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

Cobb, Marcus P A

August 2018

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

Marcus P. A. Cobb*

August 2018

Abstract

In some situations forecasts for a number of sub-aggregations are required for analysis in addition to the aggregate itself. In this context, practitioners typically rely on bottom-up methods to produce a set of consistent forecasts in order to avoid conflicting messages. However, using this approach exclusively can mean that forecasting accuracy is negatively affected when compared to using other methods. This paper presents a method for increasing overall accuracy by jointly combining the forecasts for an aggregate, any sub-aggregations, and the components from any number of models and measurement approaches. 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 from the empirical application suggest that the method is successful in allowing the strengths of the better-performing approaches to contribute to increasing the performance of the rest.

Keywords: Bottom-up forecasting; Forecast combination; Hierarchical forecasting; Reconciling forecasts

JEL codes: C53, E27, E37

*This research was produced while studying at the School of Economics and Finance, Queen Mary University of London and the author acknowledges and is grateful for their financial support.

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 some additional information to the public. 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

[1]This is supported theoretically by Lütkepohl (1987) and 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.

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 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

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

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 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

[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 after.

[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).

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.

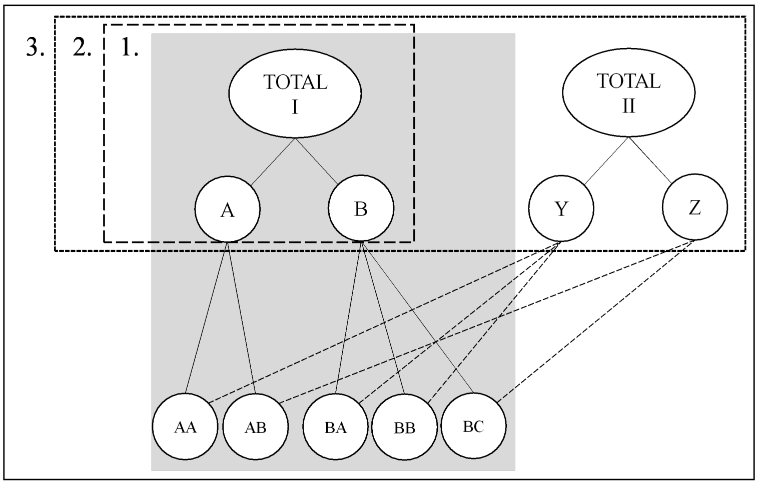
The rest of the chapter is organized as follows: Section 2 develops the framework that allows series from different levels of aggregation from any number of measurement approaches to be combined. Section 3 presents an empirical implementation using CPI data for France, Germany and the United Kingdom. Section 4 summarizes the conclusions.

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.

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 aggregate forecasts resulting 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 their initial estimates.

Figure 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).

To have a notion of the forecasting setting under consideration, Figure 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 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]

2.1 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

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

and Gysting, 2004). In a recent paper, Rodrigues (2014) cast the whole problem of balancing statistical economic data into a Bayesian framework. They suggest treating the data as stochastic processes, modelling their 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 their implementation is specific to balancing economic data, the principle behind their 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. They establish that a number of the conventional reconciliation methods are in fact particular cases of their general framework and show that there is a one-to-one correspondence. Based on this correspondence, they argue that it is possible to identify the conventional method's underlying assumptions and go on to suggest using least squares approaches when uncertainty estimates are available.

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. In particular 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 program of the form:

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

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]

2.1.2 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:

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

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 and 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 used. As pointed out by Pavia-Miralles (2010), this is one of the most successful methods in the area, 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.[7]

Using this loss function and minimizing it subject to the restriction that the aggregate has to be equal to the sum of the components, $(1+\alpha_t)y_t - \sum_{j=1}^{N} w_{j,t}(1+\beta_{j,t})q_{j,t}$, produces

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

[7] Their approach is more general in that it considers minimizing the h-th differences so as to allow for movement preservation if necessary.

