Forecasting Economic Aggregates Using Dynamic Component Grouping

Cobb, Marcus P A

September 2017

Forecasting Economic Aggregates Using Dynamic Component Grouping

Marcus P. A. Cobb*

September 2017

Abstract

In terms of aggregate accuracy, whether it is worth the effort of modelling a disaggregate process, instead of forecasting the aggregate directly, depends on the properties of the data. Forecasting the aggregate directly and forecasting each of the components separately, however, are not the only options. This paper develops a framework to forecast an aggregate that dynamically chooses groupings of components based on the properties of the data to benefit from both the advantages of aggregation and disaggregation. With this objective in mind, the dimension of the problem is reduced by selecting a subset of possible groupings through the use of agglomerative hierarchical clustering. The definitive forecast is then produced based on this subset. The results from an empirical application using CPI data for France, Germany and the UK suggest that the grouping methods can improve both aggregate and disaggregate accuracy.

Keywords: Forecasting economic aggregates; Bottom-up forecasting; Hierarchical forecasting; Hierarchical Clustering;

JEL codes: C38, C53, E37

*The author is grateful to Andrea Carriero and Marco Mariotti for their valuable comments and support. 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.

Non-technical Summary

When forecasting economic aggregates, practitioners are faced with many options even when only the level of disaggregation is considered. These include forecasting at the level of disaggregation that is required to answer a particular question, disaggregating further or forecasting at a more aggregate level and reconciling the lower levels of disaggregation if necessary. The usual argument behind using the components is that allowing for different specifications across disaggregate variables may capture more precisely the dynamics of a process that becomes too complex through aggregation. Favouring forecasting directly is that it would be less affected by disaggregate misspecification, data measurement error and structural breaks. Ultimately, whether it is better to forecast components together or separately depends on the particular forecasting models and data. An option to improve forecasting performance in this setting, is to work on the modelling and another is to look for data transformations that allow existing models to perform better. This paper presents a framework to do the latter.

Grouping components together can produce new series with characteristics that differ quite significantly from those of the originating series. In this context, it might be possible to find specific groupings that avoid the problems associated with disaggregate forecasting while still allowing for distinct disaggregate dynamics to be picked up in the process. With this objective we develop a two-stage method that combines statistical learning techniques and traditional economic forecasting evaluation. In the first stage, we use agglomerative hierarchical clustering to reduce the dimension of the problem by choosing a subset of feasible groupings based on the commonality among the different components. In the second stage, we try different selection procedures on the resulting hierarchy to produce the final aggregate forecast. These selection procedures include choosing a single grouping based on some criterion and combining the whole subset of groups.

The results from an empirical application using CPI data for France, Germany and the UK show that the grouping method can improve overall accuracy. The results show that some of the methods that selected a unique grouping performed better than the best performing non-grouping method, both in terms of aggregate and disaggregate accuracy. They also show that the forecast combination methods performed well overall. This suggests that expanding the pool of forecasts by trying different combinations of components with the same forecasting approach may have a similar effect to that of expanding the pool by trying different models.

1 Introduction

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

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

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

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

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

An option to improve forecasting performance in this setting, is to work on the modelling, like Hendry and Hubrich (2011) that include disaggregate information in a direct

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

aggregate approach or Bermingham and D'Agostino (2014) that include common factors in a bottom-up approach. Another less obvious way, is to look for data transformations that allow existing models to perform better.

As mentioned before, adding components together results in new series with characteristics that may differ quite significantly from those of the originating ones. In this context, it may be possible to purposefully find specific groupings that show more desirable properties than those of the individual components and the aggregate.

Some authors have proposed using purpose-built groupings to increase overall forecasting accuracy, but it would seem that, at least in economic forecasting, it has had little impact (Duncan et al., 2001). A reason for this may be that the number of possible groupings grows exponentially with the number of components meaning that traditional methods, that would usually rely on evaluating all possible outcomes, are really only usable for problems with relatively few components.[2] For larger problems, a different approach becomes necessary.

One that has been relatively successful recently, particularly given the increase in popularity of methods for Big Data, is one that performs grouping conditional on some feature of the original data. These have been in use for a while in the context of electricity price forecasting (Weron, 2014) and, with the relatively recent surge in computational power, computer intensive methods and availability of high-frequency data, they have expanded to other areas of research. For example, Yan et al. (2013) report significant improvements in the context of wind power prediction, Jha et al. (2015) for inventory planning in retail and Gao and Yang (2014) for forecasting stock market returns.

The success of these methods, however, depends on the chosen feature being useful in obtaining the desired outcome. The assumption upon which many of these models are built on, is that by grouping series that behave in a similar way, the idiosyncratic errors within groups will tend to offset each other while the more relevant individual dynamics will be retained to be modelled.

Although these problems are set in a different context, the purpose of the methods are very similar to those of grouping components to increase the forecasting accuracy of an economic aggregate. They belong, however, to an area of research of statistical learning that has focused almost exclusively on extracting information from very large datasets. Many relevant economic aggregates, like GDP and CPI, do not fall in this category and it is unclear whether these methods will work with relatively small samples.

[2] With three components the feasible set is five: the aggregate, full disaggregation and three options where one component is forecasted on its own and the other two together. With four components the possibilities grow to fifteen and with five components to 52.

In this context, we develop a method to forecast economic aggregates based on purpose built groupings of components using statistical learning techniques. The two-stage method consists of trying to find the grouping of components at each point in time that produces the best aggregate forecast. In the first stage, we use agglomerative hierarchical clustering to reduce the dimension of the problem and, in the second, we use a selection procedure on the resulting hierarchy to produce the final aggregate forecast.

The rest of the paper is organized as follows. Section 2 presents the component grouping framework. Section 3 presents an empirical implementation using CPI data for France, Germany and the United Kingdom. Section 4 summarizes the conclusions.

2 A purpose driven grouping framework for aggregate forecasting

As pointed out by James et al. (2013), Statistical Learning refers to a broad set of tools for understanding data. These include some approaches that are intended for prediction among other objectives. They usually require computing the input and output for each event which may be undesirable in problems that are very large. Other methods try to learn relationships and structure from a dataset without a clear objective. They work directly and produce results based on the features of the original data and require, therefore, significantly less computation. The challenge of using these methods lies in tuning the algorithms so that they achieve a desired purpose.

Although the implementations and techniques differ, the assumption on which many of the models intended to forecast time-series are built on, is that forecasting series that behave similarly as a group will tend to produce more accurate aggregate forecasts than if they are modelled separately. This assumption would also seem reasonable within the context of forecasting economic aggregates, given that the relevant literature shows that accounting for commonality among components is key to forecasting accuracy and, in particular, that ignoring it would be detrimental for the bottom-up approach (Duarte and Rua, 2007; Espasa and Mayo-Burgos, 2013; Bermingham and D'Agostino, 2014).[3]

Regarding the method that performs the grouping, within the area of unsupervised learning there are many.[4] One that seems well suited for the particular setting is Hierarchical Clustering. The method is concerned with discovering unknown subgroups in

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

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

data. The most commonly used method is the agglomerative alternative, that starts with a set of groups, or clusters, that contain a single element each and proceeds by grouping the data into fewer units with more elements each.[5] The only thing the algorithm needs to work is some sort of dissimilarity measure between each pair of observations and then one for each cluster that is formed. For the fused clusters, those other than those containing a single original observation, typically the dissimilarity measures are calculated from the original dissimilarity measures following a procedure referred to as linkage. The result of running the algorithm is always a hierarchical structure that has exactly as many levels as the number of initial components, with the individual components as the lowest level and the aggregate as the highest. In the context of grouping for forecasting, this means that the direct aggregate and bottom-up approaches are always available as options to be chosen to produce the definitive forecast.

