	0.83	0.85	0.88
OSP	0.79	0.81	0.82	0.85	0.77	0.79	0.80	0.82
OPT	0.86	0.94	0.97	0.99°	0.79	0.79	0.80	0.82

Note: Mean square forecasting error of each combination method relative to that of the direct approach using the random walk model for each horizon. The combination weighting schemes are the simple average (EQ.W), in-sample fit (ISP), out-of-sample performance (OSP) and optimal weights (OPT). For the aggregate optimal weights we use the approach in Conflitti et al. (2015) that impose that weights should be non-negative and sum up to one. $^{\circ}$ and $^{\circ\circ}$ denote that the respective forecast is statistically worse than the best single model within the sample according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. Calculated over the 2001-2015 period excluding 2008 and 2009.

Table 2.11: Cumulative Disaggregate Forecasting Errors Excluding Crisis

	France				Germany				United Kingdom			
Horizon	1	2	3	4	1	2	3	4	1	2	3	4
Sub-agg.1												
Single Model Median	1.15	1.23	1.29	1.31	1.10	1.16	1.26	1.26	1.09	1.13	1.15	1.16
Traditional Comb.												
Eq.W.	1.05	1.08°	1.09	1.09°	1.03	1.08°°	1.10°°	1.13°°	0.98	0.98	0.98	0.99
ISP	1.11°	1.16°°	1.19°°	1.19°°	1.06	1.08°	1.12°	1.15°°	1.02	1.05	1.06	1.08
OSP	1.04	1.08°	1.08	1.09	1.03	1.06°	1.09°°	1.09°°	0.98	0.98	0.98	1.01
Multi-level Comb.												
Eq.W.	1.00	1.05	1.07	1.07	1.02	1.07°°	1.10°°	1.12°°	0.92	0.96	0.97	0.97
ISP	1.07	1.13°	1.16°°	1.16°°	1.05	1.07°°	1.11°°	1.14°°	0.94	1.02	1.04	1.05
OSP	1.01	1.05	1.04	1.05	1.02	1.06°°	1.09°°	1.09°°	0.90	0.96	0.95	0.96
OPT	1.01	1.06	1.08	1.07	1.03	1.08°°	1.12°°	1.13°°	0.93	0.96	0.96	0.95
Sub-agg.2												
Single Model Median	1.16	1.26	1.30	1.32	1.11	1.17	1.27	1.26	1.09	1.14	1.15	1.15
Traditional Comb.												
Eq.W.	1.03	1.07	1.07	1.09	1.05	1.09°°	1.12°°	1.13°°	0.98	0.96	0.94	0.93
ISP	1.10	1.14°°	1.15	1.17°	1.09°°	1.10°°	1.14°	1.15°	1.01	1.01	0.99	0.99
OSP	1.03	1.05	1.04	1.05	1.05	1.07°	1.11°°	1.09°°	0.98	0.97	0.94	0.93
Multi-level Comb.												
Eq.W.	0.98	1.03	1.03	1.05	1.02	1.08°°	1.12°°	1.12°	0.82	0.87*	0.87*	0.86
ISP	1.05	1.10	1.11	1.14	1.05	1.08°°	1.15°°	1.15°	0.84	0.90	0.91	0.91
OSP	0.98	1.03	1.00	1.02	1.02	1.07°°	1.12°°	1.10°	0.81	0.86**	0.85**	0.84*
OPT	1.00	1.03	1.03	1.05	1.02	1.09°°	1.14°°	1.14°	0.83	0.86**	0.86*	0.85
Sub-agg.3												
Single Model Median	1.21	1.34	1.39	1.33	1.09	1.14	1.19	1.18	1.03	1.05	1.10	1.09
Traditional Comb.												
Eq.W.	1.10°	1.13°°	1.11	1.08	1.02	1.06	1.06	1.08	0.96	0.92	0.94	0.93
ISP	1.18°°	1.22°°	1.21°	1.19	1.04	1.06	1.08	1.11°	1.02	0.98	1.02	1.02
OSP	1.10°	1.11°	1.08	1.05	1.01	1.04	1.03	1.03	0.95	0.92	0.94	0.94
Multi-level Comb.												
Eq.W.	1.08	1.11	1.09	1.08	1.01	1.05	1.05	1.07	0.94	0.90*	0.92	0.95
ISP	1.17°°	1.22°°	1.20	1.20	1.03	1.05	1.07	1.10	0.97	0.94	0.98	1.01
OSP	1.08	1.11°	1.07	1.07	1.01	1.04	1.03	1.03	0.92	0.89**	0.91	0.93
OPT	1.08	1.11	1.08	1.07	1.01	1.05	1.06	1.07	0.94	0.89**	0.91	0.93
Components												
Single Model Median	1.16	1.21	1.24	1.26	1.06	1.09	1.11	1.12	1.08	1.11	1.13	1.13
Traditional Comb.												
Eq.W.	1.05	1.06	1.05	1.06	1.01	1.03	1.01	1.02	1.00	1.00	1.00	1.00
ISP	1.11°°	1.12°	1.11	1.12°°	1.04°	1.04°	1.03	1.05	1.05	1.05	1.05	1.06
OSP	1.04	1.05	1.03	1.04	1.01	1.03	1.00	1.00	1.00	0.99	0.98	0.98
Multi-level Comb.												
Eq.W.	1.05	1.06	1.04	1.04	1.02	1.04	1.01	1.01	1.01	0.99	0.98	0.97
ISP	1.11°°	1.11°	1.10	1.11°°	1.04°	1.05°	1.04	1.05	1.04	1.02	1.01	1.02
OSP	1.04	1.04	1.01	1.02	1.01	1.03	1.01	1.00	0.98	0.96	0.95	0.95
OPT	1.06°	1.06	1.04	1.05	1.02	1.04	1.02	1.02	1.02	1.00	0.98	0.97

Note: Cumulative mean square forecasting error of the forecast that results from the combination approaches for each method relative to the minimum achievable from the single models for each horizon. The combination weighting schemes are the simple average (EQ.W), in-sample fit (ISP), out-of-sample performance (OSP) and optimal weights (OPT). ° and °° denote that the respective forecast is statistically worse than the best model for that country according to the Modified Diebold-Mariano statistic at a 10 and 5% significance level. * and ** denote that the respective forecast is statistically better than the best model for that country according to the same statistic and significance levels. Calculated over the 2001-2015 period excluding 2008 and 2009.

Chapter 3

Forecasting Aggregates Using Dynamic Component Grouping

3.1 Introduction

When forecasting economic aggregates, practitioners are often faced with the choice of either forecasting them directly or forecasting their components and then summing them up. Sometimes the decision may be determined by considerations other than accuracy, such as when a question cannot be answered just by looking at the aggregate, or an underlying scenario for the aggregate forecast is needed. Nevertheless, aggregate forecasting accuracy is usually also a concern in these cases too (Esteves, 2013).

The options available for forecasting are many, even when the only aspect considered is the level of disaggregation. These include forecasting at the level of disaggregation required to answer a particular question, disaggregating further, or forecasting at a more aggregate level and reconciling the lower levels of disaggregation if necessary.

The usual argument behind using the components to forecast an aggregate is that allowing for different specifications across disaggregate variables may capture more precisely the dynamics of a process that becomes too complex through aggregation (Barker and Pesaran, 1990). In support of this view, Granger (1990) shows that the sum of many simple stationary processes can produce a fractional integrated aggregate, while Bermingham and D'Agostino (2014) show that the dispersion of the persistence of individual series has an accelerating effect on the increase of complexity in the aggregate.

A point in favour of forecasting the aggregate directly is that, in practical applications, it is likely that the disaggregate models may suffer from misspecification. For example, if the disaggregate models neglect the fact that a number of components share common factors, the forecasting errors will tend to cluster, producing a negative effect on the aggregate forecast (Granger, 1987). The direct aggregate forecast would be less affected by these features in the data and other problems, such as those resulting from data measurement error and structural breaks (Grunfeld and Griliches, 1960; Aigner and Goldfeld, 1974).

The theoretical literature supports using the disaggregate forecasts, or bottom-up approach, but the results in the empirical literature are mixed.[1] Ultimately, whether the magnitude of the aggregation error compensates the specification errors in the disaggregate model depends on the particular forecasting models and data (Pesaran et al., 1989).

An option to improve forecasting performance in this setting is to work on the modelling, like Hendry and Hubrich (2011), who include disaggregate information in a direct aggregate approach or Bermingham and D'Agostino (2014) who include common factors in a bottom-up approach. Another less obvious way is to look for data transformations that allow existing models to perform better. As mentioned before, adding components together results in new series with characteristics that may differ quite significantly from those of the originating ones. In this context, it may be possible to search deliberately for specific groupings that show more desirable properties than those of the individual components and the aggregate. On this line, Espasa and Senra (2017) argue that for disaggregation to be useful for forecasting, it should not only follow economic and institutional criteria, but result in sub-aggregations that possess properties that allow for proper modelling. Some authors have proposed using purpose-built groupings to increase overall forecasting accuracy, but it would seem that, at least in economic forecasting, this has had little impact (Duncan et al., 2001). A reason for this is that the number of possible groupings grows exponentially with the number of components. This means that traditional methods, that would usually rely on evaluating all possible outcomes, are really only usable for problems with relatively few components (Espasa and Senra, 2017).

An exception to this apparent lack of research is Espasa and Mayo-Burgos (2013) who, in the context of inflation forecasting, propose a method that groups components by searching for common trends and cycles. They do however, use a limited pairwise testing procedure due to computational concerns. In other fields of research, however, methods to deal with large problems of the kind have been developed. One that has been relatively successful recently, particularly given the increase in popularity of methods for dealing with large datasets, is one that performs grouping conditional on

[1] Examples of these comparisons are Espasa et al. (2002), Benalal et al. (2004), Hubrich (2005) and Giannone et al. (2014) for inflation in the Euro area; Bermingham and D'Agostino (2014) for inflation in the U.S. and the Euro area; Marcellino et al. (2003), Hahn and Skudelny (2008), Burriel (2012) and Esteves (2013) for European GDP growth; and Zellner and Tobias (2000), Perevalov and Maier (2010) and Drechsel and Scheufele (2013) for GDP growth in specific industrialized countries.

some feature of the original data. These have been in use for a while in the context of electricity price forecasting (Weron, 2014) and, with the relatively recent surge in computational power, computer intensive methods and availability of high-frequency data, they have expanded to other areas of research. For example, Yan et al. (2013) report significant improvements in the context of wind-power prediction, Jha et al. (2015) for inventory planning in retail, and Gao and Yang (2014) for forecasting stock-market returns.

The success of these methods, however, depends on the chosen feature being useful for obtaining the desired outcome. The assumption upon which many of these models are built is that, by grouping series that behave in a similar way, the idiosyncratic errors within groups tend to offset each other, while the more relevant individual dynamics are retained to be modelled. Building on their promising results, this chapter develops a method to forecast economic aggregates based on purpose-built groupings of components, using statistical learning techniques. The two-stage method consists of trying to find the grouping of components that produces the best aggregate forecast at each point in time. In the first stage, Agglomerative Hierarchical Clustering is used to reduce the dimension of the problem and, in the second, a final single aggregate forecast is selected following a procedure on the previously chosen hierarchy.

The rest of the chapter is organized as follows: Section 3.2 presents the component grouping framework, Section 3.3 presents an empirical implementation using CPI data for France, Germany and the United Kingdom, and Section 3.4 summarizes the conclusions.

3.2 A Purpose Driven Grouping Framework for Aggregate Forecasting

As pointed out by James et al. (2013), statistical learning refers to a broad set of tools for understanding data. These include some approaches that are intended for prediction based on the relation between exogenous variables and the variable of interest within a training sample. Other methods try to do this by exploiting useful relationships and structure in the data. The latter work directly and produce results based on features of the original data and are therefore relatively inexpensive in terms of computation. The challenge of using these methods lies in tuning the algorithms so that they achieve a desired result.

Although the implementations and techniques differ, the assumption on which many of the models intended to forecast time-series are based is that forecasting series that behave similarly as a group will tend to produce more accurate aggregate forecasts than those modelled separately. According to Espasa and Senra (2017), this assumption would also seem reasonable within the context of forecasting economic aggregates. In terms of producing useful sub-aggregations, they advocate selecting groupings that posses very different distributional properties. This is also in line with the literature which shows that accounting for commonality among components is key to forecasting accuracy (Duarte and Rua, 2007; Bermingham and D'Agostino, 2014).[2]

In terms of actually performing the grouping based on the features of the data, there are many methods in the statistical learning literature.[3] One that seems well-suited for the particular setting is Hierarchical Clustering. This method is concerned with discovering unknown subgroups in data. The most commonly-used method is the agglomerative alternative, which starts with a set of groups, or clusters, that contain a single element each and proceeds by grouping the data into fewer units with more elements in each one.[4] The only thing the algorithm needs in order to work is some sort of dissimilarity measure between each pair of observations and then one for each cluster that is formed. For the fused clusters, those containing more than a single observation, the dissimilarity measures are typically calculated from the original dissimilarity measures, following a procedure referred to as linkage. The result of running the algorithm is always a hierarchical structure that has exactly as many levels as the number of initial components, with the individual components as the lowest level and the aggregate as the highest.

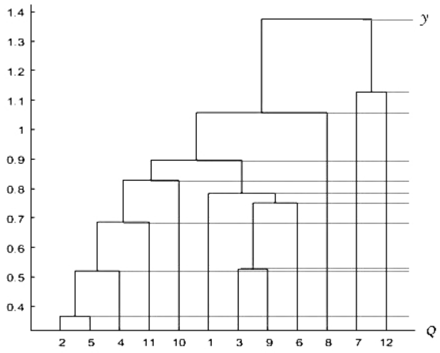
As the hierarchical clustering proceeds by fusing two observations or series at a time, it produces an intuitive tree-based representation of the final structure. This representation is called a dendrogram and an example of one is presented in Figure 3.1. At the bottom are all the individual elements. Moving up, some of the elements are paired with similar observations producing a number of clusters. Higher up, the clusters themselves fuse, either with single elements or other clusters. Each new fusion generates a new aggregation level that differs from the last by having one new element more and two existing elements less. The dissimilarity is measured on the vertical axis and, therefore, choosing a grouping based on some specific dissimilarity level is equivalent to drawing a horizontal line across the dendrogram at the desired level and using the groupings that are formed below that line. Doing so at each fusion in Figure 3.1 highlights all possible aggregation alternatives. It also makes it clear that each suggested aggregation corresponds to an interval of dissimilarity levels, with the direct aggregate and bottom-up approaches being always available as options to be chosen to

[2] This view goes beyond the direct versus bottom-up debate. The success of the dynamic factor models, proposed initially by Geweke (1977) and extended by Stock and Watson (2002) and Forni et al. (2005) among others, is just an example.

[3] For example, Yan et al. (2013) use Support Vector Machines, Gao and Yang (2014) use Hierarchical Clustering and Support Vector Regression and Jha et al. (2015) use Self Organizing Maps.

[4] The less popular divisive approach starts from one large group that contains all the elements and divides it up accordingly.

Figure 3.1: Aggregation Levels on a Dendrogram



produce the definitive forecast.

At first sight, it could seem that Hierarchical Clustering might be the solution to the grouping problem. However, the method provides no guidance as to whether the groupings in the structure are meaningful, nor if one grouping is better than another in any particular sense (Murphy, 2012).[5] The problem with identifying an appropriate grouping right away is that, even if one exists, the particular dissimilarity threshold below which components should be grouped in order to obtain the most accurate aggregate forecast is unknown. By narrowing down the set of groupings, however, the clustering process reduces the initial problem to a manageable size that can then be tackled with direct evaluation methods, including those that are common in the traditional forecasting literature.

In this context, for an aggregate with n components and an evaluation sample going from $t = 1$ to T, the two-stage dynamic grouping procedure is described by the following sequence:

For period t= 1 to T:

1. Grouping subset selection through Hierarchical Clustering[6]

    - Start with each of the n series in their own cluster

    For $s = 0$ to $n - 1$:

    - Obtain pairwise dissimilarity between the $n - s$ clusters
    - Aggregate the two clusters with the lowest dissimilarity to form a new series in a new cluster.

[5] This is the case for the widely used deterministic approach. Heller and Ghahramani (2005) develop a probabilistic approach that does provide guidance from within the clustering process.

[6] Detailed descriptions of Agglomerative Hierarchical Clustering may be found in standard Statistical Learning texts and surveys like Hastie et al. (2009), Murtagh and Contreras (2012) or James et al. (2013).

- Remove the two clusters used in the fusion of the previous step.

Next s

2. Produce definitive aggregate forecast

   - Select one or more grouping alternatives out of the subset of size n
   - Produce forecasts for all components and sub-aggregations in the chosen grouping alternatives.
   - For each grouping alternative aggregate component and sub-aggregation forecasts to produce an aggregate forecast.
   - Combine the available aggregate forecasts if more than one grouping alternative was chosen,

Next t

3.3 Empirical Application

As an empirical application of the method, a forecasting exercise using CPI data from France, Germany and the United Kingdom is performed. Univariate autoregressive and Bayesian multivariate methods are used to forecast the data and evaluate the aggregate forecasting accuracy of the grouping procedure by comparing the results with those of the direct forecast, the corresponding bottom-up approach and a combination of both.[7]

3.3.1 Data

For the exercise, CPI data is used for France, Germany and the United Kingdom disaggregated to twelve components. The data is quarterly and seasonally adjusted, spanning from 1991 to 2015 and available from the OECD statistics database.[8]

The breakdown of the aggregate is the following:

[7] That is, the exercise involves comparing the improvement of the grouping against the corresponding direct and bottom-up approach as opposed to finding the best aggregation from the whole pool of alternative forecasts.

[8] No inconsistencies arise from the seasonal adjustment given that the aggregates are adjusted indirectly, that is as the sum of the seasonally adjusted components.

Table 3.1: Components Breakdown for Empirical Application

1. Food and non-Alcoholic beverages
2. Alcoholic beverages, tobacco and narcotics
3. Clothing and footwear
4. Housing, water, electricity, gas and other fuels
5. Furnishings, household equipment and maintenance
6. Health
7. Transport
8. Communication
9. Recreation and culture
10. Education
11. Restaurants and hotels
12. Miscellaneous goods and services

3.3.2 Forecasting Models

Autoregressive Model of Order One (AR1)

Many of the aggregate-disaggregate forecasting competitions use univariate autoregressive methods and therefore seem like a sound choice to start off with. Regardless of the numerous developments in econometric modelling, they continue to perform well (Marcellino, 2008; Chauvet and Potter, 2013). In particular, the exercise uses an autoregressive model of order one, $x_{i,t} = a_i + \rho_i x_{i,t-1} + \epsilon_{i,t}$, for the variables made stationary through differentiation according to unit root tests.[9] The forecasts are then produced using:

$$\hat{x}_{i,t+1|\mathrm{t}} = \hat{a}_i + \hat{\rho}_i x_{i,t}$$

Bayesian VAR (BVAR)

Acknowledging that interdependencies among components could play an important role, Bayesian Vector Autoregressive models (BVARs) are also used for forecasting. This is done by following the implementation in Banbura et al. (2010). In practice, the whole multivariate process is forecasted using five lags and, as in Banbura et al. (2010), the overall tightness that produces the same in-sample fit as that of the direct aggregate forecast is chosen.

The estimated model is

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}_1\mathbf{X}_{t-1} + \ldots + \mathbf{A}_5\mathbf{X}_{t-5} + \epsilon_t$$

and the forecasts are produced using

$$\hat{\mathbf{X}}_{t+1|\mathrm{t}} = \hat{\mathbf{c}} + \hat{\mathbf{A}}_1\mathbf{X}_t + \ldots + \hat{\mathbf{A}}_5\mathbf{X}_{t-4}$$

[9] The differentiation for each series is presented in section 3.B of the Appendix

3.3.3 Alternatives for Measures of Dissimilarity

Dissimilarity measures and linkage methods have a defining impact on results, and the relevant literature provides many alternatives to choose from. As James et al. (2013) point out, the choice of which alternative to use depends on the type of data and question at hand. In the statistical learning literature it is not unusual to use simple correlation as the dissimilarity measure for time-series. The forecasting literature, however, points towards the notion of commonality. The problem is that there is not one single way of measuring it. For this reason six different possibilities are used based on what has been suggested in the literature.

Alternative 1: Pearson's Correlation

In the statistical learning literature there are many alternatives, but in the context of time-series the most obvious are measures for correlation. Probably the best-known is Pearson's correlation coefficient, which measures the strength of the linear relationship between two variables. Although its limitations are many, its widespread use make it an obvious benchmark for the rest of the measures.

The correlation coefficient between x_i and x_j is defined as $\rho_{x_i x_j} = \frac{cov(x_i,x_j)}{\sigma_{x_i}\sigma_{x_j}}$, where $cov(x_i, x_j)$ is the covariance between x_i and x_j and σ_{x_i},σ_{x_j} are the respective standard deviations. As a higher correlation, in absolute terms, is associated with similarity, the corresponding dissimilarity measure is defined as:

$$PC_{x_i,x_j} = 1 - \text{abs}\left(\frac{cov(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}}\right)$$

Alternative 2: Spearman's Correlation

As pointed out by Hauke and Kossowski (2011), sometimes Pearson's correlation coefficient can produce results that are undesirable or misleading. This may be because it is restricted to linearity or requires variables to be measured on interval scales.

Spearman's rank correlation coefficient is a non-parametric rank statistic that assesses how well an arbitrary monotonic function can describe the relationship between two variables. Therefore it is not affected by non-linearity. In practice, however, it is just the Pearson's Correlation coefficient, in which the data are converted to ranks before calculating the coefficient.

The rank correlation coefficient between x_i and x_j is defined as $r_{x_i x_j} = \frac{cov(x_i^{rank},x_j^{rank})}{\sigma_{x_i^{rank}}\sigma_{x_j^{rank}}}$, where x_i^{rank} and x_j^{rank} are the ranks of x_i and x_j respectively. Again, as a higher correlation, in absolute terms, is associated with similarity, the corresponding dissimilarity measure is defined as:

$$SC_{x_i,x_j} = 1 - \text{abs}\left(\frac{cov(x_i^{rank}, x_j^{rank})}{\sigma_{x_i^{rank}}\sigma_{x_j^{rank}}}\right)$$

Alternative 3: Latent Factor

In the context of measuring commonality in applications with financial data, Adrian (2007) and Bussière et al. (2015) use the variance explained by the first principal component to measure the commonality among a set of variables. As they explain, the decomposition transforms the original variables into a new set that are orthogonal and in which they are ordered so that the first retains most of the variation present in all of the original variables while the last has the least. This is in line with the approaches in the Dynamic Factor Models literature that try to capture the common factors using Principal Component Analysis (Stock and Watson, 1998, 2002).

As explained by Hastie et al. (2009), for n series of length T, the sample's covariance matrix $\frac{1}{T}\mathbf{X}^T\mathbf{X}$ can be rewritten using the eigen decomposition as $\mathbf{V}\mathbf{D}^{2}\mathbf{V}^T$. The columns of $\mathbf{V}$, the eigenvectors, are the principal component directions of $\mathbf{X}$ and $\mathbf{z}_1 = \mathbf{X}v_1$, with v_1 being the first column of $\mathbf{V}$, is the first principal component. The values on the diagonal of $\mathbf{D}^{2}$ are the eigenvalues associated with each eigenvector, that is d_1^2 for v_1.

It can be shown that $\text{Var}(\mathbf{z}_1) = \text{Var}(\mathbf{X}v_1) = \frac{d_1^2}{T}$. Then the total variance explained by the first principal component is $d_1^2/\sum_{l=1}^{n} d_l^2$. As a higher total explained variance is associated with similarity, the corresponding dissimilarity measure is defined as:

$$LF_{x_i,x_j} = 1 - \left(\frac{d_1^2}{\sum_{l=1}^{n} d_l^2}\right)$$

Alternative 4: Persistence

Bermingham and D'Agostino (2014) point out that series that have very different persistence will tend to suffer more from omitted variable bias if they are forecasted together than series with a similar persistence. They advocate forecasting series with different persistence separately.