a solution that the definitive aggregate forecast is:

$$\tilde{y}_t = \tilde{Q}_t = \frac{Q_t^2 + y_t \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)}{Q_t + \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \tag{3}$$

and the definitive forecast for any given component is:

$$\tilde{q}_{n,t} = \left(1 + \frac{\varphi_t}{\phi_{n,t}} \cdot \frac{y_t - Q_t}{Q_t + \sum_{j=1}^{N} \left(\frac{\varphi_t}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \right) q_{n,t} \tag{4}$$

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 (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 (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.[8]

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

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

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 component's 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:

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

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

$$\tilde{y} = \frac{y + \displaystyle\sum_{k=1}^{K}\left(Q_k \cdot \frac{Q_k}{\chi_k}\right)}{1 + \displaystyle\sum_{k=1}^{K}\frac{Q_k}{\chi_k}} \tag{5}$$

and the definitive forecast for any given component being:

$$\tilde{q}_{k,n} = \left(1 + \frac{\varphi}{\phi_{k,n}} \cdot \frac{\tilde{y} - Q_k}{\chi_k}\right) q_{k,n} \tag{6}$$

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.

2.1.4 Bounds for and Response to Reliability Weight Values

In this section the bounds for the weights are explored in order to establish the feasible region in which they guarantee a unique solution for the minimization problem. The sensitivity of the final outcome to the choice of weights is also explored.

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. As regards to finding a single solution, both extremes for the reliability of a forecast are examined.[9]

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. Theoretically all forecasts cannot be certain, but in practice the weights have to be given a finite number.

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 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 (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:

$$\tilde{y}_t = \left(1 + \frac{k}{k - (k-1)s_i}\right)^{-1} \left(\frac{k}{k - (k-1)s_i} Q_t + y_t\right) \tag{7}$$

[9]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 Multi-Level Hierarchies and Alternative Sub-aggregations

This section presents the general framework for multi-level combination. The method involves deriving for each forecast for the aggregate and sub-aggregations a set of consistent component forecasts and then combining them to produce a definitive bottom-up forecast. The method 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.[10]

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 (3) and (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:

$$\hat{q}_{n,t}^{(y)} \quad = \quad \left(1 + \frac{y_t - Q_t}{\phi_{n,t} \cdot \sum_{j=1}^{N} \left(\frac{1}{\phi_{j,t}} w_{j,t} q_{j,t} \right)} \right) q_{n,t} \tag{8}$$

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 the

[10] The derivation is shown in detail in section B of the Appendix,

original and y-consistent forecasts:

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

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 (8). That is:

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

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:

$$\tilde{q}_{n,t} = \left[1+\frac{1}{\phi_{n,t}+\sum\limits_{s=1}^{S}\varphi_{s,k,t}}\cdot\sum_{s=1}^{S}\left(\frac{\varphi_{s,k,t}}{\phi_{n,t}}\cdot\frac{y_{s,k,t}-Q_{s,k,t}}{\chi_{s,k,t}}\right)\right]q_{n,t} \tag{9}$$

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

$$\tilde{y}_t = \frac{\phi_t Q_t+\sum\limits_{s=1}^{S}\varphi_{s,t}Y_{s,t}}{\phi_t+\sum\limits_{s=1}^{S}\varphi_{s,t}} \tag{10}$$

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 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.

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.[11] 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.

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 1.

Table 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

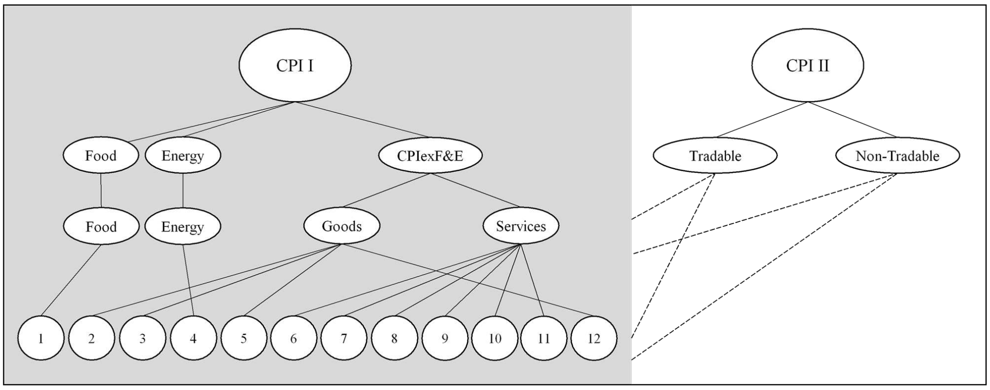
Regarding the sub-aggregations, three are chosen. The first two are in line with the extensive literature considering core measures for inflation. Aron and Muellbauer (2012) make a relatively extensive survey of studies that measure the benefits of removing certain components for forecasting, most of 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.[12] Taking all these factors into consideration, the

[11] This is only a pseudo real-time forecasting exercise given that historical data revisions and vintages are ignored.