At first sight, it could seem that hierarchical clustering might be the solution to the grouping problem. However, the method provides no guidance on whether the groupings in the structure are meaningful nor if one grouping is better than another in any particular sense (Murphy, 2012).[6] This could be seen as a drawback, but, in the context of forecasting the economic aggregate, it might work out as an advantage.

The problem with identifying an appropriate grouping right away, is that, even if there is one, the particular dissimilarity threshold below which components should be grouped so as to obtain the most accurate aggregate forecast is unknown. By narrowing down the set of groupings, however, the clustering process reduces the initial problem to a manageable size that can then be tackled with evaluation methods that are common in the traditional forecasting literature.

In what follows, we present a two-stage grouping framework to forecast economic aggregates, that consists of defining the hierarchy, based on the commonality among components, and then choosing how to produce the definitive aggregate forecast based on that hierarchy.

2.1 Guided selection of a subset of groupings

Dissimilarity measures and linkage methods have a defining impact on the results and the relevant literature provides many alternatives to choose from. As James et al. (2013) point out, the choice of what alternative to use depends on the type of data and question at hand.

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

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

In the statistical learning literature it is not unusual to use simple correlation as the dissimilarity measure for time-series. The forecasting literature, however, points towards the notion of commonality. The problem is that there is not a unique way of measure it. For this reason we present six different possibilities based on what has been suggested in the literature.

All but one of the measures are used within the context of the traditional hierarchical clustering approach that is deterministic. The exception is set within a probabilistic framework. In nature they are very similar given that both have a hierarchy as the outcome. The fundamental difference is that the more common deterministic method needs to be provided with dissimilarity measures. The probabilistic method, on the other hand, works out the dissimilarity from the data itself. It therefore makes sense to present them separately.

2.1.1 Deterministic grouping algorithm

The implementations of deterministic agglomerative hierarchical clustering are relatively simple.[7] In the context of an aggregate with n components, the algorithm proceeds by calculating the pairwise commonality between the n series and aggregating the two with the highest commonality. This leaves $n-1$ series. The traditional approach would involve calculating the pairwise commonality of the new cluster with the remaining components using a particular linkage method and proceed to aggregate the next two series with the highest commonality. The process is repeated until only the aggregate is left.

In a departure from the standard clustering algorithm, for our implementation, at each step, we calculate the pairwise commonality between the newly formed cluster and the remaining components by computing the dissimilarity measures between the new series instead of using linkage.[8] This makes the approach slower, but, by not using a linkage method, it does not make any assumptions regarding how the commonality transmits from the components to the aggregate.

For the dissimilarity measures, five measures are evaluated:

Pearson's Correlation

In the machine learning literature there are many alternatives, but in the context of time-series the most obvious are measures for correlation. Probably the best known is Pearson's correlation coefficient that measures the strength of the linear relationship

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

[8] Proceeding in this way is equivalent to restarting the traditional algorithm after every fusion.

between two variables. Although its limitations are many, its widespread use make it an obvious benchmark for the rest of the measures.

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

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

Spearman's Correlation

As pointed out by Hauke and Kossowski (2011), sometimes the Pearson's correlation coefficient can produce results that are undesirable or misleading. This can be a result of being restricted to linearity or requiring variables to be measured on interval scales.

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

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

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

Latent factor

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

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

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

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

Persistence

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

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

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

Forecast-error clustering

Bermingham and D'Agostino (2014) also state that ignoring the common factor and interdependencies will tend to make forecasting errors cluster instead of cancelling out.

Having this phenomenon in mind, we again fit AR(1) models to each component but this time we use as the dissimilarity measure the correlations of the out-of-sample forecasting errors for the most recent periods.

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

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

2.1.2 Probabilistic grouping algorithm

As pointed out by Murphy (2012), it would be desirable for a clustering method to provide some insight into the quality of the groupings. However, as traditional clustering methods are deterministic, this is not possible. Probabilistic algorithms have been proposed, but until recently their increased complexity have hindered their implementation.

One that does compare favourably to the traditional methods is the Bayesian Hierarchical Clustering method by Heller and Ghahramani (2005). The main idea, is that, through empirical Bayesian methods, it performs the grouping based on the probability of two observations being generated from the same underlying function.

The essence of the method can be seen from the explanation in Murphy (2012).[9] Let $D = \{x_1, \ldots, x_n\}$ represent all the data and D_i the data at subtree T_i. Then, at each step, subtrees T_i and T_j are compared to see if they should be merged together. The hypothesis to be evaluated, is that x_i and x_j come from the same probabilistic model $p(x \mid \theta)$ of unknown parameters θ. Then define D_{ij} as the merged data, and let M_{ij} equal one if they should be merged and zero if they should not. The probability of a merge is given by

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

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

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

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

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

With this, all the elements to build the hierarchy are available.

The algorithm starts with each observation in its own cluster. It calculates all the pairwise merge probabilities and proceeds to merge the clusters with the highest posterior merge probability. It then recalculates the pairwise merge probabilities. It continues in

[9] A complete description can be found in Savage et al. (2009).

this way, merging the pairs with the highest merge probability until only the aggregate is left.

The method is developed for cross-section, but Cooke et al. (2011) extend it to time-series in the context of gene expression measurement. Through the introduction of Gaussian process regression, an equivalent grouping process is performed based on the probability of two observations having the same latent function.

2.2 Producing a unique aggregate forecast

The outcome from the clustering algorithm is a complete hierarchy and because of the way the algorithm works it will offer a number of levels of aggregation equal to the number of original components. As the hierarchical clustering proceeds by fusing two observations or series at a time, it produces an intuitive tree-based representation of the final structure. This representation is called a dendrogram. Figure 1 shows two different examples for twelve components. At the bottom are all the individual elements. Moving up some of the elements are paired with similar observations producing a number of clusters. Higher up, the clusters themselves fuse, either with single elements or other clusters.

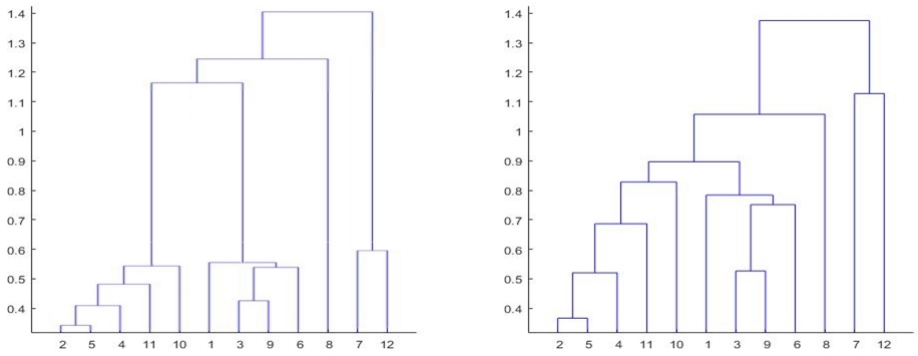
As mentioned before, the algorithm by itself does not provide any advice with regards to what grouping to use.[10] On the dendrogram, however, the vertical axis presents the level of dissimilarity and therefore visual inspection can provide some guidance.

