UNIVERSITY OF TWENTE.

KLM engineering & maintenance

Hierarchical forecasting of engineering demand at KLM Engineering & Maintenance

Master Thesis, February 2019

Author:
Ian Breed

Supervisors:
University of Twente:
Dr. E. Topan
Dr.ir. L.L.M. van der Wegen

KLM Engineering & Maintenance:
H. Lucas
R. Steenkist

University of Twente

Industrial Engineering and Management

Production and Logistic Management

Preface

This master thesis was written in order to complete my study Industrial Engineering and Maintenance at the University of Twente. I had the good fortune to perform my research at KLM Engineering and Maintenance and some acknowledgments and thanks are in order.

First of all I would like to extend my gratitude towards KLM specifically my supervisors Hans Lucas and Rik Steenkist for their guidance, trust and patience. The process took longer than expected but I never had the feeling that trust in my capabilities had vanished. Your loose approach to guidance allowed me to define my own path and ask for advice where necessary, which I always promptly got. Additionally, their loose approach allowed me to make the most of my graduation period at KLM, gaining experience in the board of the KLM intern society 'Taxibaan' as well as participate in other interesting, but unrelated, projects. The experience I have gained over the past period is invaluable and will help me in future, which I hope to prove in the time to come. A large part of me enjoying my time was also attributable to the colleagues of Cabin engineering where I spent my days.

Finally, I would like to extend my thanks to my university supervisors. My first supervisor, Engin Topan, proved to possess substantial patience and good advice for me to improve and finish my thesis. Lastly, I would like to thank my secondary supervisor, Leo van der Wegen, for his insights and participation in our productive discussions with Engin. Both of your advice substantially improved the end result.

I thank you all for your support and KLM for the opportunity to further prove myself in time to come.

Ian Breed

February 2019

UNIVERSITY
OF TWENTE.

KLM engineering & maintenance

Management summary

In this research we have focussed on designing and testing a forecasting framework for engineering demand of KLM Engineering & Maintenance (E&M). E&M consists of a collection of diverse and specialized teams and engineers. They perform a variety of tasks in support of KLM or external customers. Their tasks range from repair development and technical documentation, requiring a single hour, to overhaul projects requiring multiple fulltime employees for more than a year. Because the tasks and activities are so diverse E&M has struggled with forecasting demand. Currently they apply a simple but intuitive method, taking the mean of the past seven months in order to forecast the entire coming year. This forecast is adjusted based on expert opinion and judgment without much regard for the statistical method used. In this approach a lot of trust and responsibility is placed on the human forecaster to produce accurate results.

There is a desire to gain more insight in demand behaviour as a means to increase control over capacity. More accurate forecasts can lead to more accurate budgets as well as help on when to make decisions regarding more operational capacity. As a result the research question we tried to answer was:

"How to accurately forecast uncertain demand for KLM engineering with quantitative and qualitative methods?"

A system and data analysis showed that demand could be categorized in five descriptive characteristics:

- Was it demand from an Internal (KLM) or external customer
- For which specific division/customer
- Which aircraft type
- Was it a routine or non-routine task
- Which specific tasks was it demand for

Each of these characteristics provides us with a possibility of looking at a specific part of total demand, we can look at total demand for a certain type for instance, or combine characteristics and look at demand for a specific task on non-routine basis for a specific customer. Any of the characteristics can be combined to create a subset of demand that contains unique behaviour, and thus information, possibly of value to forecasting. A combination of all 5 characteristics presents the most disaggregate and specific demand, we call this the bottom time series (BTS). From the BTS we have multiple options for aggregating demand, e.g. all demand for specific aircraft types, which are called groups. Aggregating all BTS results in the total demand not specific to any characteristic. Defining all different combinations of the characteristics leads to 16 different groups containing 1716 different demand subsets.

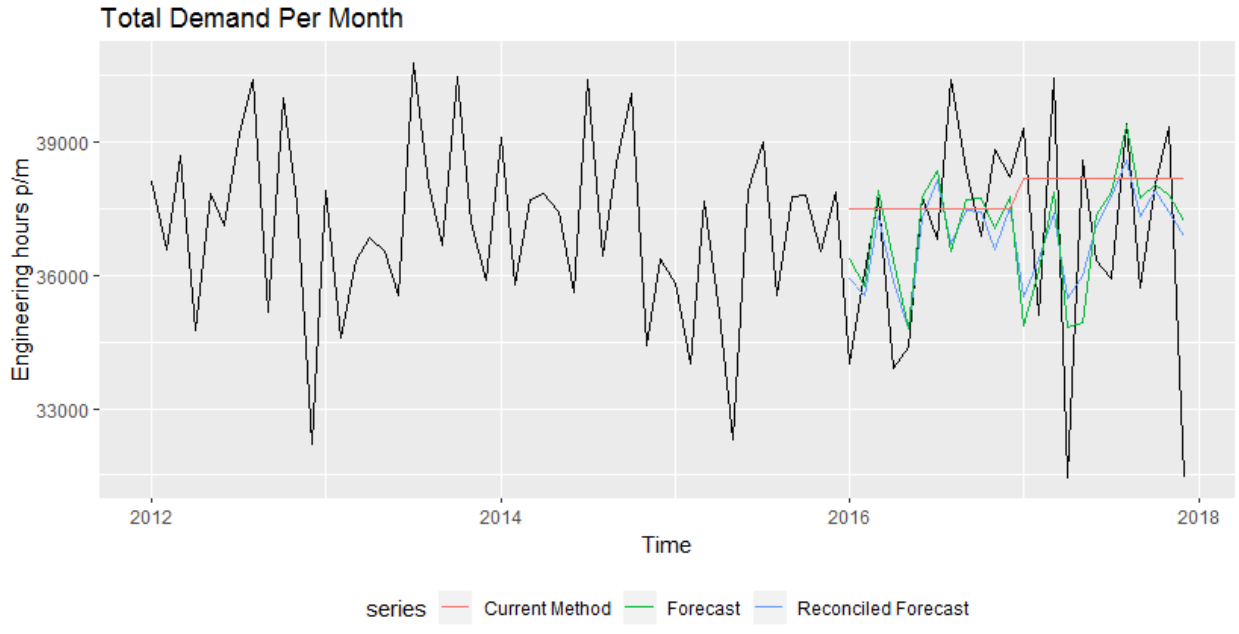
Each of the subsets contains information and requires a forecast. But because each has its own behaviour and no single forecasting model can be expected to correctly forecast all the different subsets. We propose a forecasting framework that leverages the capabilities of multiple models through forecast combination and reconciliation.

By applying ten different models and calculating all their combinations we define a collection of forecasts that uses all the information extracted by the different models. Using the mean absolute scaled error (MASE) we are then able to select the most accurate approach and provide a trustworthy statistical forecast. This significantly improves accuracy over the current forecasting approach, a 20% accuracy gain, in terms of the MASE, is achieved over both 2016 and 2017.

The MASE also functions as an indicator for series that require additional judgmental input. A MASE of <1 implies performance better than a naive forecast, >1 implies worse performance. In some series demand behaved in such unpredictable manners that MASE of >1000 were observed. These high values serves as an outlier detection system and help to identify series that require judgemental forecasts.

To properly apply judgemental forecasting clear rules and guidelines should be followed. Initial trust should be in a statistical forecast that is capable of producing accurate forecast without human input. This statistical forecast is provided by our framework. Then bad performance should be identified, for instance through identifying large MASE values. Finally, experts should be consulted and provide adjustments to the initial forecast while documenting and justifying their decisions.

As a result of the previous steps all demand subsets have forecast that are as good as it they can with the available information, both data and organizational knowledge have been used. But inconsistencies occur between the forecasts of different demand subsets. Each subset is part of a group but also aggregates to higher level subsets, all BTS that include the characteristic for the 77 type sum to its demand. Yet due to different information presented in their data the forecasts for the relevant BTS will not sum to the forecast for total demand of the 777. This discrepancy can be mitigated through reconciliation, minimally changing all the involved forecasts in order to obtain one coherent forecast. This is important to ensure that decisions on different levels, e.g. budgeting on tactical and planning on operational, are based on the same numbers and align. The method slightly decreases overall accuracy but the gained alignment is useful in an organizational sense. The following figure illustrates forecasts for total demand by the current method, the most accurate and reconciled forecast.



By applying a more sophisticated forecasting approach we improve accuracy significantly. Transforming the gained accuracy to actual demand numbers we find that the current method overestimates demand over 2017 equivalent to 9,3 FTE. The proposed framework was able to reduce this overshoot to 0,9 FTE freeing 8,4 FTE from the forecast. Such knowledge allows the organization to make better and more flexible decisions on how to apply capacity. The proposed method was able to accurately forecast total yearly with less than 1% error margin, reducing the absolute error compared to the current approach by 55% and 89% for 2016 and 2017 respectively.

We recommend that KLM engineering take the following steps to improve their forecasting accuracy:

- KLM needs to decide what kind of forecasting accuracy they require for what subsets of demand and implement a suitable forecasting method accordingly.
- Implement the presented framework for a comprehensive approach capable of handling all different subsets.
- Clean the source data and analyse its function as a proxy for demand.
- Define rules and guidelines on how to apply judgemental forecasting and investigate its effectiveness.

Throughout our research the focus has been on a comprehensive statistical forecasting model. By opting for a more complex framework other forecasting options fell out of our scope. Dynamic regression where external variables are used to explain variation is an interesting step for further research to more accurately forecast specific demand streams, however initial tests did provide mixed results. A more thorough framework for judgmental forecasting should be investigated and tested to maximize its accuracy. Finally, the forecasts and results were created by using work hours as a proxy for demand, this assumption needs to be tested and further more specific data analysis might provide more information about the underlying demand patterns.

Table of Contents

1 Introduction

Throughout this thesis we investigate how demand for skilled engineering labour hours should be forecast in a complex setting with diverse tasks. We do so to provide the engineering department of KLM Engineering and Maintenance (E&M) more insight in what to expect. As a result they will be able to exert more control in matching their skilled labour capacity to demand, leading to a better utilization of skills and knowledge. The goal is to develop a forecasting model that can provide them with more accurate results and information. In this chapter we focus on the background of the problem and the research goals. First, by introducing the organizational context in Section 1.1 and then the research problem and goal in Section 1.2. Lastly, this leads us to the research questions and scope in Section 1.3.

1.1 Organizational context

1.1.1 KLM

The 'Koninklijke Luchtvaart Maatschapij' (KLM) is the national airline of the Netherlands. It was granted a royal title by the queen during the founding in 1919 and has been the 'Royal Dutch Airline' ever since. It still operates under its original name making it the oldest airline that does so. It is a global network carrier and operates from Amsterdam airport Schiphol. KLM does so with a fleet of 119 different aircraft consisting of Boeing 737's, 747's, 777's, 787's and Airbus A330's to about 150 different destinations. If the wholly owned subsidiaries KLM Cityhopper and Transavia are included they add around 100 additional aircraft to the network and 160 additional routes.

The network is used primarily for transporting passengers and secondarily for cargo. In order to effectively operate the entire network, the fleet needs to be used efficiently and effectively, requiring constant upkeep and repairs to ensure conforming to safety standards. This is where the maintenance division of KLM, Engineering & Maintenance (E&M) is key.

1.1.2 KLM E&M

As the maintenance division of KLM airlines, Engineering & Maintenance (E&M) is responsible for performing the necessary maintenance to keep the aircraft and its components safe and up to the required technical standards. These standards are defined by the European Aviation Safety Agency (EASA) in EU-OPS, the regulations for commercial passenger and cargo aviation. They describe training, documentation, procedure and compliance requirements for different aviation related subjects like Aircraft maintenance and other related subjects such as performance and operational procedures. In essence, the regulations define a set of rules that have to be properly followed before commercial operation is allowed. These rules create a need for due diligence in designing, documenting, executing and evaluating maintenance.

E&M is certified by EASA to undertake such activities through a design and maintenance organization approval (DOA and MOA). Also called part-21 a DOA regulates that an organization has fulfilled the requirements to design and certify changes to aircraft, repairs, and parts and appliances for aircraft. MOA, or part-145, regulates the physical execution of maintenance according to such designs, to ensure continued airworthiness of the aircraft. As such, E&M is allowed to design, certify and produce/execute maintenance and repairs, for both KLM and other customers. The design, certification and related support, but not the physical execution of maintenance, is the responsibility of the central engineering (CE) department. Engineering is responsible for a various collection of

tasks requiring different skills. In order to illustrate the type of work conducted, a selection of the different engineering activities:

- Provide production (i.e. the physical execution of maintenance) support:
  - With the necessary documentation and instructions to execute tasks according to regulation.
  - Provide direct support on questions unclarified by instructions.
- Design and certify repairs.
- Evaluate and certify/approve service bulletins from original equipment manufacturers (OEM).
- Provide expert/specialist support on unconventional issues.
- Evaluation, certification and approval on the phase-in and -out of, aircraft, parts and modifications
- Project management for major overhauls and modifications to aircraft.
- Keep repair and maintenance manuals up to date
- Guarantee proper documentation practice of technical documentation

Provide such a wide range of activities requires a collection and mix of specialists in both engineering and support activities. Engineering specialists are divided in departments and teams focusing on specific areas of aircraft. For instance, the cabin department focuses on the internal areas of an aircraft, the passenger area but also the crew spaces and the cargo hold. The department is subdivided into different teams such as avionics (electrical components), seats and mechanical (e.g. storage bins, toilets). In these teams there often is an additional division of knowledge and responsibility per different type of aircraft. Support specialists focus on tasks including project management, documentation of aircraft related information such as repair manuals and regulatory oversight.

1.2 The research problem and goal

1.2.1 The problem

To effectively perform the different engineering and support tasks, the number of available skilled engineering hours' (capacity) should match the demand for different skilled engineering work (demand). To create capacity that can match demand a budget is necessary to employ engineers with the necessary knowledge. Budgeting accurately therefore requires demand forecasts, which are made every August for the coming year. The forecasted demand is then translated to a required capacity and the budget reflects the costs associated with said capacity

A more accurate forecast therefore leads to more accurate capacity. However, forecasting this demand is not straightforward. The range of different activities, customers (KLM, partners and others) and aircraft types creates a diverse and variable demand. To illustrate, an administrative tasks might only need 1 hour. A large overhaul/modification project can require a team of multiple fulltime employees for more than 1.5 years. This variation, and the uncertainty it causes, makes accurate forecasting, and thus budgeting, difficult. Difficulty increases when smaller subsets of total demand are considered because the relative variability increases. We illustrate this by showing different demand subsets starting with the total demand in hours per month in Figure 1.1. In Figure 1.2 we highlight two additional subsets, demand specific to the aircraft type (Boeing) 777 and demand not specific to any aircraft type. The figure implies that total demand is most affected by the

variability of demand not linked to a type as the 777 demand appears fairly stable. However, as we can see in Figure 1.3, the relative variability in the 777 is large, mostly due to a major increase in demand from 2014 to 2016. In order to have the required capacity to fulfil the 777 demand, which requires specific knowledge, it is important to notice and predict this, yet the increase was nearly imperceptible on the scale of total demand.