To take this point into consideration, an AR(1) model is fitted to each component, $x_{i,t} = a_i + \rho_i x_{i,t-1} + \epsilon_{i,t}$, and the difference in the estimated persistence parameter is used as a measure for dissimilarity:

$$PE_{x_i,x_j} = \text{abs}\left(\text{abs}\left(\hat{\rho}_i\right) - \text{abs}\left(\hat{\rho}_j\right)\right)$$

Alternative 5: Forecast-error Clustering

Granger (1987) states that ignoring the common factor and interdependencies between components will tend to make forecasting errors cluster instead of cancelling out. With this phenomenon in mind, AR(1) models are fitted to each component and the correlations of the resulting out-of-sample forecasting errors for the most recent periods are used as the dissimilarity measure.

Specifically, for each component i, $x_{i,t-p+1} = a_i + \rho x_{i,t-p} + \epsilon_{i,t}$, where p is the number of periods that are evaluated for the measure. With the model, forecasts are generated from $t-p+1$ to t and the corresponding forecasting errors are calculated as $\hat{x}_{i,s|\text{s-1}} - x_{i,s}$ for $s = t-p+1$ to t. They are then collected in $\hat{\mathbf{e}}_i^t$. With this, the dissimilarity measure is defined as:

$$EC_{x_i,x_j} = 1 - \text{abs}\left(\frac{cov(\hat{\mathbf{e}}_i^t, \hat{\mathbf{e}}_j^t)}{\sigma_{\hat{\mathbf{e}}_i^t}\sigma_{\hat{\mathbf{e}}_j^t}}\right)$$

Alternative 6: Bayesian Hierarchical Clustering

The final alternative uses a procedure developed by Cooke et al. (2011) based on a method by Heller and Ghahramani (2005). The criterion for commonality is given by the probability of two different time-series having been generated from the same underlying function and therefore belonging to the same cluster. In the framework, this probability is referred to as the probability that two series should be merged.

The essence of the method can be seen from the explanation in Murphy (2012).[10] Let $D = \{x_1, \ldots, x_n\}$ represent all the data and D_i the data at subtree T_i. Then, at each step, subtrees T_i and T_j are compared to see if they should be merged together. The hypothesis to be evaluated is that x_i and x_j come from the same probabilistic model $p(x \mid \theta)$ of unknown parameters θ. Then define D_{ij} as the merged data, and let M_{ij} equal one if they should be merged and zero if they should not. The probability of a merge is given by

$$r_{ij} = \frac{p(D_{ij} \mid M_{ij}=1)p(M_{ij}=1)}{p(D_{ij} \mid M_{ij}=1)p(M_{ij}=1) + p(D_{ij} \mid M_{ij}=0)p(M_{ij}=0)}$$

$p(M_{ij} = 1)$ is the prior probability of a merge and can be computed from the data (Heller and Ghahramani, 2005). If M_{ij} is equal to one, the data is assumed to come

[10] A complete description can be found in Savage et al. (2009). They present the R/BHC computational package to perform Bayesian Hierarchical Clustering. Their package is used for the empirical application.

from the same model meaning

$$p(D_{ij} \mid M_{ij} = 1) = \int \left[\prod_{x_n \in D_{ij}} p(x_n \mid \theta) \right] p(\theta \mid \lambda) d\theta$$

with λ being a hyperparameter that can be provided or estimated from the data. If M_{ij} is equal to zero, the data is assumed to have been generated independently and

$$p(D_{ij} \mid M_{ij} = 0) = p(D_i \mid T_i) p(D_j \mid T_j)$$

3.3.4 Alternatives for Producing the Definitive Aggregate Forecast

The outcome from the clustering algorithm is a complete hierarchy and, because of the way the algorithm works, it will offer a number of levels of aggregation equal to the number of original components. As mentioned before, the algorithm by itself does not provide any advice with regard to which grouping to use. This means that an exogenous criterion is required. For this purpose, six different alternatives are presented, separating the methods into those that seek to select a single level of disaggregation and those that use a combination of the different groupings.

3.3.4.1 Disaggregation Level Selection

Scheme 1: In-sample Fit

Probably the approach most commonly used to judge a model in the forecasting literature is in-sample fit. It has some known drawbacks, but its widespread use makes it a natural choice. For this particular case the in-sample forecasting error is used. To choose the level of aggregation for forecasting period $t+1$, for each level of aggregation within the proposed hierarchy at time t, the forecasting models and parameters calculated using data up to period t are used, to calculate the one-step-ahead mean square forecasting error (MSFE) for the sample up to period t.

With this, the in-sample fit for disaggregation level i, at time t is:

$$ISF_{i,t,v} = \frac{1}{v} \sum_{s=t-1-v}^{t-1} \left(\hat{x}_{i,s+1|\mathrm{t}} - x_{i,s+1} \right)^2$$

where v determines how much data is included in the measure.

The level of aggregation with the lowest in-sample forecasting error is then used to forecast period $t+1$.

Scheme 2: Past Out-of-sample Forecasting Performance

One of the drawbacks of in-sample criteria is that they will tend to over-fit the data (Eklund and Karlsson, 2007). It is therefore very common to use out-of-sample evaluation as an alternative. The out-of-sample criterion, for forecasting period $t+1$, is calculated in this case using a recursive out-of-sample forecasting exercise. For each level of aggregation within the proposed hierarchy at time t, the parameters with data up to period $t-v$ are estimated and period $t-v+1$ is forecasted. The parameters with data up to period $t-v+1$ are then estimated and $t-v+2$ forecasted, continuing in the same way and stopping with the forecast for period t. Then, the MSFE using these forecasts is calculated.

With this, the out-of-sample performance for disaggregation level i, at time t is:

$$OOS_{i,t,v} = \frac{1}{v} \sum_{s=t-1-v}^{t-1} \left(\hat{x}_{i,s+1|\mathrm{s}} - x_{i,s+1} \right)^2$$

where v determines how much data is included in the measure.

The level of aggregation with the lowest out-of-sample forecasting error is then used to forecast period $t+1$.

Scheme 3: Lowest Average Error Dissimilarity Threshold

Unsupervised learning, of which the clustering method used to produce the subset of groups is part, is often challenging because there is no response variable. In this context, however, the ultimate objective is to find the level of aggregation at which the resulting aggregate forecast error is lowest. For this purpose, a supervised method can be used to try to learn the best grouping for the purpose of forecasting.

For this particular measure, the degree of commonality, as measured by the value of the corresponding dissimilarity measure, is assumed to be related to the forecasting error. The method to estimate this relation proceeds by calculating the average forecasting error conditional on the level of dissimilarity for the training sample. In practice, this involves calculating the forecasting errors associated with the values on the vertical axis of all the dendrograms for the sample up to period t and averaging the results. To make the averaging over different periods possible, a smoothing spline is used to interpolate the forecasting errors for each period. To forecast period $t+1$ the level of aggregation associated with the dissimilarity that is closest to the minimum average error is chosen.

Scheme 4: Probabilistic Criterion

The Bayesian Hierarchical Clustering method proceeds by building the hierarchy based on the estimated probability of two observations coming from the same underlying function. Heller and Ghahramani (2005) suggest that a natural decision rule for groupings in this context is only to perform fusions that have a posterior merge probability greater than 50%. This criterion, however, can only be applied to hierarchies produced by the probabilistic algorithm.

3.3.4.2 Disaggregation Level Averaging

A popular way of dealing with the choice between two or more competing forecasts is to avoid the decision altogether and combine them. The idea of forecast combination has been around for a long time and deals with the issue of exploiting the information contained in each individual forecasts in the best possible way. The literature on it is extensive and the surveys by Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002) and Timmermann (2006) not only bear witness to it but also highlight the robustness of the gains in forecasting accuracy due to its use.

Scheme 5: Equal Weights Among Aggregate Forecasts

A very attractive feature of forecast combination is that simple combination schemes are surprisingly effective (Timmermann, 2006). In fact, the equal-weighted forecast combination performs so well that researchers have tried to explain why this should occur (Smith and Wallis, 2009; Elliott, 2017). In view of this, given that each level of the hierarchy produces an aggregate forecast, the simplest thing to do is to average the aggregate forecasts for all levels.

Scheme 6: Equal Weights Among Distinct Forecasts

In this context, however, averaging the aggregates is not the same as assigning equal weights to each distinct forecast. To see why, it is helpful to look back at the dendrogram in Figure 3.1. The last-but-one fusion of the algorithm involves components 7 and 12. If the forecasts are generated independently of each other, for all of the groupings below their fusion, the corresponding aggregate forecasts include those for these two individual components. So when all aggregate forecasts are averaged, the forecasts for these two components are implicitly given a weight that is ten times larger than the forecasts of the components that are fused in the first step.[11]

An alternative approach to averaging the aggregate forecasts is to give equal weights to each single forecast. That means including each individual component forecast, each intermediate aggregate forecast and the aggregate forecast only once.[12]

[11] This is not the case for the multivariate forecasting models.

[12] To do this it is necessary to combine forecasts from multiple levels of aggregation and this is done using the method described in section 3.A of the Appendix. This approach can also be used to obtain a consistent underlying forecasting scenario.

3.3.5 Forecasting Accuracy Comparison

3.3.5.1 Set-up of the Evaluation Exercise

The evaluation exercise is performed over the 2001-2015 period, leaving the first ten years of data to estimate the models. It is set up in a quarterly rolling scheme using a ten-year window, where in each period the models are re-estimated and a one-step-ahead forecast is generated.

The forecasting accuracy is presented by means of the model's MSFE relative to that of a benchmark model. For variable i and using model m, the relative MSFE is

$$\text{RelMSFE}^{(i,m)} = \frac{\text{MSFE}^{(i,m)}_{T_0,T_1}}{\text{MSFE}^{(i,0)}_{T_0,T_1}}$$

with

$$\text{MSFE}^{(i,m)}_{T_0,T_1} = \frac{1}{T_1 - T_0 + 1} \sum_{t=T_0}^{T_1} \left(y^{(m)}_{i,t+1}|t - y_{i,t+1} \right)^2$$

where $y^{(m)}_{i,t+1}|t$ is the forecasted value for $t+1$ at time t and T_0 is the last period of actual data in the first sample used for the evaluation and T_1 is the last period of actual data in the last sample. As usual, a RelMSFE lower than one reflects an improvement over the benchmark model for which $m = 0$. To evaluate the significance of the differences, the forecasts are compared using the Modified Diebold-Mariano test for equality of prediction accuracy proposed by Harvey et al. (1997).

3.3.5.2 Benchmark Forecasting Approaches

The objective of the whole exercise is to find out whether there are successions of intermediate aggregations that may improve forecasting accuracy, as opposed to restricting oneself to using either the direct or the full bottom-up approach. As a starting point, these two approaches are therefore obvious comparison points.

As mentioned before, there is a lot of literature supporting the gains in forecasting accuracy achieved by using forecast combination and particularly the good performance of the unsophisticated equal-weighted variant. Therefore, as a second set of benchmarks, the results of the grouping procedure are measured up against the equal-weighted combination of the direct and bottom-up approaches.

Bermingham and D'Agostino (2014) find that the performance of the bottom-up approach may improve if common features among components are accounted for. To see whether the univariate models in this exercise are negatively affected by not incorporating this point, a factor-augmented autoregressive model is presented alongside the benchmark models. This is done by extending each univariate autoregressive model from the bottom-up approach to include one factor:

$$x_{i,t} = a_i + \rho_i x_{i,t-1} + \gamma_i F_{t-1} + \epsilon_{i,t}$$

The factor, F, is estimated with the first principal component from the whole set of components following Stock and Watson (2002). The corresponding forecast for each component is generated using:

$$\hat{x}^{FAAR}_{i,t+1|t} = \hat{a}_i + \hat{\rho}_i x_{i,t} + \hat{\gamma}_i \hat{F}_t$$

3.3.6 Results

3.3.6.1 Aggregate Forecasting Performance Evaluation

The exercise seeks to determine the potential benefits of using dynamic grouping methods. A first step is therefore to evaluate how the benchmark models perform. Table 3.2 shows what would be a traditional aggregate-disaggregate comparison for the three series by presenting the MSFE of the direct and bottom-up approaches and their combinations. It also presents the results for the factor-augmented AR models to have a notion of whether the suggestion by Bermingham and D'Agostino (2014) is able to improve the univariate bottom-up methods in these particular settings.

In terms of comparing the direct and bottom-up approaches, the results for the benchmark models show that in five out of six cases the respective bottom-up approach performs better. It is only for the BVAR for the United Kingdom that the direct approach performs better. In terms of comparing the AR(1) and BVAR, the univariate approach tends to do better with improvements going from 5 to 12%. In the cases of France and the United Kingdom, the AR(1)s show improvement over the direct method by 9 and 12% respectively. For Germany, on the other hand, the BVAR is marginally better than the AR(1) improving over the direct approach by 6%. Although in some cases the magnitude of the improvements seems large, none are statistically significant. In the case of the combinations, on the other hand, the resulting forecasts from the AR(1)s are statistically better for France and the United Kingdom even when the size of the improvements are the same as those of the respective bottom-up approaches. Similarly, for the BVARs the results for France and Germany are statistically better, with the magnitudes being slightly different, lower for France and higher for Germany. In the case of the United Kingdom, the use of combination means an improvement in accuracy over both the direct and bottom-up approaches by 6 and 23 percentage points respectively, but neither is statistically significant. As mentioned before, in their

Table 3.2: Aggregate Forecasting Performance of Benchmark Methods

	Factor aug.	AR(1)		BVAR	
	AR(1)	Bottom-Up	Combination	Bottom-Up	Combination
France	0.91	0.91	0.91**	0.95	0.93**
Germany	0.98	0.95	0.96	0.94	0.95*
United Kingdom	0.88	0.88	0.88**	1.17	0.94

Note: Mean square forecasting error relative to the direct method. Benchmark models are the bottom-up and equal-weight combination of the direct and bottom-up approaches of the AR(1) and BVAR and a factor augmented AR(1) following Bermingham and D'Agostino (2014). * and ** denote significance of the forecasting performance difference based on the modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015.

implementation Bermingham and D'Agostino (2014) find that using factor-augmented univariate models for the components shows improvements over the univariate models alone. In this case, their method does not seem to have the intended impact.

Moving on to the grouping framework, Table 3.3 presents the MSFE of the grouping methods relative to the direct approach for the three countries and both models.[13] Figures in bold denote that the corresponding grouping method shows an improvement over the best benchmark model. Looking at the details by country and model shows that for France and the univariate models, almost all grouping methods are better than the direct approach and in many cases they are an improvement over the best benchmark. The maximum improvement over the direct approach is 11% with the best methods being the Bayesian/in-sample (B/IS), persistence/dissimilarity threshold (P/DT) and many of the methods that involve the combination of all the aggregation levels. All of these show improvements that are statistically significant. For the BVARs on the other hand, many of the grouping methods are better than the direct approach and some are an improvement over the best benchmark. The maximum improvement over the direct approach is also 11%, with the best methods being the P/DT and many of the combinations. For Germany and the univariate models, most grouping methods perform better than the direct approach, but by a smaller magnitude than for France. In only a few cases the grouping methods are an improvement over the best benchmark, but in these cases the improvement is comparatively large: 11% over the direct approach and 6% over the best benchmark. In this case, the best methods are P/DT and both combination choice methods that use the persistence dissimilarity measure. For these three methods, the improvements are statistically significant. For the BVARs on the other hand, less than half the grouping methods show an improvement over the direct approach and none over the best benchmark. The best methods are Bayesian/dissimilarity threshold and again P/DT with a maximum improvement of 5%. The latter is the only method for which the difference is statistically significant. For the United Kingdom and the univariate models, as with France, almost all grouping methods perform better than the direct approach and in many cases they are

[13] In this context and for the rest of this section, grouping method refers to the pairing of a dissimilarity measure and a choice method.

Table 3.3: Aggregate Forecasting Errors of Grouping Methods by Country and Model

	AR(1)						BVAR					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
France												
Pearson's corr.	0.92	0.96	0.92*		**0.89****	0.92**	1.01	0.98	0.96		**0.89****	**0.91****
Spearman's corr.	**0.91**	0.93	0.93		**0.89****	0.92**	1.06	1.08	1.06		0.94	0.93*
latent factor	0.96	0.96	0.99		0.92**	0.93**	1.00	1.03	0.98		**0.93***	**0.92****
persistence	**0.91**	0.93	**0.90****		**0.90***	**0.90****	1.04	0.98	**0.90****		0.94	**0.92****
f-error clustering	**0.91**	0.93	**0.91**		**0.89****	**0.91****	1.04	1.04	0.98		0.94	0.94*
Bayesian	**0.89***	0.93	0.92	1.02	0.92	0.94	1.00	1.00	0.98	1.03	0.95	0.95
Germany												
Pearson's corr.	0.98	1.02	1.00		0.99	0.98	1.05	1.11	1.06		1.00	0.99
Spearman's corr.	0.98	1.01	1.02		0.99	0.98**	1.06	1.12	1.05		1.00	0.98
latent factor	0.99	1.01	1.01		0.99	0.99	1.05	1.01	1.04		1.00	0.99
persistence	0.97	0.97	**0.89****		**0.93****	**0.94****	1.07	1.02	0.96**		0.97	0.96
f-error clustering	0.99	1.01	0.97		0.99	0.98	1.11	1.06	0.98		0.97	0.98
Bayesian	0.98	0.99	0.96	1.00	0.96*	0.97*	0.98	1.08	0.95	1.02	0.97	0.97
United Kingdom												
Pearson's corr.	0.90	0.90	0.95		0.88	**0.86****	**0.91**	**0.88**	**0.93**		0.95	**0.90**
Spearman's corr.	0.89	0.95	**0.87**		0.90	0.89*	1.00	**0.91**	1.00		0.98	**0.91**
latent factor	**0.86**	0.94	**0.86**		**0.86***	0.88*	0.96	0.99	**0.89**		**0.89**	**0.90**
persistence	0.94	0.94	1.00		0.94	0.90	**0.93**	1.02	0.95		1.01	0.97
f-error clustering	0.96	0.99	**0.86**		0.89	**0.86****	0.96	1.00	1.04		0.94	**0.91**
Bayesian	**0.86**	0.91	**0.88**	1.11	0.89*	0.90*	**0.87**	0.94	1.16	1.18	0.95	0.95

Note: Mean square forecasting error relative to the direct method. Grouping method dissimilarity measures: Pearson's correlation, Spearman's correlation, latent factor given by the variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1) and Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, out-of-sample criterion, dissimilarity threshold, probabilistic criterion, forecast combination that assigns equal weights to the aggregate forecasts (FC1) and forecast combination that assigns equal weights to each distinct forecast (FC2). In bold MSFE lower than the lowest of either the respective full bottom-up approach or the direct approach. * and ** denote significance of the forecasting performance difference based on the Modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015.

an improvement over the best benchmark. The maximum improvement over the direct approach is 14% with many different methods achieving the best performance. Only in the case of some of the combination methods, however, are the improvements statistically significant. Finally, for the BVARs most of the methods are better than the direct approach and many show an improvement over the best benchmark. The best method is B/IS with an improvement of 13% over the direct approach and 7 percentage points over the best benchmark. None of the differences, however, come out as statistically significant.

These results seem to show some aspects that are common to all the different cases and others that are not. In order to draw some overall conclusions, an absolute and a relative summarizing measures are presented. The former uses the average of the deviation of the respective MSFE from that of the corresponding best overall performing

Table 3.4: Overall Aggregate Forecasting Performance of Grouping Methods

	Average Percentage Deviation from Best Method						Average Rank Difference with Best Method					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
Pearson's corr.	7.4	9.0	8.9		4.5	**3.8**	15.3	19.8	18.8		9.3	**8.3**
Spearman's corr.	10.1	11.9	10.5		6.4	4.6	17.8	23.7	22.0		13.7	9.5
latent factor	8.4	11.1	7.5		**4.3**	4.8	17.3	23.3	15.7		9.5	11.7
persistence	9.5	9.6	4.9		6.2	**4.5**	16.3	18.8	**9.0**		10.3	**8.3**
f-error clustering	11.4	12.5	7.4		5.0	**4.2**	20.8	25.2	12.3		9.8	**8.5**
Bayesian	**4.2**	9.1	9.3	18.9	5.5	6.1	**7.0**	20.3	12.0	25.5	11.3	13.5

Note: Relative performance of the grouping methods as measured by the average deviation of the respective MSFE relative to that of the best performing grouping method by category and as the average difference in rank according to MSFE over the six sets of forecasts. Grouping method dissimilarity measures: Pearson's correlation, Spearman's correlation, latent factor given by the variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, out-of-sample criterion, dissimilarity threshold, probabilistic criterion, forecast combination that assigns equal weights to the aggregate forecasts (FC1) and forecast combination that assigns equal weights to each distinct forecast (FC2). In bold the five best performers in each category. Calculated over 2001-2015.

grouping method over all six sets of forecasts. The latter presents the average difference in rank of the grouping methods, where the most accurate in the MSFE sense is ranked first and the least accurate is ranked last, that is 31st. For both measures a smaller number means a better performing model. The outcome for the measures is presented in Table 3.4 and both tell a similar story. In terms of selecting a single level for obtaining the definitive forecasts, the B/IS and P/DT methods come out as best by a fair margin. In terms of the combination methods, the performance of most methods is good, relative to the remaining single-level methods. When comparing combination approaches, although the differences are relatively small, the approach that gives equal weight to each distinct forecast comes out better, with the best results being observed for the Pearson's Correlation, persistence and forecast-error clustering dissimilarity measures. The two summarizing measures do differ, however, in identifying the ordering of the best methods. The absolute measure identifies the Person's/distinct forecast combination as the best method, while the relative measure puts the B/IS first. The difference in magnitudes in the scores used for the specific rankings, however, is relatively small.

All this suggests that the B/IS, P/DT and combination methods are better on average than the rest in terms of aggregate accuracy. The evaluation sample, however, includes the financial crisis, meaning that the results could be affected considerably by this episode alone. After removing the crisis years from the sample, however, the overall results remain and in some cases improve slightly.[14] The better performing grouping methods used with the univariate model improve around 2 percentage points across countries. For the BVARs the results are similar except for the United Kingdom, where the gap between the best grouping method and the best benchmark broadens

[14] The corresponding tables are presented in section 3.B of the Appendix.

to 13 percentage points. In terms of the relative accuracy of the aggregation level choice methods, the best single-level methods do not change but fall behind most of the combination methods, all this in a context where univariate benchmark models are hardly affected, while the performance of the BVARs benchmarks deteriorate slightly both for the bottom-up approach and combination.[15]

Taking all of this into consideration, the overall results suggest that the use of the dynamic grouping method can result in improvements in forecasting accuracy. They also show, however, that the outcome depends greatly on the specification, dataset and models used. Looking into the workings of each grouping method in more detail could provide some insight into the differences between methods and why some perform better than others.

3.3.6.2 Grouping Method Performance Comparison

In this particular exercise, 31 different specifications for the dynamic grouping method are evaluated, using six different sets of forecasts. As it can be seen from Table 3.3, the performance of the different specifications is quite heterogeneous. It can also be seen that there is neither a dissimilarity measure nor a choice method that outperforms the rest outright. There are, however, certain pairings that seem to work better. Examining some aspects of the performance of the different specifications could result in a better understanding of the differences between them. As the grouping methods involve two steps, a way of approaching further analysis is to look at the subset selection and level choice methods separately and try to establish the merits of each one of them.

To evaluate the grouping subset selection process alone, it is first necessary to establish some way of measuring the performance of each dissimilarity measure independently of the choice method. One way of comparing them is in terms of the best possible improvements for each dissimilarity measure. Table 3.5 presents the MSFE relative to the direct method of the succession of aggregation level selections that lead to the lowest MSFE over the whole sample by dissimilarity measure. Considering all scenarios, the improvements range from 25 up to 57 percentage points. Common to all countries is that the improvements over the direct method are larger for the BVARs than for the univariate models by about 5 percentage points and that the differences between measures are small relative to the potential improvement. This suggests that the good performance of one grouping method or another is not solely due to the grouping subset selection process. In fact, the selection processes associated with the two best performing grouping methods, that is the persistence and Bayesian dissimilarity measures, do not come out on top in the comparison.