Choosing a grouping based on some specific dissimilarity level is equivalent to drawing a horizontal line across the dendrogram at that desired level and using the groupings that are formed below that line. In Figure 1, for example, the dendrogram on the left suggests that there are four distinct groups based on the distance between the fusions. This, because the four groups form relatively close to the bottom and are only fused again relatively near to the top. More often than not, however, visual inspection is not enough to learn appropriate groupings (Murphy, 2012; James et al., 2013). That is, it is not uncommon that no obvious cutting points are revealed. The hierarchy depicted on the right of Figure 1, serves as an example. In these cases it is necessary turn to an exogenous criterion.

For this purpose, we present six different alternatives separating the methods in those that seek to select a single level of disaggregation and those that use a combination of the different groupings.

[10] This is the case for the traditional deterministic approach.

Figure 1: An example of dendrograms



2.2.1 Disaggregation level selection

In-sample fit

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

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

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

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

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

Past out-of-sample forecasting performance

One of the drawbacks of the in-sample criteria is that it will tend to over-fit the data. Therefore, it is very common to also use out-of-sample evaluation. For our case, the out-of-sample criterion, for forecasting period $t+1$, is calculated using a recursive out-of-sample forecasting exercise. That is, for each level of aggregation within the proposed hierarchy at time t, we estimate the parameters with data up to period $t-v$ and forecast $t-v+1$, then estimate the parameters with data up to period $t-v+1$ and forecast

$t - v + 2$ and continue in the same way stopping with the forecast for period t. Then, we calculate the RMSFE using these forecasts.

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

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

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

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

Lowest average error dissimilarity threshold

Unsupervised learning, of which the clustering method used to produce the subset of groups is part of, is often challenging because there is no response variable. In our context, however, the ultimate objective is to find the level of aggregation at which the resulting aggregate forecast error is lowest. For this purpose, we can use a supervised method to try to learn the best grouping for the purpose of forecasting. We do this by relating the degree of commonality, as measured by the corresponding dissimilarity measure, with the forecasting error.

The way in which we do this is by calculating for the training sample the average forecasting error conditional on the level of dissimilarity. This corresponds to calculating the forecasting error associated with the values on the vertical axis of all the dendrograms for the sample up to period t and averaging the results.[11] To make the averaging over different periods possible, we use a smoothing spline to interpolate the forecasting errors for each period. To forecast period $t + 1$ we choose the level of aggregation associated with the dissimilarity that is closest to the minimum average error.

Probabilistic criterion

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

[11] On the dendrogram, the height of the first fusion of any two observations indicates how different the two observations are. Observations that fuse at the very bottom are quite similar to each other, whereas observations that fuse close to the top will tend to be quite different.

2.2.2 Disaggregation level averaging

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

Equal-weights among aggregate forecasts

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

Equal-weights among distinct forecasts

In this context, however, averaging the aggregates is not the same as assigning equal-weights to each distinct forecast. To see why, it is helpful to look back at the dendrograms in Figure 1. On the one on the rights, the last-but-one fusion of the algorithm involves components 7 and 12. If the forecasts are generate independently of each other, for all of the groupings below their fusion, the aggregate forecast involves including the forecast for these two individual components. Then, when all aggregate forecasts are averaged, the forecast for both components are implicitly given a weight that is ten times larger than the forecasts of the components that are fused in the first step.[12]

An alternative approach is to give equal weights to each unique forecast. That means only including each individual component forecast, each intermediate aggregate forecast and the aggregate forecast once.[13]

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

[13] To do this it is necessary to combine forecasts from multiple levels of aggregation and we do so by extending the method for combining two different aggregation levels proposed in Cobb (2017). This is presented in the Appendix in section A.1.

3 Empirical Application

As an empirical application of the method we perform a forecasting exercise using CPI data from France, Germany and the United Kingdom. We use univariate autoregressive and Bayesian multivariate methods to forecast the data and evaluate the aggregate and overall forecasting accuracy of the grouping procedure by comparing the results with that of the direct forecast and that of the corresponding bottom-up approach[14].

3.1 Data

For the exercise we use the CPI data for France, Germany and the United Kingdom disaggregated to twelve components. The data is quarterly and seasonally adjusted, spanning from 1991 to 2015 and available from the OECD statistics database.[15]

The breakdown of the aggregate is the following:

Table 1: Components Breakdown

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

3.2 Forecasting models

Autoregressive model of order one (AR1)

Many of the aggregate-disaggregate forecasting competitions mentioned in the literature review use univariate autoregressive methods and therefore we do so too. Regardless of the numerous developments in econometric modelling, they continue to perform well (Marcellino, 2008). In particular, we use an autoregressive model of order one,

[14] That is, we compare the improvement of the grouping against the corresponding direct and bottom-up approach as opposed to finding the best aggregation from the pool of alternatives for both AR(1)'s and BVAR's.

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

$x_{i,t} = a_i + \rho_i x_{i,t-1} + \epsilon_{i,t}$, for the variables made stationary through differentiation according to unit root tests.[16] The forecasts are then produced using:

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

Bayesian VAR (BVAR)

We do acknowledge, however, that interdependencies among components could play an important role, so we also use Bayesian Vector Autoregressive models (BVARs) following the implementation in Banbura et al. (2010). In practice, we forecast the whole multivariate process using five lags and the choice of overall tightness, as in Banbura et al. (2010), that produces the same in-sample of that of the direct aggregate forecast.

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

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

3.3 Forecasting Accuracy Comparison

3.3.1 Set-up of the Evaluation Exercise

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

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

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

with

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

[16] The differentiation for each series is presented in section B.1 of the Appendix

where $y_{i,t+1}^{(m)}|t$ is the forecasted value for $t+1$ at time t and T_0 is the last period of actual data in the first sample used for the evaluation and T_1 is the last period of actual data in the last sample. As usual a RelMSFE lower than one reflects an improvement over the benchmark model for which $m = 0$. To evaluate the significance of these differences, we compare the forecasts using the modified Diebold-Mariano test for equality of prediction mean squared errors proposed by Harvey et al. (1997).[17]

Regarding measuring the overall forecasting accuracy of the components we do so by comparing the cumulative absolute errors in the contribution to the aggregate level. For this purpose we define the cumulative absolute root mean square forecasting error for an aggregate with N components q_n and using model m as

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

where $q_{n,t+1}^{(m)}|t$ is the forecasted value for $t+1$ at time t and T_0 is the last period of actual data in the first sample used for the evaluation and T_1 is the last period of actual data in the last sample.