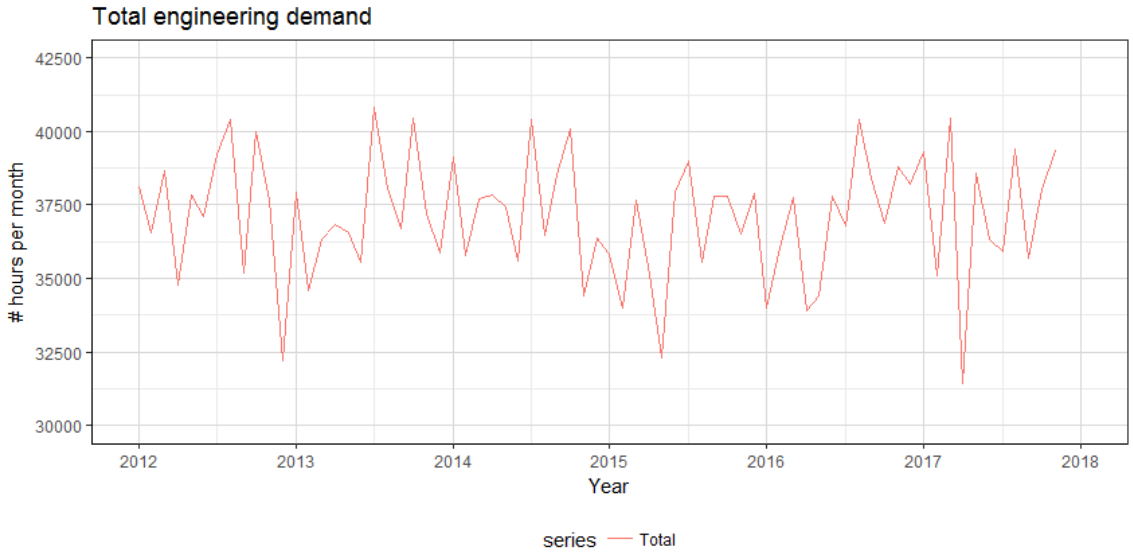


Figure 1.1 Total demand work p/m

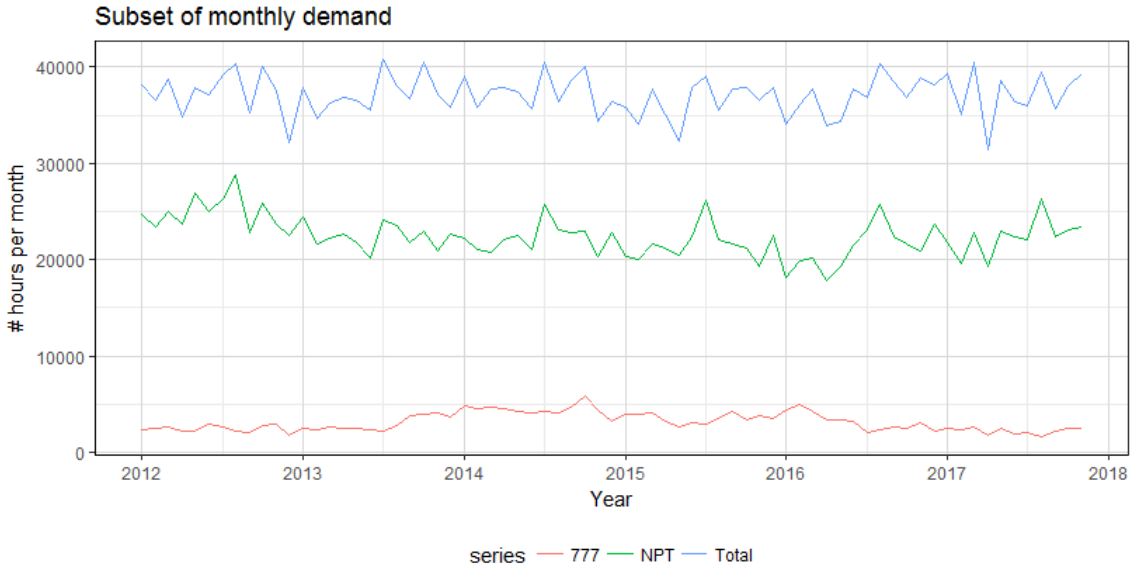


Figure 1.2 Total demand, No-Plane Type (NPT) specific demand, Boeing 777 demand, all p/m

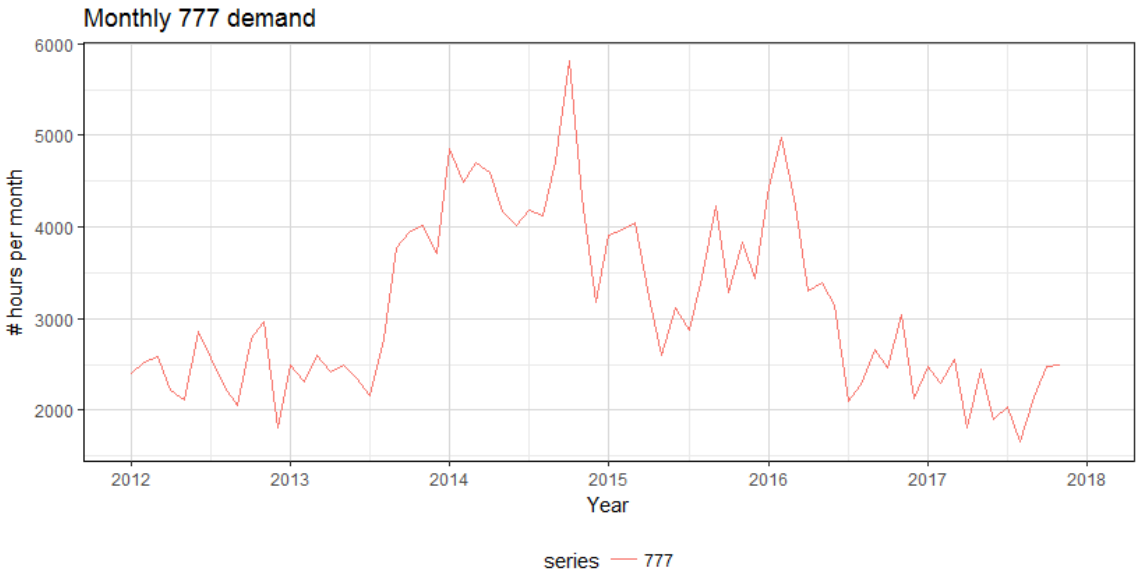


Figure 1.3 Total demand p/m for the Boeing 777 with a volume increase from 2014 to 2016

The previous figures show us how different subsets of demand have different behaviour. Because of this, the organization finds it difficult to obtain insight in the behaviour and uncertainty of demand. This has led to two pressing issues in matching capacity to demand:

- Budgeting accurately for coming periods is difficult. The highest levels of demand exhibits substantial variation and demand subsets are relatively even more variable. As a result forecasting demand is challenging and accurate budgets for the necessary capacity becomes difficult.
- Difficulties in timely up- or down-scaling of specific capacity. Current forecasts look at the entire year for high level planning/budgeting. But demand can be divided into more specific categories, e.g. teams, tasks and aircraft types, and each has its own characteristics. These subsets of demand are not fully considered in the forecasts and when they are, only on a yearly basis. As a result, it becomes more difficult to match capacity to specific demand. Adding the fact that a training period can take up to 6 months makes it even more difficult to effectively match the two.

So demand characteristics make it hard to forecast demand, especially for more specific subsets. Currently these difficulties are mostly handled by relying on experience, opinion and limited insight in historical data. There is a desire to gain more control over forecasting, instead of the gut-based, subjective, decisions. Preferably leading to more accurate forecasts and as a result more control over matching capacity with demand, Section 1.2.2 further discusses the research goal.

1.2.2 The research goal

The presented problem is a lack of control over matching capacity with demand. The goal of research should therefore focus on improving said control. We have seen that the issues come from the variable behaviour of demand, if we knew the amount of demand before it occurs then matching capacity to it becomes straightforward. So an accurate forecast of demand before it occurs is desired to enable more informed decisions on capacity.

Several things are necessary for an accurate forecast of demand. First, insight in past demand is important in order to identify general trends or other behaviour of importance. Secondly, some future effects on demand cannot directly be learned from historical data. Large projects are often unique for the specific subset of demand it belongs to (see Figure 1.3) and have no or little relation to past behaviour. External information, such as expert opinion, is therefore assumed to be of importance as they can better predict how specific demand subsets might behave under extraordinary circumstances. Finally, this should be combined in a model where statistical forecasts and expert opinions can work together resulting in a more fact driven method to match capacity to demand. Based on this we define the following research goal that when achieved helps reach the desired situation:

Analyse historical demand to gain insight in its behaviour and variation and use this information to accurately forecast demand while including necessary external information.

In conclusion, we defined the research problem and goal based on the experienced issues and the desired situation. The demand for engineering work is diverse and its behaviour and variability change depending on which subset of demand is regarded. This makes it difficult to control capacity, leading to mismatches with demand and possibly to missed opportunities because of over- or under-capacity. Control despite these uncertainties is desired to make better decisions and more accurate budgets. Such an increase in control requires insight in past and future demand behaviour. Forecasting could provide this insight and make budgeting, as well as capacity planning more fact based and accurate. Defining the research problem led to a clear goal: increase the control over matching demand with capacity by more accurately forecasting demand. Section 1.3 builds on this goal to construct questions that will guide the research to the necessary answers and insights.

1.3 Research questions and scope

The research problem and goal clearly describe the desired direction of a more accurate forecasting model. To achieve this goal within reasonable time and other restriction we require clear direction and boundaries. The research is guided by the research questions in Section 1.3.1 and the scope is defined in Section 1.3.2.

1.3.1 Research questions

The main scope of the research is derived from the research goal in Section 1.2.2. The main goal is a more accurate forecasting of demand based on past data and expert knowledge. As such the main research question becomes:

"How to accurately forecast uncertain demand for KLM engineering with quantitative and qualitative methods?"

Sub questions

In order to answer the main research question we defined seven sub-questions that will provide partial answers, leading to an overall answer:

1. What does the engineering demand consist of and how does it cause uncertainty?
2. What are the characteristics of the available data and is it suitable for forecasting?
3. What is the current demand forecasting practice?

Questions 1, 2, and 3 analyse the current situation. They explore the details of (historical) demand, its uncertainty and current forecasting practice. Question 1 identifies the structure of demand, its subsets and possible drivers of uncertainty. Question 2 is answered by analysing the demand data providing insight in what data is available and its characteristics. Finally, Question 3 evaluates the current forecasting practice indicating current methods and possible limitations. All three questions are addressed in Chapter 2 System analysis.

4. What forecasting methods are suitable according to literature?
5. How can forecasting performance and validity be measured?

Questions 4 and 5 look at literature and determine suitable methods for forecasting and measuring its performance both quantitative and qualitative models are considered for forecasting to work with data and incorporate expert judgment. Performance and validity measures provide means to compare the forecasts to benchmarks and organizational performance. Both Questions 4 and 5 are answered in Chapter 3 Literature review.

6. How should forecasting be applied for engineering demand?
7. How does the proposed method perform?

Questions 6 and 7 guide the implementation of the model and the evaluation of the results, respectively implementing the knowledge gained for Questions 4 and 5. First, applying theoretical knowledge to the organization, leads to a fit between theoretical ideas and a realistic application in Chapter 4. Then performance and validity measures are used to evaluate the results in Chapter 5.

1.3.2 Scope

In order to realize results within the given restrictions, the scope of the research needs to be further demarcated with additional choices and assumptions. The focus is on providing boundaries for the choices and activities that take place throughout the research.

What to forecast?

The forecasts focus on demand, defined as the total number of hours necessary for all skilled engineering work.

- All levels of demand relevant to the organization will be considered. From total demand to subsets for departments, aircraft types, (non-)routine tasks and combinations thereof (see Section 2.3 for elaboration).
- All levels of demand are considered and the lowest levels are indicative of required skills. We assume that from there matching skill to demand is trivial for the organization and as such capacity planning is not considered.

Forecasts should be able to accurately forecast at least 1 year ahead in monthly steps.

- Budgeting requires a forecast for the entire coming fiscal year. Additionally, when used periodically a year ahead forecast serves as an evaluation for the current budget and if it is still on track.
- Monthly forecasts are useful on an operational level, for instance to anticipate seasonal changes and possibly helping to prioritize certain tasks or projects.

Model choices and desires

If possible explanatory variables should work with the model.

- Demand is expected to be influenced by external factors (e.g. the # of aircraft in the fleet) so (a part) of the model should to accept this information to possibly produce better forecasts/explanation of variation, Section 3.4.2.5 elaborates on using external variables.

Expert opinion and judgements are necessary and including them should be possible in the model.

- The airline business is complicated and sometimes suspect to sudden shifts. Experts have the experience to make/adjust forecasts to include these events (see Section 3.4.1 for elaboration). They can adjust the actual statistical forecast where necessary to anticipate on effects unforeseeable by the data.

There is a preference for an interpretable model.

- The input should be relatable to the output, given a certain set of data, with or without external variables, a 'readable' output of the model fit and is preferred as it can give insight into the dynamics of the underlying demand process. A model that explicitly fits seasonality can tell us what kind of change to expect each season.
- Black box methods (e.g. neural networks) are therefore out of the scope. The predictive power might be good but no meaningful way is available to relate the output to the input.

Modelling and analysis will be done with R and Excel.

- Software with university/external license could be (more) suitable but then no implementation in the organization can take place, R and Excel are greenlit by the organization and have sufficient capabilities for statistical analysis and forecasting.
- R has packages/capabilities for several different forecasting methods, shortening the time and limiting the complexity necessary for modelling.

Organizational assumptions

Workhour administration is correct.

- Realistically there will be contamination, because hour administration is done by hand (see Section 2.2). However, we assume that errors will happen either up or down somewhat equally and has the same chance to happen on every entry. Under these assumptions the contamination should even out over all activities equally.

Tasks performed outside of the relevant engineering unit are out of scope.

- Engineering tasks are what is relevant to the department, any other work such as physical maintenance is out of scope.

Worked man hours are representative of actual demand.

- In situations where demand exceeds the available capacity, tasks can be rejected (outsourced) and, preferably, postponed. The selection of which demand to postpone is based on importance, high priority tasks are done first. Postponed demand is then counted in the period in which it is performed. As a result the total man hours worked represents the total demand apart from only a small portion of cancellations.

1.4 Chapter conclusion

Throughout this chapter we focussed on assessing the organizational context, the preferred situation and what the research goals should be in order to move toward the desired situation. From the context we find that the engineering tasks are diverse and range from singular tasks to large projects. This leads to the organization experiencing problems with making accurate forecasts for demand, in turn causing issues with accurately budgeting for capacity. We conclude that the desired situation is one where forecasts are based on objective, more accurate, methods leading to more control over capacity due to better forecasts. As a result the research goal is defined to focus on analysing past demand and using that knowledge to build a forecasting model.

From the context and within constraints a path was set out by the research questions on how to realize such a model. First, the focus is on an analysing the current situation which increases the knowledge on demand behaviour and how it manifests in the data. Additionally, the current forecasting method is evaluated. Then, suitable methods to forecast and measure its performance measures are collected from literature. Finally, the methods are applied to the data within the organizational context, after which the results are evaluated. Hopefully leading us to fulfil the research goal as defined in Section 1.2.2. Chapter 2 describes our first step towards the goals and provides answers to research questions 1, 2, and 3.

2 System analysis

In this chapter we look at the current situation in relation to reaching the goals as defined in Section 1.2. The characteristics of demand are explored and the current way of forecasting demand is analysed. Section 2.1 looks at the different elements that demand consists of. Sections 2.2 and 2.3 take the available data and provide an analysis on what information and characteristics of demand it contains. Section 2.4 focusses on how this data and characteristics are currently used in the forecasting process.

2.1 Analysing demand

As previously explained in Section 1.1.2 engineering work consists of a diverse set of tasks. All these tasks serve to support maintenance for either KLM, its partners or other customers. Under different approvals of the European Aviation Safety Agency (EASA), Engineering and Maintenance (E&M) is licensed to perform maintenance because it does so according to certain regulations. As a result E&M is allowed to not only perform the actual physical maintenance on aircraft and their components it is also allowed to design, certify and perform other related support tasks, which falls under the responsibility of engineering. Their tasks, not physical maintenance, is the demand relevant to our research. Section 2.1.1 elaborates on the categorization of demand in different characteristics; the kind of task performed, for which customer, which aircraft type it concerned and the routineness of tasks. These characteristics follow from the information available in registered demand data. With a clearer overview of demand and the present characteristics, several external factors that influence the uncertainty of demand are addressed in Section 2.1.2.

2.1.1 Demand characteristics

2.1.1.1 *Engineering tasks*

The tasks performed by engineering are ranging from developing repairs, project management, planning and certifying overhauls, to performing administrative and documentation tasks. There are 62 different tasks identified by the engineering department. The list of 62 tasks and their descriptions are presented in 0. To structure these tasks they are divided by into different categories. Table 2.1 shows the 12 different categories into which they are divided.

Table 2.1 Categories of engineering task

ID	*Category*	*# of sub tasks*
1	AMP Management	6
2	Fleet Performance Management	4
3	Data Management	3
4	Operator Support	11
5	Production Support	5
6	Maintenance Package	2
7	Maintenance Consulting	4
8	Data Management	2
9	Design Engineering	3
10	Transaction Services	4
12	Internal	16
13	Absenteeism	2

Note. ID 11 is a legacy code and thus omitted

As shortly touched upon in Section 1.1.2 these categories contain a wide array of different engineering tasks, a few examples are:

- Provide production (the physical execution of maintenance) support:
  - With the necessary documentation and instructions to execute tasks according to regulation.
  - Direct support on question unclarified on instructions.
- Design and certify repairs.
- Evaluate and certify/approve service bulletins from original equipment manufacturers.
- Provide expert/specialist support on unconventional issues.
- Evaluation, certification and approval on the phase-in and -out of aircraft, parts and modifications
- Project management for major overhaul/modification to aircraft.

Some of these tasks might fall into one specific category, repair development for instance falls under ID 12: Design engineering and has a dedicated product cod RD. Larger projects require multiple tasks from different categories. Each task describes a specific activity, requiring different skills/knowledge and is therefore suspected to have different demand behaviours.

2.1.1.2 *Customers*

Engineering tasks are performed for KLM, its partners and other clients. In a sense they are all customers of the engineering product and they are generally referred to as customers from here onward, unless otherwise specified. The tasks, in their core, are comparable regardless whether it is performed for KLM or external customers. In order to differentiate between the different parties they are all classified with specific codes. KLM E&M divisions (internal customers) are specified to department levels, external customers are classified under a general code ZZ and a unique identifier. Column 1 and 2 of Table 2.2 illustrate the different customers and their codes, due to confidentiality external customers will not be explicitly named in this report. Overall, work for KLM (E&M) is the largest portion of demand at nearly 95%. But demand can be very different per each customer, each customer has different agreements on tasks for engineering to perform. Some might require daily operational support, others might only assign one specific project. Each customer's demand can therefore be assumed to behave differently.