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

Figure 2: Aggregation Structure of the Empirical Application



aggregation structure for the empirical application is presented in Figure 2.

3.2 Forecasting Models

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:

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

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 specifications are used. The two first sets of VARs include Gross Domestic Product (GDP) and the respective set of series: i.e. Headline CPI alone, and each of the three sub-aggregations and components separately. The first of the set of VARs is estimated with all variables in first differences. In the second, the variables are differentiated according to a unit root test.[13]

Following the notion in Hendry and Hubrich (2011), VARs that include all CPI series and GDP in the same model are also estimated. Similarly, the VARs are estimated in first differences and differentiated according to the tests.

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.

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

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.

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

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 as:

$$RMSPE_{i,u,p,h,v} = \sqrt{\frac{1}{v}\sum_{s=u-h-v}^{u-h}\left(\frac{x_{i,s+h}|p}{x_{i,s+h}} - 1\right)^2} \tag{11}$$

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:[14]

$$\omega_{i,t,h,v}^{ISP} = \frac{1}{RMSPE_{i,t,t,h,v}} \tag{12}$$

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 MSE 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:

$$\omega_{i,t,h,v}^{OSP} = \frac{1}{RMSPE_{i,t,s,h,v}} \tag{13}$$

[14] In the context of forecast combination using predictive measures, Eklund and Karlsson (2007) raise awareness regarding the possibility of distorting weight distribution due to overconfidence in models that over-fit data. Aiolfi and Favero (2005) for example use the model's R^2 to decide on the combination of forecasts.

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 (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.[15] As the weights from this method depend on the aggregation structure and not on the reliability of the forecasts, the combination weights do not change from one period to the next.

3.4 Forecasting Accuracy Evaluation

The forecasting accuracy is presented for different horizons by means of the model's mean square forecasting error (MSFE) relative to that of a benchmark model. That is, for variable i, horizon h and using model m, the relative MSFE is:

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

with

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

where $y^{(m)}_{i,t+h}|t$ 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 $m = 0$.

[15] The derivation is presented in section C of the Appendix

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:

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

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).[16]

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 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 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 being not 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.

[16]Original test proposed by Diebold and Mariano (1995).

Table 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: Aggregate 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. 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.

Table 3: Combination Aggregate Forecasting Error

Horizon	Aggregate 1	2	3	4	Multi-level 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.

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 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 also presented. 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 MSE 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, an approximation to the weighting scheme by Hyndman et al. (2011) is used.[17]

[17] The approximation is presented in section C of the Appendix.

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 multi-level approaches seem negligible.

As regards disaggregate accuracy, Table 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, that denote improvements over the best singe 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

Table 4: Cumulative Disaggregate Forecasting Error

	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.