3.3.2 Benchmark forecasting approaches

The objective of the whole exercise is to evaluate if there are successions of intermediate aggregations that can improve overall forecasting accuracy as opposed to restricting oneself only to using either the direct or the full bottom-up approach. These two approaches are, therefore, the obvious comparison points.

We also acknowledge that Bermingham and D'Agostino (2014) find that the performance from the bottom-up approach could improve if the common features among components are accounted for. To see how our application measures up to an alternative approach we also compare it to a factor augmented autoregressive model. Following their implementation, we extend each univariate autoregressive model from the bottom-up approach to include one factor

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

The factor, F, is estimated with the first principal component following Stock and Watson (2002) and computed over all components. The corresponding forecast for each

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

Table 2: Benchmark Forecasting Performance

	France	Germany	UK
Bottom-Up AR(1)	0.91	0.95	0.88
Bottom-Up BVAR	0.95	0.94	1.17
Factor augmented AR(1)	0.91	0.98	0.88

Note: Root mean squared forecasting error relative to the direct method. * and ** denote significance of the forecasting performance difference based on the modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015.

component is generated using

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

3.4 Results

3.4.1 Forecasting Performance Comparison

A first step to look at the results of the grouping methods is to evaluate how the benchmark models perform. In particular, Table 2 shows what would be a traditional aggregate-disaggregate comparison for the three series by presenting the root mean squared forecasting error of the direct and bottom-up approaches. It also presents the results for the factor augmented AR models to have a notion of whether the suggestion by Bermingham and D'Agostino (2014) can improve the univariate bottom-up methods in these particular settings.

We see that in five out of six of the cases the respective bottom-up approach performs better than the direct approach. In particular, the univariate approach tends to do better than the BVARs with improvements going from 5 to 12%, while the BVAR's improve for France and Germany, about 5%, but do quite a bit worse than the direct method for the UK. In regards to the factor augmented AR, it does not seem to give any advantage to the simple AR. Although some of the differences could seem quite large, it is worth noting that they are not statistically significant.

Moving on to the grouping framework, Table 3 presents the root mean squared forecasting error of the grouping methods relative to the direct approach for the three countries. The first thing that can be said from an overall assessment is that they are heterogeneous among series, dissimilarity criteria and choice methods. In many cases, the grouping methods improve over the best non-grouping method but, although the maximum gain is 13%, in most cases the improvement rarely goes over 5%. Of those that do not improve over the best non-grouping method, most lie somewhere between

Table 3: Relative Forecasting Errors

	AR(1)						BVAR					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
France												
Pearson corr.	0.92	0.96	0.92*		**0.89****	0.92**	1.01	0.98	0.96		**0.89****	**0.91****
Spearman corr.	**0.91**	**0.91**	0.98		**0.87****	**0.90****	1.06	1.09	0.99		**0.88****	**0.89****
1st princ.comp	0.96	0.96	0.99		0.92**	0.93**	1.00	1.03	0.98		**0.93***	**0.92****
persistence	0.91	0.93	**0.90****		**0.90***	**0.90****	1.04	0.98	**0.90****		**0.94**	**0.92****
f-error clustering	0.92	0.95	**0.94***		**0.88****	**0.91****	1.04	1.08	**0.93**		**0.92***	**0.94***
Bayesian	**0.89***	0.93	0.92	1.02	0.92	0.94	1.00	1.00	0.98	1.03	**0.95**	**0.95**
Germany												
Pearson corr.	0.98	1.02	1.00		0.99	0.98	1.05	1.11	1.06		1.00	0.99
Spearman corr.	0.98	1.01	1.02		0.99	0.98**	1.06	1.12	1.05		1.00	0.98
1st princ.comp	0.99	1.01	1.01		0.99	0.99	1.05	1.01	1.04		1.00	0.99
persistance	0.97	0.97	**0.89****		**0.93****	**0.94****	1.07	1.02	0.96**		0.97	0.96
f-error clustering	0.97	1.00	0.98		0.98	0.98*	1.14	1.08	1.00		1.01	0.98
Bayesian	0.98	0.99	0.96	1.00	0.96*	0.97*	0.98	1.08	0.95	1.02	0.97	0.97
UK												
Pearson corr.	0.90	0.90	0.95		0.88	**0.86****	**0.91**	**0.88**	**0.93**		**0.95**	**0.90**
Spearman corr.	0.89	0.95	**0.87**		0.90	0.89*	1.00	**0.91**	1.00		**0.98**	**0.91**
1st princ.comp	**0.86**	0.94	**0.86**		**0.86***	**0.88***	**0.91**	**0.90**	**0.88***		1.01	**0.99**
persistance	0.94	0.94	1.00		0.94	0.90	1.00	**0.99**	**0.88***		1.01	**0.99**
f-error clustering	0.96	0.99	**0.86**		0.89	**0.86****	**0.96**	1.00	1.04		**0.94**	**0.91**
Bayesian	**0.86**	0.91	**0.88**	1.11	0.89*	0.90*	**0.87**	**0.94**	1.16	1.18	**0.95**	**0.95**

Note: Root mean squared forecasting error relative to the direct method. Grouping method dissimilarity measures: Pearson correlation, Spearman correlation, Variance explained by the first principal component, similarity in persistance measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic criterion, Forecast Combination method 1 and Forecast Combination method 2. In bold RMSFE lower than the lowest of either the respective full Bottom-Up approach or the direct approach. * and ** denote significance of the forecasting performance difference based on the modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015.

the direct and full bottom-up approaches, but in some cases the performance is worse than that of either non-grouping methods.

If we go into the details, we find that for France the forecast combination choice methods perform well overall. They provide improvements for most dissimilarity measurement choices and, although not necessarily large in magnitude, these improvements are statistically significant. In regards to the other choice methods, the coupling of the persistence dissimilarity measure and the dissimilarity threshold choice method performs well. All this is true for both the AR and BVARs. A difference, however, arises for the other choice methods between the forecasting models. For the AR all but the probabilistic choice improve on the direct method, while for the BVAR many methods do worse.

For Germany, the assessment is rather different. Few methods improve on the best non-grouping method and many are worse than either the direct or bottom-up approaches. However, even if the overall performance is poor, the forecast combination choice methods still perform better than most of the alternative methods that goes to show their robustness. The exception to this poor performance are the methods that use the persistence dissimilarity measure where some statistically significant improvements are obtained. Again, the dissimilarity threshold choice method performs well. Regarding differences between the forecasting models, for the BVARs most methods perform worse than the direct approach .

For the UK the outcome for the two forecasting models is quite different so it is worth looking at them separately. First, the results for the ARs look similar to the previous cases. The magnitudes of the gains in accuracy are relatively small, but again the forecast combination choice methods produce statistically significant improvements. However, in this case the dissimilarity threshold choice method performs well with all dissimilarity measure choices except the one using persistence as the dissimilarity measure. For the BVARs, on the other hand, there are many methods that show larger gains, around 10% over the direct method. The combination of the persistence dissimilarity measure and the dissimilarity threshold choice method again shows improvements that are statistically significant, but, in this case, many of the other dissimilarity measure choices also show relevant improvements for one or more choice methods.