2.1.1.3 *Aircraft types*

Tasks are also be performed for different types of aircraft, in general the tasks are always performed for types that KLM has in their own fleet. KLM operates a fleet consisting of 5 different types in December 2017:

Boeing 737	50 aircraft	Average age of 10 years
Boeing 747	18 aircraft	Average age of 22 years
Boeing 777	29 aircraft	average age of 9 years
Boeing 787	10 aircraft	Average age of 1.5 year
Airbus A330	13 aircraft	Average age of 8 years
Fleet (total)	120 aircraft	Average age of 10 years

From here on the aircraft will be referred to by type number (737,747, 777, 787, and 330).
The age and type of use (e.g. European or intercontinental) per type have an effect on the amount or frequency of maintenance. In effect demand is different in both required tasks and amount per

aircraft. To illustrate, the 747 is being prepared to be phased out in the coming years. As such no large overhauls are planned but phase-out tasks and documentation become more important. On the other hand, the 787 is fairly new and is expected to have "infant mortality" failures leading to more work. We can conclude from this that different aircraft can be expected to have different demand behaviour. Additionally, a lot of work is not explicitly connected to a certain type, often involving more general support activities.

2.1.1.4 *(Non-) routine demand*

In order to perform tasks for customers there are agreements on what work to perform and the amount of work in hours. These predefined agreements result in 'routine' engineering tasks that can freely be charged to the customer when they request support and if the expended time falls within the agreements. What routine demand consists of differs per customer, production support might be routine for E&M subdivisions but not for an external customer. Additionally, these routine tasks are usually allowed to be within certain hourly boundaries per request without the number of requests being agreed on. This can result in variable amount of work and as such these agreements do not serve as a good forecasts.

When tasks fall outside standard agreements, non-routine tasks are issued. Such tasks are called sales orders (SO) as an additional 'sale' outside of current agreements was made. Non-routine tasks vary largely in their scope and extensiveness. They can range from a feasibility study of an hour to modification/overhaul projects that require dedicated teams for more than a year. As such they can have a sizeable effect on the division of work within departments. Their diversity is expected to create demand behaviour specific to that demand.

In order to differentiate between the types of work two classifications are used. Routine tasks are classified by their customer code as explained in 2.1.1.2. If the customer is a KLM division it is complemented with an aircraft type when applicable. To illustrate, following the codes from Table 2.2 a routine task for KLM regarding a 777 is identified by CW/777 and a task for Air France as ZZ/AFI. Non-routine tasks are given a unique identifier of 6 numbers that can be translated to the customer that requested it.

Table 2.2 Customer and type codes

Customer	Customer code	Type code (5 most common)
Central engineering (E&M)	*CE*	
Engine services (E&M)	*TM*	
Component services (E&M)	*VA, VC, VI, VR*	*330, 73N, 744, 777, 787*
Airframe H (E&M)	*TL*	*330, 73N, 744, 777, 787*
Airframe P (E&M)	*TZ, TT, TF, TG*	*330, 73N, 744, 777, 787*
E&M other	*TA, TP, TQ*	*330, 73N, 744, 777, 787*
KLM	*CW*	*330, 73N, 744, 777, 787*
Maintenance control center (KLM)	*TO*	*330, 73N, 744, 777, 787*
External customers	*ZZ*	*16 External customer codes*

2.1.1.5 *Total demand*

In the previous subsections we evaluated 4 different characteristics that make up demand for Engineering. We are able to conclude that each of the characteristics have a sufficiently different nature that demand for any of them is likely to exhibit characteristics not necessarily shared with others. Section 2.1.1.1 explained how demand consists of different tasks and that demand for categories, let alone individual tasks, can differ a lot. A similar conclusion was made in Section 2.1.1.2 where it became clear that different customers have different needs resulting in varying demand. Aircraft types in Section 2.1.1.3 are significantly different in design, age and amount, all of which influences the demand. Additionally, work can either be routine or non-routine. Section 2.1.1.4 concluded that the definition is customer dependable as it is defined on what work is pre agreed upon (routine) and what is done outside of that scope (non-routine), again resulting in variable demand. This leads us to see that the experienced uncertainty in demand can be attributed the multiple differently behaving parts making up the whole. Especially considering that the 4 characteristics are not mutually exclusive. Subsets of demand exist for many of their combinations, each again with specific behaviour of demand. There is validity to claim that the uncertainty in demand is caused by, external, changes to these characteristics which we explore in the following section.

2.1.2 Drivers of uncertainty

The previous Section, 2.1.1, focussed on what engineering demand consists of and how it is divided in subsets based on those characteristics. Here we identify the external aspects that drive the uncertainty of demand through those characteristics. This helps us to identify information that we require demand data to have. Through organizational knowledge, context and analysis of the demand process, several possible drivers of uncertainty were identified. Engineering work, as described in Section 1.1.2 and 2.1.1, consists of continuous technical support for customers in the shape of designing, certifying, documenting and evaluating maintenance. In part this means that there are certain routine tasks to support the customer and non-routine tasks related to requests out of the predefined scope (as described in Section 2.1.1.4). Thus all engineering demand is caused by performing some form of support as required by the customer, be it expected (Routine) tasks or unexpected (non-routine) tasks.

Focussing on routine tasks we assume that changes in intensity of airline operation affects the required amount of maintenance on those aircraft. Maintenance is expected to correlate with certain engineering tasks. A different utilization leads to different (frequency of) faults to maintain in turn affecting engineering demand. While this assumption seems straightforward there is no/limited insight in the seasonal effects on engineering demand. As a result, it is unclear how a change in operation actually affects engineering work. Airline operation is observed to change overtime and varies per season leading two the first 2 drivers of uncertainty:

- Limited insight in the effect of seasonal change in engineering demand.
- A lack of understanding and knowledge on change and growth in demand for maintenance and how it affects demand.

Further uncertainty is caused by changes in the aviation market and their customers' expectations. Low cost airlines are growing rapidly and passenger expectations are always changing (for instance on comfort and connectivity). To make sure that their product is still attractive airlines need to adapt where necessary leading to changes in engineering demand. This presents uncertainty driver 3:

- Uncertainty in how changes in the market and passenger needs affect demand.

Apart from the changing market, customers also need to adapt their aging fleet of aircraft. Some types have been flying for decades and experience has provided knowledge on what to expect. Introductions of types and phasing out old ones could cause shifts in demand. For instance the 787 is from a completely new generation where much more of its workings are electrical instead of mechanical. This might require different kinds of knowledge and support. This is uncertainty driver 4:

- The fleet of aircraft is developing which has an uncertain effect on engineering demand.

The fifth and final driver of uncertainty stems from the wide range of non-routine tasks. As stated in Section 2.1.1.4, non-routine tasks can vary from a task of a single hour to a team of 8+ FTE for more than a year. This makes it difficult to provide reliable forecasts for non-routine tasks apart from expert opinions. Leading to the uncertainty driver 5:

- Diverse non-routine tasks with limited insight in their demand and its patterns.

These drivers are expected to cause/influence the variability observed in demand. Figure 2.1 illustrates how total demand is affected by the 5 identified causes of uncertainty.

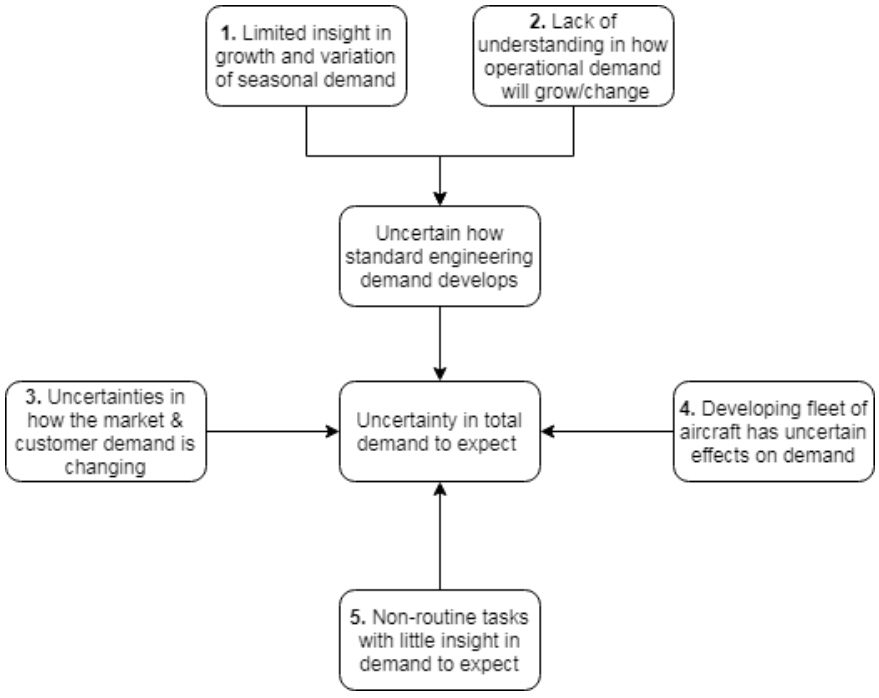


Figure 2.1 Drivers of uncertainty in the total engineering demand

We expect these drivers to cause part of the variability in demand and in order to forecast and gain insights we need the data to have information on the drivers. If it does not contain this information we suspect that no proper forecast can be made as relevant information is not available. The drivers follow from the characteristics in Section 2.1.1 which are deducted from available data. Therefore, the data should contain most of the required information, the following information linked to each driver in order should be available:

1. Historical data spanning multiple seasons to show if and what demand (subsets) have different behaviour depending on seasons.
2. Sufficiently long historical demand to provide information on general changes in demand over time.
3. Customers per task to provide information on their specific demand.
4. Aircraft types per task to show the different effect of new and old craft in the fleet, indicating the effect of fleet development.
5. A distinction between Routine and non-routine tasks to analyse demand behaviour for both.

Analysing the demand, how it is structured and how this might drive its uncertainty has provided insights for answering research Question 1: "What does the engineering demand consist of and how does it cause uncertainty?" Demand for engineering consists of a wide range, and combinations of tasks for customers and aircraft types, both on routine and non-routine basis executed by different suitable departments and teams. From these building blocks of demand and aided by organizational insight and context we were able to define 5 drivers of variability in demand that possibly cause uncertainty. The forecast should be able to use information on these and in order to do so it requires the available data to contain relevant information, which will be discussed in Section 2.2.

2.2 Historical demand data

The general structure of engineering demand has been identified in Section 2.1 along with how it causes variability. This section focusses on how that structure is documented in data, its quality and characteristics answering research question 2 in the process. First, the general structure of the data is analysed in Section 2.2.1. Then, some necessary and useful data transformations will be discussed in Section 2.2.2.

2.2.1 The data structure

The available data uses the demand characteristics as described in Section 2.1. It contains the registered engineering activities per employee per day, these work hours are used as a proxy for demand as described in Section 1.3.2. Currently, the available data ranges from 2012-2017 and Table 2.3 shows a sample. The task codes from 2.1.1 (and 0) can be seen in the prd.code column. The customer code and type as described in Section 2.1.1.2 and 2.1.1.3 are found in the IVS code column as is the sales order (SO) number of a non-routine tasks (see the first row under IVS code) as described in 2.1.1.4.

Table 2.3 Source data

Rec. Order	OpAc	Description	Pers.No.	Number	IVS code	Date	Prd cod
3005123	10	WR.14.020: Seat density 737-700	xxxxx	10	176106	2-1-2015	MO
3005276	10	PFO PH-BVN/BVO TR.14.777.005	xxxxx	8	TL/777	2-1-2015	MO
3002111	10	Repair Development KLM A330	xxxxx	5	CW/330	3-1-2015	RD
3004302	10	Verlof (vakantie, ATV)	xxxxx	8	CE	3-1-2015	VA
3004456	10	Repair Development KLM 744	xxxxx	3	CW/744	3-1-2015	RD
3004682	10	NDO werkzaamheden in H11 aan 777	xxxxx	3	TL/777	3-1-2015	XH

Apart from the previously described demand characteristics the sample contains additional details. In order of the table:

Rec. Order	Unique numerical identifier for a set of tasks for a job/project.
OpAc	Line number of the rec. order, specifying a certain task.
Description	Written description of the task.
Per. No	Numeric Identifier of 5 characters for person that performed the task (anonymized).
Number	The number of hours worked on the task on that day.
IVS code	Either a SO or the customer and type code as explained in 2.1.1.
Date	The date on which the number of hours were worked.
Prd Cod	The engineering task as explained in Section 2.1.1.

The presented data contains the data characteristics as described in 2.1.1. Additionally, we can identify the necessary information for forecasting and insight, as stated in Section 2.1.2. For each of the driver of uncertainty we show how the data contains the information.

- Information to see the effect of change in demand over time, drivers 1 and 2, need sufficient historical data to identify trends and season throughout the years. As data ranges 6 years from 2012 through 2017 it contains the required information.
- Information on the effect of the changing market and passenger's needs, driver 3, is identified by separate analysis of customer demand, the IVS code provides this.
- The effect of different types in the fleet and their development, driver 4, is addressed by information specifying for which type work was performed, this is again provided by the IVS field.
- The distinction in the IVS field between a regular code and a SO number also provides information on routine vs non-routine tasks required by driver 5 to identify their respective behaviour.

Even though, no analysis has taken place yet we can conclude that the data should contain the information that we require. Its structure, frequency of registration and history going back 6 years should allow us to forecast into the future with information from the past. Some improvements to the clarity and density of the information can be improved in order to enhance both the quality and ease of analysis, to do so we enrich the data in Section 2.2.2 before moving on to a data analysis in Section 2.3.

2.2.2 Enriching the data

The source data from 2.2.1 provides the information that we desire but we can make it more information dense and accessible by splitting, adding and transforming some of the data. In order to illustrate the steps taken to enrich the data we show the effects of the adjustments on the data of Table 2.3. Table 2.4 shows the resulting adapted dataset with additional information. Where possible each characteristic as defined in Section 2.1.1 is given its own field in order to increase clarity. Additional information was extracted from the description field to further enrich the available information. Section 4.2 elaborates on the steps taken to adjust the data and make it more manageable.

Table 2.4 Enhanced dataset

Rec.Order	OpAc	description	Pers.no	Hours	Date	Prd code	IVS	IVS enriched	Type	SO?	Cust code	KLM?
3005123	10	WR.14.020: Seat density 737-700	xxxxx	10	2-1-2015	MO	176106	176106	73N	TRUE	KLM	TRUE
3005276	10	PFO PH-BVN/BVO TR.14.777.005	xxxxx	8	2-1-2015	MO	TL	316017	777	TRUE	VOH	TRUE
3002111	10	Repair Development KLM A330	xxxxx	5	3-1-2015	RD	CW	CW	330	FALSE	FS	TRUE
3004302	10	Verlof (vakantie, ATV)	xxxxx	8	3-1-2015	VA	CE	CE		FALSE	CE	TRUE
3004456	10	Repair Development KLM 744	xxxxx	3	3-1-2015	RD	CW	CW	744	FALSE	FS	TRUE
3004682	10	NDO werkzaamheden in H11 aan 777	xxxxx	3	3-1-2015	XH	TL	TL	777	FALSE	VOH	TRUE

Per column of Table 2.4 we provide a short description of what was done to increase usability, unchanged up to prd.code:

Prd Cod	Where applicable legacy codes were replaced by current standard (e.g. TB now TL).
IVS	Contained code and type identifier for either plane or customer now split.
IVS enriched	The rec.order or description was matched to a SO and documented (row 2).
Type	Type taken from original IVS or derived from description (row 1, from description).
SO?	Sales order identifier, if true the task was part of a non-routine job.
Cust.code	Customer identifier, for internal divisions or external customers. Either derived from IVS code, SO number or the description. Divided in all customers and subdivisions.
KLM	KLM identifier, true when customer code is part of KLM. Divided in KLM vs external customers.

Adding these columns allows subsets of demand data to analyse the drivers of uncertainty as discussed in 2.1.1. Demand can now be easily split on aircraft, customer, tasks, routine characteristics or combinations thereof, creating a large number of different subsets. Section 2.3 explores the different demand subsets and their behaviour of the data in more depth.

2.3 Demand data behaviour

Section 2.2 presented the available data and the information it contains that will help in determining the behaviour of different demand subsets of interest. In Section 2.1 we stated that there are several demand streams of interest, complexity arises because these characteristics are not mutually exclusive. Tasks can be done for any type and customer on both routine and non-routine jobs, likewise a customer can have different types requiring different tasks. We call a combination of different demand characteristics a *subset* as it only a part of total demand. This section will serve to provide initial insights in the behaviour of demand subsets. We do so by shortly introducing relevant data behaviour in Section 2.3.1 and subsequently identify these in selected subsets in Section 2.3.2.

2.3.1 Some relevant data characteristics

To compare the different demand we need an objective way of looking at them. We provide initial insights with two relevant characteristics, trend and seasonality, both are described as well as a way to detect them. These are relevant for now as they can provide information on past demand behaviour and which we need to predict how it might behave in the future.