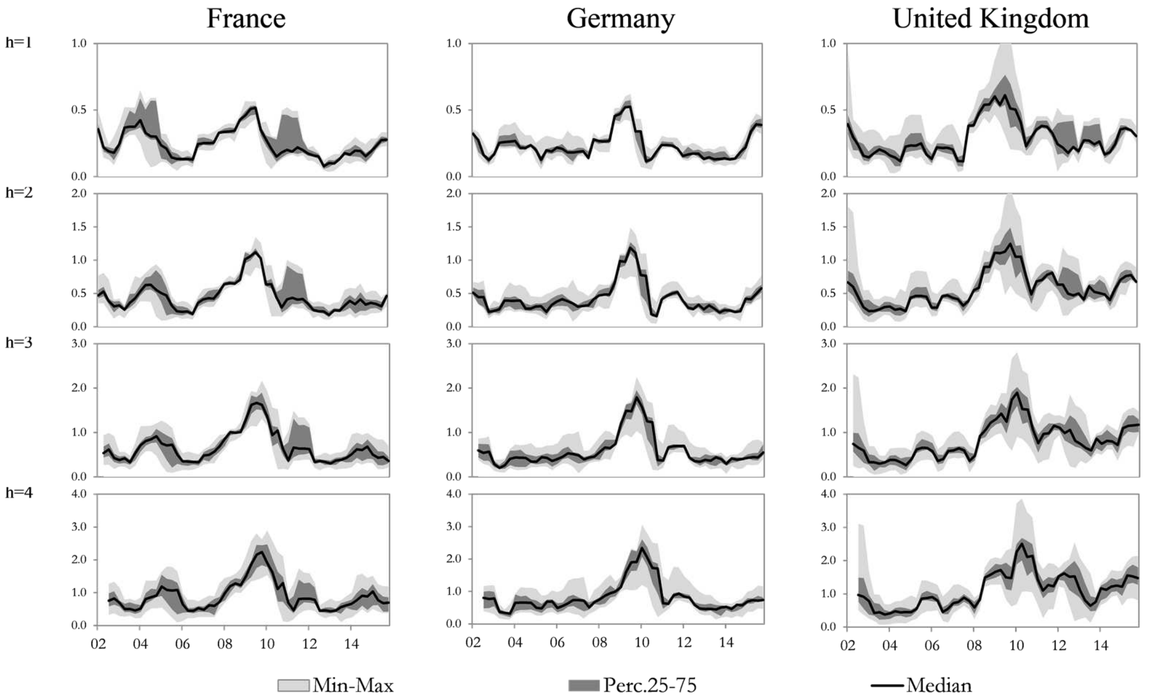
horizon. In terms of relative performance of the different weighting schemes, one result from the aggregate outcome that is also present at the disaggregate level for all three countries is that the in-sample weighting scheme comes out worst of all. For France, in fact, the differences with the best 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.

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 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 3. It presents the four-quarter rolling MSFE for the aggregate for all models

Figure 3: Dispersion of the Rolling Forecasting Error



Note: Four-quarter rolling mean square forecasting error (MSFE) 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.

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 or 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 favourable and unfavourable circumstance, 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 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 performed well in stable times failed completely with the increase of volatility and that models perform differently in expansions and recessions. This last point had been previously documented in Marcellino (2008) who finds that in recessions their more sophisticated models showed a marked deterioration, making the simple random walk the best performer. 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 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.[18]

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 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.

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.

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

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 bottom-up only 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

A Derivation for a One-level Combination

This section presents detailed derivation of the proposed method, including certain useful auxiliary results. The premise revolves around the desirable properties that the loss function is expected to fulfil, particularly that of the solution to the multi-level combination being equivalent to that of traditional single-variable methods if the conditions are comparable. The 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 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_t, comes from forecasting the aggregate directly, while the second one, Q_t, is the simple sum of the forecasts of its components $q_{n,t}$.

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:

$$\min_{\alpha,\beta} \left[(1+\alpha)\, y - y\right]^2 + \left[(1+\beta)\, Q - Q\right]^2 + 2\lambda \left[(1+\alpha)\, y - (1+\beta)\, Q\right] \tag{14}$$

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

$$(1+\alpha)y = Q - \alpha y \tag{15}$$

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

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

$$\min_{\alpha,\beta_n} (\alpha y)^2 + \textstyle\sum_{j=1}^{N} (\beta_j q_j)^2 + 2\lambda \left[(1+\alpha)y - \textstyle\sum_{j=1}^{N}(1+\beta_j)q_j\right] \tag{16}$$

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:

$$\tilde{y} = \frac{N \cdot y + \sum_{j=1}^{N} q_j}{N + 1} = \frac{1}{N + 1}\left(N \cdot y + Q\right) \tag{17}$$

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:

$$\min_{\alpha,\beta} \left[\frac{(1+\alpha)\, y - y}{y}\right]^2 + \left[\frac{(1+\beta)\, Q - Q}{Q}\right]^2 + 2\lambda \left[(1+\alpha)\, y - (1+\beta)\, Q\right] \tag{18}$$