[15] A more noticeable difference is that the improvements over the direct method of the combination using BVARs cease to be statistically significant.

Table 3.5: Lowest Achievable Forecasting Errors by Dissimilarity Measure

	France		Germany		United Kingdom	
Choice method	AR(1)	BVAR	AR(1)	BVAR	AR(1)	BVAR
Pearson corr.	0.66	0.59	0.75	0.67	0.54	0.48
Spearman corr.	0.67	0.61	0.75	0.69	0.53	0.49
1st princ.comp	0.69	0.58	0.75	0.69	0.53	0.43
persistence	0.67	0.63	0.75	0.69	0.56	0.51
f-error clustering	0.66	0.63	0.75	0.69	0.55	0.50
Bayesian	0.69	0.64	0.75	0.69	0.53	0.49
Min - Max	0.02	0.06	0.01	0.02	0.03	0.08

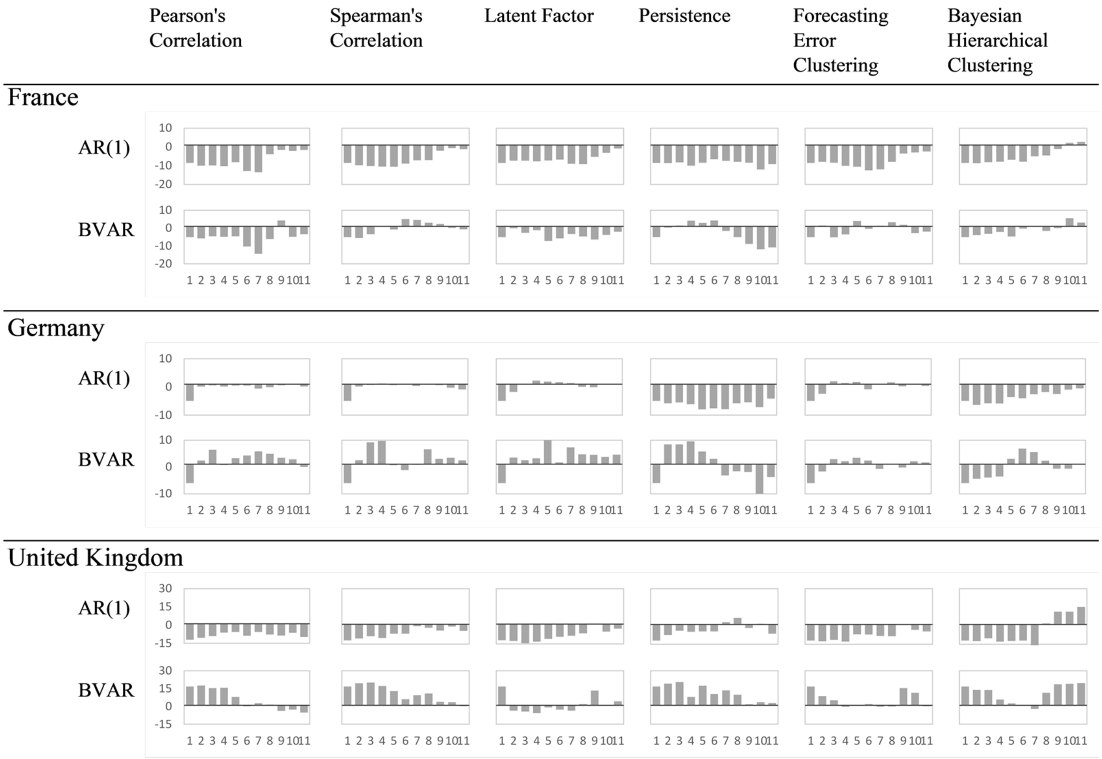
Note: Mean square forecasting error relative to the direct method of the succession of aggregation level selections that lead to the lowest MSFE over the whole sample by dissimilarity measure. Dissimilarity measures: Pearson correlation, Spearman correlation, variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1) and Bayesian Hierarchical Clustering. Calculated over 2001-2015.

Using the best possible performance as a benchmark alone, however, could be misleading with regards to the role of the dissimilarity measure. It could be that some measures perform well consistently, even though they do not lead to the lowest possible MSFE. To have a notion of whether this could be the case, Figures 3.2 and 3.3 present absolute and relative summary measures for the aggregation levels by dissimilarity measure, country and model. The former presents the percentage deviation in MSFE relative to the direct method for each aggregation level while the latter presents the distribution by aggregation level that results from ranking each one according to MSFE over the whole sample.

From what can be observed in Figure 3.2, the performance of the different dissimilarity measures is not as homogeneous as the comparison with the best possible improvement might suggest. Some cases do look similar, as in the case of France using the AR(1)s for example, but in most cases there are substantial differences. Probably the most outstanding one is for the German univariate models. In this case, four of the dissimilarity measures show a negligible impact, but the other two achieve relatively large gains through their particular selection of groupings. In this case, as it can be seen from the results in Table 3.3, the P/DT method improves 11% over the direct method as opposed to only 4% of the next best method. All this suggests that, although not wholly responsible, the choice of dissimilarity measure has a defining impact on the outcome.

In terms of analysing the choice methods alone, the task is not straightforward either, given that their performance is conditional on the dissimilarity measures. In this context, two perspectives are examined. The first consists of comparing the relative performance of the choice methods across dissimilarity measures and the second involves assessing how the methods perform compared to simply selecting a given aggregation

Figure 3.2: Mean Square Forecasting Error by Aggregation Level



Note: Mean square forecasting error relative to the direct method for each aggregation level by dissimilarity measure, country and model. Presented as the percentage deviation. Negative numbers reflect a lower MSFE and therefore an improvement. The horizontal axis presents the eleven aggregation levels other than the aggregate that is used as the benchmark and with one being the full bottom-up approach. Dissimilarity measures: Pearson correlation, Spearman correlation, variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1) and Bayesian Hierarchical Clustering. Calculated over 2001-2015.

level for the entire exercise. In terms of the relative performance, from the results in Table 3.3, it is clear that the probabilistic choice criterion does not work as intended in this context. In all scenarios it almost always performs worse than the other grouping methods. Although not so extreme, another method that does not perform well relative to others is the past out-of-sample performance choice criterion. In most cases, it is beaten either by the in-sample, persistence and combination criteria. With regard to these methods, their relative performance depends on the dissimilarity measure and dataset.

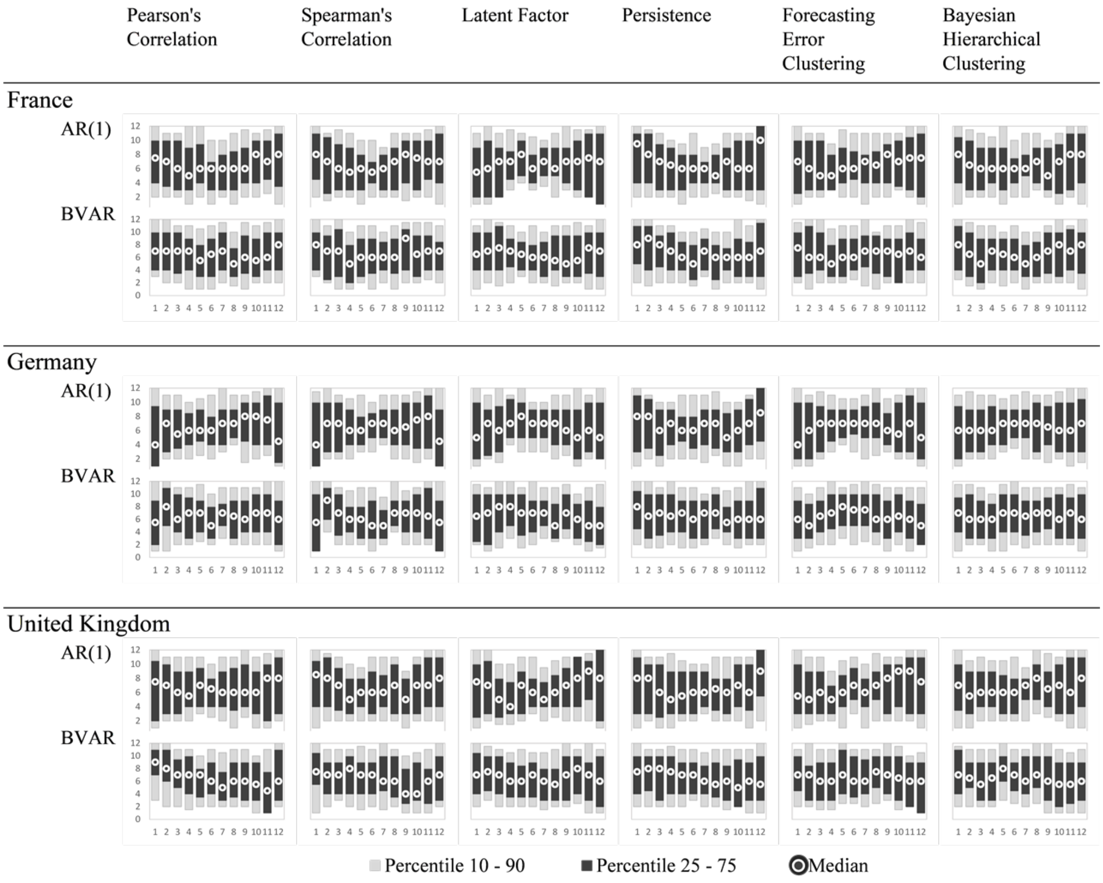
In terms of examining the performance of the choice methods for a given dissimilarity measure, the performance of different aggregation levels is usually quite heterogeneous. This becomes apparent from looking in more detail at the charts in Figure 3.2, where within the same hierarchy in many cases specific aggregation levels exhibit a significantly lower MSFE than others. This can be seen most clearly for the BVARs for France using Pearson's Correlation and Germany using persistence. This raises the question of whether these aggregation levels consistently perform well relative to the other aggregation levels. In that case, choosing the appropriate grouping level for forecasting should be relatively straightforward. A lower MSFE over a given period, however, does not mean that an aggregation level is necessarily better all the time. To have a notion of the consistency with which the aggregation levels perform, Figure 3.3 presents the distribution by aggregation level that results from ranking each one by MSFE for each period over the whole sample. If an aggregation level were to perform better than the rest on a regular basis, one would observe a short bar close to the bottom, and if this happened all of the time, the distribution would collapse on one.

From what can be observed, this does not occur. There are some cases where the distribution is relatively concentrated, but all of these are located in the middle of the ranking. Although based on the medians alone there are aggregation levels that rank better than the rest in most of the scenarios, once the distribution is considered it is clear that the aggregation level rankings fluctuate considerably in all scenarios. In most cases the 10 to 90% interval covers nearly the whole ranking and for some of them the 25 to 75% interval does so too. In this context, it would seem that dynamically choosing a succession of aggregation levels that lead to improvements could be quite challenging.

Focusing only on the choice methods that seek to determine a single aggregation level for forecasting, some evidence of improvements actually occurring can be found by comparing the MSFE by aggregation level with the outcome of the choice methods presented in Table 3.3. The clearest example is seen in the case of the BVAR for the United Kingdom. Here, the grouping methods improve over the direct and bottom-up approaches by as much as 13%. The best of the aggregation levels presented in Figure 3.2, however, improve less than 6% with most of the aggregation levels actually being worse then the direct approach. To a lesser extent, the same is true for the univariate models for Germany, where the P/DT method shows an improvement of 3 percentage points over the best performing aggregation level. This improvement in performance, however, is not generalized. That of the other choice methods greatly depends on the dissimilarity measure being used.

The performance of the combination methods, on the other hand, are less heterogeneous and are good overall. This is not entirely surprising in light of the preceding analysis that suggests that the persistence of the forecasting performance of the aggregation levels is relatively low. This phenomenon is not uncommon in the forecasting literature where, in the context of forecasting model comparison, models that perform well up to a point suddenly stop doing so (Aiolfi and Timmermann, 2006). In the given scenario, dynamically choosing successions of aggregation levels that lead to improvements could prove to be very challenging. In this context, Hubrich and Skudelny (2017) find that forecast combination helps with hedging against bad forecast performance. This seems to be the case here.

Figure 3.3: Ranking Distribution of Forecasting Performance by Aggregation Level



Note: Distribution of ranking according to MSFE of the resulting aggregate forecast for each level of aggregation when the whole sample is considered. The horizontal axis presents each level and the vertical the ranking, from first to 12th. The lightly shaded area represents the 10 to 90% percentile interval, the dark area the 25 to 75% and the circles denote the median. Calculated over the 2001-2015 sample.

In spite of the apparently unfavourable characteristics of the forecasting scenario, two of the grouping methods that seek to dynamically select a single aggregation level exhibit results that are comparable to those of the combination methods. In order to help identify whether a specification is appropriate or not, one might ask whether these two methods share any common characteristics that are lacking in the rest. Nothing obvious appears in the previous analysis and comparing the succession of aggregation level choices of the different methods provides no additional insight either.[16] One possible source for this better performance is that both methods attempt to determine the measure for commonality from a function of the data instead of the data itself. Data often includes idiosyncratic noise that affects the estimated coefficients and causes unpredictable effects on sub-aggregations. Both the Bayesian and the persistence methods estimate the dissimilarity measures and therefore include an error term. This could potentially lead both methods to be less affected by idiosyncratic noise that is present in the data. Corroborating whether this is in fact the case is left for further research.

[16] The information for all aggregation choices is presented in section 3.B of the Appendix.

3.3.6.3 Overall Assessment

The overall evaluation is that the dynamic grouping methods are capable of increasing forecasting accuracy, but that improvements depend critically on specification. The analysis of the data reveals that the aggregation levels that are formed in the subset selection procedure exhibit low performance persistence. One would expect this to contribute to making the choice of the best aggregation level for forecasting particularly hard. Nevertheless, two of the grouping methods that seek to select a single aggregation level for forecasting every period perform relatively well. These are the Bayesian/in-sample and persistence/dissimilarity threshold methods. The reason why they outperform many of the other methods is not clear, but a possible explanation could come from the fact that they both take into account the idiosyncratic noise in the data in the grouping process. The combination methods, on the other hand, perform well overall. In line with what has been observed in the forecasting literature that deals with model comparison in similar situations, this suggests that using them in this context can provide a way of introducing the robustness of forecasting combination into the procedure without having to introduce different forecasting models. The evaluation sample includes the most recent financial crisis, but the overall results do not change when it is removed from the sample.

3.4 Conclusions

This chapter presents a framework to forecast economic aggregates based on purpose-built groupings of components. The idea underpinning this approach is that there are reasons to support both forecasting an aggregate directly and as the sum of its components. In particular, the literature emphasises the importance of accounting for commonality among components, so the focus is put on this feature. To produce the groupings, a two-stage approach is followed. First, the dimension of the problem is reduced by selecting a subset of possible groupings through the use of Agglomerative Hierarchical Clustering. The second step involves producing the definitive forecast either by choosing the appropriate grouping from the subset or combining all of the options in the subset.

The results from the empirical application support the thesis that grouping methods can improve overall accuracy. On the one hand, some of the methods that select a single grouping perform significantly better than the best performing non-grouping methods. On the other hand, the forecast combination choice methods perform well overall. The analysis of the performance of the grouping methods highlights the fact that the persistence of the performance of the aggregation levels for these datasets is low. In this context, the chosen sub-aggregations have a defining impact on forecasting accuracy. These results also suggest, although unintentionally, that institutionally imposed aggregations could be inefficient in terms of aggregate forecasting accuracy. Another conclusion that can be extracted from the results of the empirical exercise is that the biggest overall improvements are observed in the case where the bottom-up approach is less accurate than the direct approach. However, maybe because the exercise contemplates only moderate disaggregation for the bottom-up approaches, in five out of the six scenarios the bottom-up approach out-performs the direct method. This could suggest, in line with Espasa and Mayo-Burgos (2013) and Bermingham and D'Agostino (2014) who encourage using the maximum disaggregation possible in order to benefit from the disaggregate dynamics, that further gains could be obtained from using these methods with higher degrees of disaggregation.

In terms of further research, two directions seem natural. The first relates to extending the grouping method to incorporate information from more periods than just the one in question. Currently, the process approaches each period independently. This setting could be affected by sudden jumps in classification resulting from unusual shocks. A possible extension could be to implement smooth transitioning between hierarchical structures or cross-validation of the incidence of specific data. The second points at adding robustness to the choice of dissimilarity measures. This would include developing more sophisticated dissimilarity measures and choice methods on which to base the grouping methods and also explore using the combination of many features simultaneously, instead of having to choose a single one. The motivation for the latter stems from the good performance of the combination methods and the fact that the literature recommends trying many different parameters and comparing results (Hastie et al., 2009; James et al., 2013).

Appendix:

3.A Equal Weighting of Distinct Forecasts in a Multi-level Combination

In the context of combining forecasts from the different aggregation levels resulting from a hierarchical clustering process, combining the different aggregate forecasts with equal weights implicitly gives higher reliability weights to components that are fused later in the algorithm, if the component forecasts are generated independently. This can be visualized by looking at Figure 3.1. Horizontal lines are drawn on this dendrogram every time a fusion occurs to illustrate the twelve options for groupings and the corresponding dissimilarity thresholds. At the very top is the aggregate and at the very bottom all of the components in their own cluster. Let an aggregate forecast for a given grouping, Q, be denominated by the superscript of the number of fusions that have occurred in that grouping and for simplicity that aggregation weights are all equal to one. The full bottom-up forecast, that is the sum of the forecasts of the n individual components q_n, is then written as $Q^{(0)} = \sum_{n=1}^{N} q_n^{(0)}$. The aggregate forecasts for the grouping after the first fusion is $Q^{(1)}$, the one after the second fusion $Q^{(2)}$ and so on.

For all other aggregation levels, the multi-level combination procedure presented in chapter 2 can be used to obtain forecasts of components that are consistent with each level of aggregation. For the example shown in Figure 3.1, the forecast for the aggregation level that includes only the first fusion is produced from $Q^{(1)} = \sum_{n\neq\{2,5\}}^{N} q_n^{(1)} + Q_{\{2,5\}}$ where $Q_{\{2,5\}}$ is the forecast of the sum of components 2 and 5 and $\sum_{n\neq\{2,5\}} q_n^{(1)}$ is the sum of the forecasts of all the remaining components. To have a complete disaggregate scenario, components 2 and 5 need to be inferred from the forecast for their sum. Using the aforementioned method, $q_2^{(0)}$ and $q_5^{(0)}$ are reconciled with $Q_{\{2,5\}}$ obtaining $\tilde{q}_2^{(1)}$ and $\tilde{q}_5^{(1)}$. With this $Q^{(1)} = \sum_{n\neq\{2,5\}}^{N} q_n^{(1)} + \tilde{q}_2^{(1)} + \tilde{q}_5^{(1)}$. The same procedure can be used for every level, finishing with the direct aggregate forecast given by $y = Q^{(11)} = \sum_{n=1}^{N} \tilde{q}_n^{(11)}$.

The whole set of forecasts can be written as:

$$\mathbf{q} = \begin{bmatrix} q_1^{(0)} & q_2^{(0)} & q_3^{(0)} & q_4^{(0)} & q_5^{(0)} & q_6^{(0)} & \cdots & q_{12}^{(0)} \\ q_1^{(1)} & \tilde{q}_2^{(1)} & q_3^{(1)} & q_4^{(1)} & \tilde{q}_5^{(1)} & q_6^{(1)} & \cdots & q_{12}^{(1)} \\ q_1^{(2)} & \tilde{q}_2^{(2)} & q_3^{(2)} & \tilde{q}_4^{(2)} & \tilde{q}_5^{(2)} & q_6^{(2)} & \cdots & q_{12}^{(2)} \\ & & & & \vdots & & & \\ \tilde{q}_1^{(11)} & \tilde{q}_2^{(11)} & \tilde{q}_3^{(11)} & \tilde{q}_4^{(11)} & \tilde{q}_5^{(11)} & \tilde{q}_6^{(11)} & \cdots & \tilde{q}_{12}^{(11)} \end{bmatrix}$$

where $\sim$ denotes component forecasts that are the result of reconciling the individual forecasts, $q_n^{(0)}$, with a forecast for an intermediate aggregate.

The equal-weighted aggregate forecast combination can be written as:

$$\overline{Q} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} Q^{(0)} \\ \vdots \\ Q^{(11)} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} \mathbf{q} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{q}_1 & \cdots & \bar{q}_{12} \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

where $\bar{q}_n$ denotes the equal-weighted forecast combination for component n.

If the forecasts are produced using a univariate model, all those that do not involve reconciling will be the same, unless the model is changed deliberately for each level of aggregation. The same is true of fusions that are not fused again with other clusters at that aggregation level. In the example it is clear that for the first fusion, $q_n^{(0)} = q_n^{(1)}$ for all n except for 2 and 5 and this can be expanded to the whole hierarchy. Replacing the unaltered forecasts with the original ones produces a particular pattern, in that forecasts only change after fusions at that levels:

$$\mathbf{q}^{\text{Indep.}} = \begin{bmatrix} q_1^{(0)} & q_2^{(0)} & q_3^{(0)} & q_4^{(0)} & q_5^{(0)} & q_6^{(0)} & \cdots & q_{12}^{(0)} \\ q_1^{(0)} & \tilde{q}_2^{(1)} & q_3^{(0)} & q_4^{(0)} & \tilde{q}_5^{(1)} & q_6^{(0)} & \cdots & q_{12}^{(0)} \\ q_1^{(0)} & \tilde{q}_2^{(2)} & q_3^{(0)} & \tilde{q}_4^{(2)} & \tilde{q}_5^{(2)} & q_6^{(0)} & \cdots & q_{12}^{(0)} \\ & & & & \vdots & & & \\ \tilde{q}_1^{(11)} & \tilde{q}_2^{(11)} & \tilde{q}_3^{(11)} & \tilde{q}_4^{(11)} & \tilde{q}_5^{(11)} & \tilde{q}_6^{(11)} & \cdots & \tilde{q}_{12}^{(11)} \end{bmatrix} \tag{3.1}$$

From this, it is clear that if one were to average $Q^{(0)}$ and $Q^{(1)}$, it would implicitly give higher reliability weights to the forecasts from all the components other than 2 and 5, meaning that equal-weight aggregate combination in this case does not give equal weight to each distinct forecast. In fact, the later in the process that the component is fused, the greater the implicit additional reliability. This can be seen from what the definitive forecast would be for component twelve, the last single component to be fused in the example from Figure 3.1. This would be $\bar{q}_{12} = \frac{1}{12}(10q_{12}^{(0)} + \tilde{q}_{12}^{(10)} + \tilde{q}_{12}^{(11)})$.

However, given the structure of equation 3.1, it is quite easy to comply with giving each distinct forecast the same weight by simply removing the repeated forecasts from the component averages. In the case of component twelve, $\bar{q}_{12}^{*} = \frac{1}{3}(q_{12}^{(0)} + \tilde{q}_{12}^{(10)} + \tilde{q}_{12}^{(11)})$.

3.B Additional Information from Empirical Application

Table 3.6: Differentiation for Empirical Application

	France	Germany	United Kingdom
1. Food and non-alcoholic beverages	2	2	1
2. Alcoholic beverages, tobacco and narcotics	2	2	1
3. Clothing and footwear	1	1	1
4. Housing, water, electricity, gas and other fuels	1	2	2
5. Furnishings, household equipment and maintenance	2	2	1
6. Health	1	1	1
7. Transport	1	1	1
8. Communication	1	2	1
9. Recreation and culture	1	1	2
10. Education	2	1	2
11. Restaurants and hotels	2	1	2
12. Miscellaneous goods and services	2	2	1

Note: Number of times the series is differentiated to make it stationary according to the parametric unit root test in Gomez and Maravall (1996).