Trend (-cycle)

The trend is a long term in- or decrease in the data which can change over time (Hyndman & Athanasopoulos, 2018, p. 2.3). This can be observed in data as a non-zero slope. In reality it cannot be expected that a certain slope will stay truly constant over time and the intensity of the in- or decrease will change. A lot of time series actually exhibit periods of up and down changes in a cyclical manner over time without a fixed frequency (Hyndman & Athanasopoulos, 2018, p. 2.3). Because they are hard to distinguish from one another the trend and cycle are often considered together as the trend-cycle.

Seasonality

Seasonality is a periodic change of in- and decreases depending on the 'season' it is in (Hyndman & Athanasopoulos, 2018, p. 2.3). It differs from a cycle as it has a fixed and known frequency and can thus be anticipated, e.g. ice cream sales peak in summer months and weekends experience less traffic.

Remainder

If trend and seasonality are removed from the data set the remainders are what is left. It represents the unexplained variation of the data. This variation can be due to inherent uncertainty or in effect to other unidentified (external) factors.

Decomposing a time series

Time series can be broken up into its respective parts by decomposing. Figure 2.2 by Hyndman et al (2018, p. 6.6) shows an example of a decomposed data series, split into its seasonal, trend and remainder components through STL decomposition as developed by (Cleveland, Cleveland, McRae, & Terpenning, 1990). By evaluating the parts separately their effect can be seen much more clearly. We conclude this to be a sufficient first step in the data exploration. Section 2.3.2 will provide insight in some of the possible demand subsets and their characteristics.

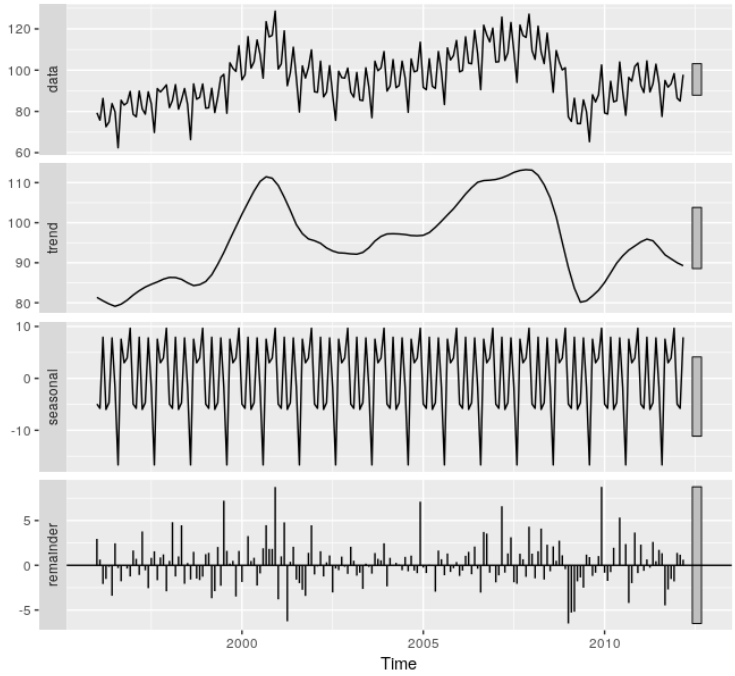


Figure 2.2 Decomposed time series (Hyndman & Athanasopoulos, 2018, p. 6.6)

2.3.2 Demand behaviour in the data

Many different subsets of demand can be considered and analysed in the data. To gain some initial understanding a few were selected to inspect their trend and seasonality. Some top level subsets are considered before moving down to a more specific overview to see how their behaviour and uncertainty change. We decompose the demand using STL (Cleveland, Cleveland, McRae, & Terpenning, 1990) to visualize the characteristics with an assumed seasonality of 12, relating months in different years to each other.

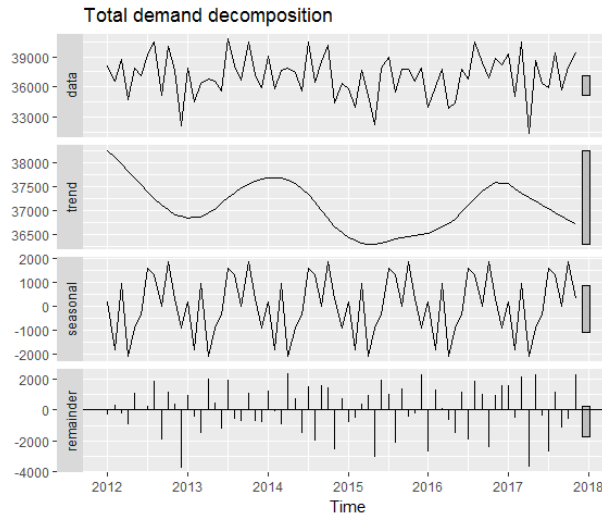


Figure 2.3 Total demand decomposition

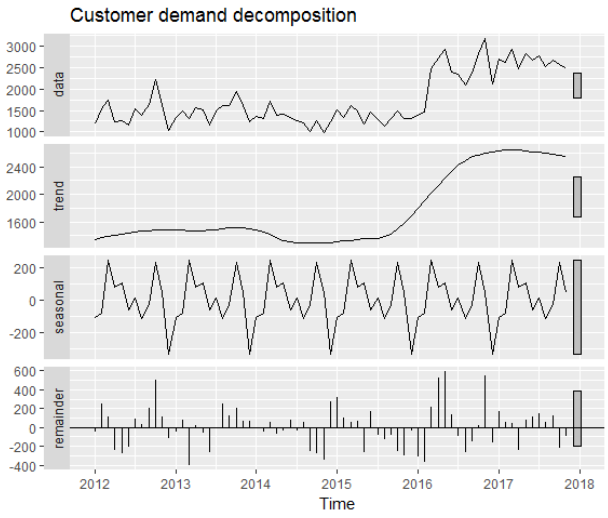


Figure 2.4 Customer demand decomposition

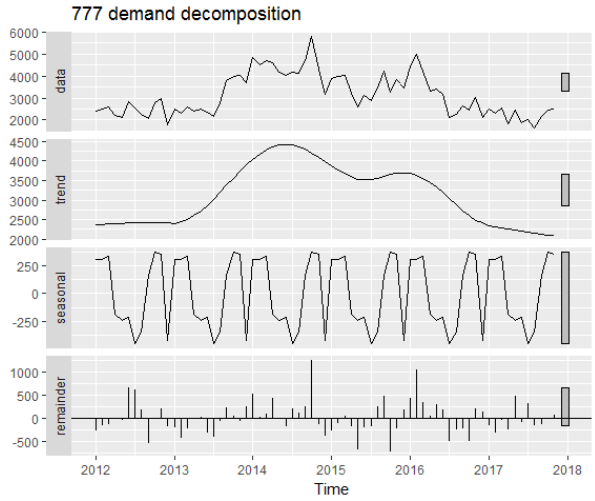


Figure 2.5 777 demand decomposition

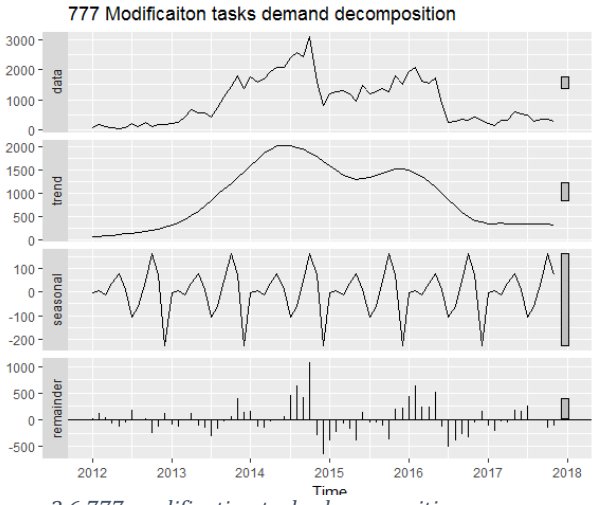


Figure 2.6 777 modification tasks decomposition

Total demand is decomposed in Figure 2.3, it shows a trend-cycle but it has a small effect compared to seasonality and the remainder when the scale is taken into account. With the characteristics from Section 2.1.1 we can make any desired subset of demand. Our first choice is take a subset of demand for all external, non-KLM, customers decomposed in Figure 2.4. A clear level shift can be seen around 2016 where the average demand appears to jump to a higher level, seasonality plays a smaller role. Demand for the 777 is decomposed in Figure 2.5 it experiences a clear rise in demand from 2013 through 2016 and experiences small seasonal effects. We take a further subset of the 777 demand,

focussing on the biggest contributor of demand over time, modification tasks (MO in Appendix A). Figure 2.6 decomposes the modification tasks for the 777 and a similar pattern to that of the total 777 can be seen. The trends of the total 777 demand and that of the specific modification tasks look nearly identical in shape and range. The behaviour of the total 777 demand is affected strongly by a modification project which influences the characteristics of the top level.
These decompositions clearly show the different dynamics experienced by different subsets of demand. Each possible combination of the characteristics described in Section 2.1.1 (customer, type, task and routine) has demand with different behaviour presenting different information. The trend in Figure 2.5 and Figure 2.6 coincide because of a large cabin modification project to the 777 type aircraft during that time. If only the total 777 demand was considered it would not have been clear that the change was due to one type of tasks/project. It is reasonable to assume that this holds true for any subsets of demand, more specific combinations of characteristics produce demand information unique to that level. From this we conclude that all subsets contain potentially useful information and that all of them should be considered and forecasted

Concluding that all different combinations and resulting subsets are of interest has a substantial impact. The number of combinations is high and it is not realistic to assess series manually, yet as seen in this section we can conclude that each shows different behaviour. This provides a high level answer to research question 2: "What are the characteristics of the available data and is it suitable for forecasting?" The characteristics of demand change with each different subset considered. As there is no singular way to describe the demand we can assume that the forecast suitability of the data will depend on how the model can handle the different characteristics. With this conclusion we note that the forecast model needs the ability to handle a wide range of demands with different behaviour and characteristics. Additionally the model will need to be in line with the forecasting desires and practice of the organization analysed in Section 2.4 by looking at the current forecasting practice.

2.4 Current forecasting practice

The demand characteristics and behaviour have been addressed in the previous sections of Chapter 2. This helps us to choose what kind of model to consider in the coming chapters. However, to be successful we want to reach the desired situation from Section 1.2.2, with more accurate and fact based forecasts as the goal. As stated by Duffuaa and Raouf (2015, pp. 20-21) and, Hyndman and Athanasopoulos (2018, p. 1.6), in order to properly leverage forecasting a forecasting goal is necessary. This requires us to ask why and what we are going to forecast. Chapter 1 and the previous sections of Chapter 2 already provided answers to these questions from a research perspective. Yet, further insights are possible by addressing how forecasting is currently leveraged in the organization; which tools and techniques are used? No information from the data or forecast is leveraged apart from budgeting which can be split into 2 steps, a quantitative (statistical) forecast and a qualitative (judgemental) adjustment.

2.4.1 Quantitative forecasts

Once a year, after August, all the engineering work done that year is categorized into type of work and the relevant customers as previously explained in Section 2.1.1. This collection of data is then used for two goals, evaluating this year's budget and determining a budget for the coming year. 2 methods are used:

Mean method

An extrapolation of the average hours spend is used to forecast the following year. The hours spend up to august are divided by 7 (# months passed) and multiplied by 12 (# months total) to get an indication for a yearly total.

$$Total\ exp\ \#\ hours\ current\ year = \frac{\#hours\ up\ to\ July}{7} * 12$$

This total is assumed to be close enough to the total for the coming year. As a result the average monthly demand is taken, applied to the entirety of the coming year and used as an indication for the budget. This extrapolation disregards any seasonality or trend present in the data but is intuitive and easy to apply.

Trend method

A slightly less simple but still intuitive method is the trend method. The same 7 months of data are used, but instead of taking the average, a linear trend is fit on the data. The trend is then extrapolated to 12 months and the sum gives the forecasted demand for a year. This number is used to evaluate if the current budget is on track and to see whether demand will in- or decrease.

Both methods have issues, the average or the trend is assumed to stay constant for the coming periods by these methods. In itself this is not a terrible assumption, forecasting requires us to assume that at least some part of the behaviour will be stable over time. However, the data used for this forecast make this an unwise assumption. First, the data used to forecasts are only of a period of 7 months. This period is shorter than the desired forecast horizon and will fail to capture any dynamics experienced in a longer timeframe. Especially when we consider that using 7 months of this year to forecast the entirety of next year means that we are foresting 17 months ahead (Aug. this year until next Dec.), a period nearly 2,5 times as long as the used information. The assumption that the average/trend will hold over this entire period needs to be correct for the forecast to make sense. Second, because the data only reflect 7 months of demand from this year no information on any long term averages or trends is available.

These methods have been chosen to forecast due to their intuitiveness. No expertise on more complex forecasting methods is available and has therefore been applied within the available means. Additionally, the methods are said to have worked well in the past, reportedly reaching r-square values of 95%. However, this is not substantiated by the data, fitting a linear trend yields an average r-squared of about 30% implying a poor job of explaining variation in the demand. This is to be expected as the historical data used is limited and variable, which a straight line fit will not capture.

2.4.2 Qualitative adjustments

The quantitative forecast produce a base forecast used by those responsible for making the budget. Their experience is that the quantitative forecast provides a ballpark estimate and need to be adjusted based on experience and gut feeling. This process has been jokingly called “the big yearly Sudoku”, meaning that it feels like moving the numbers until they makes sense. In order to produce proper estimates relevant parties, like team leads, are consulted to reflect on the demand and what they expect for next year. In this way subsets for each team can be considered and adjusted according to expert knowledge and opinion. Routine and non-routine jobs are both taken into account through adjustments and then aggregated to arrive at a yearly forecast, two judgmental adjustments are highlighted.

Regular/routine tasks

To come to a final budget, management asks relevant team leaders who discuss with team members on a realistic number for the coming year. The aggregate of these teams is put next to the mean method Section 2.4.1. A concession between the two is made, usually, to the side of the judgemental forecast, because the expert opinion is valued over the mean. As discussed further in Section 3.4.1, this is not best practice for applying judgemental forecasts and clear rules should be in place while preferably using and trusting statistical forecasts. On the other hand, teams usually have a good feeling on out of the ordinary events and work for the coming year, like the introduction of a new type of aircraft or a large scale modification. These kind of unexpected variations in demand cannot be derived from historical data.

Projects

As with the routine-tasks the forecast for projects (a substantial part of non-routine tasks) is done on gut feelings. Project managers are in charge of the large scale projects and also for a forecast on the expected hours required. Based on inside information, their own expectations and rumours they have some leads on what kind of work and projects to expect, forming the basis for their forecasts. Experience has taught them that more often than not projects are 'spontaneously' requested, often in response to market developments. This could result in unforeseen hours necessary for projects. As these usually have high priority, hours of regular work might then be 'cannibalised' to perform project tasks. Experience with these projects has taught them to overestimate the required hours in order to compensate.

Apart from budget forecasts they produce forecasts for individual projects. This is done based on their experience and through considering analogous projects from the past. The main goal of this is to provide a planning of when what kind of work will be necessary. While doing so no models are used apart from applying a factor over the hours from comparable previous projects. In short they base their decisions on their experience and opinion combined with an analogy forecast where they consider past projects. The principles of forecasting in Section 3.4.1 would require them to use data as much as possible and use structured rules when applying adjustments.

2.4.3 Final forecast

Section 2.4 has answered research question 3: "What is the current demand forecasting practice?" Simple quantitative forecasts and substantial qualitative adjustments are used to arrive at a strategic level forecast for the total budget required to fulfil all hours. In the end this budget performs fairly adequate because of its high level. Over- and underestimates effectively cancel each other out when everything is aggregated for the total forecast. This leads to hours originally forecast for certain tasks being borrowed/used for other tasks, deficits in one area are filled by a surplus in another. As a result the aggregate gives a reasonable indication but the further you drill down in the forecast the less accurate it becomes. More sophisticated forecasts are not applied due to a lack of expertise and reliance on expert knowledge. Analysing the current forecasting practice completes the system analysis, Section 2.5 will summarize our findings.

2.5 Conclusion

Research questions 1, 2, and 3 focussed on understanding the current situation through a system analysis. The subjects were, respectively, demand, data and current forecasting practice.

Research question 1: *"What does the engineering demand consist of and how does it cause uncertainty?"* was answered in Section 2.1. We found that demand is constructed from a wide range of tasks performed for a collection of customers and types on routine and non-routine basis. We conclude that these different parts of demand are responsible for steering the demand and driving its uncertainty. We found that the causes of uncertainty could be identified in historical demand, given that it contains the characteristics driving demand.

Research question 2: *"What are the characteristics of the available data and is it suitable for forecasting?"* was answered in Sections 2.2 and 2.3. We focussed on the data, whether it contained the characteristics necessary to differentiate between the different subsets of demand. Analysis showed that different subsets of demand contain different information that explains variability, be it demand for a type, customer, task or combination thereof. The different behaviour depending on the subset also implies that the forecast suitability changes per subset. From this we conclude that all different subsets are of possible importance, implying that the forecasting model will need a way to handle the different characteristics that the demand behaviour exhibits. In order to handle all different demand subsets we believe an automated process to be necessary as manual tuning for each subset would require too much time.