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:

$$\tilde{y} = (1+\alpha)\, y = (1+\beta)\, Q = \left(\frac{y \cdot Q}{y^2 + Q^2}\right)(y + Q) \tag{19}$$

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 (16) results in:

$$\min_{\alpha,\beta_n} (\alpha y)^2 + \sum_{j=1}^{N} (\beta_j \eta_j)^2 + 2\lambda \left[(1+\alpha)y - \sum_{j=1}^{N} (1+\beta_j) q_j \right] \tag{20}$$

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 (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:

$$(\alpha y)^2 + \sum_{n=1}^{N} q_j Q \, (\beta_j)^2 \tag{21}$$

Result 4: Incorporating Multiple Component Forecasts

If more than one set of forecasts for the same components are included in equation (21), a bias similar to that of equation (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 (21) to a setting with D sets of disaggregate forecasts for the N components. The minimization problem may be written as:

$$\begin{aligned} &\min_{\alpha,\beta_n} (\alpha y)^2 + \textstyle\sum_{d=1}^{D} \sum_{j=1}^{N} q_{d,j} Q_d \beta_{d,j}^2 \\ &\quad + 2 \textstyle\sum_{d=1}^{D} \left(\lambda_d \left[(1+\alpha)y - \sum_{j=1}^{N} (1+\beta_{d,j}) q_{d,j} \right] \right) \\ &\quad +2 \textstyle\sum_{d=2}^{D} \sum_{j=1}^{N} \left(\delta_{d,j} \left[(1+\beta_{1,j}) q_{1,j} - (1+\beta_{d,j}) q_{d,j} \right] \right) \end{aligned} \tag{22}$$

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

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

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

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

$$\begin{aligned} \tilde{y} &= \Phi(Q_1 + Q_2 - \alpha y) \\ &= \tfrac{\Phi}{1+\Phi}(y + Q_1 + Q_2) \end{aligned} \tag{23}$$

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 (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 (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:

$$y_t = \frac{1}{\Gamma_t}\sum_{i=1}^{A}\gamma_{i,t}y_{i,t} \qquad \text{and} \qquad q_{n,t} = \frac{1}{\Delta_{n,t}}\sum_{d=1}^{D}\delta_{d,n,t}q_{d,n,t} \tag{24}$$

with $\gamma_{i,t}$ and $\delta_{d,n,t}$ being the reliability weights, $\Gamma_t = \sum_{i=1}^{A}\gamma_{i,t}$ and $\Delta_{n,t} = \sum_{d=1}^{D}\delta_{d,n,t}$.

The combination procedure remains unchanged except for the weights φ_t and $\phi_{n,t}$, which are set to reflect the reliability of the combined forecasts y_t and $q_{n,t}$ as opposed to the initial preliminary forecasts $y_{i,t}$ and $q_{d,n,t}$.

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 $\varphi_t = A$ and $\phi_{n,t} = D$ making the solution for the aggregate forecast:

$$\tilde{y}_t = \frac{1}{A+D}\left(A \cdot y_t + D \cdot \sum_{j=1}^{N} w_{j,t}q_{j,t}\right) \tag{25}$$

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

$$\begin{aligned}\tilde{y}_t &= \tfrac{1}{A+D}\left(A \cdot \tfrac{1}{A}\textstyle\sum_{i=1}^{A} y_{i,t} + D \cdot \sum_{j=1}^{N} \tfrac{1}{D} w_{j,t} \sum_{d=1}^{D} q_{d,j,t}\right) \\ &= \tfrac{1}{A+D}\left(\textstyle\sum_{i=1}^{A} y_{i,t} + \sum_{d=1}^{D}\sum_{j=1}^{N} w_{j,t}q_{d,j,t}\right)\end{aligned} \tag{26}$$

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:

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

The first order conditions are:

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

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:

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

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

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

Then using 1. and dividing by φ:

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

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:

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

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

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

$$\tilde{y} = \frac{y + \sum_{k=1}^{K} \left(Q_k \cdot \frac{Q_k}{\chi_k} \right)}{1 + \sum_{k=1}^{K} \frac{Q_k}{\chi_k}} \tag{27}$$