From the results that are common among the different cases we can draw some overall conclusions. One is that the forecast combination choice methods performed well with most dissimilarity measure choices and, in particular, in most cases the improvements were statistically significant. The other is that the persistence dissimilarity measure combined with the dissimilarity threshold choice method performed best overall.

Table 4: Relative Performance of Grouping Methods

	Average Percentage Deviation From Best Method						Average Rank Difference With Best Method					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
Pearson corr.	7.0	8.5	8.3		**4.4**	**3.8**	15.7	19.8	19.2		**9.7**	**8.7**
Spearman corr.	9.4	11.1	9.5		4.8	**3.5**	18.0	22.2	21.0		11.2	**7.0**
1st princ.comp	7.2	8.8	7.1		6.2	6.1	16.2	20.5	16.0		13.7	14.0
persistance	9.9	8.4	**3.3**		5.8	4.7	18.8	17.3	**6.7**		11.2	9.2
f-error clustering	11.0	12.7	6.8		4.7	**4.1**	20.7	26.0	13.7		9.8	**9.0**
Bayesian	**4.1**	8.5	8.6	17.0	5.2	5.8	**7.3**	20.2	12.0	25.8	11.3	13.3

Note: Relative performance of the grouping methods as measured by the average deviation of the respective root mean squared forecasting error (RMSFE) relative to that of the best performing grouping method by category and as the average difference in rank according to RMSFE over the six sets of forecasts. Grouping method dissimilarity measures: Pearson correlation, Spearman correlation, Variance explained by the first principal component, similarity in persistance measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion, Forecast Combination method 1 and Forecast Combination method 2. In bold the best performers in each category. Calculated over 2001-2015.

To evaluate these findings, Table 4 presents the relative performance of the 31 grouping methods for the two forecasting models and three countries.[18] Two summarizing measures are presented. The first calculates the average over all six sets of forecasts of the deviation of the respective root mean squared forecasting error (RMSFE) from that of the best overall performing grouping method. The second, calculates the average difference in rank of the grouping methods, where the most accurate, in the RMSFE sense is ranked first and the least accurate is ranked last, 31st in this case. For both measures a smaller number means a more accurate model.

Both measures support the assessment made in the previous paragraphs. The method based on the persistence dissimilarity measure and the dissimilarity threshold choice criterion comes out best overall. Also, the forecast combination choice method performed better for all dissimilarity measure criteria, particularly the combination approach that gives equal weight to each distinct forecast. Both measures, however, also point to the good performance of the combination of the Bayesian Hierarchical Clustering and the in-sample choice criterion, something that is not obvious at first sight from Table 3.

Regarding the accuracy of the components, Table 5 presents the median, minimum and maximum cumulative errors for each choice method relative to those of the bottom-up approach.[19] For the first five sets of forecasts there is little difference between the cumulative forecasting errors of the grouping methods and the non-grouping methods and, in fact, some look marginally worse. On the contrary, for the case of the BVAR for the UK data, that happens to be the only case where the direct approach beats the bottom-up approach, the cumulative errors are reduced by as much as to 11% depending

[18] Results conditional on dissimilarity and choice methods are found in the Appendix in section B.2.
[19] The full results are presented in the section B.3 in the Appendix.

Table 5: Relative Cumulative Forecasting Errors

	AR(1)					BVAR				
	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC
France										
Median	1.00	1.01	1.01	-	1.02	1.01	1.02	1.02	-	1.01
Min	1.00	**0.99**	1.00	1.00	1.01	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Max	1.04	1.03	1.04	1.01	1.05	1.04	1.04	1.04	1.01	1.05
Germany										
Median	1.01	1.02	1.01	-	1.03	1.02	1.02	1.03	-	1.04
Min	1.01	1.01	1.00	1.02	1.02	1.01	1.01	1.00	1.03	1.03
Max	1.02	1.02	1.04	1.03	1.04	1.04	1.03	1.06	1.04	1.06
UK										
Median	1.02	1.05	1.01	-	1.09	**0.97**	**0.95**	**0.98**	**-**	**0.94**
Min	1.00	1.01	1.00	1.06	1.06	**0.90**	**0.90**	**0.90**	**0.93**	**0.89**
Max	1.03	1.06	1.07	1.07	1.11	**0.99**	**0.98**	1.00	**0.95**	**0.97**

Note: Cumulative root mean squared forecasting error relative to the direct method. Median, minimum and maximum values obtained from all grouping method dissimilarity measures and multilevel forecast combination methods. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion. In bold CumRMSFE lower than that of the respective full Bottom-Up approach. Calculated over the 2001-2015.

on the grouping and choice method.

All this suggests that the grouping methods can improve overall accuracy. However, no dissimilarity measure for grouping nor aggregation level choice method by themselves clearly dominated the rest. From the individual and average results, however, in terms of disaggregation level selection, the dissimilarity threshold criterion used with either the first principal component, persistence or forecasting error clustering dissimilarity measures tended to outperform the others. For the forecast combination choice methods, all dissimilarity measure choices performed relatively well.

As it is the case in most empirical applications, the impact of the grouping methods depends on the specific dataset. In particular, improvements in disaggregate accuracy were obtained only in the case where the direct approach was better than the bottom-up approach. It was also in this case that relatively more non-combination grouping methods improved aggregate accuracy. This could suggest that it is in settings like this, where the methods have a better chance of producing improvements. Such a result would not be entirely surprising, given the motivation for using dynamic grouping in the first place; that is to capture disaggregate dynamics in cases where full disaggregation could introduce to much noise.

Having said that, the use of the grouping methods could increase aggregate accuracy even in cases where full disaggregation is better than the direct approach. The overall good performance of the forecast combination choice methods suggests that the grouping methods can provide a way of introducing the robustness of forecasting combination into the procedure without having to introduce different forecasting models. Although,

in terms of disaggregate accuracy there were hardly any gains, in many cases the accuracy was similar to that of the best non-grouping method.

4 Conclusions

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

The results from the empirical application support that grouping methods can improve overall accuracy. On the one hand, some of the methods that selected a unique grouping performed better than the best performing non-grouping method. On the other hand, the forecast combination choice methods performed well overall. The exercise, however, contemplated only moderate disaggregation for the bottom-up approaches in which the biggest improvements were observed in the case where the bottom-up approach was less accurate than the direct approach. Espasa and Mayo-Burgos (2013) and Bermingham and D'Agostino (2014) encourage using the maximum disaggregation possible in order to benefit from the disaggregate dynamics. All this suggests, that the method could perform well in a context of higher disaggregation.

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

Appendix

A Empirical Framework

A.1 Multilevel combination where each unique forecast is given equal weights

In this section we show how we implement the multilevel combination of the hierarchy, where each unique forecast is given equal weights. To do this we first show that, for the case of equal weights, combining the aggregate forecasts produced from different aggregation levels can be equivalent to deriving a set of component forecasts that are consistent with different aggregate forecasts combining them to produce a definitive bottom-up forecast. With this, each distinct combined component forecasts can be used to produce the combination where each unique forecast is given equal weights.