Research question 3: "*What is the current demand forecasting practice*?" found an answer in Section 2.4 which evaluated the current forecast process. Current quantitative forecasting uses two simple, but intuitive ways of extrapolating the demand to coming periods. By using limited historical data and a simple method potential important information in the data is disregarded. The statistical forecast is then used to judgmentally produce a final forecast mainly based on experience and gut feeling. The reliance on expert opinion creates potential for forecasting bias and errors. But the experience used in the judgmental adjustments is of importance as expert knowledge can anticipate events that the data cannot predict. The proposed forecasting model should therefore have some way of incorporating this expert knowledge.

3 Literature review

Chapters 1 and 2 found that forecasting can provide the insight necessary to increase control over capacity. Assuming we could perfectly forecast demand planning, and thus control, are easy. However, forecasting is not straightforward and some things are easier to forecast than others. In Section 2.3 we learned that the behaviour of demand varies depending on the subset considered and the model should be able to handle this. "Good forecasts capture the genuine patterns and relationships which exist in the historical data, but do not replicate past events that will not occur again" (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018). The goal is to capture change through time and use this to predict the future state. In this section theory and literature on different methods are explored to answer research questions 4 and 5. Section 3.1 starts with a general description of forecasting. Section 3.2 introduces preferred forecasting processes. Section 3.3 discusses data transformations beneficial and sometimes required for forecasting. Section 3.4 describes a collection of both quantitative and judgemental forecasting methods. Section 3.5 describes ways to measure forecasting performance. Finally, Section 3.6 proposes ways to reduce uncertainty by combining forecasts.

3.1 Forecasting in general

A forecast aims to predict what the state of something will be like at a future point in time. To do so it looks at the past and projects an expected value to the desired date. This is trivial if the subject to forecast is constant through time. Realistically things are never as stable in an organizational context. The subjects to forecast usually do not exhibit stable characteristics and the best one can do is make the forecast as good as possible. But that raises the question, what is a 'good' forecast? A good forecasts utilizes the relations and patterns present in historical data while ignoring events that will not happen again (Hyndman & Athanasopoulos, 2018). When successful a forecast becomes useful for organizations in decision making, how useful the forecast depends on how accurate the forecasts are. How well something can be forecast depends on different factors, Hyndman and Athanasopoulos (2018, p. 1.1) identify three:

1. How well we understand the factors that contribute to it;
2. How much data is available;
3. Whether the forecasts can affect the thing we are trying to forecast.

They go on to state that dependent on these factors forecast accuracy can range from high to lower than 50%. Leading to the statement that an important step when forecasting, is recognizing whether something can be accurately forecast at all before producing the forecast itself. This might help to accurately set goals and expectations for the forecast, on the other hand complicated processes might not be easy to understand making it difficult to state beforehand what to expect. As the complexity in organizations is usually high it becomes key to structurally work toward a forecast to ensure its quality.

In our context we comply with points 2 and 3. Several years of demand data, a multiple length of our forecast horizon, is available and our forecast are independent of actual demand. Point 1 is more problematic, in Section 2.3 we concluded that all subsets of demand have potentially important information. As a result we do not fully understand each subset or their contributing factors to the total. We can only conclude that the forecast ability will depend and change with the behaviour of each demand subset.

3.2 The forecasting process

Several authors have constructed frameworks for producing forecasts. Duffuaa and Raouf (2015, pp. 20-21) present a flow chart, see Appendix B, starting with defining the characteristics of the forecasting problem. They branch of on the question whether data is available and then move on to quantitative analysis. Hyndman and Athanasopoulos (2018, p. 1.6) agree with this, by first defining the forecasting problem and goal before moving on to the analytical steps involved in exploring the data. Silver, Pike and Thomas (2017) present a similar framework, Figure 3.1, which captures the process from the analytical phase onward.

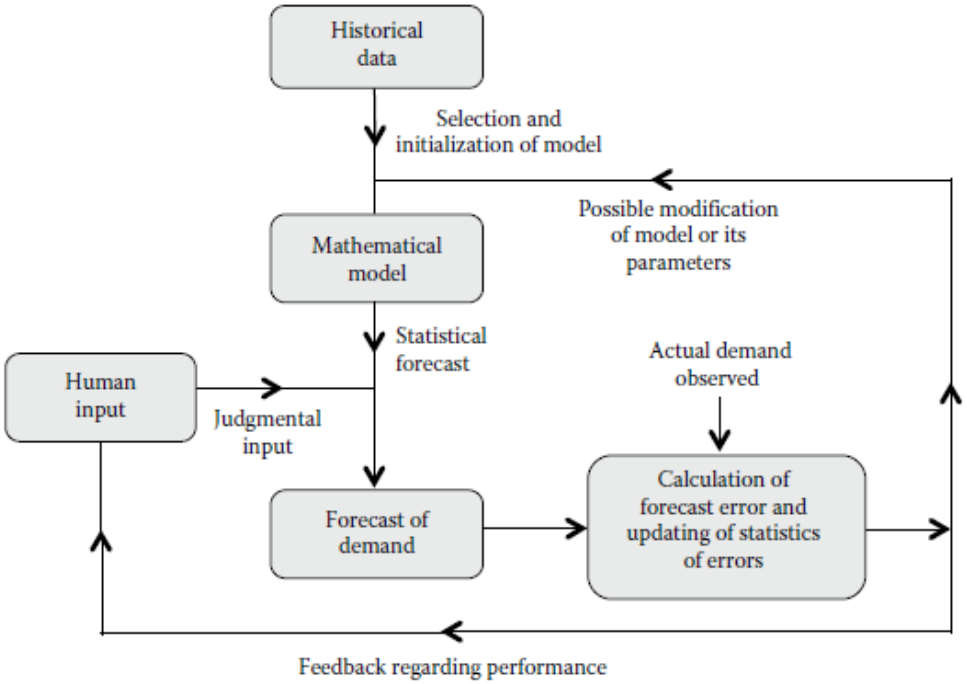


Figure 3.1 Forecasting framework (Silver, Pyke, & Thomas, 2017, p. 74)

It provides general steps on how to approach a forecast, its cyclical nature encourages evaluation and improvements over time. The first step identified after collecting data is selecting a model that fits its characteristics and describes the behaviour. In order to do so it is relevant to study the available data and extract the characteristics that explain how and why it behaves like it does. As seen in Section 2.3 manual inspection of all different demand streams would take too much time and either selection of important streams needs to take place or an automated method is required.

3.3 Data transformations

In order to make forecasts the information in the data needs to be extracted and continued into the future. To effectively extract this information, data transformations might be necessary. Depending on the method several assumptions on the statistical characteristics might need to hold. One such assumption is the data being "stationary", having stable statistical characteristics (e.g. the mean, variance, etc.) that do not change with time. Some methods require data to be stationary and most series can be made (approximately) stationary through appropriate transformations, some of which are discussed.

3.3.1 Stationarity

Non-stationarity implies that the statistical properties of data are dependent on the moment of observation. For instance, an upward trend causes the mean of the data to increase over time and seasonality might cause different variance at different moments. This change over time can be problematic, mainly because they make predictions more difficult and methods that assume certain statistical properties will not work (completely) as intended. Stable characteristics are preferable because "[i]f we wish to make predictions, then clearly we must assume that something does not vary with time" (Brockwell & Davis, 2016). As a result transformation of the data is sometimes necessary to remove or reduce sources of variation, leading to simpler patterns in the data and usually in better forecasts (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018).

3.3.2 De-trending and de-seasonalizing

Trend and seasonality, as explained in Section 2.3.1, cause data to have different values at moments of observation. Adjusting for this transforms the data to a more stable form which makes it easier to observe the underlying characteristics. Figure 3.2 shows a dataset with a slight upward trend and seasonality. Removing the trend would result in the data stabilizing along a horizontal line, effectively leaving the seasonal effect + the remainder. De-seasonalizing will remove the periodic fluctuations leaving the trend + the remainder. See Appendix C for the individual examples.

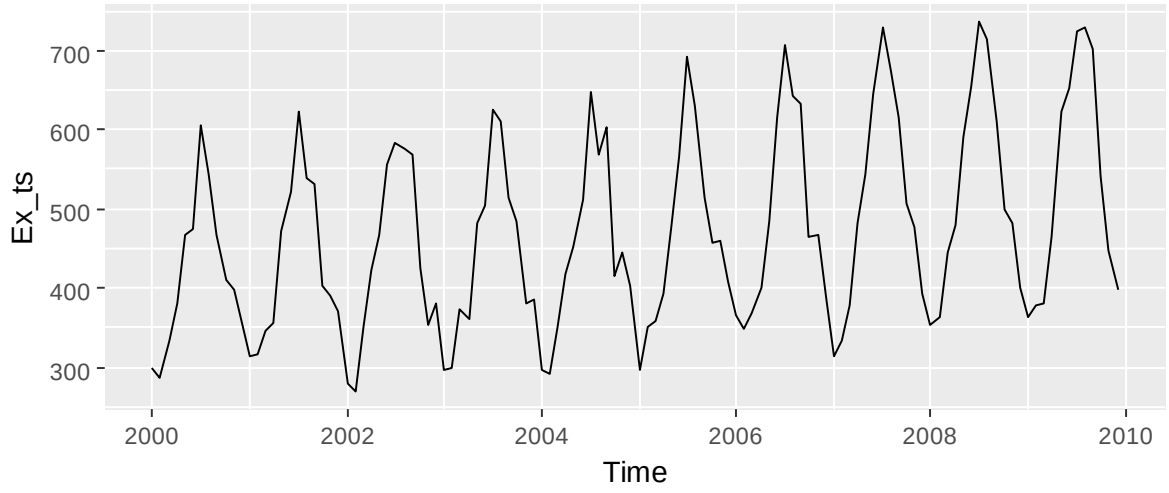


Figure 3.2 Example plot with dummy data

By isolating the terms the characteristics become much clearer. The STL (Cleveland, Cleveland, McRae, & Terpenning, 1990) decomposition in Figure 3.3 shows all the isolated parts. The trend is a clear upward line, the seasonality is symmetrical and the remainder appears random. The parts are stationary and methods that rely on this can accurately fit a model to the data. The example data was easily split into different parts because the trend was linear, the seasonality constant and the whole was an additive combination leading to a constant variance around the mean. If the series had multiplicative properties the variance or the trend, changes overtime requiring a different approach.

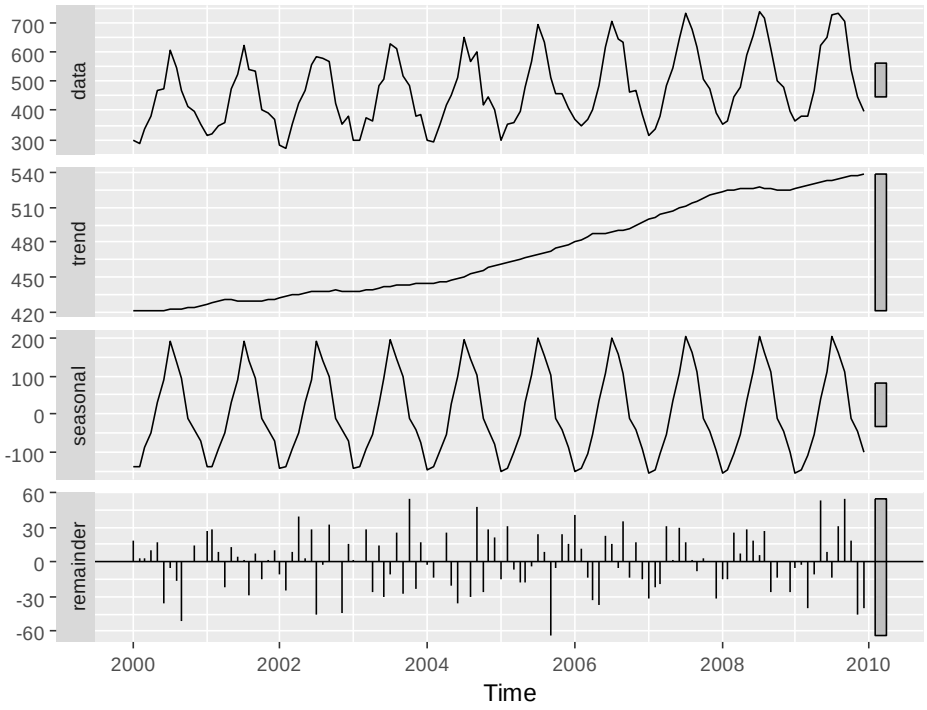


Figure 3.3 Decomposition of example data with trend and seasonality

3.3.3 Box-cox transformations

As we previously saw, data can contain different characteristics making up the original set. Trend, seasonality and the remainder all add up to the original data. The combination of these characteristics can be additive or multiplicative, this respectively shows as a constant variance through time versus a changing variation through time, as seen in Figure 3.4.

Graph A shows an additive series with constant variance while B shows a multiplicative one that increases with time. This changing variation violates assumption on normality of the data as well as stationarity. Box-Cox (BC) transformations (Box & Cox, 1964), can help to make data stationary by making the trend more linear and coercing a constant variation. BC transformations require nonzero observations y_t, where x_t is the transformed observation:

$$x_t = \begin{cases} \ln(y_t) & \lambda = 0 \\ (y_t^{\lambda} - 1)\,/\lambda & \lambda \neq 0 \end{cases}$$

It applies a (natural) log function when lambda is 0 and applies a power function otherwise. As a result the shape of the data changes as lambda changes, illustrated in Figure 3.5 for different values of lambda. Lambda should be set at a number where the variation in the data becomes as stable as possible, for Figure 3.5 anywhere between 0 and 0.4 seems reasonable. The transformed data can then be forecast more accurately due to more suitable characteristics. After the forecast it requires back transformation to show the result in the original units. Automatic estimation of lambda by minimizing the coefficient of variation was developed by Guerrero (1993) eliminating the need for manual adjustments. For the data in Figure 3.5 this method recommended a lambda of 0.085 fitting neatly in the range of the visual estimation.

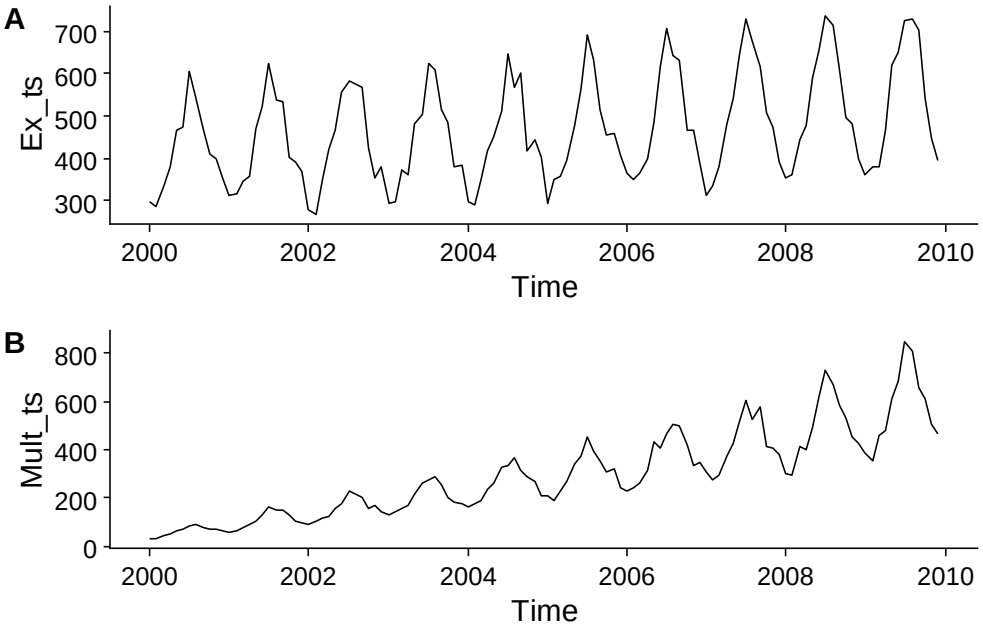


Figure 3.4 Additive vs Multiplicative graph. A: additive constant variation. B: Multiplicative increasing variance