Table 3.7: Benchmark Aggregate Forecasting Performance Excluding Crisis

	Factor aug.	AR(1)		BVAR	
	AR(1)	Bottom-Up	Combination	Bottom-Up	Combination
France	0.88*	0.90	0.91**	0.98	0.94
Germany	0.96	0.95	0.96	0.97	0.96
United Kingdom	0.90	0.90	0.88*	1.25	0.96

Note: Mean square forecasting error relative to the direct method. Benchmark models are the bottom-up and equal-weight combination of the direct and bottom-up approaches of the AR(1) and BVAR and a factor augmented AR(1) following Bermingham and D'Agostino (2014). * and ** denote significance of the forecasting performance difference based on the modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015 excluding from 2008.III to 2009.II.

Table 3.8: Aggregate Forecasting Errors by Country and Model Excluding Crisis

	AR(1)						BVAR					
	In-		Diss.	Prob.			In-		Diss.	Prob.		
Choice method	samp.	O-o-S	Thres.	crit.	FC1	FC2	samp.	O-o-S	Thres.	crit.	FC1	FC2
France												
Pearson's corr.	0.92	0.97	0.94		**0.88****	0.90**	1.03	0.97	1.00		**0.91***	**0.93****
Spearman's corr.	**0.90***	0.93	0.90*		**0.88****	0.91**	1.10	1.11	1.09		0.95	0.95
latent factor	0.94	0.95	0.98		**0.90****	0.92**	1.05	1.08	1.04		0.96	0.95
persistence	**0.88***	**0.87****	**0.89****		**0.87****	**0.88****	1.03	0.97	**0.93**		**0.94**	**0.93***
f-error clustering	0.90*	0.91**	0.90*		**0.88****	**0.90****	1.06	1.06	1.02		0.97	0.96
Bayesian	**0.88****	**0.90****	0.91	0.98	**0.89****	0.91**	1.03	1.05	1.03	1.04	**0.96**	**0.97**
Germany												
Pearson's corr.	0.99	1.01	1.01		1.00	0.99	1.11	1.16	1.09		1.01	1.00
Spearman's corr.	0.99	1.00	1.02		1.00	0.98	1.14	1.15	1.06		1.02	0.99
latent factor	0.99	1.00	1.00		1.00	0.99	1.10	1.01	1.06		1.00	0.99
persistence	0.97	**0.95**	**0.88****		**0.92**	**0.93***	1.11	1.05	0.97*		0.96	**0.95**
f-error clustering	0.99	0.99	0.98		0.99	0.99	1.19	1.12	1.01		0.98	0.99
Bayesian	0.99	0.98	0.96	1.00	0.97	0.98	1.03	1.12	0.99	1.03	0.99	0.99
United Kingdom												
Pearson's corr.	**0.87**	**0.88**	0.91		**0.87**	**0.84***	**0.88**	**0.83**	**0.86**		**0.95**	**0.88**
Spearman's corr.	**0.87**	0.95	**0.88**		0.90	0.89	**0.96**	**0.87**	0.99		1.00	**0.92**
latent factor	**0.84**	0.93	**0.84**		**0.84***	**0.88**	0.98	1.00	**0.88**		**0.91**	**0.93**
persistence	0.97	0.98	1.04		0.98	0.93	**0.92**	1.01	0.99		1.07	1.02
f-error clustering	0.92	0.96	**0.87**		**0.88**	**0.85***	**0.91**	**0.92**	1.04		**0.95**	**0.92**
Bayesian	0.89	0.90	0.89	1.19	0.93	0.93	**0.89**	**0.95**	1.24	1.25	1.00	0.99

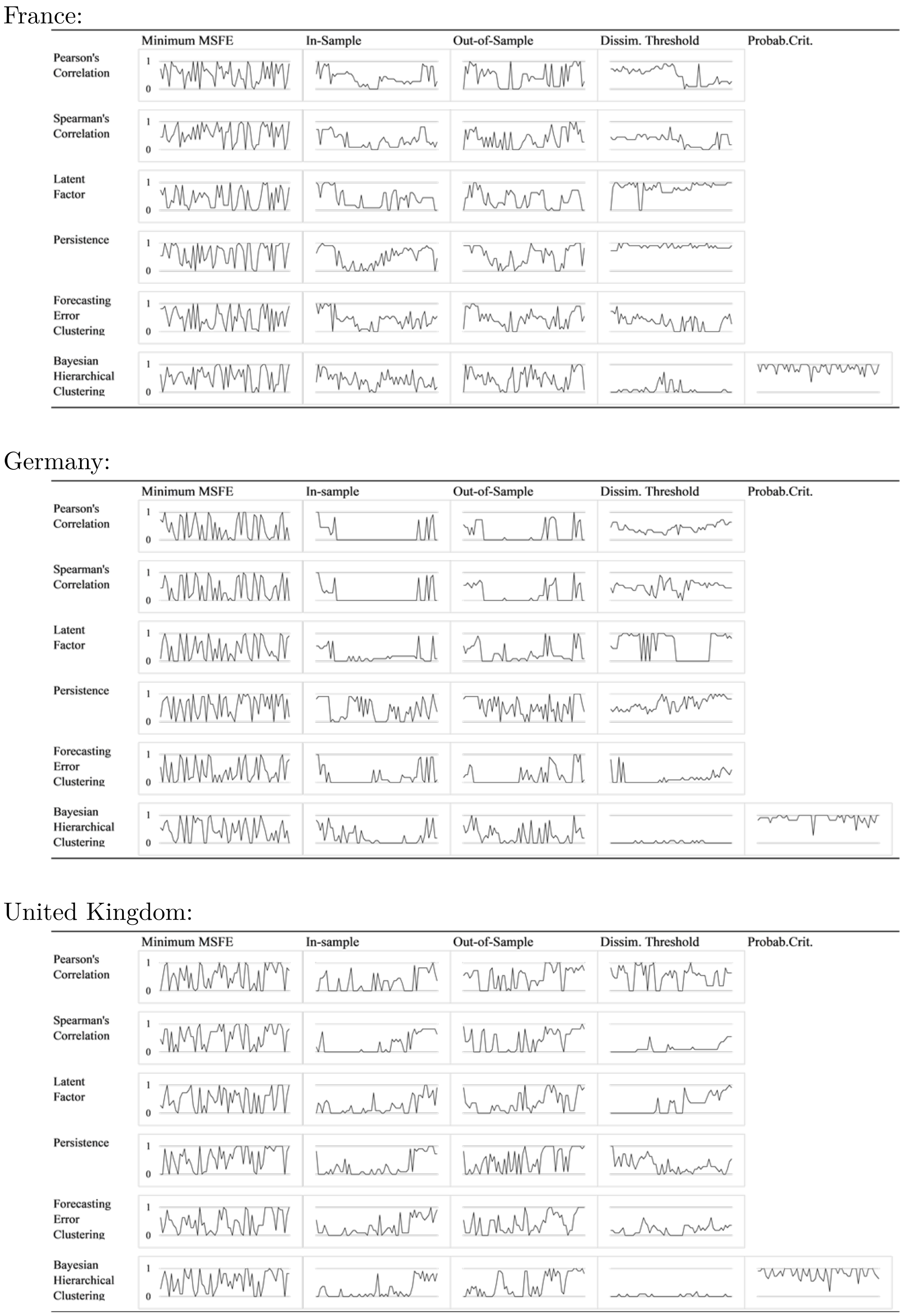
Note: Mean square forecasting error relative to the direct method. Grouping method dissimilarity measures: Pearson's correlation, Spearman's correlation, latent factor given by the variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1) and Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, out-of-sample criterion, dissimilarity threshold, probabilistic criterion, forecast combination that assigns equal weights to the aggregate forecasts (FC1) and forecast combination that assigns equal weights to each distinct forecast (FC2). In bold MSFE lower than the lowest of either the respective full bottom-up approach or the direct approach. * and ** denote significance of the forecasting performance difference based on the Modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over the 2001-2015 excluding from 2008.III to 2009.II.

Table 3.9: Overall Aggregate Forecasting Performance Excluding Crisis

	Average Percentage Deviation from Best Method						Average Rank Difference with Best Method					
	In-		Diss.	Prob.			In-		Diss.	Prob.		
Choice method	samp.	O-o-S	Thres.	crit.	FC1	FC2	samp.	O-o-S	Thres.	crit.	FC1	FC2
Pearson's corr.	9.3	9.9	9.8		**6.4**	**4.6**	14.5	18.0	18.0		**9.5**	**7.3**
Spearman's corr.	12.6	13.5	12.6		8.8	**6.7**	17.3	22.5	20.8		15.0	**10.2**
latent factor	11.7	13.1	9.6		**6.1**	6.9	18.0	23.2	17.3		10.5	12.3
persistence	11.4	10.3	8.4		8.9	7.1	15.2	14.7	10.3		10.2	**8.7**
f-error clustering	12.9	12.7	10.3		7.0	**6.3**	19.3	20.8	14.3		10.7	**10.0**
Bayesian	7.8	11.9	14.2	23.2	8.8	9.2	12.3	18.0	15.0	24.7	12.3	14.2

Note: Relative performance of the grouping methods as measured by the average deviation of the respective MSFE relative to that of the best performing grouping method by category and as the average difference in rank according to MSFE over the six sets of forecasts. Grouping method dissimilarity measures: Pearson's correlation, Spearman's correlation, latent factor given by the variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, out-of-sample criterion, dissimilarity threshold, probabilistic criterion, forecast combination that assigns equal weights to the aggregate forecasts (FC1) and forecast combination that assigns equal weights to each distinct forecast (FC2). In bold the five best performers in each category. Calculated over the 2001-2015 excluding from 2008.III to 2009.II.

Figure 3.4: Empirical Aggregation Level Choices for AR(1) Models



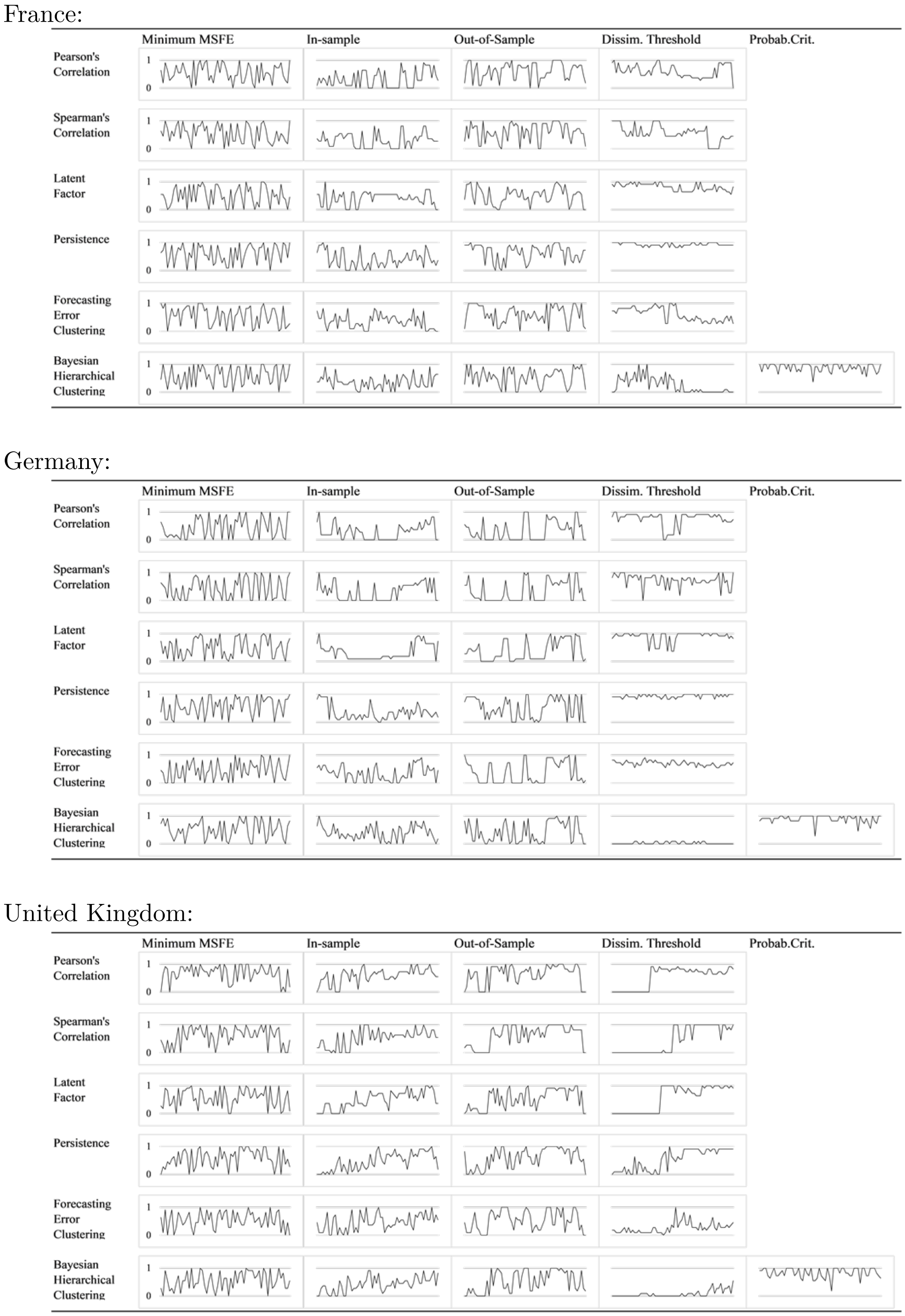
France:

Germany:

United Kingdom:

Note: Succession of aggregation level choices for the hierarchy resulting of the particular clustering process by disimilarity measure and choice method. The vertical axis measures the aggregation level. Path of minimum MSFE included for comparison. Calculated over the 2001-2015 sample.

Figure 3.5: Empirical Aggregation Level Choices for BVAR Models


France:

Germany:

United Kingdom:

Note: Succession of aggregation level choices for the hierarchy resulting of the particular clustering process by disimilarity measure and choice method. The vertical axis measures the aggregation level. Path of minimum MSFE included for comparison. Calculated over the 2001-2015 sample.

Chapter 4

Aggregate Density Forecasting from Disaggregate Components Using Large Bayesian VARs

4.1 Introduction

Economic aggregates play an important role in assessing the current state of the economy and a lot of effort is therefore put into understanding and forecasting them. As they are built from disaggregate information, there is an ongoing debate as to whether and how to incorporate disaggregate information in order to improve aggregate forecasts. For point forecasts there is sufficient evidence, in terms of aggregate accuracy, to support the benefits of including disaggregate information in the forecasting process (Brüggemann and Lütkepohl, 2013). Therefore, much attention has been paid over the years to determining whether forecasting the aggregate as the sum of its components' forecasts achieves better results than forecasting the aggregate directly. This may be due partly to the fact that policy-making institutions often require a consistent underlying scenario for their forecasts and the bottom-up approach provides them with one (Esteves, 2013; Ravazzolo and Vahey, 2014). In terms of how the bottom-up approach performs compared with other methods, Lütkepohl (1987) show that it depends on the disaggregate processes and the aggregation matrix of the particular problem. The differing results from many practical comparisons confirm that identifying the best method is an empirical matter.[1]

[1] Examples of these comparisons are Espasa et al. (2002), Benalal et al. (2004), Hubrich (2005) and Giannone et al. (2014) for inflation in the Euro area; Marcellino et al. (2003), Hahn and Skudelny (2008), Burriel (2012) and Esteves (2013) for European GDP growth; and Zellner and Tobias (2000), Perevalov and Maier (2010) and Drechsel and Scheufele (2013) for GDP growth in specific industrialized countries.

The amount of research for point forecasts contrasts with that for density forecasting. It would seem that making use of disaggregate components in a probabilistic setting has remained a relatively unexplored area. This is hard to explain, given that probability forecasting is being used increasingly in both finance and economics to assess the uncertainty surrounding forecasts (Mitchell and Hall, 2005). What is more, Bache et al. (2010) argue that using the components could allow a wide range of uncertainties to be explored and provide some useful insights into tail events, for example. Pre-crisis models are often criticized as being ill-equipped for this purpose (Del Negro et al., 2016).

Nevertheless, there are some exceptions to this relative scarcity. Proietti et al. (2017) produce forecasts for an aggregate monthly GDP measure, using density forecasts from both production and expenditure components. In their framework they forecast each component using a different dynamic factor model, use the original GDP weights to produce an aggregate forecast for each measurement approach, and then combine them to produce the definitive forecast. Mazur (2015), on the other hand, forecasts each component by means of a small Vector Autoregression (VAR) constructed following a variable selection procedure. In both cases, however, by treating each component independently they ignore important interactions between components. This could lead to poor calibration of the resulting aggregate density forecasts (Ravazzolo and Vahey, 2014). For this reason, Bache et al. (2010) and Ravazzolo and Vahey (2014) follow an alternative path in an attempt to benefit from the components. They use ensemble forecasting, a method adapted from the meteorology literature, where univariate autoregressive models are used for the components and aggregation weights are estimated in order to produce a well-calibrated aggregate forecast. In this method, however, the link between the aggregate and the components is broken, meaning that no consistent underlying scenario is available to provide support for the assessment if needed.

The reasoning behind forecasting each component independently includes the fact that univariate models are indisputably popular among practitioners (Bache et al., 2010; Ravazzolo and Vahey, 2014) and considerations regarding the computational burden traditionally associated with large specifications (Mazur, 2015). Over the last decade, however, the advances in forecasting methods have meant that alternatives capable of dealing effortlessly and efficiently with large multivariate processes have been made available. As pointed out by Carriero et al. (2015), much attention has concentrated on using Bayesian Vector Autoregressions (BVAR) with large datasets for forecasting, particularly since Banbura et al. (2010) developed their intuitive and computationally cheap implementation. It would seem, however, that the research that has developed from these advances has gone beyond answering questions that traditional VARs cannot handle and centred around forecasting economic aggregates in a data-rich environment.

Many empirical applications suggest that they are a worthy alternative to the popular Dynamic Factor Models (Gupta and Kabundi, 2011). On the other hand, the potential of BVARs as a means for producing bottom-up density forecasts seems to have been overlooked.

This chapter picks up on this point and explores whether BVARs can be used to produce well-calibrated bottom-up density forecasts. The novelty of this approach is that by modelling the whole multivariate process, the main misspecification concerns present in the current literature could be avoided. For this purpose, a framework is presented to estimate the aggregate density forecast based on the outcome of a large BVAR. It is worth mentioning that using BVARs is not the only way of accounting for the interaction between components. An alternative could be to follow an approach similar to that of Bermingham and D'Agostino (2014) in which they forecast each component using a factor-augmented autoregressive models (FA-AR) where the factor is extracted from all of the components. The appeal of alternative methods is left for further research.

As regards specifying the BVARs, different alternatives that relax the constraints of the univariate framework are explored. These include considering both fixed and time-varying parameter BVARs and allowing for stochastic volatility. The appeal of permitting these two features to be modelled is that there is ample evidence for the frequent instabilities of financial and economic time-series meaning that the more recent applied econometric literature has focused on accounting at least for changes in the estimated coefficients and variance (Rossi, 2013; González-Rivera and Sun, 2017). Considering these aspects seems particularly relevant for density forecasting given that it is the whole distribution that is being forecasted and not a single moment as is the case with point-forecasting. To make the framework easy to implement, an empirical method that is relatively simple and computationally cheap is chosen to estimate the BVARs. Then, subscribing to the view that different models may play a complementary role in forecasting an unknown process, the possibility of benefiting from both direct and bottom-up approaches through forecast combination is also explored. In doing so, different common weighting schemes are compared and the impact of evaluation window length on the estimated combination weights is examined. The results from the proposed framework and those from the subsequent combination with the direct approach are evaluated by performing an empirical application using GDP and CPI data from France, Germany and the United Kingdom.

The contributions of this research are three-fold. First, the model it presents produces bottom-up forecasts that, in terms of calibration, are similar to or better than those of the direct aggregate approach, in circumstances where a univariate bottom-up approach performs poorly. Second, the proposed framework provides a consistent underlying scenario for the aggregate forecasts, given that the original aggregation weights are used. This could be especially useful at policy-making institutions. Third, it investigates empirically how it can serve as a complement to the direct method to increase its overall performance. The evaluation with GDP and CPI data shows that the performance of the combination can result in a significant improvement over the direct method alone.

The rest of the chapter is organized as follows: Section 4.2 presents the methodology for estimating multivariate bottom-up density forecasts. This section also evaluates how they perform relative to the direct approach in an empirical implementation using GDP and CPI data for France, Germany and the United Kingdom. Section 4.3 presents the combination of the direct and bottom-up approaches and, using the same data, evaluates empirically its improvement over the individual methods. Section 4.4 summarizes the conclusions.

4.2 Disaggregate Density Forecasting Methodology

Over the last decade there has been a growing interest in Bayesian methods for policy analysis and forecasting. As pointed out in Carriero et al. (2015), much attention has concentrated on using BVARs with large datasets for point and density forecasting. The idea behind the BVAR is that prior information is imposed on the VAR coefficients to avoid overparametrization. Until relatively recently, most approaches were technically and computationally demanding. This was seen as a stumbling block for their adoption in contexts where the production of forecasts is subject to very tight time constraints (Koop, 2013; Carriero et al., 2015). In the last decade, however, alternatives that avoid the more intensive simulation have become available. The implementation suggested by Banbura et al. (2010) has received considerable attention since it was first presented. They suggest a relatively simple way of using Bayesian shrinkage to overcome the dimensionality problem in traditional VARs. They do, however, only contemplate using constant coefficients and assume homoskedastic errors. The forecasting literature has established that accounting for evolving conditions can be very important if forecasting methods are to be successful. Koop and Korobilis (2013) build on the success of Banbura et al. (2010) and take things a step further by developing a methodology that also allows implementing time-varying parameters and stochastic volatility without increasing computational demands. Due to this extra flexibility, and other convenient features of the implementation, it is the method chosen for producing the bottom-up density forecasts in this framework.

4.2.1 Large Time-varying Parameters VARs

4.2.1.1 The Model

Koop and Korobilis (2013), hereafter K&K, formulate the problem of estimating the BVAR in state-space form:

$$\begin{aligned} y_t &= X_t\beta_t + \varepsilon_t \qquad \varepsilon_t \sim \text{i.i.d.}N(0, \Sigma_t) \\ \beta_{t+1} &= \beta_t + u_t \qquad u_t \sim \text{i.i.d.}N(0, Q_t) \end{aligned} \tag{4.1}$$

where ε_t and u_s are independent of one another for all s and t. y_t for $t = 1, ..., T$ is an $M \times 1$ vector containing observations on M time series and X_t is an $M \times k$ matrix defined so that each equation contains an intercept and p lags of each of the M variables.

To estimate the model, instead of proceeding in a standard Bayesian manner by using Markov Chain Monte Carlo methods, K&K suggest replacing Q_t and Σ_t with estimates. They proceed in this manner because they argue that even for relatively small problems the computational burden could be quite significant. To overcome this problem, while still retaining time-varying parameters and stochastic volatility, they use forgetting factors to produce their approximation at each point in time. This means estimating empirically the desired parameters, but in a way that downplays, to a chosen extent, the contribution of less recent data.

With regard to the time-varying parameters, they start by noting that Q_t only appears in one place in the Kalman filtering process, specifically in the prediction step. Then, following the forgetting factors approach, they replace Q_t for $(\lambda^{-1} - 1)V_{t-1|t-1}$, where $V_{t-1|t-1}$ is the variance of $\beta_{t-1}|y_{t-1}$, resulting in $V_{t|t-1} = \frac{1}{\lambda}V_{t-1|t-1}$. The forgetting factor λ is restricted to being strictly positive and less than one, being the constant coefficient specification achievable by setting $\lambda = 1$. Similarly, to avoid using a posterior simulation algorithm to model volatility, they use an Exponentially Weighted Moving Average estimator for the measurement error covariance matrix. This is done by making $\hat{\Sigma}_t = \kappa\hat{\Sigma}_{t-1} + (1-\kappa)\hat{\varepsilon}_t\hat{\varepsilon}_t'$ with $\hat{\varepsilon}_t = y_t - X_t\beta_{t|t-1}$. Here the forgetting factor κ is also restricted to being between zero and one.