A.1.1 Joint combination using the lowest level of aggregation

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. Cobb (2017) present a framework for multilevel forecast combination, where the combined aggregate forecast is given by:

$$\tilde{y} = \frac{Q^2 + y \sum_{n=1}^{N} \frac{\varphi}{\phi_n} w_n q_n}{Q + \sum_{n=1}^{N} \frac{\varphi}{\phi_n} w_n q_n} \tag{1}$$

where $Q = \sum_{n=1}^{N} w_n q_n$.

They show that equation 1 is equal to the result of the equal-weight combination when all forecasts are assigned the same reliability. In this framework, the components are obtained from:

$$\tilde{q}_n = \left(1 + \frac{\varphi}{\phi_n} \frac{y - Q}{Q + \sum_{n=1}^{N} \frac{\varphi}{\phi_n} w_n q_n}\right) q_n \tag{2}$$

With the objective of reconciling a set of components to an aggregate, equation 2 can

be rewritten as follows:

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

Then, if we simply want to have a disaggregate scenario that is consistent with the original forecast y, we take q_n, for $n = 1$ to N, as the best guesses by setting the aggregate reliability to infinity, $\varphi \to \infty$. With this, we have:

$$\hat{q}_n^{(y)} = \left(1+\frac{y-Q}{\phi_n\cdot\sum_{n=1}^{N}\frac{1}{\phi_n}w_n q_n}\right)q_n \tag{3}$$

If we now combine the original forecast for the components with the consistent-with-y forecast for the component we get:

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

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

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

and by summing up the components we get that the aggregate is:

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

that is the same that you get from setting $\phi_n = \phi$ for the standard result for the aggregate forecast.

This is a useful result for only one level of 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 the equation 3. 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 $\chi_{s,k} = \sum_{q_n \in y_{s,k}} \frac{1}{\phi_n} w_n q_n$ and $Q_{s,k} = \sum_{q_n \in y_{s,k}} w_n q_n$.

Then the definitive forecasts for the component is:

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

Then, for the case where all forecasts within the same grouping have the same reliability, we have that the aggregate is given by:

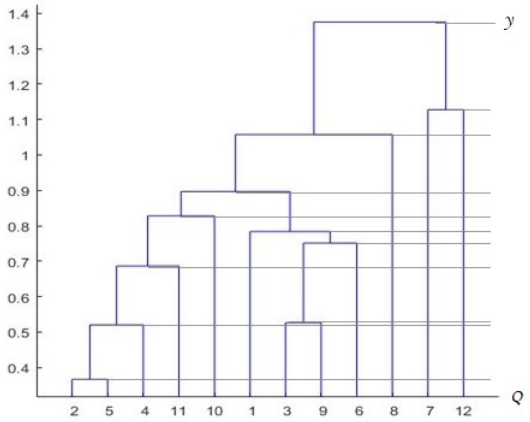
$$\begin{aligned}
\tilde{y} &= \sum_{n=1}^{N} w_n \left[1 + \frac{1}{\phi + \sum^{S} \varphi_s} \cdot \sum^{S} \left(\varphi_s \cdot \frac{y_{s,k} - Q_{s,k}}{Q_{s,k}}\right)\right] q_n \\
&= Q + \frac{1}{\phi + \sum^{S} \varphi_s} \cdot \sum^{S} \varphi_s \cdot \sum_{n=1}^{N} w_n \left(\frac{y_{s,k}}{Q_{s,k}}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

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

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

This shows that, for the case of equal weights, combining the aggregate forecasts produced from different aggregation levels can be 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.

Figure 2: Aggregation levels on a dendrogram



A.1.2 Multilevel Combination counting each distinct forecast only once

The previous section shows the desired equivalence between taking the simple average of all the aggregate forecasts and the simple average for the components that have been reconciled with the original forecasts for the different levels of aggregation. This is useful because it permits building a best guess underlying scenario for the aggregate forecasts, but it implicitly gives higher reliability weights to components that are fused later in the algorithm if the component forecasts are generated independently. This can be visualized by looking at the dendrogram in Figure 2.

On this dendrogram we have drawn horizontal lines every time a fusion occurs to illustrate the twelve options for groupings and the corresponding dissimilarity thresholds. At the very top we have the aggregate and at the very bottom all of the components in their own cluster. Let us denominate an aggregate forecast for a given grouping by the number of fusions that have occurred in that grouping. Using the nomenclature from the previous sections and assuming for simplicity that the aggregation weights, w, are all equal to one, the full bottom-up forecast is the sum of the forecasts of the individual components q_n for $n = 1$ to N. That is $Q = Q^{(0)} = \sum_{n=1}^{N} q_n^{(0)}$.

For all other aggregation levels, we can use the procedure described in A.1.1 to obtain forecasts of components that are consistent with the level of aggregation. For example, the forecast for the aggregation level that includes only the first fusion is produced from $Q^{(1)} = \sum_{n\neq\{2,5\}}^{N} q_n^{(1)} + Q_{\{2,5\}}$ where $Q_{\{2,5\}}$ is the forecast of the sum of components 2 and 5 and $\sum_{n\neq\{2,5\}} q_n^{(1)}$ is the sum of the forecasts of all the remaining components. In this case we only reconcile the forecasts $q_2^{(0)}$ and $q_5^{(0)}$ with $Q_{\{2,5\}}$ obtaining $\tilde{q}_2^{(1)}$ and $\tilde{q}_5^{(1)}$. With this $Q^{(1)} = \sum_{n\neq\{2,5\}}^{N} q_n^{(1)} + \tilde{q}_2^{(1)} + \tilde{q}_5^{(1)}$. We can then proceed in this fashion for every level, finishing with the direct aggregate forecast given by $y = Q^{(11)} = \sum_{n=1}^{N} \tilde{q}_n^{(11)}$.

We can denote the whole set of forecasts as:

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

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

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

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

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

If the forecasts are produced using a univariate model, all those that do not involve conciliation will be the same unless the model is changed purposefully for each level of aggregation. The same is true of fusions that are not fused again with other clusters at that aggregation level. In this case the set of forecasts would show a particular pattern in that the forecasts are unaltered in the next aggregation level unless there is a fusion:

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

In this context, it becomes clear that the simple average of all aggregate forecasts implicitly gives higher reliability weights to components that are fused later in the process. If we look at component twelve, we have that $\bar{q}_{12} = 10q_{12}^{(0)} + \tilde{q}_{12}^{(10)} + \tilde{q}_{12}^{(11)}$. A false sense of certainty is given to the original forecast meaning that the aggregate equal-weight combination does not satisfy the condition of giving equal weights to the distinct forecasts.

It is, however, quite straightforward to comply with the condition simply by removing

the repeated forecasts from the component averages. This can be done based on the dendrogram. For component three, for example, the component forecast for the aggregate combination would be given by $\bar{q}_3 = 3q_3^{(0)} + 2\tilde{q}_3^{(3)} + \tilde{q}_3^{(5)} + 2\tilde{q}_3^{(6)} + \tilde{q}_3^{(8)} + 2\tilde{q}_3^{(9)} + \tilde{q}_3^{(11)}$ and the distinct forecast combination comes simply from removing the repetitions. Therefore, the definitive forecast for any component becomes the average of the number of distinct forecasts that is equal to the number of fusions the component is involved in independently or as part of a group plus one.