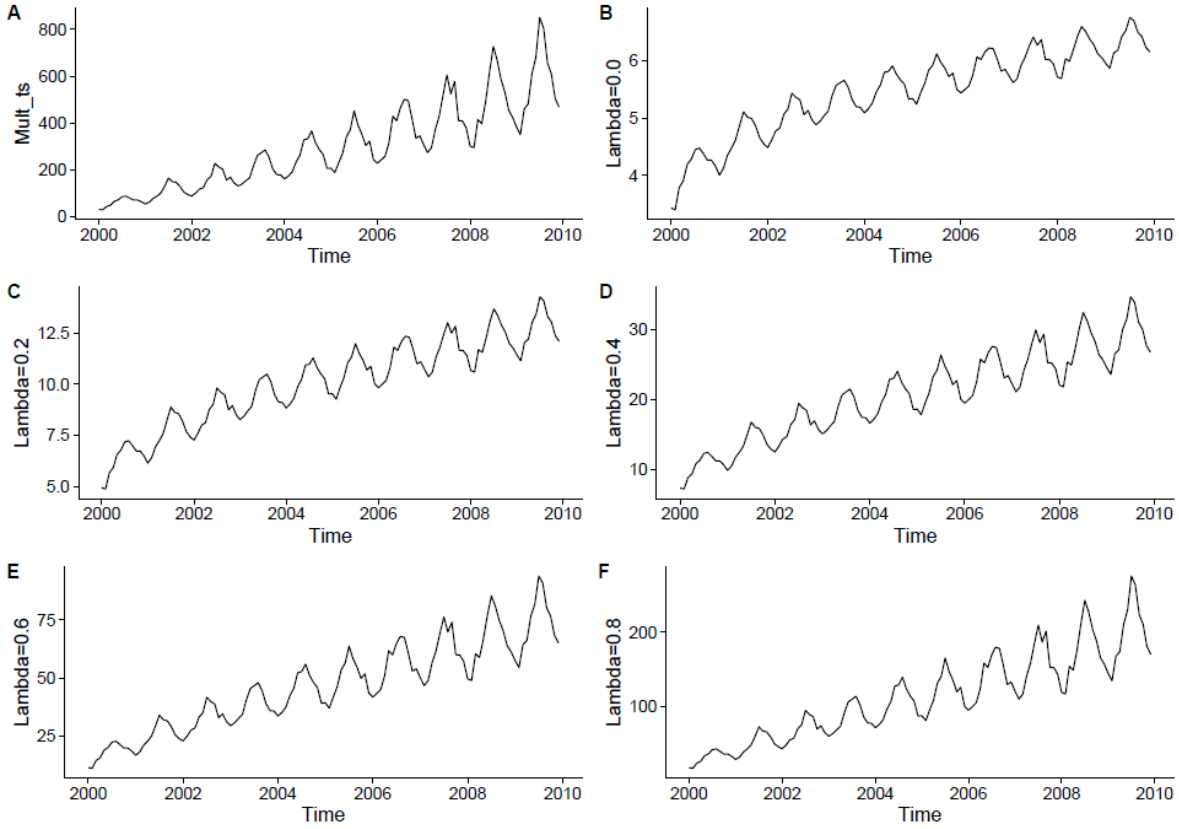


Figure 3.5 Box-Cox with different Lambdas. A: original series. B: lambda of 0.0, C: lambda of 0.2, D: lambda of 0.4, E: lambda of 0.6, F: lambda of 0.8

3.3.4 Differencing

Another approach to stabilize a time series is differencing. This method reduces or removes both the trend and the seasonality from the data by taking the difference between observations. Where a Box-Cox transform stabilizes the variance, differencing stabilizes the mean. By differencing the information conserved is the relative change between each observation according to the following formula:

$$y'_t = y_t - y_{t-m}$$

Here y_t is the original observation at time t, y'_t the transformed variable and m is the number of periods to difference. Regular differencing assumes a relationship with the preceding observation, m equals 1. Seasonal differencing assumes a relationship based on the seasonal period so m = frequency of the season (e.g. monthly seasonality means m = 12). Figure 3.6 shows the multiplicative example of Section 3.3.3 being differenced to stabilize the series. First the original (A), then the logged data (B) to stabilize the variance (a BC transform with lambda=0) and finally the differenced log series (C) to show the resulting stationary series with a stable mean and constant variance. This series had to be differenced twice, once for the seasonality in the data and once for the previous observation. An objective way to determine the number of differences needed is a unit root test like the KPSS test as determined by Kwiatkowski, Phillips, Schmidt, and Shin (1992).

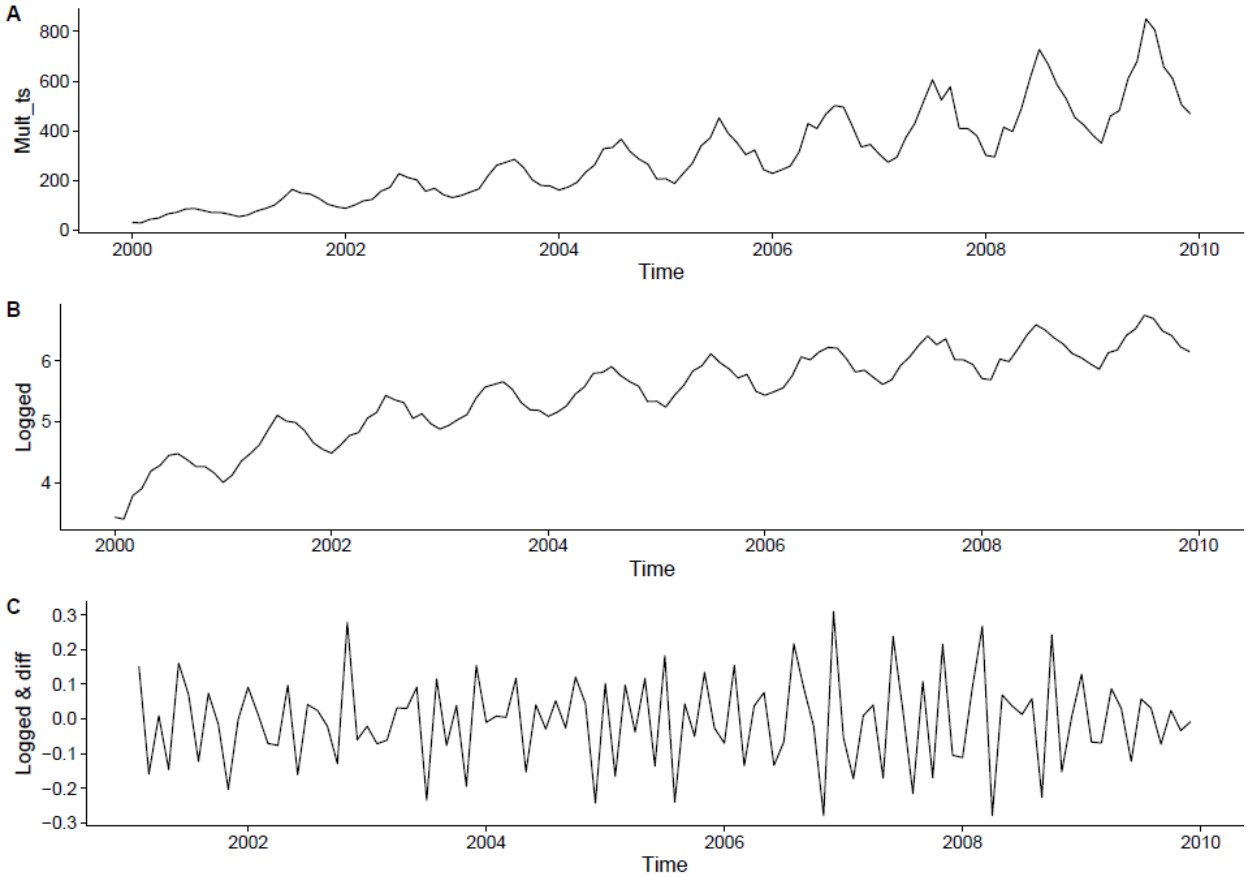


Figure 3.6 Observation to differenced stationary. A: original observations, B: Logged observations, C: Logged and differenced.

3.3.5 Data aggregation

Aggregation combines data from different levels or groups to produce an overview on a higher level. Aggregation can happen on any form of hierarchy in the time series for instance time or product groups (e.g. day → week → month and milk → dairy → fresh goods). Aggregation over such dimensions helps discover additional information in the data.

Time is an example of a hierarchy, there is a unique and defined order to it. 60 seconds in a minute, 60 minutes in an hour, 24 hours in a day, etc. This means there is a unique way to aggregate the data. Figure 3.7 by Athanasopoulos, Hyndman, Kourentzes, & Petropoulos (2017) aggregates a monthly time series to different levels, showing the difference in behaviour for different aggregations. The monthly data is highly variable and seems to have seasonal effects. Each further aggregation of the data smooths the variation further until the annual level shows only the trend. In general different aggregation levels present different information contained in the data. For instance, with temporal aggregation, seasonality present in lower levels is smoothened when moving up, while higher levels low frequency components like trend/cycles become clearer (Athanasopoulos, Hyndman, Kourentzes, & Petropoulos, 2017). Section 3.6.2 further illustrates different hierarchies and their effect.

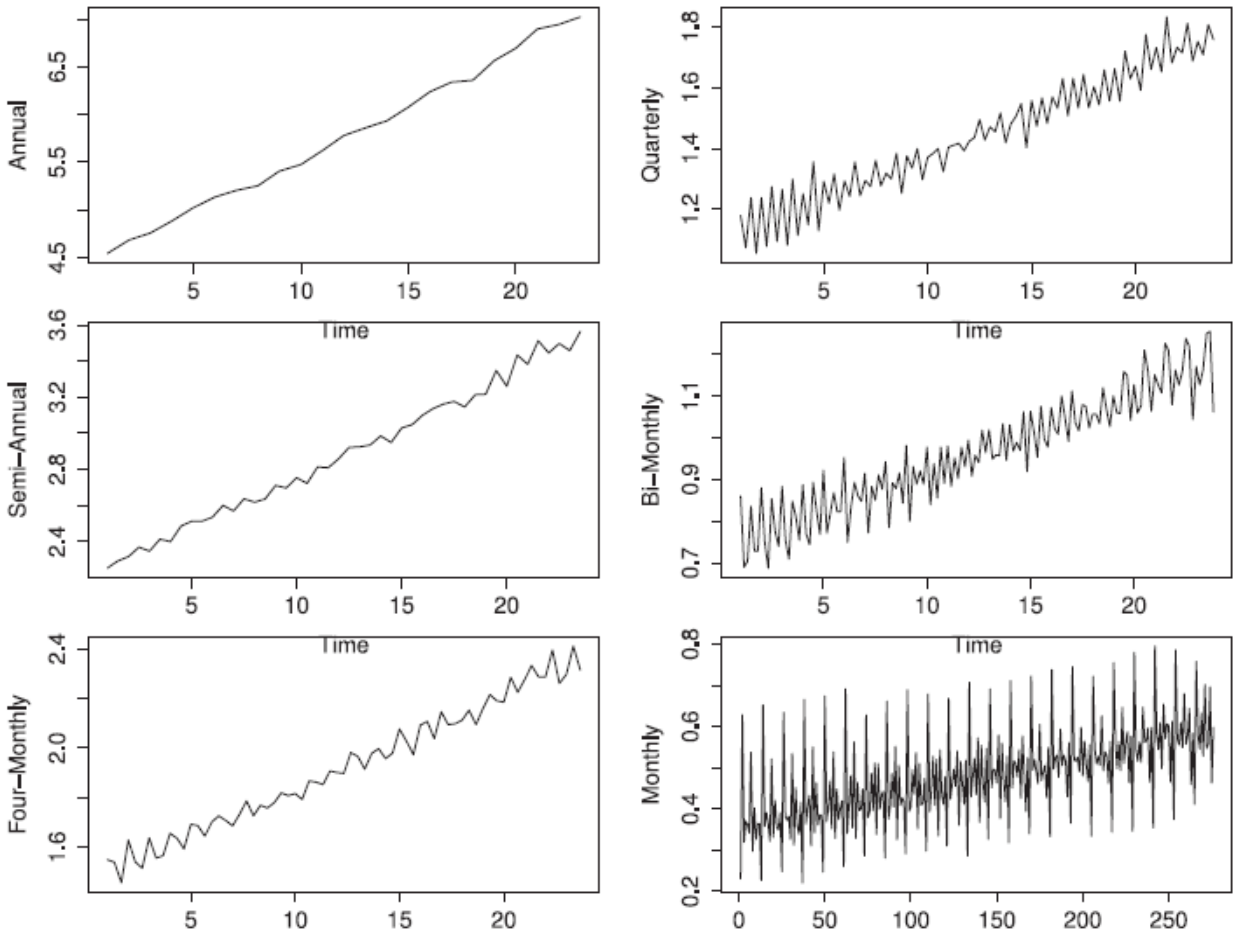


Figure 3.7 Example of data aggregation over time (Athanasopoulos, Hyndman, Kourentzes, & Petropoulos, 2017, p. 68)

One of the key decisions in aggregation is the choice of the lowest level to consider. For instance, if daily data is available the forecaster needs to decide if this level of detail will provide valuable information, maybe months will suffice. This choice has to be based on the goal of the forecast and its user. Strategic planners will most likely be interested in high level, long term forecasts while operational planners are more likely to require short term forecasts on individual products. Regardless of the chosen level Carson et al. (2011) found that" independent of the forecast horizon, the aggregate approach is always the worst performing approach, suggesting that there are always gains in terms of forecasting ability from using information which is available at the disaggregate level". Giving a strong argument for considering lower level aggregations when the goal is a higher level forecast. This gives strength to our conclusion in Chapter 2 where we stated that every combination of the available characteristics presents unique information and should therefore be considered. If we only evaluate high level demand we would lose information from the lower levels.

3.3.6 Conclusion

We first touched on the subject of data transformations in section 2.3.1 where we described the value of decomposing time series. This allowed us to obtain insight in the underlying components of data, making it easier to understand its behaviour. In Section 3.3 we focussed on why these and other transformations are sometimes necessary. The (quantitative) models that we will evaluate in Section 3.4 have certain assumption on how data behaves while this is often not the case in reality. We are bound to violate some of these assumptions, like stationarity, considering the different behaviour that occurs when we take all different demand subsets into account. Box-cox transformations and differencing provide tools for stabilizing the variance and the mean if required. Additionally, their parameters can be estimated allowing for an automated approach. Important because we concluded in Chapter 2 that the data under consideration is too large for manual tuning. Regarding theory on aggregation we conclude that 2 relevant choices were already made. First, the available data has a daily frequency but we choose to aggregate to a monthly level. This level of aggregation is suitable because it presents demand behaviour on a level relevant for capacity decisions. The potential information lost is not of use in our scope. Our second choice regards the lowest level of aggregation for the demand characteristics. In Section 2.3 we decided that all possible combinations and aggregations should be considered because the information in each level is potentially useful, this is in agreement with theory.

3.4 Forecasting models

Depending on the available data and its characteristics a suitable model for forecasting needs to be applied. This model should preferably be a mathematical representation of all characteristics in the data. However, this is only applicable when there is sufficient data available. There are also situations where no or limited data is available to fit a mathematical model to it, in these situations judgemental forecasts can provide a good alternative. In this section we will look at both judgemental and quantitative methods, their advantages and drawbacks providing an answer to research question 4.

3.4.1 Judgemental forecasting

Judgmental forecasting uses experience from forecasters and other experts to make subjective predictions. This can range from simple opinion to structured methods, while "judgment can lead to significant improvements in forecasting accuracy, it can also be biased and inconsistent" (Fildes & Goodwin, 2007).

3.4.1.1 *Principles of judgemental forecasting*

The limitations of judgemental forecasting stem from non-structured ways of working that introduce additional uncertainty into the forecasts. Hyndman and Athanasopoulos (2018, p. 4.1) identify several limitations:

- Judgmental forecasts are inconsistent due to limitations and inconsistency of the forecaster;
- Judgement can be clouded by personal or political agendas;
- Judgmental forecasts are prone to anchoring, where the forecaster is influenced by previously observed reference points.

They go on to state that "using a systematic and well-structured approach in judgmental forecasting helps to reduce the adverse effects of the limitations of judgmental forecasting" (Hyndman & Athanasopoulos, 2018, p. 4.2). Fildes and Goodwin (2007) constructed forecasting principles that help guide the application of judgment in forecasting, see Appendix D. Hyndman and Athanasopoulos (2018, p. 4.2) also provide a set of key principles that are comparable with slight differences. Silver, Pyke, and Thomas (2017, p. 119) agree wholeheartedly with Fildes and Goodwin's (2007) principles. The principles are summarized by us to the following four points.

- Where possible start with a quantitative method.
- Set clear rules on when and how to apply judgemental forecasting.
- Document and justify all decisions in the forecasts.
- Thoroughly evaluate the judgement against other methods and with different measures.

Following the principles will allow for more accurate and judgemental forecasting. This is also true if there are factors that are hard to predict by quantitative models and where judgement could fare better. Silver, Pyke, and Thomas (2017, pp. 118-119) identify factors that are not normally or easily included in statistical models and human judgment can be required to identify them.

3.4.1.2 *Judgemental models*

Adhering to the principles of forecasting is a good way to ensure quality of judgment yet they are independent of judgmental models. Different judgemental models are available, of which a few relevant alternatives are worked out.

Expert opinion

Forecasting by expert opinion is the most straightforward judgemental forecasting. Involve an expert on the subject and ask them to evaluate the situation and what their expectation for the desired forecast horizon is. The biggest advantage of this method is the ease of use and leveraging knowledge from the organization directly. However, an individual is always prone to subjectivity and bias, this makes relying solely on one experts opinion risky.

Panel of experts

Applies the principal of expert opinions to a group of them. Each individual member of the panel will provide a forecast and the results of the entire panel can be combined/averaged to produce a less biased forecast. It improves upon the singular expert opinion by mitigating subjective bias. Yet, it does not follow clear rules on how to forecast, each panel member is free to use his own rules.