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

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

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

with the final result being:

$$\tilde{q}_{k,n} = \left(1 + \frac{\varphi}{\phi_{k,n}} \cdot \frac{\tilde{y} - Q_k}{\chi_k} \right) q_{k,n} \tag{28}$$

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 (27) and is:

$$\tilde{y} = \frac{Q^2 + y \sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j q_j}{Q + \sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j q_j} \tag{29}$$

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:

$$\tilde{y} = \frac{\left(\sum_{j=1}^{N} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)^2 + \left(\frac{1}{\Gamma} \sum_{i=1}^{A} \gamma_i y_i \right) \left(\sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)}{\left(\sum_{j=1}^{N} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right) + \left(\sum_{j=1}^{N} \frac{\varphi}{\phi_j} w_j \sum_{d=1}^{D} \frac{\delta_{d,j}}{\Delta_j} q_{d,j} \right)} \tag{30}$$

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 solutions 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.

B Foundation for the Multi-level Combination Method

This section presents the derivation of how to present any aggregate forecast in terms of a consistent set of component forecasts and how, in the case of equal weights, it is equivalent to combining the aggregate forecasts produced from different intermediate aggregation levels.

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:

$$\tilde{y} = \frac{Q^2 + y\sum_{j=1}^{N}\frac{\varphi}{\phi_j}w_j q_j}{Q + \sum_{j=1}^{N}\frac{\varphi}{\phi_j}w_j q_j} \tag{31}$$

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

$$\tilde{q}_n = \left(1 + \frac{\varphi}{\phi_n}\frac{y - Q}{Q + \sum_{j=1}^{N}\frac{\varphi}{\phi_j}w_j q_j}\right) q_n \tag{32}$$

With the objective of reconciling a set of components to an aggregate, equation (32) can be rewritten as follows:

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

Then, if the objective is 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 and setting the aggregate reliability to infinity, $\varphi \to \infty$, results in:

$$\hat{q}_n^{(y)} = \left(1 + \tfrac{y-Q}{\phi_n \cdot \sum_{j=1}^{N}\frac{1}{\phi_j}w_j q_j}\right) q_n \tag{33}$$

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

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

that is slightly different from $\tilde{q}_n$ in equation (32). For equal weights among components, however, that is $\phi_n = \phi$:

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

and by summing up the components the aggregate is:

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

that is the same that is obtained from setting $\phi_n = \phi$ for the standard result in equation (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 (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.

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

$$\mathbf{Y}_t = \mathbf{S}\mathbf{Y}_{K,t} \tag{34}$$

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 $\mathbf{Y}_{2,t} = \begin{bmatrix} y_{2,1,t} & y_{2,2,t} & y_{2,3,t} & y_{2,4,t} \end{bmatrix}'$, the vector for all observations would be $\mathbf{Y}_t = \begin{bmatrix} y_{0,t} & y_{1,1,t} & y_{1,2,t} & \mathbf{Y}'_{2,t} \end{bmatrix}'$ 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:

$$\widetilde{\mathbf{Y}}_h = \mathbf{S}\mathbf{P}\hat{\mathbf{Y}}_h \tag{35}$$

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

$$\widetilde{\mathbf{Y}}_h = \mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}'\hat{\mathbf{Y}}_h \tag{36}$$

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, 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 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 5 under the Final heading.

Table 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

D Additional Information from the Empirical Application

Table 6: Differentiation for Empirical Application and Sub-aggregation Distribution