B Empirical Application

B.1 Data Transformation

Table 6: Differentiation for Empirical Application

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

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

B.2 Relative Performance of Grouping Methods

Table 7: Relative Performance of Grouping Methods

	Average Percentage Deviation From Best Method						Average Rank Difference With Best Method					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
Conditional on Dissimilarity Measure												
Pearson corr.	4.6	6.0	5.8		1.9	**1.3**	2.0	3.0	3.0		1.3	**0.7**
Spearman corr.	6.8	8.5	6.9		2.3	**1.0**	2.5	3.2	2.5		1.3	**0.5**
1st princ.comp	3.5	5.2	3.5		2.6	**2.4**	2.2	3.0	2.0		1.5	1.3
persistance	8.4	6.9	**1.8**		4.3	3.2	3.2	3.0	**1.0**		1.7	1.2
f-error clustering	7.8	9.6	3.7		1.6	**0.9**	2.3	3.7	1.7		**1.2**	**1.2**
Bayesian	**1.7**	6.1	6.1	14.6	2.8	3.3	**1.5**	3.5	**1.5**	4.8	**1.5**	2.2
Conditional on Choice Method												
Pearson corr.	3.2	**3.1**	7.5		**2.4**	**1.8**	2.2	2.2	3.0		**1.7**	2.2
Spearman corr.	5.5	5.7	8.7		2.8	**1.5**	3.2	3.3	3.8		2.5	**1.3**
1st princ.comp	3.3	**3.5**	6.3		4.3	4.0	2.5	2.3	2.8		3.2	3.7
persistance	6.0	**3.0**	**2.5**		3.8	2.7	3.0	1.8	**1.0**		2.8	2.2
f-error clustering	7.1	7.3	6.0		2.7	**2.0**	3.5	3.7	2.2		2.0	2.3
Bayesian	**0.3**	**3.1**	7.8		3.2	3.8	**0.7**	**1.7**	2.2		2.8	3.3
Overall												
Pearson corr.	7.0	8.5	8.3		4.4	3.8	15.7	19.8	19.2		9.7	8.7
Spearman corr.	9.4	11.1	9.5		4.8	3.5	18.0	22.2	21.0		11.2	7.0
1st princ.comp	7.2	8.8	7.1		6.2	6.1	16.2	20.5	16.0		13.7	14.0
persistance	9.9	8.4	**3.3**		5.8	4.7	18.8	17.3	**6.7**		11.2	9.2
f-error clustering	11.0	12.7	6.8		4.7	4.1	20.7	26.0	13.7		9.8	9.0
Bayesian	4.1	8.5	8.6	17.0	5.2	5.8	7.3	20.2	12.0	25.8	11.3	13.3

Note: Relative performance of the grouping methods as measured by the average deviation of the respective root mean squared forecasting error (RMSFE) relative to that of the best performing grouping method by category and as the average difference in rank according to RMSFE over the six sets of forecasts. Grouping method dissimilarity measures: Pearson correlation, Spearman correlation, Variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion, Forecast Combination method 1 and Forecast Combination method 2. In bold the best performers in each category. Calculated over 2001-2015.

B.3 Relative Cumulative Forecasting Errors

Table 8: Relative Cumulative Forecasting Errors

	AR(1)					BVAR				
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC
France										
Type 1 Combination										
Pearson corr.	1.00	1.00	1.00		1.01	1.00	0.99	1.00		0.99
Spearman corr.	1.00	1.00	1.01		1.01	1.02	1.03	1.02		1.02
1st princ.comp	1.04	1.02	1.04		1.04	1.04	1.04	1.04		1.04
persistence	1.00	1.01	1.02		1.02	1.03	1.02	1.01		1.01
f-error clustering	1.00	1.01	1.00		1.02	1.01	1.02	1.02		1.02
Bayesian	1.00	0.99	1.00	1.00	1.01	0.99	1.00	0.99	0.99	0.99
Type 2 Combination										
Pearson corr.	1.00	1.00	1.00		1.02	1.00	1.00	1.00		1.01
Spearman corr.	1.01	1.01	1.02		1.02	1.01	1.02	1.02		1.02
1st princ.comp	1.04	1.03	1.04		1.05	1.03	1.04	1.04		1.05
persistence	1.00	1.01	1.02		1.02	1.02	1.02	1.02		1.01
f-error clustering	1.01	1.01	1.01		1.04	1.01	1.03	1.03		1.04
Bayesian	1.00	0.99	1.00	1.01	1.02	1.01	1.01	1.00	1.01	1.00
Germany										
Type 1 Combination										
Pearson corr.	1.01	1.02	1.01		1.03	1.02	1.02	1.05		1.04
Spearman corr.	1.01	1.02	1.02		1.04	1.04	1.03	1.05		1.05
1st princ.comp	1.02	1.02	1.04		1.04	1.03	1.02	1.06		1.06
persistence	1.01	1.01	1.01		1.02	1.02	1.03	1.04		1.04
f-error clustering	1.01	1.01	1.00		1.03	1.03	1.01	1.03		1.05
Bayesian	1.01	1.01	1.00	1.02	1.03	1.01	1.03	1.00	1.03	1.03
Type 2 Combination										
Pearson corr.	1.01	1.02	1.01		1.03	1.01	1.01	1.04		1.04
Spearman corr.	1.01	1.02	1.02		1.03	1.02	1.01	1.03		1.04
1st princ.comp	1.01	1.02	1.04		1.04	1.01	1.02	1.06		1.06
persistence	1.01	1.01	1.00		1.02	1.01	1.02	1.02		1.03
f-error clustering	1.01	1.01	1.01		1.04	1.02	1.01	1.03		1.05
Bayesian	1.02	1.02	1.00	1.03	1.03	1.01	1.03	1.00	1.04	1.04
UK										
Type 1 Combination										
Pearson corr.	1.02	1.05	1.06		1.07	0.98	0.96	0.98		0.96
Spearman corr.	1.01	1.03	1.01		1.08	0.97	0.95	0.96		0.94
1st princ.comp	1.03	1.06	1.03		1.09	0.98	0.98	1.00		0.95
persistence	1.02	1.04	1.04		1.06	0.90	0.90	0.90		0.89
f-error clustering	1.02	1.05	1.01		1.09	0.95	0.93	0.96		0.91
Bayesian	1.00	1.01	1.00	1.06	1.07	0.95	0.93	1.00	0.93	0.93
Type 2 Combination										
Pearson corr.	1.03	1.06	1.07		1.09	0.99	0.96	0.98		0.97
Spearman corr.	1.02	1.05	1.01		1.11	0.98	0.96	0.96		0.95
1st princ.comp	1.02	1.05	1.01		1.11	0.99	0.98	0.99		0.96
persistence	1.01	1.04	1.04		1.07	0.95	0.91	0.93		0.90
f-error clustering	1.03	1.05	1.01		1.11	0.97	0.95	0.99		0.90
Bayesian	1.00	1.01	1.00	1.07	1.10	0.97	0.95	1.00	0.95	0.95

Note: Cumulative root mean squared forecasting error relative to the direct method. Grouping method dissimilarity measures: Pearson correlation, Spearman correlation, Variance explained by the first principal component, similarity in persistence measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion, The type of combination method refers to the multilevel forecast combination procedure. Calculated over the 2000-2015 period.