Delphi method

The Delphi method takes forecasts from a panel and combines them to improve over individual accuracy. Additionally, it defines ground rules for the process and introduces a feedback stage

Hyndman and Athanasopoulos (2018, p. 4.3) summarize the method into the following stages:

1. A panel of experts is assembled;
2. Forecasting tasks/challenges are set and distributed to the experts;
3. Experts return initial forecasts and justifications. These are compiled and summarised in order to provide feedback;
4. Feedback is provided to the experts, who now review their forecasts in light of the feedback. This step may be iterated until a satisfactory level of consensus is reached;
5. Final forecasts are constructed by aggregating the experts' forecasts.

This way of forecasting is already much more in line with the principles set out in 3.4.1.1. There is more structure, justifications for the forecast are required and the results are evaluated and changed based on feedback.

Scenario forecasting

Scenario forecasting is to not only consider the most likely outcome but specifically what could happen under differing circumstances. By forecasting based on different criteria and possibilities of the future different outcomes and risks can be identified using "what if…?" type questions. Examples of scenarios to consider could be: What would happen when the biggest competitor goes bankrupt, or the effect of a financial crisis. Even small scenarios can be useful, what if we break or beat a deadline by an x amount of time? Quantitative methods are not suited for these questions as they make predictions based on past occurrences. Human minds are more flexible and can consider a wide range of possibilities and their effects which could provide new insights if properly used.

Judgmental adjustments/interventions

One of the most effective applications of judgement in forecasting comes from judgemental adjustments. A statistical forecast provides the initial forecast which can be adjusted if necessary through judgement. This is especially useful in situations where there are in- or external factors that will influence the forecast in a way that the statistical model cannot foresee, e.g. a sales promotion or a general decline in economic prosperity. Silver, Pyke, and Thomas state that "small adjustments to forecasts tend to hurt forecasting performance while large adjustments improve the performance. This is primarily due to the fact that large adjustments are made only when a manager has significant and relevant information" (2017, p. 119). So it seems wise to consult experts on a forecast produced by a statistical model yet limit the number and size of changes they can make.

3.4.1.3 *Conclusion*

In this section we have learned about several different techniques to apply judgemental forecasting. The most important lesson is that when used judgemental forecasts should apply the key principles of forecasting from Section 3.4.1.1 to reduce bias and improve accuracy. The more complex methods presented all have measures in place to validate and evaluate the forecasts, the Delphi method can be seen as the most thorough. Judgmental adjustments are most in line with the forecasting process as described in Section 3.2, first producing a statistical forecast and only adjusting it when necessary.

3.4.2 Quantitative models

When there is data available that is of sufficient quality and quantity a mathematical model can be used. The goal of such a model is that to capture the information in the data and then extrapolate to the future. A simple example is the naive method which assumes that the last observed value will persist to the next observation. While intuitive it lacks the depth to capture more complex patterns in the data. On the other hand simple methods have often been observed to outperform complex models. Additionally, they can serve as a benchmark for more complex methods. A set of established forecasting techniques will be discussed and compared on their capabilities.

3.4.2.1 *Simple methods*

Simple methods often perform well in forecasting, often providing more accurate results than complex ones. Green and Armstrong (2015) state that no matter what method is used complexity rarely improves accuracy. They consider a model complex if it is not simple enough for the forecaster users to understand, advocating 'sophisticatedly simple' procedures. "Complexity increases errors for forecasts from judgmental, extrapolative, and causal methods by an average of more than 25 percent" (Green & Armstrong, 2015). While not an exact definition of what a simple model is it is a convincing arguments to always take simpler models into account. Even when a simple model does not end up performing most accurately it can act as a benchmark and provide understanding for forecast users. The following methods are adopted from (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018) and (Duffuaa & Raouf, 2015).

(Weighted moving) average method

The average method assumes all future values to be equal to the historical average of the data. By looking at a subset of more recent data it can be modified to a moving average. This avoids taking aged observations into account when the underlying process might have been different. The regular average assigns the same weight to each observation. If a different distribution of weights is chosen a weighted average can provide a means to assign different importance to different observations. The weighted moving average combines weights and a subset of the most recent observations. Both tactics lead to a forecast that is more affected by recent observations which can be seen as more important.

The following formulas denote the methods with y_t as the observation at time t, h the forecast horizon, w_j the weights and m the number of periods to average over.

Average: $\hat{y}_{t+h} = \frac{\sum_{j=1}^{t} y_j}{t},\ for\ any\ h$

Moving average: $\hat{y}_{t+1} = \frac{1}{m}(y_t + y_{t-1} + \cdots + y_{t-m+1}), m = \#\ of\ periods$

Weighted average: $\hat{y}_{t+1} = \sum_{j=1}^{t} w_j y_j\ ,\ \sum_{j=1}^{t} w_j = 1$

Weighted moving average: $\hat{y}_{t+1} = \sum_{j=1}^{m} w_j y_{t-j+1}$

From these methods we can learn that the current forecasting practice described in Section 2.4.1 is essentially the average method with limited observations of y_t. The merits of this forecast is that it places the most importance on the most current observations. However, it has no ability to extract any information about seasonality or changes overtime.

Naive method

The naive method assumes that the last observation will persist into the future. Such a forecast will assume that the next observation will be equal to the last. Drift can be applied to the naive method to account for long term changes over time. Including drift "is equivalent to drawing a line between the first and last observations, and extrapolating it into the future" (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018).

Naive: $\hat{y}_{t+h} = y_t, \ for\ any\ h$

Naive + drift: $\hat{y}_{t+h} = y_t + h\left(\frac{y_t - y_1}{t-1}\right), \ for\ any\ h$

From the simple methods we can learn that the current forecasting practice described in Section 2.4.1 is essentially the average method with limited observations of y_t. The merits of these simple techniques are their intuitiveness. However, they lack the ability to extract any information about seasonality or changes overtime. We see use for them in our model in in two ways. First, they should be included as benchmark methods to see whether more complicated methods can outperform them and are thus worth the effort. Second, their simplistic nature requires little fitting to data or assumptions. As a result the forecasts are unbiased and might perform well on erratic demand where assumptions for more complex models may fail to hold.

3.4.2.2 *Exponential smoothing (ETS)*

Simple exponential smoothing (SES) is reminiscent of the weighted moving average with weights that decay as observations get older. Instead of assigned weights, a smoothing parameter α is chosen with weights $w_j = \alpha(1-\alpha)^{j-1}$ leading to:

SES: $\hat{y}_{t+1} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1} y_{t-j+1}$

Various extensions to the method are possible making it capable of handling trend and seasonality in the data. Hyndman, Koehler, Snyder, & Grose (2002) defined an extended range of exponential smoothing methods. By doing so they provided a complete collection of models capable of handling **e**rror, **t**rend and **s**easonality (ETS). Each model can be labelled ETS (-,-,-) with different states for each term.

Error: Additive (A) or Multiplicative (M)

Trend: None (N), Additive (A) or Additive damped (A_d)

Seasonal: None (N), Additive (A) or Multiplicative (M)

Appendix E has the applicable formulas for each combination of states. One of the challenges in fitting an appropriate ETS model is estimating the necessary parameters. Several options exist for doing so but one of the most versatile is Akaike's information criterion (AIC) which measures an in sample fit of a model to the data, see Section 3.5.2. Fitting different parameters combinations a 'best' one is chosen based on the lowest AIC score allowing an automated approach.

ETS models are widely used and applied in practical forecast situations. Their base premise is still intuitive and their extensions to handle seasonality and trends make them useful for more complicated data series. As a versatile and widely used method we find that it should be considered for the model.

3.4.2.3 *Theta*

In the M3-forecasting competition (Makridakis & Hibon, 2000) a method that performed really well was the Theta method by Assimakopoulos & Nikolopoulos (2000). In their paper the approach is complicated and technical but Hyndman and Billah (2003) demonstrated that it was a special case of simple exponential smoothing (SES) with drift, which parameter equals half the slope of a linear trend fit to the data. While less flexible than the ETS approach described previously, its good performance and simplicity advocate for including this approach.

3.4.2.4 *AR(I)MA*

Apart from ETS, ARIMA models are among the most used models for time series forecasting. It takes an autoregressive (AR) model, differences data that was integrated (I) when passed to it and adds a moving average model (MA).

An AR model takes the values of past observations and linearly combines, or regresses, them to forecast the variable. "The term *auto* regression indicates that it is a regression of the variable against itself." (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018). It has an error term ε_t and an order p that defines the number of parameters ($\emptyset$), or the number of previous observations, to include and regress on.

AR(p): $$y_t = c + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + \varepsilon_t$$

The MA model does something very comparable but instead of past observations it uses past errors to perform the linear regression. It has an order q defining the number of parameters (θ).

MA(q): $$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

Together this makes an ARMA model, both AR and MA models require stationary data and the combination does as well. In practicality this is often not the case but can usually be achieved by differencing the data as discussed in Section 3.3.4. Integrated (I) data is un-differenced and ARIMA models take a degree of differencing *d* to account for this. Resulting in an ARIMA (*p, d, q*) model, with y_t differenced *d* times.

ARIMA: $$y_t = c + \emptyset_1 y_{t-1} + \cdots + \emptyset_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Seasonality can be handled by ARIMA though adding a seasonal component with terms (P, D, Q) which are similar to (*p, d, q*) but involve observations from relevant season instead of direct predecessors. Leading to an ARIMA (*p, d, q*) (P, D, Q) model. As with ETS one of the difficulties is estimating the parameters. The number of differences *d* can be determined as in Section 3.3.4 with a unit root test. The best order for *p* and *q* of an ARIMA model can be selected using information criteria like the AIC, Section 3.5.2, that measure the fit of a model to the in sample data. The possibility to estimate the parameters automatically again allows for a degree of automation.

Because ARIMA uses different components to model the information in the data it can be very flexible and produce good forecasts on different types of data. The inherent ability to handle non-stationary data by differencing the data makes it a versatile method we believe should be included in the model.

3.4.2.5 *TBATS (**T**rigonometric, **B**ox-cox transform, **A**RMA errors, **T**rend and **S**easonal components)*

One of the flaws of Both ETS and ARIMA is their rigid handling of seasonality. They assume strict seasons and not more than one can usually be taken into account. This means that series with more than one season (e.g. daily and monthly) cannot easily be modelled. TBATS was designed to overcome this in a by Livera, Hyndman, & Snyder (2011). It works by building of exponential smoothing state space models to allow for more (complex) seasonal patterns. The acronym TBATS contains all its parts, Trigonometric, Box-cox transform, ARMA errors, Trend and Seasonal components. The exponential smoothing method is extended to include these parts and extends those already present, we refer to the original paper for the formulas.

TBATS handles complex seasonality and can detect these as well. Additionally, by adding in a Box-Cox term stationarity of the data is simple to achieve. The downside of the method is the large number of parameters to estimate. In addition to the smoothing parameters regular ETS requires TBATS also has parameters for the box-cox transformation, the ARMA errors and the seasonal (trigonometric) terms. This leads to introducing more uncertainty as parameter estimation is never 100% correct as well as longer computation time. Nevertheless, we see this as a logical extension of ETS and ARIMA and it should be able to extract more information from certain data series.

3.4.2.6 *Intermittent data models*

When data is generated by a sporadic process there is no continuous stream and multiple observations of zeros occur. Most time series models have difficulty in accurately modelling these series as they usually assume some form of correlation between observations. We might expect to see intermittence at more specific demand subsets. There are methods specifically made to handle their intermittent nature, two are discussed:

Croston's method and SBA

Instead of trying to model the complete behaviour of an intermittent demand, Croston (1972) proposed to model the size of the demand and the intervals between demands separately. The two parts are both forecast with simple exponential smoothing (SES) and the ratio of the 2 is the demand to expect at time t. The forecasts are updated each time new demand occurs. Syntetos and Boylan (2005) showed that Croston's approach contains bias and they made an adaptation known as the Syntetos-Boylan approximation (SBA). Both methods were shown to outperform regular exponential smoothing, moving average or a zero forecasting on intermittent demand by Teunter & Duncan (2009). Croston's method is intuitive and simple while it provides potentially better forecasts for intermittent demand.

Imapa

Building on Croston's method and SBA, ways to automatically estimate their parameters was desired. Kourentzes (2014) developed a method to automatically classify intermittent demand according to a cost function. Petropoulos and Kourentzes went on to define a classification scheme, PK classification, which indicates whether a demand series should be forecasted with Croston's method, SBA or SES (2015). Additionally it was found that using different temporal aggregations (e.g. days → months → years) forecasting the different levels separately provided better forecasts (Petropoulos & Kourentzes, 2015). Kourentzes then combined PK classification and temporal aggregation into a forecasting framework for time series which he dubbed iMAPA (intermittent Multi aggregation prediction algorithm).

3.4.2.7 *Explanatory variables and Regression*

Explanatory variables

Explanatory variables are external information points that are correlated to a certain variable and can thus be used to describe part of its behaviour. For instance, temperature is a good predictor for gas usage, because cold temperatures lead to more people turning on heating systems. Temperature can therefore serve as an explanatory variable to gas usage, it explains part of its variation. We can forecast based on this principle by using regression models, but in order to do so we need relevant explanatory variables. What variables are relevant depend on what is being forecast. From there we can consider relevant variables from internal (organizational) variables and external variables, expert knowledge is necessary to select them. For instance, the number or age of a certain aircraft type might influence demand and serve as explanatory variables. External factors and their relevance can be more difficult to assess but still useful. Carson, Cenesizoglu, and Parker (2011) identified several variables impacting aviation, for instance GDP and oil prices.

Forecasting with explanatory variables requires those variables to be forecast as well. As a result forecast based on forecasts are made which can contain a lot of uncertainty. The sensitivity of this uncertainty can be tested by scenario forecasting where (slightly) different values of the explanatory variables are evaluated against the resulting forecasts (e.g. different # of aircraft introduced or at different times)

Explanatory variables can also be used to discover the main driving forces of a variable y. By evaluating different combinations of the explanatory variables the one with the most predictive power can be selected. The correlation of the explanatory variables with y could then be an indication of how much influence it has, but correlation is not causation. To find the true model of y it is important to evaluate the explanatory variables and their characteristics. On the other hand, predictive power of a model does not care whether it is the true model. If the goal is an accurate forecast the true correlation/causation relationship is not necessarily important.

Linear regression

In linear regression the assumption is that variable y is a linear combination of explanatory variables $x_1, \dots, x_n$. This combination provides a fitted line that represents the data. The least squares estimation method minimizes the total deviation from the fit and what remains are unexplained errors. These errors capture anything that may affect the variable other than the provided explanatory variables.

A single linear regression produces a simple straight line based on the relationship between the variable and one explanatory variable. Multiple linear regression produces a fit based on a multitude of explanatory variable where the coefficients $[\beta]$ in the regression measure their marginal effects (Hyndman & Athanasopoulos, 2018). The ability to add all different kinds of explanatory variables (x) make it a versatile method to model data.

Linear regression: $$y_t = \beta_0 + \beta_1 x_{1,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

Dynamic regression models

Linear regression models are less suited to "the subtle time series dynamics that can be handled with ARIMA models" (Hyndman & Athanasopoulos, 2018). However the inclusion of external variables can be very useful in explaining variation. Dynamic regression models bridge this gap by allowing the error terms of the regression to contain autocorrelation. Usually errors are assumed to be a white noise process, in this definition the error term is an ARIMA process with its own error term. The error

term ε_t in linear regression is now an ARIMA process able to interpret the time series characteristics like trend and seasonality which regular regression cannot.

The big advantage to this approach is that more information from the data itself can now be used while also allowing for external factors to explain variation. There are also downsides: First, to produce more accurate forecasts suitable external variables need to be identified and tested on their relevance and predictive power. Then, in order to forecast, the external variables need to be forecast as well (e.g. estimate of future GDP or temperatures), this introduces additional uncertainty as forecasts are made on information from other forecasts. Furthermore, parameter estimation becomes more difficult and the variable y and the explanatory variables need to be stationary for the ARIMA errors to work. Stationarity can be coerced through methods like Box-Cox and differencing but it introduces extra complexity.

3.4.2.8 *Quantitative model Comparison*

With several models reviewed we can get a sense of their general characteristics. Table 3.1 shows a summary of their capabilities.

Table 3.1 Comparison on general model characteristics

	Seasonality	Stationarity required	Explanatory variables
Simple methods	No, only through de- and re-seasonalizing before and after application	No, disregards most characteristics of the data anyway	No
ETS	Yes	No	No
TBATS	Yes, multiple seasons	No	No
ARIMA	Yes, by incorporating seasonal terms	Yes, but can enforce it through differencing	Yes, with extension to Arimax or dynamic regression
Regression	Yes, through dummy variables	No	Yes
Dynamic regression	Yes, through the ARIMA errors that can fit seasonality	Yes, ARIMA errors require stationarity	Yes
Intermittent methods	No	No	No

There are some clear differences between the methods, a rise in complexity seems to correlate with more capabilities. Yet even simple methods appear to be practical and as Section 3.4.2.1 concluded where possible one should favour simple methods over complex ones. Additionally, the differences in methods and their parameter estimation make that they extract different information from data.