	France	Germany	UK	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 7: 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°°	1.00°°	1.00°°	1.00	1.00°°	1.00°°	1.00°°	1.00°°	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°°	0.97°°	0.99°°	1.01
SVDIF	0.79	0.74	0.72	0.78	0.85	0.84°°	0.95°°	1.00°	0.92	0.85	0.85	0.85
SVDDIF	0.86	0.82	0.82	0.85	0.88	0.83°°	0.85°°	0.86°	0.93°	0.90	0.91	0.90
LVDIF	1.13°°	1.10°°	1.10°°	1.19°°	0.81	0.74	0.77°	0.81°	0.96°	0.95	0.99	1.04°°
LVDDIF	1.21°°	1.14°°	1.13°°	1.24°°	0.95°°	0.91°°	1.04°°	1.16°°	0.98°	1.01	1.03	1.07°°
Sub-agg.1												
RW	1.03°°	1.03°°	1.03°°	1.04	1.00°°	1.00°°	1.00°°	1.00°°	1.00	1.00	1.00	1.01
AR	0.85°°	0.78°°	0.75°°	0.82°°	0.84	0.76	0.76	0.79	0.94	0.91°°	0.95°°	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°°	0.82	0.81	0.87	0.88	0.82°°	0.85°°	0.87°	0.92°	0.89	0.93	0.94
LVDIF	1.16°°	1.13°°	1.13°°	1.22°°	0.80	0.74	0.76	0.81°	0.95°	0.96	1.00	1.07°°
LVDDIF	1.25°°	1.19°°	1.17°°	1.28°°	0.95°°	0.91°°	1.04°°	1.16°°	1.00°°	1.03	1.06	1.11°
Sub-agg.2												
RW	1.02°°	1.02°°	1.02°°	1.03	1.00°°	1.00°°	1.00°°	1.00°°	1.27	1.18	1.16	1.16
AR	0.81°°	0.74°°	0.73°	0.80	0.83	0.75	0.76	0.79	1.09	0.96°°	0.98°°	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°°	0.83	0.82	0.88	0.90	0.83°°	0.85°°	0.87°	1.10	0.95	0.95	0.94
LVDIF	1.18°°	1.13°°	1.14°°	1.23°°	0.80	0.74	0.77	0.81°	1.11°	1.02	1.01	1.08°°
LVDDIF	1.25°°	1.19°°	1.18°°	1.30°°	0.95°°	0.91°°	1.05°°	1.16°°	1.17°°	1.09°	1.07	1.10°°
Sub-agg.3												
RW	1.00°°	1.00°°	1.00°°	1.00	1.00°°	1.00°°	1.00°°	1.00°°	**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°°	0.86°°	0.73	0.83	0.85	0.81
LVDIF	1.13°°	1.10°°	1.10°°	1.19°°	0.80	0.74	0.76	0.81°	0.91°	0.94	0.97	1.00°
LVDDIF	1.20°°	1.14°°	1.14°°	1.24°°	0.95°°	0.90°°	1.04°°	1.16°°	0.96°°	1.06	1.08	1.09°°
Components												
RW	1.01°°	0.99°	0.98°	0.99	1.00°°	1.01°°	1.01°°	1.01°°	0.81	0.90	0.91	0.91°°
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°°	0.85	0.84	0.91	0.86	0.81°°	0.82°°	0.84	0.87	0.91	0.97	1.02
LVDIF	1.12°°	1.11°°	1.12°°	1.21°°	0.81	0.75	0.77	0.80°	0.87	0.93	0.99	1.04
LVDDIF	1.18°°	1.13°°	1.12°°	1.22°°	0.96°°	0.90°°	1.00°°	1.10°°	0.91°	1.00	1.04	1.10°

Note: Aggregate 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 Table **??**. 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). ° 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 excluding 2008 and 2009.

Table 8: Combination Aggregate Forecasting Error 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°°	0.89°	0.87	0.93	0.97°°	0.90°°	0.88	0.94
OSP	0.87°	0.79	0.75	0.80	0.89°	0.81	0.77	0.82
OPT	1.25°°	1.12°°	1.11°°	1.16°°	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°°	0.92°°	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. ° 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 excluding 2008 and 2009.

Table 9: Cumulative Disaggregate Forecasting Error 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.

References

Aiolfi, M. and C. A. Favero (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting 24*(4), 233–254.

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.

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.

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

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.

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

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.

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

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

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. 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.

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.

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.

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.

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.

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. 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.

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.

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.

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.

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.

Johnson, N. (2017, 05). Tradable and nontradable inflation indexes: Replicating new zealand's tradable indexes with bls cpi data. *2017*.

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.

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

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.

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.

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.

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

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

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 (1998). Diffusion indexes. Working Paper 6702, NBER.

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.

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