With this, following the exposition in Koop and Korobilis (2012), the algorithm proceeds by estimating the problem recursively for periods t=1,...,T starting from the initial conditions given by $\beta_0 \sim N(b_0, V_0)$ and Σ_0. The prediction step sets:

$$\beta_{t|t-1} = \beta_{t-1|t-1} \quad \text{and} \quad V_{t|t-1} = \tfrac{1}{\lambda}V_{t-1|t-1}$$

Then the updating step estimates $\hat{\varepsilon}_t = y_t - X_t\beta_{t|t-1}$ and $\hat{\Sigma}_t = \kappa\hat{\Sigma}_{t-1} + (1-\kappa)\hat{\varepsilon}_t\hat{\varepsilon}_t'$. With this,

$$\beta_{t|t} = \beta_{t|t-1} + V_{t|t-1}X_t'\left(\hat{\Sigma}_t + X_tV_{t|t-1}X_t'\right)^{-1}\hat{\varepsilon}_t$$

and

$$V_{t|t} = V_{t|t-1} - V_{t|t-1}X_t'\left(\hat{\Sigma}_t + X_tV_{t|t-1}X_t'\right)^{-1}X_tV_{t|t-1}$$

With regard to the estimation of the coefficients of the BVAR, that is the β's, they use a Normal prior. Given their choice of variable transformation, for β_0 they set the prior mean to zero and the covariance matrix to be diagonal. Specifically, for $\text{var}(\beta_0) = \underline{V}$, with $\underline{V}_i$ being its diagonal elements, they define $\underline{V}_i = \gamma/r^2$ for coefficients on the r-th lag and for the intercepts use a noniformative prior. This results in having a single hyperparameter γ to control the shrinkage of the coefficients. In this case $0 \leq \gamma < \infty$.

4.2.1.2 Parameter Selection

The model proposed by K&K is relatively simple and capable of incorporating many features, despite being governed by only three parameters. These parameters, however, have to be provided by the researcher.

As to the values governing the time-varying parameters and stochastic volatility, they provide values that would be consistent with previous literature in those areas. They do acknowledge, however, that a method that determines them from the data would be very appealing, and they therefore go on to develop one based on dynamic model selection (DMS).

They set the problem up as one of selecting one model definition from a set of models that are the same in terms of explanatory variables, but differ in terms of parameter values.[2] Their criterion is to choose the specification with the highest probability of being the appropriate one for forecasting at any given time. They estimate this probability by implementing a recursive algorithm developed by Raftery et al. (2010) that can be run conveniently within the normal Kalman filtering process used to produce the forecasts.[3]

In this context, the prediction step is extended slightly with the additional equation:

$$\pi_{t|t-1,j} = \frac{\pi^{\alpha}_{t-1|t-1,j}}{\sum_{l=1}^{J}\pi^{\alpha}_{t-1|t-1,l}} \tag{4.2}$$

[2] K&K go on to extend the approach to also allow for differing explanatory variables.

[3] The algorithm by Raftery et al. (2010) is explained in detail in Section 2.3 of Koop and Korobilis (2013).

where $\pi_{t|t-1,j}$ is the probability that model j should be used to forecast at time t given the information up to $t-1$, α is a forgetting factor and J is the number of specifications being considered. The updating step is extended by adding:

$$\pi_{t|t,j} = \frac{\pi_{t|t-1,j} p_j \left(y_t \mid y^{t-1}\right)}{\sum_{l=1}^{J} \pi_{t-1|t-1,l} p_l \left(y_t \mid y^{t-1}\right)} \tag{4.3}$$

where $p_j \left(y_t \mid y^{t-1}\right)$ is the predictive likelihood.

The idea behind the algorithm is that good performance in the recent past increases the probability of the model being the appropriate one to forecast for the following period. The predictive likelihood serves as the measure of forecast performance and the forgetting factor α to define what is understood as recent past. In this case, an α close to zero leads approximately to the equal weighting for all time periods, while setting $\alpha = 1$ corresponds to using the marginal likelihood. The procedure is sufficiently general, so K&K also use it as the method to estimate the prior hyperparameter which controls coefficient shrinkage for the VAR equations.

4.2.2 Multivariate Bottom-up Aggregate Density Estimation

As pointed out by Ravazzolo and Vahey (2014) practitioners often rely on univariate models because of the difficulties involved in modelling the dependencies between components. Ignoring these dependencies, however, means that using a traditional bottom-up approach could yield poor aggregate density forecasts. If the multivariate process is modelled well, on the other hand, using the index weights would be appropriate and should produce well-calibrated aggregate forecasts. Determining the distribution of a sum of random variables is generally quite complicated, but in this case the task is simplified greatly by the fact that the densities for the components are produced using a sampling algorithm. As any given draw describes the whole multivariate process, the aggregate forecast for that draw can be produced simply by summing the components' forecasts using the appropriate index weights. Doing this for all draws provides the aggregate bottom-up density forecast.

As pointed out by K&K, the one-step-ahead predictive density depends only on quantities that are known at time t and is given directly by:[4]

$$p\left(y_{t+1}|y_t\right) \sim N(X_{t+1}\beta_{t+1|t}, \hat{\Sigma}_{t+1} + X_{t+1}V_{t+1|t}X'_{t+1})$$

Then, following Proietti et al. (2017), let $y_{i,t+1}^{(r)}$ denote the rth draw of the one-step-ahead forecast of the ith component, $i = 1, ..., N$, at time t. Then, the rth draw from

[4] For forecasts beyond one period, predictive simulation is required.

the bottom-up aggregate forecast is given by:

$$Q_{I,t+1}^{(r)} = \sum_{i=1}^{N} w_{i,t+1} y_{i,t+1}^{(r)} \tag{4.4}$$

where $w_{i,t+1}$ is the aggregation weight for component i at time $t+1$.

4.2.3 Density Forecast Evaluation

A popular way of assessing the calibration of the density forecasts is testing the sequence of probability integral transform (PIT) values. These are defined as $p_t = F_t(x_t)$, where F_t is the predictive cumulative distribution function and x_t is the observed realization. If F_t coincides with the true data-generating process, the PITs are uniform $U(0,1)$ for any forecast horizon and i.i.d. for one-step-ahead forecasts (Diebold et al., 1998). Geweke and Amisano (2010) describe this approach as comparing the distribution of the observed data with the distribution that would have resulted if the model under consideration had been used to generate the data. Mitchell and Hall (2005) point out, however, that testing in this context is not straightforward, given that the impact of dependence on uniformity tests and vice versa is unknown. The empirical literature has relied therefore on using a number of tests simultaneously. Following Mitchell and Wallis (2011) and Ravazzolo and Vahey (2014) this application uses Pearson's chi-squared test to assess the goodness-of-fit of the PIT histogram to a univariate distribution and the Anderson-Darling test to evaluate the uniformity of the empirical cumulative distribution function of the PITs. Independence is tested directly using a Ljung-Box test using autocorrelation of up to four lags. Finally, the test proposed by Berkowitz (2001) is used to test goodness-of-fit and independence.[5]

A problem with testing the calibration alone is that it is a dichotomous assessment. That is, a density is either well-calibrated or not. In this context, it would not be uncommon to find that two or more competing forecasts are equally well-calibrated, in which case practitioners that are looking to pick a single model are only slightly better off (Gneiting et al., 2007). An alternative approach is to use scoring rules. These assign a numerical score based on the predictive likelihood and the realization of the variable. Scores are affected by both the location and dispersion of a density (Bjørnland et al., 2011). They will suffer both for over-dispersed distributions, because probability mass is too high for infrequent values, and for under-dispersed distributions, because too many observations carry a low probability. Therefore, based on the difference in their scores, models can effectively be compared. Following Carriero et al. (2015), the average logarithmic predictive density score is used to assess the relative performance

[5] The test is in fact applied on the inverse normal transform of the PITs to test for normality. We use the three degrees of freedom version that tests against a first-order autoregressive alternative and wrong mean and variance.

of each forecast.

To evaluate whether the results of the alternative approaches are significantly different, the test of equal predictive accuracy based on the Kullback–Leibler information criterion (KLIC) proposed by Mitchell and Hall (2005) is used. The test is based on the distance between a density forecast and the true density as measured by the KLIC. For a model i the distance at time t is defined as:

$$\text{KLIC}_{i,t} = E\left[\ln g_t\left(x_t\right) - \ln f_{i,t}(x_t)\right]$$

where $g_t\left(x_t\right)$ is the true unknown density. The smaller this distance, the closer the density forecast is to the true density. Mitchell and Hall (2005) show that a test for equal-density forecast accuracy of two competing density forecasts can be constructed, based on the difference between their estimated KLIC distance measures. The null hypothesis of equal accuracy being:

$$H_0: \text{ KLIC}_i - \text{KLIC}_j = 0 \Rightarrow E\left[\ln f_{j,t}(x_t) - \ln f_{i,t}(x_t)\right] = 0$$

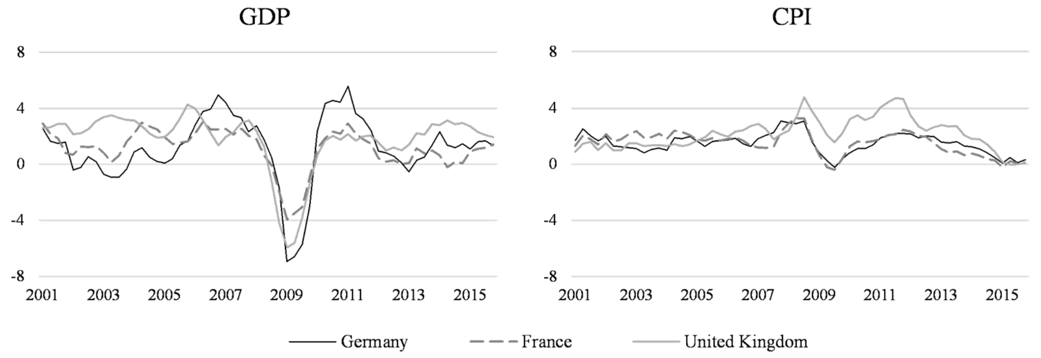
Mitchell and Hall (2005) show that a Diebold-Mariano-type test can be constructed using the sample mean of the distance measure, given by $\bar{d} = \frac{1}{T}\sum_{t=1}^{T}[\ln f_{j,t}(x_t)$ $- \ln f_{i,t}(x_t)]$, where $\bar{d}/\sqrt{S_d/T} \xrightarrow{d} N\left(0,1\right)$. This test is equivalent to the unweighted likelihood ratio test of Amisano and Giacomini (2007).

For the forecasting exercise, aggregate univariate AR models and bottom-up forecasts using univariate AR models for the components are used as benchmarks against which to compare the densities from the multivariate methods. One to four lags are considered for both methods and the best performer according to the logarithmic score is chosen.

4.2.4 Empirical Application

The success of the proposed method depends on two factors. The first is whether it performs well in circumstances where the univariate bottom-up approach fails to produce a well-calibrated aggregate forecast. The second, which may be more relevant in practical settings, is how it performs relative to other methods. The extent to which this can be measured depends fundamentally on the properties of the data used. For this reason, to have a broader assessment, more than one dataset is considered. Specifically, an out-of-sample forecasting exercise using GDP and CPI data from Germany, France and the United Kingdom is performed. The evaluation exercise data spans from 1991 to 2015 with the first ten years being used to estimate the models and from 2001 onwards for the forecasting evaluation. As the evaluation sample includes the most

Figure 4.1: Annual Growth Rate of the Quarterly Series



Source: OECD statistics database.

recent global financial crisis, one could expect the overall results of the forecasting models to be influenced by this episode. To have a notion of how this event affects the series that are considered in the evaluation exercise, Figure 4.1 presents the annual growth-rate of the quarterly series for both GDP and CPI on the same scale. As can be seen, the effect of the crisis on GDP is considerably larger than on CPI. In the light of this, it becomes apparent that their time-series characteristics over the sample are fundamentally different and very different results are therefore to be expected. Finally, the application is set up in a quarterly rolling scheme using a ten-year window where in each period the models are re-estimated and the density forecasts are generated.[6]

Regarding the forecast horizon, the scope of the exercise is restricted to one-step-ahead. The reason for this is that, in the context of this exercise, as the series considered are produced using either a fixed-base or annual overlap chain-linking method, the definitive weights for the one-step-ahead forecast are always available at the time of forecasting. For longer horizons, however, they are not. This means that for longer horizons forecasts for the weights would also be necessary. In this context, the simple method for estimating the bottom-up aggregate forecast, presented in section 4.2.2, would not be appropriate. One option would be to use the previous period's weights, as practitioners often do (Ravazzolo and Vahey, 2014), but Lütkepohl (2011) and Hendry and Hubrich (2011), among others, discuss the problems that arise from assuming weights to be unchanging and emphasize that, if the actual weights change through time, forecast performance can deteriorate quickly, with longer horizons being affected the most.

[6] It is worth noting, that the exercise does not mimic real-time forecasting, given that data revisions are not accounted for.

Table 4.1: Components Breakdown for Empirical Application

GDP		
	1. Agriculture, forestry and fishing	7. Financial and insurance activities
	2. Manufacturing	8. Real estate activities
	3. Industry and energy, excluding manufacturing	9. Professional, administrative and support service activities
	4. Construction	10. Public adm., defence, social security, education and health
	5. Trade, transport, accommodation and food services	11. Other service activities
	6. Information and communication	12. Taxes less subsidies
CPI		
	1. Food and non-alcoholic beverages	6. Health
	2. Alcoholic beverages, tobacco and narcotics	7. Transport
	3. Clothing and footwear	8. Communication
	4. Housing, water, electricity, gas and other fuels	9. Recreation and culture 10. Education
	5. Furnishings, household equipment and maintenance	11. Restaurants and hotels 12. Miscellaneous goods and services

4.2.4.1 Data

The exercise uses data for GDP from the production approach and CPI for Germany, France and the United Kingdom. The data is quarterly and seasonally adjusted, spanning from 1991 to 2015 and available from the OECD statistics database.[7] The breakdown of the aggregates is the one presented in Table 4.1.

4.2.4.2 BVAR Specifications

In terms of the specification for the BVARs, four combinations for λ and κ are used.[8] First, a homoskedastic VAR that is obtained by setting both λ and κ equal to one and in which case $\hat{\Sigma}_t$ is estimated by $\frac{1}{1-t}\sum_{i=1}^{t-1}\hat{\varepsilon}_i\hat{\varepsilon}_i'$. Second, a homoskedastic time-varying parameter VAR (TVP-VAR) with $\lambda = 0.99$, a value that K&K argue is equivalent to what has previously been used in the relevant literature and for which, in the case of quarterly data, observations five years back receive approximately 80% as much weight as last period's observation.[9] They argue that such a value leads to a gradual change in coefficients and stable models. Based on this, the third model is a heteroskedasic VAR with $\kappa = 0.99$. Finally, to allow for both features, the fourth model is a heteroskedasic

[7] For the United Kingdom the production data on the OECD database starts in 1995. The first four years of the sample are obtained by splicing backwards the historical reference tables available from the Office for National Statistics. No inconsistencies arise from the seasonal adjustment given that the aggregates are adjusted indirectly; that is, as the sum of the seasonally-adjusted components.

[8] Four lags are used for all models.

[9] The closer to zero the less consideration is given to older information.

Table 4.2: Tests on PITs for One-step-ahead Forecasts

		Germany				France				United Kingdom			
	Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
GDP													
	Bottom-Up AR	0.02	0.00	**0.21**	0.01	0.00	**0.28**	**0.37**	**0.16**	0.00	0.00	**0.14**	**0.16**
	Direct AR	**0.41**	0.00	0.01	**0.14**	**0.06**	**0.05**	**0.40**	**0.08**	**0.06**	**0.15**	**0.26**	0.02
	Homsk. VAR	**0.46**	0.02	**0.13**	**0.91**	**0.12**	**0.57**	**0.72**	**0.59**	0.03	**0.11**	**0.33**	**0.35**
	Homsk. TVP	**0.51**	0.01	**0.14**	**0.89**	**0.12**	**0.38**	**0.67**	**0.59**	0.03	**0.06**	**0.92**	**0.45**
	Hetsk. VAR	0.00	0.00	0.00	**0.77**	**0.40**	0.02	**0.31**	**0.86**	0.00	0.00	**0.34**	**0.21**
	Hetsk. TVP	0.00	0.00	0.00	**0.52**	**0.30**	0.03	0.03	**0.68**	0.00	0.00	**0.33**	**0.21**
	DMS	**0.72**	**0.08**	**0.30**	**0.72**	**0.17**	**0.54**	**0.89**	**0.70**	0.01	0.02	**0.11**	**0.29**
CPI													
	Bottom-Up AR	0.00	**0.08**	0.04	**0.10**	0.00	**0.06**	**0.07**	**0.52**	0.00	**0.93**	**0.79**	0.03
	Direct AR	**0.22**	**0.41**	**0.50**	**0.71**	**0.09**	**0.39**	**0.59**	**0.23**	**0.85**	**0.66**	**0.12**	**0.35**
	Homsk. VAR	**0.81**	0.01	**0.07**	**0.62**	**0.44**	**0.10**	**0.09**	**0.69**	**0.68**	**0.66**	**0.40**	**0.69**
	Homsk. TVP	**0.91**	**0.36**	**0.69**	**0.55**	**0.68**	**0.06**	0.02	**0.84**	**0.49**	**0.66**	**0.55**	**0.42**
	Hetsk. VAR	**0.42**	**0.11**	0.04	**0.47**	0.00	**0.18**	**0.41**	**0.08**	**0.09**	**0.71**	**0.43**	**0.12**
	Hetsk. TVP	**0.84**	**0.06**	**0.26**	**0.56**	0.00	**0.16**	**0.44**	**0.58**	0.03	**0.85**	**0.07**	**0.15**
	DMS	**0.67**	**0.59**	**0.86**	**0.65**	**0.66**	**0.05**	0.04	**0.64**	**0.43**	**0.93**	**0.98**	**0.13**

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecasts for each model for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). P-values in bold signify that the null of the respective test are not rejected at 5%. Calculated over the 2001-2015 period.

TVP-VAR with both λ and κ equal to 0.99.[10] With regard to setting the value for the overall shrinkage of the coefficients, the parameter selection algorithm described in section 4.2.1.2 is used over a wide grid for all specifications.[11]

K&K argue that the TVP-VARs are well-suited for modelling gradual evolution of coefficients. To accommodate more sudden changes, they advocate using DMS over a whole array of model specifications. Given that the sample includes the years of the financial crisis, allowing for abrupt changes in parameters could be particularly relevant. Then, as the final approach, a procedure is implemented that consists of choosing at each point in time the forecast from the model that has the highest probability of being appropriate, according to the aforementioned algorithm.[12]

4.2.4.3 Results

Table 4.2 presents the p-values for the tests on the PITs for the one-step-ahead forecasting exercise for GDP and CPI for all three countries. As Ravazzolo and Vahey (2014)

[10] K&K also choose λ and κ empirically over a grid. In this particular exercise the results are not significantly different from those obtained from setting both parameters to 0.99.

[11] Specifically, $\gamma = e^i$ selecting i from {-7, -6, ... , -1}.

[12] That is, the model with the highest value for $\pi_{t|t-1,j}$ from equation (4.2).

Table 4.3: Log Scores for One-step-ahead GDP Forecasts

	GDP			CPI		
Model	Germany	France	United Kingdom	Germany	France	United Kingdom
Direct AR	7.4	22.9	7.0	4.7	0.2	-0.4
Homsk. VAR	**12.0**	**26.1**	**10.1**	2.8	**2.9**	**1.8**
Homsk. TVP	**15.7**	**25.8**	**10.1**	3.0	**4.5**	**1.5**
Hetsk. VAR	2.4	**23.0**	5.2	**4.8**	**4.0**	0.0
Hetsk. TVP	4.3	**24.1**	5.3	**5.2**	**1.3**	-0.1
DMS	**19.3**	**25.9**	**8.3**	4.6	**4.8**	**0.6**

Note: Log predictive density scores of the one-step-ahead forecasts for each model for the three countries expressed in terms of the percentage improvement over the bottom-up univariate model (Bottom-up AR). The models are the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Log scores in bold denote improvement over the direct univariate model. * and ** denote significance of the forecasting performance difference with respect to the direct univariate model based on the KLIC test at a 10 and 5% significance levels. Calculated over the 2001-2015 period.

put it, well-calibrated forecasts should give high probability values for all four tests. The overall impression from the results for GDP, however, is that few specifications pass all four diagnostic tests.[13] For Germany, for example, only the DMS model appears to be well-calibrated; while none does for the United Kingdom. This is not surprising, however, given that the evaluation sample includes the financial crisis. The comparable performance of the Direct AR suggests that there is more to poor calibration than merely a generalized shortcoming in the bottom-up approach. On the contrary, for CPI, the forecasts from most models appear to be well-calibrated according to the tests. The univariate bottom-up model, however, fails at least one test in each case. Where multivariate models are concerned, those that include stochastic volatility are similarly well-calibrated to those that do not. The results for all datasets tend to support the claims of Ravazzolo and Vahey (2014) regarding the likelihood of obtaining poorly-calibrated forecasts from using a univariate bottom-up approach.

As mentioned before, the outcome of the PITs tests is binary and it is therefore inadequate as a way of ranking different forecasting methods. The average logarithmic predictive density scores (log scores) are used for this purpose. Table 4.3 presents them expressed in terms of the percentage improvement over the bottom-up univariate model. As one might expect after looking at the PITs, the multivariate models perform better than the univariate bottom-up approach, but the improvements are heterogeneous. In the case of GDP, overall, it is the homoskedastic models that show the best performance, improving over the aggregate univariate model by as much as eight percentage points. From the performance of the four different BVARs, it would seem that most of the

[13] That is that the null hypothesis of no calibration failure cannot be rejected at the 5% significance level. The tests are conducted on an individual basis which imply a Bonferroni-corrected (joint) p-value of 1.25%.

gains come from allowing the process to be modelled using a multivariate model and that further gains can be obtained by allowing for the coefficients to vary over time. In contrast, incorporating stochastic volatility has a negative effect. This is consistent with the results from DMS. For Germany it performs very well, showing the highest accuracy with an improvement of nearly 12% over the aggregate AR. It is also the only model for which uniformity and independence are not rejected by any of the tests. For France it performs virtually the same as the homoskedastic multivariate models both in terms of calibration according to the PITs tests and log score. For the United Kingdom it performs better than the Direct AR but worse than the homoskedastic multivariate models. In the case of CPI, on the other hand, the improvements of the multivariate models are smaller than for GDP, and heterogeneous between countries. For example, for Germany, the methods improve over the univariate bottom-up approach but are only marginally better than the direct AR if stochastic volatility is included. For the other two countries, there is little difference in accuracy between the univariate methods, but in the case of France the multivariate methods improve by as much as 5% while for the United Kingdom these are below 2%.