B.4 Forecasting Accuracy Excluding Financial Crisis

Table 9: Benchmark Forecasting Performance Excluding Crisis

	France	Germany	UK
Bottom-Up AR(1)	0.95	0.90	0.90
Bottom-Up BVAR	0.97	0.98	1.25
Factor augmented AR(1)	0.96	0.88*	0.90

Note: Root mean squared forecasting error relative to the direct method. * and ** denote significance of the forecasting performance difference based on the modified Diebold-Mariano test at a 10 and 5% significance level. Calculated over 2001-2015 excluding from 2008.III to 2009.II.

Table 10: Relative Forecasting Errors Excluding Crisis

	AR(1)						BVAR					
Choice method	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC1	FC2
France												
Pearson corr.	0.92	0.97	0.94		**0.88****	0.90**	1.03	**0.97**	1.00		**0.91***	**0.93***
Spearman corr.	**0.89***	**0.90****	0.98		**0.87****	**0.89****	1.09	1.11	1.01		**0.89****	**0.91****
1st princ.comp	0.94	0.95	0.98		**0.90****	0.92**	1.05	1.08	1.04		**0.96**	**0.95**
persistance	**0.88***	**0.88****	**0.89****		**0.87****	**0.88****	1.03	**0.97**	**0.93**		**0.94**	**0.93***
f-error clustering	0.93*	0.94	0.93		**0.87****	**0.90****	1.08	1.11	0.98		**0.94**	**0.96**
Bayesian	**0.88****	**0.90****	0.91	0.98	**0.89****	0.91**	1.03	1.05	1.03	1.04	**0.96**	**0.97**
Germany												
Pearson corr.	0.99	1.01	1.01		1.00	0.99	1.11	1.16	1.09		1.01	1.00
Spearman corr.	0.99	1.00	1.02		1.00	0.98	1.14	1.15	1.06		1.02	0.99
1st princ.comp	0.99	1.00	1.00		1.00	0.99	1.10	1.01	1.06		1.00	0.99
persistance	0.97	**0.95**	**0.88****		**0.92**	**0.93***	1.11	1.05	**0.97***		**0.96**	**0.95**
f-error clustering	0.98	0.98	0.97		0.98	0.98	1.22	1.12	1.05		1.00	0.99
Bayesian	0.99	0.98	0.96	1.00	0.97	0.98	1.03	1.12	0.99	1.03	0.99	0.99
UK												
Pearson corr.	**0.87**	**0.88**	0.91		**0.87**	**0.84***	**0.88**	**0.83**	**0.86**		**0.95**	**0.88**
Spearman corr.	**0.87**	0.95	**0.88**		0.90	**0.89**	**0.96**	**0.87**	**0.99**		1.00	**0.92**
1st princ.comp	**0.84**	0.93	**0.84**		**0.84***	**0.88**	**0.90**	**0.83**	**0.85***		1.04	1.01
persistance	0.97	0.98	1.04		0.98	0.93	1.00	0.91	**0.85***		1.06	1.03
f-error clustering	0.92	0.96	**0.87**		**0.88**	**0.85***	**0.91**	**0.92**	1.04		**0.95**	**0.92**
Bayesian	**0.89**	**0.90**	**0.89**	1.19	0.93	0.93	**0.89**	**0.95**	1.24	1.25	1.00	**0.99**

Note: Root mean squared forecasting error relative to the direct method. Grouping method dissimilarity measures: Pearson correlation, Spearman correlation, Variance explained by the first principal component, similarity in persistance measured as the difference of the estimated rho for an AR(1), forecasting error clustering for AR(1), Bayesian Hierarchical Clustering. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion, Forecast Combination method 1 and Forecast Combination method 2. In bold RMSFE lower than the lowest of either the respective full Bottom-Up approach or the direct approach. Calculated over the 2001-2015 excluding from 2008.III to 2009.II.

Table 11: Relative Cumulative Forecasting Errors

	AR(1)					BVAR				
	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC	In-samp.	O-o-S	Diss. Thres.	Prob. crit.	FC
France										
Median	1.00	1.00	1.01	-	1.01	1.01	1.02	1.01	-	1.01
Min	1.00	**0.99**	1.00	1.00	1.00	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
Max	1.02	1.02	1.03	1.00	1.04	1.03	1.03	1.04	1.00	1.05
Germany										
Median	1.02	1.02	1.01	-	1.03	1.02	1.02	1.03	-	1.04
Min	1.01	1.00	1.00	1.03	1.02	1.01	1.00	1.00	1.03	1.02
Max	1.02	1.02	1.04	1.04	1.05	1.04	1.03	1.06	1.04	1.06
UK										
Median	1.02	1.05	1.01	-	1.09	**0.97**	**0.95**	**0.98**	**-**	**0.94**
Min	1.00	1.02	1.00	1.07	1.07	**0.91**	**0.91**	**0.91**	**0.94**	**0.90**
Max	1.03	1.06	1.07	1.08	1.12	**0.99**	**0.99**	1.00	**0.96**	**0.96**

Note: Cumulative root mean squared forecasting error relative to the direct method. Median, minimum and maximum values obtained from all grouping method dissimilarity measures and multilevel forecast combination methods. Choice methods: In-sample criterion, Out-of-sample criterion, dissimilarity threshold, Probabilistic Criterion. In bold CumRMSFE lower than that of the respective full Bottom-Up approach. Calculated over the 2001-2015 excluding from 2008.III to 2009.II.

References

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

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

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

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

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.

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

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

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

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

Cobb, M. P. A. (2017, February). Joint forecast combination of macroeconomic aggregates and their components. MPRA Paper 76556, University Library of Munich, Germany.

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

Diebold, F. X. and J. A. Lopez (1996). Forecast evaluation and combination. in Maddala and Rao, eds., Handbook of Statistics (Elsevier, Amsterdam).

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

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

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

Duncan, G. T., W. L. Gorr, and J. Szczypula (2001). *Forecasting Analogous Time Series*, pp. 195–213. Boston, MA: Springer US.

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

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).

Gao, Z. and J. Yang (2014, September). Financial time series forecasting with grouped predictors using hierarchical clustering and support vector regression. *International Journal of Grid & Distributed Computing 7*(5), 53–64.

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

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.

Granger, C. (1990). *Aggregation of time series variables: a survey*, Chapter 2. Routledge.

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

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

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.

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

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

Hauke, J. and T. Kossowski (2011, 06). Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *30*, 87–93.

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

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

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

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

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

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.

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

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

Newbold, P. and D. I. Harvey (2002). Forecast combination and encompassing. M.P. Clements and D.F. Hendry, eds., A Companion to Economic Forecasting (Blackwell, Oxford), 268–283.

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

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

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

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

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

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

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

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

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

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