The improvements over the bottom-up univariate model in some cases seem quite substantial, up to 26% in the case of French GDP, so an obvious question to ask is whether the differences in predictive accuracy are significant or not. To assess whether they are, the KLIC test is used. As mentioned before, it compares two loss differential series in a way that is analogous to the point forecast accuracy test popularized by Diebold and Mariano (1995). Although the improvements seem quite large in magnitude, the differences are not significant according to the test. This would seem odd at first, but the recursive log scores that are presented in Figure 4.2 could provide some insight into why this is the case.[14]

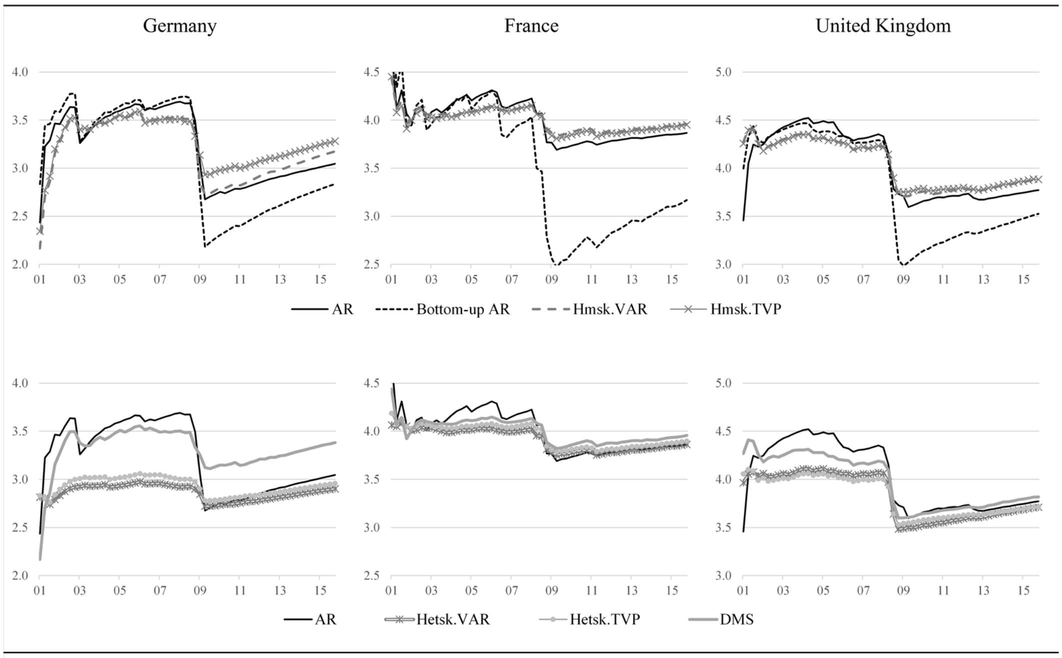
For GDP, it is immediately obvious that the crisis produces a sharp decline in the scores. Common to all three countries is that the bottom-up univariate model is by far the most affected of all models. The second most affected model, however, is the aggregate AR. Although the univariate models are among the best performers until the crisis, the multivariate models show falls that are proportionally smaller and therefore end up being better over the whole sample, at least in some cases.[15] The performance of both homoskedastic VARs is slightly worse than that of the univariate methods up to the crisis, but the comparatively smaller impact of the crisis suggests that the increased uncertainty due to the estimation of additional parameters could be worthwhile. The opposite seems to be the case with the methods that incorporate stochastic volatility. For CPI, on the other hand, it is still the case that the multivariate

[14] For each series, the homoskedastic models are presented in the top panel and the heteroskedastic models and DMS in the bottom. The aggregate AR is included in both to serve as a reference point.

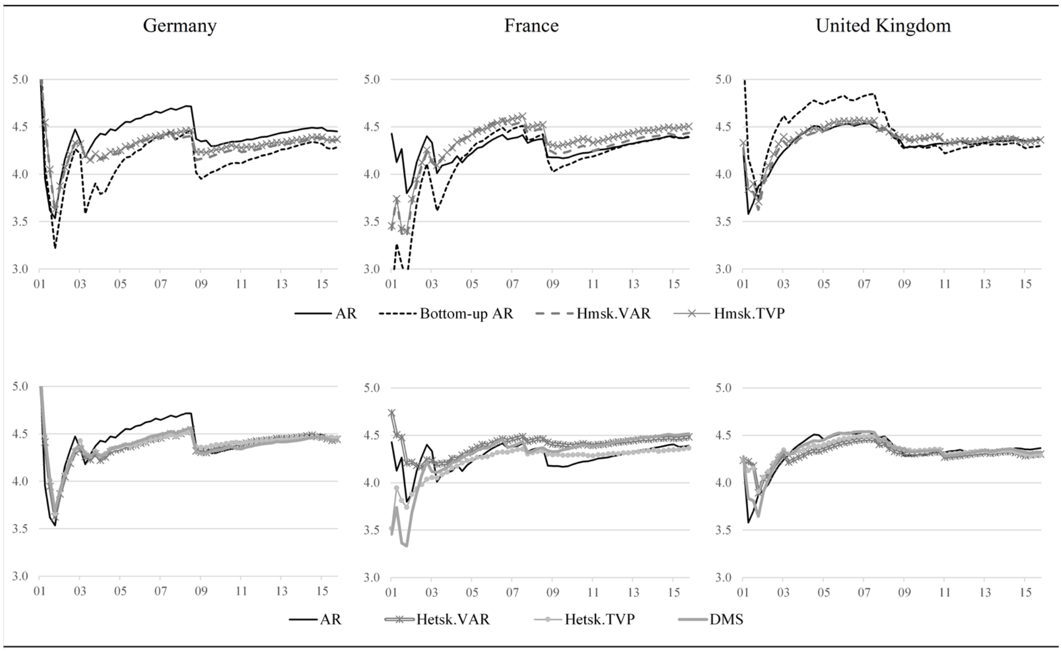
[15] The recursive log scores calculated excluding the crisis years, not reported, show that the univariate models perform very well over the restricted sample.

Figure 4.2: Recursive Log Scores for One-step-ahead Forecasts

GDP:



CPI:



Note: Recursive log scores calculated over the 2001-2015 period. The models are the aggregate univariate model (Direct AR), the bottom-up forecast using univariate AR models for the components (Bottom-Up AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS).

models are proportionally less affected than the univariate models, but the overall impact of the crisis is smaller. This translates into there being no significant difference between models in terms of performance. Unlike the case of GDP, here the added complexities do not seem to pay off.

As a summary of the results, it can be said that, although the performance of the different methods varies quite significantly depending on the dataset, there are a number of things that can be learned from them. The first is that, in line with the concerns raised by Ravazzolo and Vahey (2014), the univariate bottom-up approach would be likely to produce poorly calibrated forecasts in certain circumstances. In the same setting the proposed multivariate bottom-up methods perform similarly to the direct approach or better. The varying degrees of success of the different specifications, however, also suggest that the added complexities may not always be justified in terms of performance. It would seem that most of the improvements are attainable in the homoskedastic fixed-parameter setting. This comes as good news for practitioners, as it suggests that the more extended implementation by Banbura et al. (2010) would probably also work well in the same setting.

4.3 Combination of Direct and Bottom-up Density Forecasts

The differences between the results for GDP and CPI in the empirical application in the previous section suggest that the strengths of the multivariate methods only emerge if the interactions among variables are sufficiently strong. The more significant effects of the financial crisis on GDP, both in magnitude and persistence, would seem to be the reason behind the multivariate methods beating the univariate alternatives. The KLIC tests and the evaluation excluding the crisis years support this view. These results reinforce the idea that the model of choice will differ, depending on the situation.

In this context, where competing forecasts are available for the same variable, there is a growing literature to support combining these forecasts instead of choosing between them. For point forecasts, the idea has been around for quite some time (Bates and Granger, 1969), and many surveys like Clemen (1989), Diebold and Lopez (1996), Newbold and Harvey (2002) and Timmermann (2006) highlight the robustness of the gains in forecasting accuracy due to its use. More recently, the attention has shifted to density forecasts. The appeal of combined density forecasts over individual ones has been argued by Mitchell and Hall (2005), Ranjan and Gneiting (2010) and Mitchell and Wallis (2011) among others.

In section 4.2, the direct and bottom-up approaches are presented as competing methods to forecast an aggregate. In contrast, the aim here is to evaluate whether combining them may result in an improvement over the individual approaches. To this effect, the densities resulting from combining the forecasts from the direct and each of the bottom-up approaches of the previous section, plus pooling them all, are compared against those of the direct method.

4.3.1 Empirical Framework

4.3.1.1 Density Combination Methods

A very attractive feature of forecast combination for point forecasts is that simple combination schemes are surprisingly effective (Timmermann, 2006). In fact, the equal-weighted forecast combination performs so well that researchers have tried to explain why this should be so (Smith and Wallis, 2009; Elliott, 2017). It is not surprising, that simple weighting schemes are also popular for probabilistic forecasting. As pointed out by Ranjan and Gneiting (2010) the weighted linear combination, which is often referred to as a linear opinion pool, is probably the most popular method, and substantial empirical evidence attests to its benefits. Specifically, the combined aggregate density forecast from the linear opinion pool is given by:

$$p(Q_C)_{t+1} = \sum_{j=1}^{J} \omega_{j,t+1} p(Q_j)_{t+1} \tag{4.5}$$

where $\omega_{j,t+1}$ and $p(Q_j)_{t+1}$ are the combination weight and aggregate density forecast for approach j for time $t+1$.

There is evidence, however, that it has at least one important drawback. The problem stems from the fact that density forecasts have to be as sharp as possible, subject to being well calibrated. In some cases, however, a linear combination of well-calibrated individual predictive distributions necessarily produces poorly calibrated forecasts (Hora, 2004; Ranjan and Gneiting, 2010).[16]

An alternative to the linear opinion pool that has been proposed in the literature is the logarithmic opinion pool. This is a geometric weighted average of the individual densities given by:

$$p(Q_C)_{t+1} = K \prod_{j=1}^{J} p(Q_j)_{t+1}^{\omega_{j,t+1}} \tag{4.6}$$

where K is a constant that ensures the density adds up to one. As pointed out by

[16] Calibration measures how close conditional event frequencies are to the forecast probabilities. Sharpness describes how far away the forecasts are from the marginal event frequency (Gneiting et al., 2007).

Bjørnland et al. (2011), there are a number of differences between the linear and logarithmic opinion pools, but probably the most defining one is the fact that the latter will result in a zero combined density for any region in which a single density has zero density. Therefore, the resulting densities of the log opinion pool tend to be less dispersed than those of the linear opinion pool. Which pooling method performs better, however, depends on the particular scenario.

4.3.1.2 Estimated Combination Weights

Following Proietti et al. (2017), both the linear and logarithmic opinion pools are supplied with estimates coming from three popular weighting schemes. Based on the good performance in the point forecast setting, equal weights and weights based on the mean square error (MSE) are used. Following the likes of Hall and Mitchell (2007) and Jore et al. (2010) weights based on the log scores are also used.

The weights for model j are given by:

$$\omega_{j,t+1} = \frac{s_{j,t+1}}{\sum_{j=1}^{J} s_{j,t+1}}$$

where for equal weights $s_{j,t+1} = 1$. For the weights based on MSE $s_{j,t+1} = \frac{1}{MSE_{j,t}}$ with $MSE_{j,t} = \frac{1}{t-t_s+1} \sum_{i=t_s}^{t} (x_i - \hat{x}_{j,i})^2$ where, for time i, x_i is the observed realization, $\hat{x}_{j,i}$ is the mean of the density forecast from model j, $f_{j,i}$, and t_s is the first period considered in the evaluation window. For weights based on the log score the average over the evaluation window is used. That is, $s_{j,t+1} = \exp\left[\frac{1}{t-t_s+1} \sum_{i=t_s}^{t} \log S_i\right]$ with $S_i = \log f_{j,i}(x_i)$.

Based on the findings by Aiolfi and Timmermann (2006) regarding the weakening of forecasting performance persistence over longer evaluation windows and strong evidence of abrupt changes in relative performance of models, three different window lengths are considered and weights computed over each one:

1. Short window: the performance in the previous period alone is used, $t_s = t$.
2. Rolling window: the average performance over the last two years is used, $t_s = t-7$.
3. Expanding window: all the available information is used, $t_s = T_0$ where T_0 is the first period in the forecast evaluation sample.

4.3.1.3 Combined Density Estimation

In terms of actually producing the combined aggregate density forecast, the approach by Bjørnland et al. (2011) is followed. The cumulative distributions of the different density forecasts are represented by using piecewise linear functions, with knots at a common grid of points $g \in \{g_1, g_2, \ldots, g_M\}$, where $g_{j-1} < g_j$.

Let model j have a simulation sample $\left\{Q_j^{(1)}, Q_j^{(2)}, \ldots, Q_j^{(B)}\right\}$, then its empirical cumulative distribution at point g is given by

$$F_M \left(Q_j {\leq} g; j\right) = \frac{1}{B} \sum_{k=1}^{B} \mathbb{I}\left(Q_j^{(k)} \leq g\right) \tag{4.7}$$

where

$$\mathbb{I}\left(Q_j^{(k)} \leq g\right) = \begin{cases} 1 \text{ if } Q_j^{(k)} \leq g \\ 0 \text{ if } Q_j^{(k)} {>} g \end{cases} \tag{4.8}$$

The linear and log opinion pools are then computed at $\{g_1, g_2, \ldots, g_M\}$ using equations (4.5) and (4.6).[17]

4.3.2 Empirical Evidence

As in section 4.2, to examine how the different combination alternatives perform relative to the individual models, the log scores are compared. Likewise, the significance of the differences in predictive accuracy are assessed by means of the KLIC test. Then, to understand how the different models contribute to the combined outcome, the behaviour of the estimated combination weights is examined over the different evaluation window lengths.

4.3.2.1 Relative Forecasting Performance

Tables 4.4 and 4.5 present the logarithmic scores for the GDP one-step-ahead forecast combination using linear and logarithmic opinion pools. To keep the results comparable to those of section 4.2, the scores are expressed in terms of the percentage improvement over the corresponding bottom-up univariate model. Scores in bold highlight those that are higher than the corresponding individual models.[18]

The first thing that becomes apparent from the proportion of scores in bold is that

[17] If necessary, the components' forecasts can be reconciled with the combined aggregate forecast so as to have a consistent underlying scenario. A possible way of doing this is presented in the Appendix.

[18] The tests on the PITS are found in the Appendix.

Table 4.4: Log Scores for One-step-ahead GDP Forecast Combination

	Linear Opinion Pool							Logarithmic Opinion Pool						
	Eq.w	MSE weights			LogScore weights			Eq.w	MSE weights			LogScore weights		
Germany		Exp.w	Roll.w	Short.w	Exp.w	Roll.w	Short.w		Exp.w	Roll.w	Short.w	Exp.w	Roll.w	Short.w
Bottom-up AR	6.5	6.9	7.4	**7.7**	6.5	6.5	**10.4****	3.3	3.4	3.7	3.7	3.0	3.0	**10.0****
Homsk. VAR	**14.1**	**14.6**	**14.9**	**16.2**	**14.1**	**14.4**	**15.0**	**12.5**	**13.3**	**13.5**	**14.9***	**12.3**	**12.4***	**16.0***
Homsk. TVP	**17.5**	**18.0**	**18.2**	**19.8***	**17.4**	**17.2**	**18.2***	14.3*	14.9**	15.1**	**16.9****	14.0**	13.5**	**19.3****
Hetsk. VAR	**7.7**	**9.1**	**8.8**	**12.7**	**9.7**	**9.2**	**14.9**	**12.2**	**12.4**	**11.9**	**13.8**	**10.8**	**9.9**	**21.0**
Hetsk. TVP	**8.9**	**10.3**	**10.0**	**14.1**	**10.2**	**9.1**	**15.5**	**12.2**	**12.4**	**11.6**	**14.2**	**11.0**	**10.5**	**21.1**
DMS	17.3	17.5	18.1	**19.6***	17.1	17.1	**21.3**	15.7*	15.8**	16.6**	17.1**	15.5**	15.5**	**22.4***
All models	14.1	15.9	15.3	18.4	16.6	16.2	**21.6**	14.1*	14.9**	14.8*	17.0**	12.9**	12.5**	**23.3****
France														
Bottom-up AR	19.3	19.5	19.8	21.5	18.5	18.0	**27.5****	11.4	11.4	11.8	13.4	10.2	10.2	**26.1***
Homsk. VAR	26.1	**26.2**	**26.4**	**28.3****	26.0	26.0	**28.9***	26.1	**26.2**	**26.4***	**27.9****	25.9*	25.9*	**29.7****
Homsk. TVP	25.8	**25.8**	**26.0**	**27.5***	25.7	25.8	**28.4***	25.8	**25.7**	**26.0***	**27.1****	25.6	25.7*	**29.1****
Hetsk. VAR	**24.9**	**25.0**	**25.3**	**27.7**	**25.1**	**25.1**	**27.6**	**25.2**	**25.4**	**25.6**	**27.1***	**25.1**	**24.9**	**29.2***
Hetsk. TVP	**25.4**	**25.5**	**25.8**	**27.7***	**25.6**	**25.5**	**27.9**	**25.4**	**25.5**	**25.7**	**27.1****	**25.2**	**25.3**	**29.0***
DMS	**26.3**	**26.3**	**26.4**	**28.3****	**26.2**	**26.3**	**28.8***	**26.0**	**26.0***	**26.2***	**27.6****	25.8*	25.9*	**29.6****
All models	**26.1**	**26.4**	**26.8**	**30.1****	**26.3**	**26.2**	**29.3***	23.4	23.7	24.2	**27.0****	22.0	21.5	**30.1****
United Kingdom														
Bottom-up AR	**10.4***	**10.4***	**10.9****	**12.2****	**10.4***	**10.6***	**11.3***	5.3	5.7	6.1	**7.7**	5.5	5.8	**8.1**
Homsk. VAR	**12.1***	**12.1***	**12.5****	**14.4****	**12.1***	**12.3***	**13.9****	**10.5***	**10.4***	**10.9****	**13.2****	**10.5***	**10.7***	**14.3****
Homsk. TVP	**11.9***	**11.9***	**12.2***	**13.7****	**11.9***	**12.0***	**13.6****	**10.5***	**10.4***	**10.7****	**12.6****	**10.4***	**10.5***	**13.9****
Hetsk. VAR	**9.7**	**10.1**	**10.5**	**12.1***	**10.1**	**10.2**	**11.3**	**8.7**	**8.7**	**9.1**	**11.2***	**8.8**	**8.9**	**11.7***
Hetsk. TVP	**9.4**	**9.8**	**10.1**	**11.8**	**9.9**	**9.9**	**10.8**	**8.7**	**8.7**	**9.0**	**11.0***	**8.8**	**8.9**	**11.8***
DMS	**11.5**	**11.6***	**12.1***	**13.2****	**11.7***	**11.8***	**12.9****	**9.9**	**9.8***	**10.2****	**12.0****	**9.8***	**10.0***	**12.7****
All models	**9.6**	**10.1**	**10.4**	**12.8***	**10.0**	**10.1**	**12.3**	7.5	8.1	**8.4**	**10.5**	7.8	8.0	**10.3**

Note: Log predictive density scores of the one-step-ahead forecast combination using the weighting schemes of each model and the direct AR for the three countries expressed in terms of the percentage improvement over the bottom-up univariate model (Bottom-up AR). The models are the indirect univariate AR model (Bottom-up AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). The weighting schemes are equal weights (Eq.w) and based on recent past mean square error (MSE.w) and based on the log score (LS.w). The combination weights estimation window lengths are one quarter (Short), eight quarters (Roll.) and all the available sample (Exp.). Log scores in bold denote improvement over both the direct univariate model and the corresponding bottom-up method. The scores for the pool of all models are compared with those obtained by the DMS approach. * and ** denote significance of the forecasting performance difference with the direct univariate model based on the KLIC test at a 10 and 5% significance levels. Calculated over the 2001-2015 period.

Table 4.5: Log Scores for One-step-ahead CPI Forecast Combination

	Linear Opinion Pool							Logarithmic Opinion Pool						
	Eq.w	MSE weights			LogScore weights			Eq.w	MSE weights			LogScore weights		
Germany		Exp.w	Roll.w	Short.w	Exp.w	Roll.w	Short.w		Exp.w	Roll.w	Short.w	Exp.w	Roll.w	Short.w
Bottom-up AR	4.6	4.8	4.9	6.5	4.5	4.6	7.4**	3.9	4.2	4.3	5.7	3.8	3.9	7.7**
Homsk. VAR	6.7	7.4	7.4	9.6**	7.0	7.1	8.0*	6.5	6.7*	6.7	8.7**	6.6	6.7	8.3**
Homsk. TVP	6.3	7.1	7.0	8.7*	6.7	6.7	7.3	6.7	6.8*	6.7	8.2**	6.8	6.8	8.1**
Hetsk. VAR	6.1	6.4	6.7	8.8**	6.2	6.4	8.1*	6.6	6.7	7.1*	8.7**	6.7	6.9	8.9**
Hetsk. TVP	7.4	7.9	8.0	9.8*	7.6	7.8	8.8	7.1	7.0*	7.3*	8.7**	7.2	7.4*	9.3**
DMS	7.2	7.6	7.6	9.5**	7.5	7.5	8.2*	6.5	6.6*	6.6*	8.2**	6.6	6.7	8.1**
All models	6.3	6.8	7.0	10.3**	6.5	6.7	8.5	6.3	6.6	6.8	9.6**	6.4	6.6	9.7**
France														
Bottom-up AR	2.1*	2.4**	2.6**	4.5**	2.7**	2.8**	4.6**	2.3	2.5	2.7	4.1**	2.9*	3.0*	5.4**
Homsk. VAR	3.7*	4.0*	4.1**	6.0**	3.9**	3.9**	5.8**	3.0*	3.3**	3.5**	5.6**	3.2**	3.2**	6.0**
Homsk. TVP	4.4**	4.7**	4.8**	6.5**	4.5**	4.6**	6.2**	3.8**	4.2**	4.4**	6.1**	4.0**	4.0**	6.5**
Hetsk. VAR	5.1*	5.1	5.1	6.9**	5.1*	5.2*	6.3**	4.5**	5.4**	5.4**	6.7**	4.9**	4.9**	7.6**
Hetsk. TVP	3.5	3.9	3.9	5.4*	3.8	3.9	5.2	3.4*	3.3*	3.5*	5.5**	3.5*	3.5*	6.7**
DMS	5.0*	5.2*	5.2*	6.8**	5.1*	5.2*	6.8**	4.2**	4.6**	4.6**	6.3**	4.3**	4.4**	7.5**
All models	5.7*	6.2**	6.3**	8.2**	6.2**	6.3**	8.2**	5.6**	6.3**	6.6**	8.1**	6.3**	6.4**	9.7**
United Kingdom														
Bottom-up AR	3.4**	4.4*	4.7**	7.2**	4.1*	4.2**	6.8**	3.3	3.1	3.9	6.6**	3.3	3.5	7.8**
Homsk. VAR	3.0*	3.4*	3.7**	5.7**	3.5*	3.5*	5.1**	3.3**	3.9**	4.2**	6.3**	4.1**	4.2**	6.2**
Homsk. TVP	2.6*	3.0*	3.4*	5.2**	3.1*	3.2*	4.5**	3.0**	3.6**	4.0**	5.8**	3.8**	3.9**	5.6**
Hetsk. VAR	1.7	1.8	2.3	4.1**	2.1	2.2	3.5*	1.9	2.4	2.8*	4.6**	2.8*	2.9*	4.5**
Hetsk. TVP	1.5	1.7	2.2	3.9**	1.9	2.1	3.3*	2.0*	2.6	3.0*	4.6**	2.9*	3.0*	4.5**
DMS	2.1	2.5	2.8*	4.5**	2.6	2.7	4.1**	2.6**	3.1*	3.4**	5.2**	3.3*	3.5**	5.2**
All models	2.5	3.1	3.4	6.1**	2.9	3.0	4.6**	3.3	3.8	4.2*	6.7**	3.6	3.7	6.2**

Note: Log predictive density scores of the one-step-ahead forecast combination using the weighting schemes of each model and the direct AR for the three countries expressed in terms of the percentage improvement over the bottom-up univariate model (Bottom-up AR). The models are the indirect univariate AR model (Bottom-up AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). The weighting schemes are equal weights (Eq.w) and based on recent past mean square error (MSE.w) and based on the log score (LS.w). The combination weights estimation window lengths are one quarter (Short), eight quarters (Roll.) and all the available sample (Exp.). Log scores in bold denote improvement over both the direct univariate model and the corresponding bottom-up method. The scores for the pool of all models are compared with those obtained by the DMS approach. * and ** denote significance of the forecasting performance difference with the direct univariate model based on the KLIC test at a 10 and 5% significance levels. Calculated over the 2001-2015 period.

the effect of combining is positive in most cases. However, the magnitude and significance of the improvements vary considerably from one model to another depending on the weight estimation window length and method. One result that repeats itself for all datasets and combination methods, is that the short combination weight estimation window produces considerably higher scores than those of the longer windows. In most cases the improvements over the direct method are statistically significant according to the KLIC test while this is not necessarily so in the case of longer windows. In terms of the improvement in performance for each model, the univariate bottom-up approach improves the most. Focusing only on the results from the short weight evaluation window, for GDP, these vary between 3 percentage points for Germany and 28 percentage points for France. For CPI, they vary between nearly 4 percentage points for France and 8 percentage points for Germany and the United Kingdom. The improvements of the multivariate methods are less impressive, but only because of the good performance of the non-combined methods. The log scores of the heteroskedastic VARs increase between 3 and 15 percentage points, while the improvements of the homoskedastic VARs and DMS are comparatively smaller. In terms of significance, most of the combinations using the short window are significantly better than the direct method alone. An interesting result for the DMS combination and pooling of all models for Germany, is that, although the scores are not higher than those of the DMS approach alone, they are statistically better than the direct approach, while the DMS approach by itself is not.

When comparing the combination methods, it is not clear in terms of the magnitude of the log scores whether the linear opinion pool is better than the logarithmic opinion pool. However, the latter does produce forecasts that are statistically better than the direct approach in more cases. This probably comes as a result of the fact that the aforementioned method produces more focused densities. In terms of comparing the approach for calculating the weights, that is using MSE or log scores, neither comes out as significantly better than the other.

As regards pooling all the models together, the overall results are good. Most of the scores are relatively high and in many cases the differences with the direct approach are statistically significant. For CPI, in most cases, the forecasts are better than the other combination approaches in terms of log scores. For GDP, this is also the case, except for the United Kingdom, where they are usually beaten by at least one of the other methods.

4.3.2.2 Behaviour of the Combination Weights

Examining how the combination weights evolve over the evaluation sample can provide some insight into the performance of the different models and the influence of the weighting methods. Figures 4.3 and 4.4 present the combination weights for the corresponding bottom-up methods for GDP and CPI for the three countries over the 2001-2015 period.[19] The shaded areas highlight years 2008 and 2009 so as to be able to assess the effects of the financial crisis on the models.

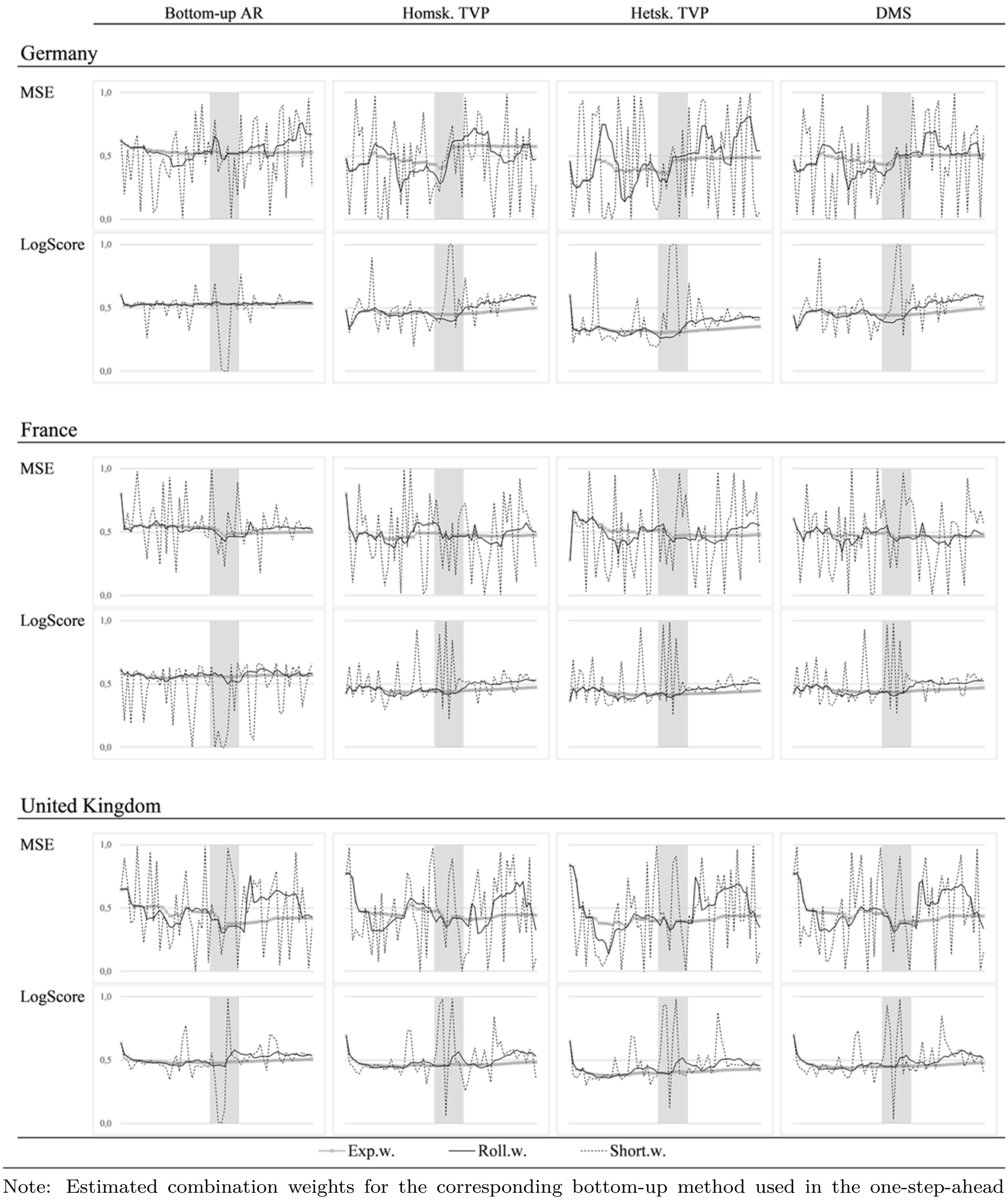
As regards comparing the equal, MSE and log score weights, it is clear that those based on the MSE are considerably more volatile than those based on the log scores. This is obvious for the short weight evaluation window, but also visible for the rolling and expanding windows. It can be observed that for MSE, it is often the case that short window weights go to the extremes, either zero or one. The rolling window, although less erratic, shows abrupt changes and the expanding window often shows visible steps. All this is not entirely surprising, given that this weight calculation method only considers the first moment of the distribution. For the log score weights on the other hand it seems that they are centred more around the equal weights, and deviate when necessary.

In terms of assessing the performance of the individual models, the weights calculated using the short evaluation window reveal some evidence of the poor calibration of the univariate bottom-up approach. For all datasets, at some point during the crisis years, it is given zero or almost zero weight in the combination, while for some datasets this happens even if the crisis years are ignored. This suggests that this method is very inferior to the direct approach. The opposite happens for the multivariate bottom-up methods. During the crisis years, it is the direct method that gets given almost zero weight, while in the rest of the sample the weights mostly wander around the 0.5 of equal weights. All this suggests that the considerably higher scores of the short evaluation windows, with regard to the equal weights, are mainly due to the better performance of the multivariate bottom-up models during the crisis years. From this perspective, it would seem that expanding and long windows are not flexible enough to pick up all the benefits of models that excel in very particular circumstances. Another thing that can be appreciated is that, in the case of GDP, adding stochastic volatility to the BVARs results in the models producing over-dispersed densities. This can be seen in that they are consistently given a lower weight than the corresponding homoskedastic versions.

In conclusion, the main findings of this section are that, for this empirical implementation the strengths of the multivariate bottom-up methods become apparent during the crisis. In this perspective, to benefit fully from these strengths when using forecast combination, it is necessary for the weighting strategy to be flexible enough to allow for some considerable switching between models. Last but not least is the fact that simply pooling all models together coupled with a short weight-evaluation window,

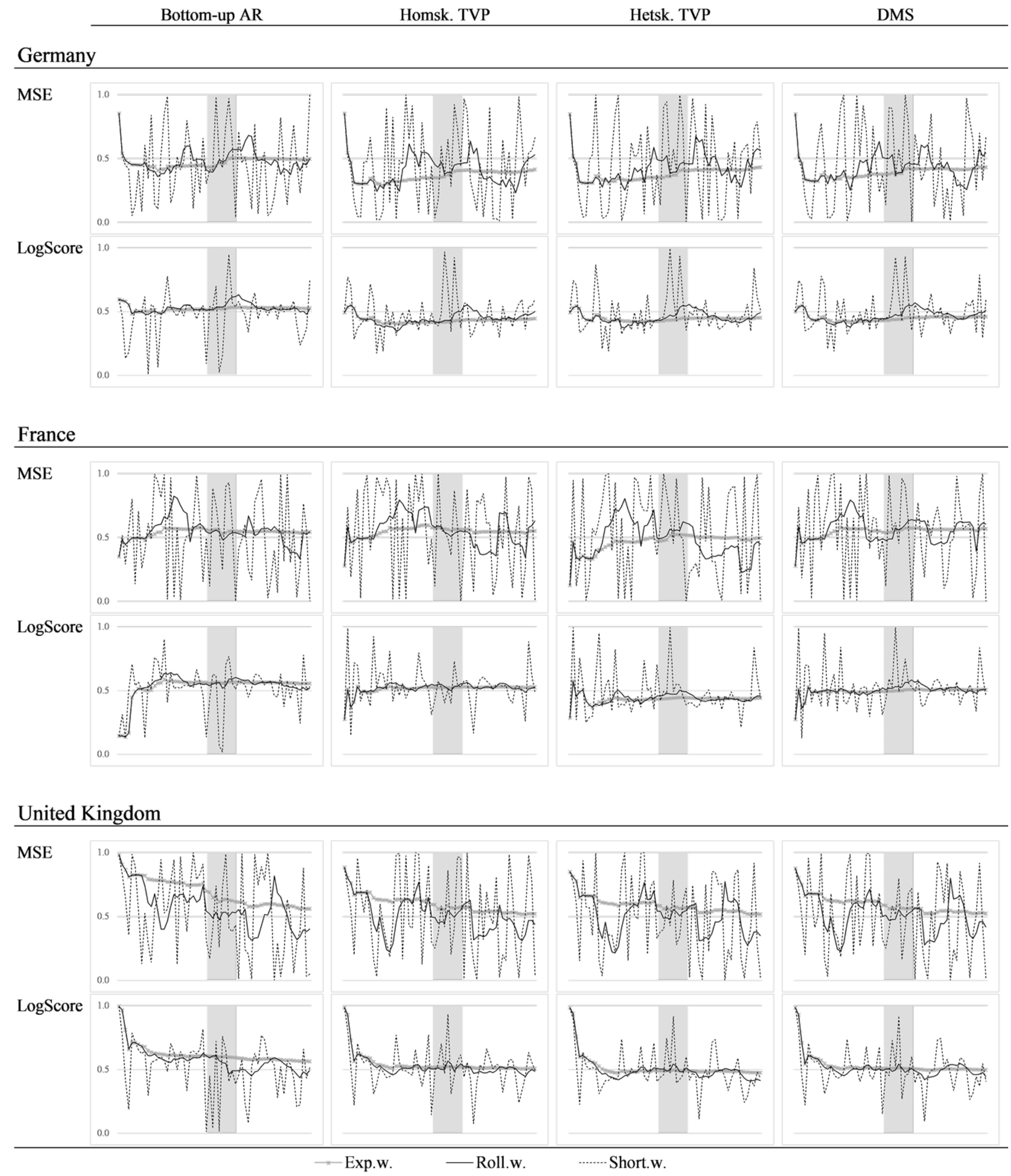
[19] That is, a weight of zero corresponds to giving all the weight to the direct approach. The figures for the fixed-parameter models have been dropped for clarity and because the main findings are present in their time-varying counterparts.

Figure 4.3: GDP Combination Weights



Note: Estimated combination weights for the corresponding bottom-up method used in the one-step-ahead forecast combination with the direct method. The estimated weighting schemes are based on mean square error (MSE) and log scores. The estimation window lengths are one quarter (Short), eight quarters (Roll.) and all the available sample (Exp.). Calculated over the 2001-2015 period. Shaded area highlights years 2008 and 2009.

Figure 4.4: CPI Combination Weights



Note: Estimated combination weights for the corresponding bottom-up method used in the one-step-ahead forecast combination with the direct method. The estimated weighting schemes are based on mean square error (MSE) and log scores. The estimation window lengths are one quarter (Short), eight quarters (Roll.) and all the available sample (Exp.). Calculated over the 2001-2015 period. Shaded area highlights years 2008 and 2009.

although not always the best performer, produces very good results.

4.4 Conclusions

This chapter presents a methodology to use the information at a component level to produce well-calibrated and competitive aggregate density forecasts. To do this, large Bayesian VARs are used to extend to a probabilistic setting the bottom-up approach used commonly for point forecasts. The implemented method is relatively simple, being both flexible and computationally cheap and able to consider both fixed and time-varying parameter VARs and stochastic volatility. Finally, the framework is extended so as to benefit from the strengths of both the direct and bottom-up approaches through forecast combination.

The empirical application using CPI and GDP data for France, Germany and the United Kingdom shows that overall the multivariate methods are capable of producing bottom-up forecasts that are calibrated and perform equally well or better than the aggregate benchmarks. The results also suggest that there are additional gains from allowing for time-varying parameters. The largest improvements, however, are obtained from combining the direct and multivariate bottom-up methods. The analysis of the results seems to suggest that this is mainly due to the impact of the financial crisis. This would tend to support the claims by Bache et al. (2010) that components are useful to explore events in the tails of the distribution.

In terms of future research, there are many possibilities. One is to produce the estimates for the time-varying and stochastic volatility parameters, using alternative methods which include using a full Bayesian approach and comparing the results with those of the approximations. A natural extension would be to couple the method with one designed to forecast the aggregation weights and use the augmented framework to forecast at longer horizons. A third direction for research could be to incorporate useful economic indicators and other relevant variables into the forecasting process in a way similar to that of Banbura et al. (2010).

Appendix:

4.A Additional Information from Empirical Application

Table 4.6: Tests on PITs for GDP Forecasts using Linear Opinion Pool Combinations

		Germany				France				United Kingdom			
	Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Equal Weights:													
	Bottom-Up AR	0.13	0.00	0.01	0.02	0.00	0.18	0.11	0.12	0.04	0.01	0.02	0.06
	Homsk. VAR	0.56	0.00	0.02	0.64	0.47	0.27	0.19	0.29	0.17	0.14	0.30	0.12
	Homsk. TVP	0.76	0.00	0.04	0.69	0.41	0.21	0.55	0.22	0.15	0.07	0.24	0.13
	Hetsk. VAR	0.05	0.00	0.00	0.81	0.91	0.04	0.53	0.35	0.05	0.01	0.31	0.06
	Hetsk. TVP	0.72	0.00	0.00	0.80	0.79	0.11	0.05	0.33	0.03	0.01	0.15	0.07
	DMS	0.72	0.00	0.01	0.44	0.53	0.35	0.78	0.33	0.13	0.04	0.27	0.10
	All models	0.36	0.00	0.01	0.76	0.66	0.16	0.35	0.55	0.02	0.01	0.23	0.18
MSE based:													
	Bottom-Up AR	0.23	0.00	0.00	0.04	0.00	0.16	0.27	0.13	0.04	0.00	0.03	0.07
	Homsk. VAR	0.62	0.00	0.00	0.53	0.48	0.02	0.03	0.19	0.08	0.00	0.07	0.05
	Homsk. TVP	0.83	0.00	0.00	0.54	0.43	0.02	0.03	0.13	0.07	0.01	0.17	0.06
	Hetsk. VAR	0.03	0.00	0.00	0.87	0.90	0.00	0.04	0.33	0.02	0.00	0.01	0.03
	Hetsk. TVP	0.62	0.00	0.00	0.62	0.64	0.01	0.19	0.27	0.01	0.00	0.03	0.04
	DMS	0.69	0.00	0.00	0.54	0.50	0.03	0.18	0.13	0.10	0.00	0.08	0.06
	All models	0.16	0.00	0.00	0.89	0.57	0.00	0.05	0.54	0.01	0.00	0.00	0.07
Log score based:													
	Bottom-Up AR	0.12	0.00	0.01	0.02	0.00	0.23	0.49	0.13	0.03	0.01	0.02	0.06
	Homsk. VAR	0.57	0.00	0.02	0.60	0.46	0.17	0.36	0.26	0.18	0.14	0.29	0.10
	Homsk. TVP	0.80	0.00	0.05	0.65	0.41	0.18	0.47	0.20	0.16	0.06	0.30	0.12
	Hetsk. VAR	0.41	0.00	0.00	0.53	0.88	0.05	0.58	0.28	0.07	0.01	0.15	0.05
	Hetsk. TVP	0.96	0.00	0.00	0.55	0.81	0.12	0.70	0.28	0.06	0.01	0.14	0.05
	DMS	0.67	0.00	0.07	0.41	0.53	0.32	0.62	0.29	0.14	0.04	0.24	0.08
	All models	0.73	0.00	0.00	0.64	0.67	0.20	0.81	0.47	0.02	0.01	0.22	0.17

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecast combination of each model with the direct AR forecast for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Calculated over the 2001-2015 period.

Table 4.7: Tests on PITs for GDP Forecasts using Log Opinion Pool Combinations

		Germany				France				United Kingdom			
	Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Equal Weights:													
	Bottom-Up AR	0.10	0.00	0.01	0.01	0.00	0.07	0.36	0.12	0.01	0.00	0.01	0.08
	Homsk. VAR	0.46	0.00	0.01	0.52	0.32	0.13	0.19	0.28	0.11	0.15	0.50	0.11
	Homsk. TVP	0.61	0.00	0.05	0.54	0.26	0.14	0.10	0.22	0.13	0.15	0.63	0.11
	Hetsk. VAR	0.95	0.00	0.00	0.44	0.26	0.00	0.04	0.24	0.11	0.02	0.15	0.04
	Hetsk. TVP	0.87	0.00	0.00	0.48	0.75	0.05	0.21	0.26	0.12	0.03	0.09	0.05
	DMS	0.51	0.00	0.03	0.34	0.35	0.19	0.09	0.31	0.14	0.06	0.21	0.07
	All models	0.58	0.00	0.02	0.54	0.01	0.00	0.02	0.35	0.01	0.00	0.09	0.17
MSE based:													
	Bottom-Up AR	0.20	0.00	0.00	0.03	0.00	0.06	0.07	0.13	0.01	0.00	0.00	0.07
	Homsk. VAR	0.55	0.00	0.00	0.42	0.39	0.01	0.09	0.18	0.07	0.01	0.15	0.04
	Homsk. TVP	0.79	0.00	0.00	0.43	0.33	0.01	0.14	0.13	0.08	0.01	0.26	0.04
	Hetsk. VAR	0.83	0.00	0.00	0.87	0.75	0.00	0.00	0.24	0.06	0.00	0.02	0.02
	Hetsk. TVP	0.95	0.00	0.00	0.73	0.81	0.01	0.01	0.21	0.03	0.00	0.25	0.03
	DMS	0.62	0.00	0.00	0.40	0.42	0.02	0.09	0.13	0.14	0.00	0.14	0.05
	All models	0.88	0.00	0.00	0.81	0.01	0.00	0.00	0.32	0.01	0.00	0.00	0.07
Log score based:													
	Bottom-Up AR	0.09	0.00	0.01	0.01	0.00	0.09	0.09	0.13	0.01	0.00	0.01	0.08
	Homsk. VAR	0.46	0.00	0.01	0.49	0.30	0.11	0.21	0.25	0.11	0.14	0.38	0.10
	Homsk. TVP	0.45	0.00	0.03	0.51	0.25	0.14	0.22	0.19	0.12	0.13	0.50	0.09
	Hetsk. VAR	0.97	0.00	0.00	0.34	0.19	0.00	0.03	0.20	0.10	0.02	0.06	0.03
	Hetsk. TVP	0.59	0.00	0.01	0.35	0.68	0.04	0.09	0.23	0.14	0.03	0.14	0.04
	DMS	0.50	0.00	0.03	0.32	0.33	0.16	0.18	0.28	0.13	0.06	0.21	0.06
	All models	0.40	0.00	0.07	0.44	0.00	0.00	0.02	0.30	0.01	0.00	0.25	0.16

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecast combination of each model with the direct AR forecast for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Calculated over the 2001-2015 period.

Table 4.8: Tests on PITs for CPI Forecasts using Linear Opinion Pool Combinations

		Germany				France				United Kingdom			
	Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Equal Weights:													
	Bottom-Up AR	0.18	0.23	0.57	0.47	0.08	0.02	0.25	0.50	0.31	0.40	0.05	0.32
	Homsk. VAR	1.00	0.04	0.19	0.84	0.52	0.02	0.20	0.45	0.27	0.55	0.45	0.49
	Homsk. TVP	1.00	0.04	0.45	0.75	1.00	0.21	0.59	0.58	0.18	0.44	0.52	0.42
	Hetsk. VAR	0.94	0.17	0.08	0.71	0.01	0.70	0.51	0.42	0.06	0.56	0.62	0.43
	Hetsk. TVP	0.98	0.05	0.25	0.68	0.50	0.06	0.31	0.20	0.04	0.65	0.32	0.43
	DMS	0.98	0.34	0.72	0.77	0.61	0.42	0.85	0.48	0.19	0.51	0.64	0.41
	All models	0.98	0.05	0.18	0.56	0.62	0.04	0.24	0.64	0.12	0.71	0.42	0.29
MSE based:													
	Bottom-Up AR	0.36	0.03	0.06	0.56	0.11	0.00	0.02	0.36	0.27	0.08	0.03	0.39
	Homsk. VAR	0.90	0.00	0.11	0.77	0.85	0.00	0.05	0.58	0.04	0.01	0.00	0.63
	Homsk. TVP	0.95	0.00	0.00	0.86	0.90	0.00	0.04	0.77	0.03	0.00	0.00	0.53
	Hetsk. VAR	0.87	0.01	0.14	0.47	0.00	0.00	0.02	0.14	0.00	0.01	0.00	0.51
	Hetsk. TVP	0.87	0.00	0.01	0.56	0.17	0.00	0.00	0.25	0.00	0.05	0.01	0.60
	DMS	0.93	0.02	0.04	0.51	0.24	0.00	0.00	0.60	0.03	0.03	0.03	0.67
	All models	0.57	0.00	0.00	0.53	0.10	0.00	0.00	0.35	0.02	0.01	0.00	0.52
Log score based:													
	Bottom-Up AR	0.16	0.20	0.31	0.46	0.07	0.02	0.17	0.54	0.22	0.55	0.36	0.24
	Homsk. VAR	1.00	0.05	0.21	0.85	0.53	0.02	0.21	0.46	0.23	0.57	0.89	0.53
	Homsk. TVP	1.00	0.05	0.47	0.76	0.99	0.19	0.54	0.59	0.15	0.50	0.96	0.44
	Hetsk. VAR	0.96	0.20	0.37	0.73	0.01	0.64	0.77	0.41	0.04	0.47	0.52	0.42
	Hetsk. TVP	0.98	0.07	0.24	0.70	0.63	0.09	0.36	0.20	0.03	0.75	0.45	0.41
	DMS	0.98	0.40	0.82	0.78	0.60	0.47	0.78	0.47	0.16	0.50	0.56	0.43
	All models	0.99	0.05	0.08	0.55	0.55	0.05	0.33	0.62	0.14	0.70	0.58	0.25

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecast combination of each model with the direct AR forecast for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Calculated over the 2001-2015 period.

Table 4.9: Tests on PITs for CPI Forecasts using Log Opinion Pool Combinations

		Germany				France				United Kingdom			
	Model	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB	Bkw.LR	AD	χ^2	LB
Equal Weights:													
	Bottom-Up AR	0.03	0.15	0.37	0.43	0.02	0.02	0.25	0.56	0.16	0.67	0.15	0.31
	Homsk. VAR	0.89	0.19	0.86	0.80	0.16	0.05	0.31	0.45	0.64	0.92	0.81	0.49
	Homsk. TVP	0.92	0.24	0.38	0.66	0.73	0.17	0.18	0.52	0.46	0.69	0.60	0.41
	Hetsk. VAR	0.88	0.37	0.54	0.72	0.83	0.37	0.59	0.29	0.29	0.67	0.94	0.41
	Hetsk. TVP	0.98	0.21	0.73	0.67	0.87	0.05	0.16	0.13	0.27	0.78	0.43	0.46
	DMS	0.87	0.64	0.91	0.79	0.83	0.14	0.43	0.46	0.54	0.78	0.43	0.42
	All models	0.82	0.06	0.01	0.50	0.43	0.00	0.09	0.65	0.36	0.62	0.80	0.29
MSE based:													
	Bottom-Up AR	0.14	0.02	0.02	0.53	0.07	0.00	0.01	0.42	0.41	0.08	0.04	0.38
	Homsk. VAR	0.95	0.00	0.01	0.75	0.66	0.00	0.05	0.58	0.09	0.01	0.00	0.65
	Homsk. TVP	0.99	0.01	0.02	0.82	1.00	0.00	0.02	0.74	0.06	0.00	0.00	0.53
	Hetsk. VAR	0.90	0.01	0.00	0.48	0.06	0.00	0.00	0.26	0.01	0.02	0.00	0.50
	Hetsk. TVP	0.97	0.00	0.00	0.58	0.75	0.00	0.00	0.13	0.02	0.06	0.00	0.61
	DMS	0.94	0.02	0.01	0.52	0.72	0.00	0.01	0.55	0.10	0.04	0.03	0.67
	All models	0.89	0.00	0.00	0.48	0.77	0.00	0.00	0.54	0.25	0.00	0.00	0.46
Log score based:													
	Bottom-Up AR	0.02	0.13	0.08	0.42	0.03	0.03	0.27	0.59	0.05	0.75	0.18	0.19
	Homsk. VAR	0.86	0.24	0.94	0.78	0.17	0.04	0.33	0.46	0.50	0.75	0.80	0.51
	Homsk. TVP	0.88	0.25	0.42	0.67	0.77	0.16	0.26	0.53	0.33	0.60	0.97	0.42
	Hetsk. VAR	0.86	0.39	0.52	0.73	0.84	0.53	0.80	0.29	0.13	0.48	0.14	0.39
	Hetsk. TVP	0.96	0.24	0.92	0.68	0.83	0.09	0.35	0.14	0.12	0.57	0.45	0.43
	DMS	0.84	0.60	0.94	0.79	0.87	0.15	0.25	0.46	0.44	0.66	0.89	0.43
	All models	0.75	0.07	0.04	0.49	0.43	0.01	0.09	0.66	0.27	0.60	0.56	0.23

Note: P-values for the calibration tests on the probability integral transform (PIT) of the one-step-ahead forecast combination of each model with the direct AR forecast for the three countries. The tests are the LR test proposed by Berkowitz (2001) (Bkw.LR), the uniformity tests by Anderson-Darling (AD) and a Pearson's chi-squared (χ^2), the Ljung-Box test (LB) for independence. The models are the bottom-up univariate model (Bottom-up AR), the direct univariate AR model (Direct AR), the homoskedastic VAR, the homoskedastic TVP-VAR, the heteroskedastic VAR, the heteroskedastic TVP-VAR and the result of dynamic model selection over the four VARs (DMS). Calculated over the 2001-2015 period.

A Procedure for Obtaining Combined Components' Density Forecast.

To produce the combined disaggregate density forecasts for an approach the procedure is as follows:

1. Simulate the combined aggregate density by producing B realizations, the same number as that used to produce the bottom-up forecasts.
2. Sort the draws from the simulated combined density, $Q_{C,t+1} = \left\{Q_{C,t+1}^{(1)}, Q_{C,t+1}^{(2)}, \ldots, Q_{C,t+1}^{(B)}\right\}$ by value from smallest to largest.
3. Sort the draws from the aggregate forecast resulting from the bottom-up approach, $Q_{I,t+1}$.
4. Use the combination procedure presented in Chapter 2 to reconcile the disaggregate forecasts corresponding to the smallest value of $Q_{I,t+1}$ with the smallest of $Q_{C,t+1}$ by assigning total confidence to the latter.
5. Do the same with the second smallest value and so on for all B realizations.

Chapter 5

Summary

This thesis centres around forecasting economic aggregates and their components. The focus is on improving overall performance through the development of applied methods that benefit from the strengths of both direct and bottom-up approaches. The research is divided into three chapters. The first two concentrate on point forecasts, while the third shifts the attention towards density forecasts.

Chapter 2 presents a framework for jointly combining forecasts for an aggregate, any intermediate sub-aggregations and the components from any number of measurement approaches, based on the subjective reliability of each of the forecasts involved. The rationale for doing this is that incorporating the coherence that is required by the aggregation structure explicitly could lead to improvements in overall accuracy. By accounting for the reliability of each forecast, the objective is for the strengths of the better-performing ones to spill over to the others. An empirical application using CPI data from France, Germany and the United Kingdom suggests that this is in fact the case. The results from the exercise show that the overall forecasting accuracy of the weaker aggregation approaches often improves without any significant effects on the other forecasts. Also, although the performance of different models varies greatly, depending on the country, aggregation approach and whether the years from the financial crisis are included in the evaluation period, the multi-level combination method often performs at least as well as the best-performing single methods and in most cases significantly better than the median of the single models, both in terms of aggregate and disaggregate accuracy. This would suggest that it retains the desirable feature from traditional forecast combination methods of safeguarding against committing large mistakes because of uncertainty regarding the best model.

In Chapter 3 a method to forecast economic aggregates using purpose-built groupings of components is presented. The objective in developing such a method is to improve the performance of any forecasting model by transforming the data in a way that avoids the problems associated with disaggregate misspecification, while still allowing for distinct disaggregate dynamics to be picked up in the process. The method relies on using Agglomerative Hierarchical Clustering to reduce the dimension of the problem by choosing a subset of feasible groupings based on the commonality among the different components. A single definitive forecast is then produced, based on this subset. An empirical application of the method using CPI data for France, Germany and the United Kingdom is performed to evaluate the outcome for many different specifications. The results show that some of the specifications often perform better than the best benchmark methods. The methods that perform well include a couple that aim to select a unique grouping and many that use a combination of all the groupings in the initial subset. The good performance of the latter suggests that expanding the pool of forecasts by trying different combinations of components provides a way of benefiting from the strengths of forecast combination without necessarily increasing the number of forecasting models. The results are robust to whether the financial crisis years are included in the evaluation sample or not.

In Chapter 4 the attention is drawn to producing density forecasts for economic aggregates based on the density forecasts for their components. The motivation for doing this is born of the idea that accounting for the dynamics of the underlying components can contribute to a better understanding of a wide range of uncertainties. With the aim of incorporating the dynamics of the components, but also the interaction between them, a framework that models the whole multivariate process is developed. An empirical application using CPI and GDP data from France, Germany and the United Kingdom shows that the multivariate methods are capable of producing bottom-up forecasts that are calibrated and perform equally well or better than comparable aggregate methods. The differences between the results for GDP and CPI, where for the former the effects of the most recent financial crisis are significantly larger both in magnitude and persistence, suggest that the strengths of the multivariate methods might only emerge if the interactions among variables are strong enough. Based on these results, which involve acknowledging that the strengths of the direct and bottom-up approaches may appear under different circumstances, the framework is extended to allow combining both. The outcome from doing so in the same empirical exercise are results that are significantly better in terms of aggregate accuracy than those of the direct method, both for GDP and CPI.

To conclude, this research shows that it is possible to benefit from alternative measurement approaches and aggregations even when these, taken separately, cannot solve the problem at hand. The gains from using these methods will depend both on how they are specified and on the particular dataset. The results from the empirical applications suggest that going for many specifications and combining the results is an effective way to avoid making large forecasting errors.

Bibliography

Adrian, T. (2007). Measuring risk in the hedge fund sector. *Current Issues in Economics & Finance 13*(3), 1 – 7.

Aigner, D. J. and S. M. Goldfeld (1974). Estimation and prediction from aggregate data when aggregates are measured more accurately than their components. *Econometrica: Journal of the Econometric Society*, 113–134.

Aiolfi, M. and A. Timmermann (2006). Persistence of forecasting performance and conditional combination strategies. *Journal of Econometrics 135*, 31–53.

Alessi, L., E. Ghysels, L. Onorante, R. Peach, and S. Potter (2014). Central bank macroeconomic forecasting during the global financial crisis: the European Central Bank and Federal Reserve Bank of New York experiences. *Journal of Business & Economic Statistics 32*(4), 483–500.

Amisano, G. and R. Giacomini (2007). Comparing density forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics 25*, 177–190.

Antipa, P., K. Barhoumi, V. Brunhes-Lesage, and O. Darné (2012). Nowcasting German GDP: A comparison of bridge and factor models. *Journal of Policy Modeling 34*(6), 864–878.

Aron, J. and J. Muellbauer (2012). Improving forecasting in an emerging economy, South Africa: Changing trends, long run restrictions and disaggregation. *International Journal of Forecasting 28*(2), 456–476.

Aruoba, S. B., F. X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song (2013). Improving US GDP measurement: A forecast combination perspective. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 1–25. Springer.

Bache, I. W., J. Mitchell, F. Ravazzolo, and S. P. Vahey (2010). Macro-modelling with many models. *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects*. Chapter 16.

Banbura, M., D. Giannone, and L. Reichlin (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics 25*(1), 71–92.

Barker, T. and M. Pesaran (1990). *Disaggregation in Econometric Modelling: An Introduction*, Chapter 1. Routledge.

Bates, J. M. and C. W. Granger (1969). The combination of forecasts. *Operations Research Quarterly 20*, 451–468.

Bell, V., L. W. Co, S. Stone, and G. Wallis (2014). Nowcasting UK GDP growth. *Bank of England Quarterly Bulletin 54*(1), 58–68.

Benalal, N., J. L. Diaz del Hoyo, B. Landau, M. Roma, and F. Skudelny (2004). To aggregate or not to aggregate? Euro area inflation forecasting. Working Paper 374, European Central Bank.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics 19*(4), 465–474.

Bermingham, C. and A. D'Agostino (2014). Understanding and forecasting aggregate and disaggregate price dynamics. *Empirical Economics 46*(2), 765–788.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2011). Weights and pools for a Norwegian density combination. *The North American Journal of Economics and Finance 22*(1), 61–76.

Brüggemann, R. and H. Lütkepohl (2013). Forecasting contemporaneous aggregates with stochastic aggregation weights. *International Journal of Forecasting 29*(1), 60–68.

Burriel, P. (2012). A real-time disaggregated forecasting model for the Euro area GDP. *Economic Bulletin*, 93–103.

Bussière, M., M. Hoerova, and B. Klaus (2015). Commonality in hedge fund returns: Driving factors and implications. *Journal of Banking & Finance 54*, 266 – 280.

Carriero, A., T. E. Clark, and M. Marcellino (2015). Bayesian VARs: specification choices and forecast accuracy. *Journal of Applied Econometrics 30*(1), 46–73.

Chauvet, M. and S. Potter (2013). Forecasting output. *Handbook of Economic Forecasting 2*(Part A), 141–194.

Clark, T. E. (2004). An evaluation of the decline in goods inflation. *Economic Review-Federal Reserve Bank of Kansas City 89*(2), 19.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting 5*(4), 559–583.

Conflitti, C., C. De Mol, and D. Giannone (2015). Optimal combination of survey forecasts. *International Journal of Forecasting 31*(4), 1096–1103.

Cooke, E. J., R. S. Savage, P. D. Kirk, R. Darkins, and D. L. Wild (2011). Bayesian hierarchical clustering for microarray time series data with replicates and outlier measurements. *BMC Bioinformatics 12*(1), 399.

Dalgaard, E. and C. Gysting (2004). An algorithm for balancing commodity-flow systems. *Economic Systems Research 16*(2), 169–190.

Del Negro, M., R. B. Hasegawa, and F. Schorfheide (2016). Dynamic prediction pools: an investigation of financial frictions and forecasting performance. *Journal of Econometrics 192*(2), 391–405.

Denton, F. T. (1971). Adjustment of monthly or quarterly series to annuals totals: An approach based on quadratic minimization. *Journal of the American Statistical Association 66*(333), 99–102.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

Diebold, F. X. and J. A. Lopez (1996). Forecast evaluation and combination. In Maddala and Rao (Eds.), *Handbook of Statistics*. Elsevier.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Drechsel, K. and R. Scheufele (2013). Bottom-up or direct? Forecasting German GDP in a data-rich environment. IWH Discussion Papers 7, Halle Institute for Economic Research.

Duarte, C. and A. Rua (2007). Forecasting inflation through a bottom-up approach: How bottom is bottom? *Economic Modelling 24*(6), 941–953.

Duncan, G. T., W. L. Gorr, and J. Szczypula (2001). Forecasting analogous time series. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Boston, MA, pp. 195–213. Springer US.

Eklund, J. and S. Karlsson (2007). Forecast combination and model averaging using predictive measures. *Econometric Reviews 26*(2-4), 329–363.

Elliott, G. (2017). Forecast combination when outcomes are difficult to predict. *Empirical Economics 53*(1), 7–20.

Elliott, G. and A. Timmermann (2005). Optimal forecast combination under regime switching. *International Economic Review 46*(4), 1081–1102.

Espasa, A. and I. Mayo-Burgos (2013). Forecasting aggregates and disaggregates with common features. *International Journal of Forecasting 29*(4), 718–732.

Espasa, A. and E. Senra (2017). Twenty-two years of inflation assessment and forecasting experience at the Bulletin of EU & US Inflation and Macroeconomic Analysis. *Econometrics 5*(4), 44.

Espasa, A., E. Senra, and R. Albacete (2002). Forecasting inflation in the European Monetary Union: A disaggregated approach by countries and by sectors. *The European Journal of Finance 8*(4), 402–421.

Esteves, P. S. (2013). Direct vs bottom–up approach when forecasting GDP: Reconciling literature results with institutional practice. *Economic Modelling 33*, 416–420.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The generalized dynamic factor model. *Journal of the American Statistical Association 100*(471).

Frale, C., M. Marcellino, G. L. Mazzi, and T. Proietti (2011). EUROMIND: A monthly indicator of the euro area economic conditions. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 174*(2), 439–470.

Gao, Z. and J. Yang (2014). Financial time series forecasting with grouped predictors using Hierarchical Clustering and Support Vector Regression. *International Journal of Grid & Distributed Computing 7*(5), 53–64.

Geweke, J. (1977). The dynamic factor analysis of economic time series. *Latent variables in socio-economic models 1*.

Geweke, J. and G. Amisano (2010). Comparing and evaluating Bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*(2), 216–230.

Giannone, D., M. Lenza, D. Momferatou, and L. Onorante (2014). Short-term inflation projections: A Bayesian vector autoregressive approach. *International Journal of Forecasting 30*(3), 635–644.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

Gomez, V. and A. Maravall (1996). Programs TRAMO and SEATS. Instructions for the User. Working paper, Banco de Espana. 9628, Research Department, Bank of Spain.

González-Rivera, G. and Y. Sun (2017). Density forecast evaluation in unstable environments. *International Journal of Forecasting 33*(2), 416 – 432.

Goodhart, C. (2004). Gradualism in the adjustment of official interest rates: Some partial explanations. Financial Markets Group special paper. Technical Report 157, London School of Economics.

Granger, C. (1990). Aggregation of time series variables: a survey. In T. Barker and M. Pesaran (Eds.), *Disaggregation in Econometric Modelling*. Routledge.

Granger, C. W. (1987). Implications of aggregation with common factors. *Econometric Theory 3*(02), 208–222.

Granger, C. W. and R. Ramanathan (1984). Improved methods of combining forecasts. *Journal of Forecasting 3*(2), 197–204.

Grunfeld, Y. and Z. Griliches (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, 1–13.

Gupta, R. and A. Kabundi (2011). Forecasting macroeconomic variables using large datasets: dynamic factor model versus large-scale BVARs. *Indian Economic Review*, 23–40.

Hahn, E. and F. Skudelny (2008). Early estimates of Euro area real GDP growth: a bottom up approach from the production side. Working Paper Series 0975, European Central Bank.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*(1), 1–13.

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

Hargreaves, D., H. Kite, B. Hodgetts, et al. (2006). Modelling New Zealand inflation in a Phillips curve. *Reserve Bank of New Zealand Bulletin 69*(3), 23–37.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting 13*(2), 281–291.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (2nd Edition ed.). Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hauke, J. and T. Kossowski (2011). Comparison of values of Pearson's and Spearman's Correlation Coefficients on the same sets of data. *Quaestiones Geographicae 30*, 87–93.

Heller, K. A. and Z. Ghahramani (2005). Bayesian Hierarchical Clustering. In *Proceedings of the 22nd international conference on Machine learning*, pp. 297–304. ACM.

Hendry, D. F. and K. Hubrich (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business & Economic Statistics 29*(2).

Higgins, P. C. (2014). GDPNow: A Model for GDP "Nowcasting". FRB Atlanta Working Paper 2014-7, Federal Reserve Bank of Atlanta.

Hoogerheide, L., R. Kleijn, F. Ravazzolo, H. K. Van Dijk, and M. Verbeek (2010). Forecast accuracy and economic gains from Bayesian model averaging using time-varying weights. *Journal of Forecasting 29*(1-2), 251–269.

Hora, S. (2004). Probability judgements for continuous quantities: Linear combinations and calibration. *Management Science 50*, 597–604.

Hsiao, C. and S. K. Wan (2014). Is there an optimal forecast combination? *Journal of Econometrics 178*, 294–309.

Hubrich, K. (2005). Forecasting Euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting 21*(1), 119–136.

Hubrich, K. and F. Skudelny (2017). Forecast combination for Euro area inflation: a cure in times of crisis? *Journal of Forecasting 36*(5), 515–540.

Hyndman, R. J., R. A. Ahmed, G. Athanasopoulos, and H. L. Shang (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis 55*(9), 2579–2589.

Jacobs, D. and T. Williams (2014, September). The determinants of non-tradables inflation. *RBA Bulletin*, 27–38.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer. Chapter 10.

Jha, A., S. Ray, B. Seaman, and I. S. Dhillon (2015). Clustering to forecast sparse time-series data. In *2015 IEEE 31st International Conference on Data Engineering*, pp. 1388–1399.

Johnson, N. (2017). Tradable and nontradable inflation indexes: Replicating New Zealand's tradable indexes with BLS CPI data. *Monthly Labor Review 5*.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics 25*(4), 621–634.

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*(1), 33–41.

Koop, G. and D. Korobilis (2012). Large time-varying parameter VARs. Working Paper Series 11-12, The Rimini Centre for Economic Analysis.

Koop, G. and D. Korobilis (2013). Large time-varying parameter VARs. *Journal of Econometrics 177*(2), 185–198.

Koop, G. M. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics 28*(2), 177–203.

Lütkepohl, H. (1987). *Forecasting aggregated vector ARMA processes*, Volume 284. Springer Science & Business Media.

Lütkepohl, H. (2011). Forecasting nonlinear aggregates and aggregates with time-varying weights. *Jahrbücher für Nationalökonomie und Statistik*, 107–133.

Marcellino, M. (2008). A linear benchmark for forecasting GDP growth and inflation? *Journal of Forecasting 27*(4), 305–340.

Marcellino, M., J. H. Stock, and M. W. Watson (2003). Macroeconomic forecasting in the Euro area: Country specific versus area-wide information. *European Economic Review 47*(1), 1–18.

Mazur, B. (2015). Density forecasts based on disaggregate data: nowcasting Polish inflation. *Dynamic Econometric Models 15*, 71–87.

Mitchell, J. and S. G. Hall (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR fancharts of inflation. *Oxford bulletin of economics and statistics 67*(s1), 995–1033.

Mitchell, J. and K. F. Wallis (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics 26*(6), 1023–1040.

Mogliani, M., O. Darné, and B. Pluyaud (2017). The new MIBA model: Real-time nowcasting of French GDP using the Banque de France's monthly business survey. *Economic Modelling 64*(Supplement C), 26 – 39.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* MIT press.

Murtagh, F. and P. Contreras (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2*(1), 86–97.

Newbold, P. and D. I. Harvey (2002). Forecast combination and encompassing. In M. Clements and D. Hendry (Eds.), *A Companion to Economic Forecasting*, pp. 268–283. Blackwell Press: Oxford.

Pavia-Miralles, J. (2010). A survey of methods to interpolate, distribute and extrapolate time series. *Journal of Service Science and Management 3*(4), 449–463.

Peach, R. W., R. W. Rich, and M. H. Linder (2013). The parts are more than the whole: separating goods and services to predict core inflation. *Current Issues in Economics and Finance 19*(7).

Perevalov, N. and P. Maier (2010). On the advantages of disaggregated data: Insights from forecasting the US economy in a data-rich environment. Working Papers 10-10, Bank of Canada.

Pesaran, M. H., R. G. Pierse, and M. S. Kumar (1989). Econometric analysis of aggregation in the context of linear prediction models. *Econometrica: Journal of the Econometric Society*, 861–888.

Proietti, T., M. Marczak, and G. Mazzi (2017). Euromind-D: A density estimate of monthly Gross Domestic Product for the Euro area. *Journal of Applied Econometrics 32*(3), 683–703.

Raftery, A. E., M. Kárnỳ, and P. Ettler (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics 52*(1), 52–66.

Ranjan, R. and T. Gneiting (2010). Combining probability forecasts. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(1), 71–91.

Ravazzolo, F. and S. P. Vahey (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics 18*(4), 367–381.

Reeves, R. and M. Sawicki (2007). Do financial markets react to Bank of England communication? *European Journal of Political Economy 23*(1), 207–227.

Rodrigues, J. F. (2014). A Bayesian approach to the balancing of statistical economic data. *Entropy 16*(3), 1243–1271.

Rosa, C. (2013). The financial market effect of FOMC minutes. *FRBNY Economic Policy Review*.

Rossi, B. (2013). *Advances in Forecasting under Instability*, Volume 2 of *Handbook of Economic Forecasting*, Chapter 21, pp. 1203–1324. Elsevier.

Savage, R. S., K. Heller, Y. Xu, Z. Ghahramani, W. M. Truman, M. Grant, K. J. Denby, and D. L. Wild (2009). R/BHC: fast Bayesian Hierarchical Clustering for microarray data. *BMC Bioinformatics 10*, 242.

Smith, J. and K. F. Wallis (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics 71*(3), 331–355.

Stock, J. and M. Watson (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. in R.F. Engle and H. White, eds., Festschrift in Honour of Clive Granger (Cambridge University Press, Cambridge) 1-44.

Stock, J. H. and M. W. Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business & Economic Statistics 14*(1), 11–30.

Stock, J. H. and M. W. Watson (1998). Diffusion indexes. Working Paper 6702, NBER.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.

Stock, J. H. and M. W. Watson (2015). Core inflation and trend inflation. Technical report, National Bureau of Economic Research.

Tallman, E. W. and S. Zaman (2017). Forecasting inflation: Phillips curve effects on services price measures. *International Journal of Forecasting 33*(2), 442–457.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting 1*, 135–196.

Wei, X. and Y. Yang (2012). Robust forecast combinations. *Journal of Econometrics 166*(2), 224–236.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting 30*(4), 1030 – 1081.

Yan, J., Y. Liu, S. Han, and M. Qiu (2013). Wind power grouping forecasts and its uncertainty analysis using Optimized Relevance Vector Machine. *Renewable and Sustainable Energy Reviews 27*(C), 613–621.

Zellner, A. and J. Tobias (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting 19*(5), 457–465.