In Section 2.3 we concluded that all possible subsets of the data should be considered as each could contains potentially relevant information. Here we observe that different models extract different information from data. Continuing our line of thought would imply we need some way of selecting the most suitable model for each subset as each method might fit a certain subset better than another. Yet there is no clear way for us to decide which method(s) would be most suited without requiring more intensive knowledge of each data series. For now we can conclude there is no way to define a singular (or set of) model(s) as most suitable.

3.4.3 Conclusion

In this section we investigated different forecasting models both, judgemental and quantitative. On judgemental models we can conclude that the most important steps to consider is adhering to the key principles of forecasting. The principles set out guidelines that ask the forecaster to start with and trust a statistical model, set clear rules on when to use judgment, document and justify their decision and evaluate the forecast and its process. These guidelines reduce the limitations of judgmental forecasting mostly introduced by, unconscious, bias. Several models exist to further guide the application of judgment where the consensus points toward consulting multiple experts and getting their forecasts in agreement. Judgmental adjustments seem most applicable, given that a qualitative forecasts produces accurate and trustworthy results.

A selection of quantitative forecasting models was reviewed. Simple to more complex models, each has different assumptions and parameters to estimate. This led us to conclude that different models can extract different information from data. As we do not know every detail of our dataset we find that no clear selection can be made based on theoretical properties before we can measure their performance on the data. In the next sections we will describe measures that we can use to compare the performance of different models.

3.5 Forecasting performance and accuracy

To evaluate the performance of a forecast a frame of reference and tools to compare the forecasts are necessary. When can a forecast be called accurate and how can they be compared. Forecasting measures typically look at either the residuals left after a model fit on the entire dataset or look at the error between the forecast and reality. Additionally, we would prefer measures that can easily be compared to others, a single measure is meaningless without a frame of reference.

3.5.1 When is a forecast accurate?

A forecast can only be deemed accurate to a certain degree, all models are a simplification of the 'true' data generation process and will never fully capture all information. George Box (1987) once said "All models are wrong but some are useful", implying that while not perfectly correct they can be practical in their results.

We can look at the quality of a model and its forecast accuracy in several ways. One type of measure looks at the amount of information captured from the source data by fitting the model, the other looks at accuracy by evaluating the forecast errors. The first tells something about the fit on the original data, which can help in choosing a model from a set of alternatives. However, a good model fit does not necessarily lead to accurate forecasts. Accuracy measures can help provide an objective way to compare forecasts to each other. No objective level of accuracy can be set as a goal because it is always case dependent. "One cannot simply take industry-specific forecasting errors as benchmarks and targets" (Kolassa, 2008) instead internal benchmarks should guide practice and the definition of accurate forecasts. Two intuitive comparisons/benchmarks are easily available. First, against simple models as discussed in 3.4.2.1, their performance provides a good indication whether a more complex model actually provides higher accuracy. Second, the comparison against the current practice and results of forecasting provides insight in the quality and performance of the model.

3.5.2 Measures of accuracy

There are a lot of different kind of accuracy measures available to gauge the performance of a model and forecast. At the core of most measures is a comparison between the forecast and the actual values. The difference between the two is the forecast error (ϵ_{t+h}), "not a mistake [but] the unpredictable part of an observation" (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018). Actual forecasts cannot be evaluated until time passes and actual values become available, so to test accuracy on data it has to be split into a training and a test set. Usually the majority of data is used to construct the training set and fit the model, the rest of the data is used to test the resulting forecasts and get the relevant forecast errors. In this way predicted values can be compared to actual values without waiting for the future. We evaluate different measures that take the forecast errors and produce a value that compares accuracy. Another method evaluates accuracy with such a measure on a rolling training and test set. Finally, an information criterion will be addressed that says something about the relative quality of the model fit to the data.

Measures in general

Hyndman and Koehler (2006) present a collection of commonly used accuracy measures. From 6 presented categories 3 possess desired statistical properties (Franses, 2016), i.e. the predictive accuracy approaches asymptotic normality with the Diebold and Mariano (1995) methodology. As a result criteria based on squared, absolute and absolute scaled errors are deemed suitable. Other measures based on relative absolute, absolute percentage and symmetric absolute percentage were found not to have these properties. With y_t the observation at time t and h the forecast horizon.

Forecast error: $\epsilon_{t+h} = y_{t+h} - \hat{y}_{t+\mathrm{h}}$

Squared

Two relevant measures use a squared error to assess accuracy, one takes the mean while another takes the root of that mean. Using the square error makes these measures vulnerable to outliers as these have disproportionate effect. As a result models with bad fits on exceptional events will score worse. Due to being based directly on the data the measures are scale dependent, "useful when comparing different methods applied to the same set of data [but not] when comparing across data sets that have different scales" (Hyndman & Koehler, 2006).

*Mean square error (MS*E): $mean(\epsilon_t^2)$
Root MSE (RMSE): $\sqrt{mean(\epsilon_t^2)}$

Added benefit of the RMSE is that it returns the measure to the original scale of the data.

Absolute

Absolute error measures are easy to interpret because they are on the same scale of the data and only ensure positivity without other transformations. Like the squared measures they are scale dependent and may only be compared with same scale data.

Mean absolute error: $mean(|\epsilon_t|)$
Median absolute error: $median(|\epsilon_t|)$

Absolute scaled

To overcome the dependency on the scale of original data Hyndman and Koehler (2006) proposed a scaled error. It compares the error produced by the current forecast against the in sample one step ahead error produced by a naive forecast. It does so by producing a ratio between the current forecast error and a scaling factor. Resulting in a value < 1 implying better and > 1 indicating worse performance than the one step ahead in-sample naive. Because the measure is based on a ratio from the same scale it becomes scale free and comparison with other methods on different data become feasible.

Absolute scaled error (ASE): $q_j = \frac{\epsilon_t}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|}$

Mean ASE (MASE): $Mean(|q_j|)$

The MASE is recommended to be used when "comparing forecast accuracy on several series with different scales when the Mean Average Percentage Error (MAPE) is inappropriate" (Hyndman R. J., 2014), which is the case when (near) zero values are to be considered.

Cross validation

Cross validation (CV) is the process of repeatedly fitting a model and testing it on a test set. The training set is all observations before the first test set observation. Then after each test the first test observation is incremented by one, as a result the training set grows by one as well. Then the model is re-estimated and retested. Figure 3.8 by Hyndman and Athanasopoulos (2018) shows the progression of training and test series in CV. The model is fit on the blue nodes and a one step ahead forecast is produced and compared to the red node. It uses all information available in the observed data and averages the forecast error over all the test and training sets, thus providing the average one step ahead performance of a model. "This implies that CV can be considered to be an estimator of predictive mean square error" (Konishi & Kitagawa, 2008, p. 242). The result is a fair and unbiased measure of a models forecasting accuracy and allows for easy comparison. The biggest downside of cross-validation is the computation it incurs. In the Figure 3.8 20 repetitions are used, meaning that the model is fit to the data 20 times, produces a one step ahead forecast 20 times and determines average performance. This process is then repeated for all combinations of models and time series of interest. Fortunately, Akaike's information criterion is an approximation and easier to calculate.

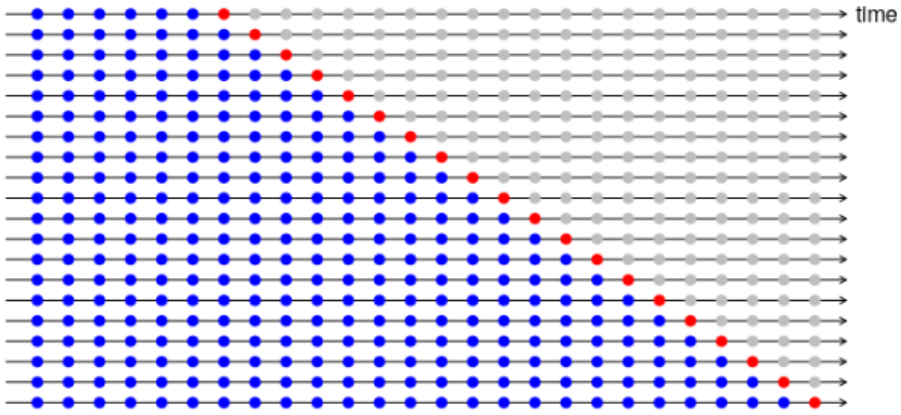


Figure 3.8 CV for one step ahead forecasting. Blue: training set. Red: test set (Hyndman & Athanasopoulos, 2018, p. 3.4)

Akaike's information criterion

Akaike's information criterion (AIC) measures the relative quality of a model by estimating the information lost after fitting to the data. As defined by Akaike (1974) the criterion measures the goodness of fit of a model and penalizes it for the number of estimated parameters, balancing in-sample fit with overfitting. Originally defined by Akaike (1974) with k as the number of estimated parameters and L as the maximum likelihood:

$$AIC = 2k - 2 * \ln(L)$$

Given that the underlying data is the same a smaller value AIC implies that more information has been captured in the model and is thus the best choice. Konishi & Kitagawa (2008, p. 245) show that the AIC is asymptotically equivalent to cross validation, making it a suitable estimator for selecting models predictive power. Additionally, the required computaion time is far lower making AIC is a useful tool for model selection and parameter estimation. For instance, fitting an Arima model requires several parameter estimations and it is difficult to determine the best values. Calculating the AIC for each different parameter combination allows an easy choice for that with the lowest AIC.

3.5.3 Uncertainty in forecasts

Uncertainty is inherent to any model fit to data as estimating such a model has implicit and explicit assumptions. For instance, the assumption of normally distributed forecast errors is common. To produce confidence intervals, to indicate uncertainty, the fit model being true must often be assumed. This regularly results in too narrow intervals as it does not take all uncertainty into account and the model can never be exactly true to reality. Prediction intervals try to take the uncertainty of model estimation into account and provide wider, but more accurate, intervals. Giving insight in the kind of values that can be expected. Some models have pre-defined prediction intervals, often assuming the residuals to be uncorrelated. A way to provide prediction intervals is without them being defined is running a bootstrap simulation with the model.

Bootstrap intervals

Bootstrapping refers to using the available data to produce a more accurate results. By decomposing a data series, as in Section 3.3, randomly sampling from the residuals and recombining with the other data components (trend, season), slightly different data series can be generated. Bergmeir et al. (2016) Developed and applied this technique on ETS models but the practice is applicable with others as well. Using these slightly different series to forecast allows prediction intervals to be made. Simulating the approach with a lot of different combinations returns slightly different forecasts each time. The sensitivity of the model is thus tested and the combination of all these results then leads to a prediction interval for the forecast. This models the uncertainty in fitting the model and will therefore result in wider, but more accurate, intervals.

Bootstrap aggregating

Bootstrapping the original series gives a better measure of forecasting uncertainty but it can also improve point forecasts (Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018). The variations in the bootstrapped time series compared to the original show how uncertain the model estimation is and what small changes can produce. It improves point forecasts by aggregating the simulated forecasts from all these different time series which will on average produce more accurate results. Section 3.6 will elaborate on why combining forecasts produces better results.

3.5.4 Conclusion

In Section 3.4 we concluded that selecting appropriate models on theoretical properties was problematic as each could be better suited to a different subset of demand. We needed a measure to test their relative performance, which we found in the mean absolute scaled error (MASE). The MASE is scale free, can handle zero values and provides a natural benchmark against the in-sample one step ahead naive forecast. These properties allows us to use it as a comparison between different models on the different demand subsets. Parameter estimation and selection of models is enabled by the AIC which measures the quality of a fit on test data. Finally we addressed the uncertainty inherent to forecasting and how that affects the prediction intervals. Some models have defined prediction intervals but others do not or produce them too narrowly based on regular confidence intervals. Bootstrapping is a way to reduce uncertainty but incurs high computational costs.

3.6 Forecast combination

Different methods are suited to different characteristics in data. Model selection is usually necessary to make sure that the capabilities of the model fit with the data characteristics. A common approach is to compare different models based on measures like described in Section 3.5. Even then deciding on a singular model introduces uncertainty, it has assumptions and estimated parameters that are not known to be correct. Section 3.5.3 shortly described how this leads to wide prediction intervals, necessary to capture a realistic spread of the forecast values. A probable approach is fitting and combining results from different models to improve their accuracy and reduce faults incurred by relying on one method. This section looks at using different models on different levels of aggregation and combining the forecasts in order to obtain more accurate forecasts.

3.6.1 Model combination

It has been recognized since Bates and Granger (1969) that combining forecasts from different models results in better accuracy. Throughout the years and in multitude of research this has been confirmed and analysed. The performance increase of forecast combination is attributed to the fact that different models are able to extract different information from the data. Through combining, this information is put together and uncertainties coming from whether the right model was defined are mitigated. Petropoulos, Hyndman, and Bergmeir (2018) concluded that small changes in data can result in different models selected, further acknowledging that model uncertainty is the issue.

It could be expected that weighting the combination of forecasts would create the best results however this often turns out not to be the case. Claeskens, Magnus, Vasnev, and Wange (2016) conclude that estimated optimal combinations are generally outperformed by a simple averaging of the forecasts. They go on to state that this 'forecast combination puzzle' is caused by the fact that the weights themselves also require estimation instead of being fixed (Claeskens, Magnus, Vasnev, & Wange, 2016). As a result the unbiased equal weights combination can often outperform it.

Just combining forecasts looks past the fact whether a certain model was actually suitable at all. Kourentzes, Barrow, and Petropoulos (2018) present a pooling method in which a selection of the available methods is made. They present proof that selecting a suitable pool of methods for combination performs at least as well or better than selecting a single method or combining all of them (Kourentzes, Barrow, & Petropoulos, 2018). Indicating that a reasonable selection will improve accuracy by removing methods that do not contain sufficient predictive power.

Combining forecasts works regardless of whether an approach is qualitative or quantitative given that the approach itself produces reasonable forecasts. Expert opinion for instance can be very helpful in indicating the coming of unforeseen demand or to estimate the duration of a sales spike after a promotion. Lawrence, Edmundson and O'Connor (1986) found that combining and averaging judgemental forecasts provided better accuracy than either one. This improvement was especially apparent in short term easy to forecast series. Apparently, because judgemental forecasts can contain information not captured by statistical models from past data and combining therefore increases accuracy.

Armstrong (Principles of Forecasting., 2001) stated that "Combining forecasts is especially useful when you are uncertain about the situation, uncertain about which method is most accurate, and when you want to avoid large errors, " Additionally, he found that forecasting errors could be reduced with an average of 12.5% by combining forecasts.

3.6.2 Hierarchical and grouped forecasting

Depending on its structure data can have *hierarchies* to aggregate over, a unique order in which the data can be summed. Figure 3.9 shows a two level hierarchy where we can regard the different branches as a unique division, applying a population example; Total level represents a country population, the A and B level the province, and the bottom level per city. Each lower level is distinctly part of its parent node. Such a structure in data can be used to forecast for all the nodes. Top-down and bottom-up forecasting are two common approaches. Top-down assumes the total forecast disaggregates to lower levels with weights. Bottom-up assumes that forecasting the lowest level allows summation to accurately forecast higher levels. Both disregard information from other different levels, these represent aggregations and Section 3.3.5 showed how that presents different characteristics and information. Another hierarchical structure to consider is temporal where the levels of the hierarchy are equal, higher or lower steps in time (Athanasopoulos, Hyndman, Kourentzes, & Petropoulos, 2017).

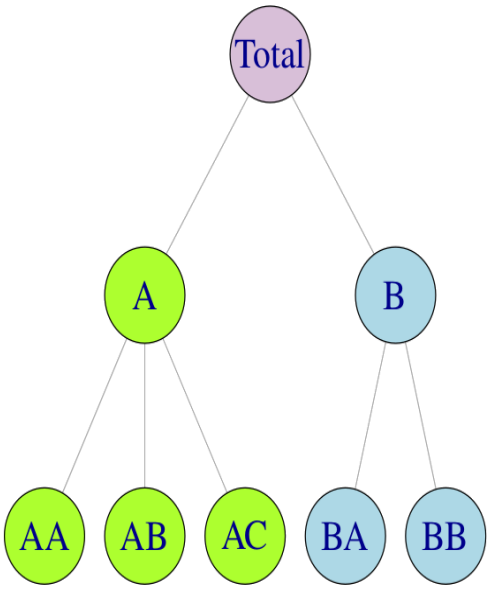


Figure 3.9 A two level hierarchy (Hyndman & Athanasopoulos, 2018, p. 10.1)

Some data is subject to more general structures than a strict hierarchy. In these instances there is no unique order in which aggregation should take place. Hyndman and Athanasopoulos (2018, p. 10.2) refer to these as grouped time series, Figure 3.10 shows an example. The nodes on the bottom level are identical in each structure but the middle level has different nodes, there are two ways to aggregate over these characteristics. To keep with the population example, A and B can be regarded as population per city and X and Y can be regarded as Male and Female population. These characteristics imply no inherent order and therefore it does not matter to which we aggregate first, however, the order does produce different results. Both the hierarchical and grouped structure allows leveraging information from the different levels while decreasing model uncertainty through forecast combination.

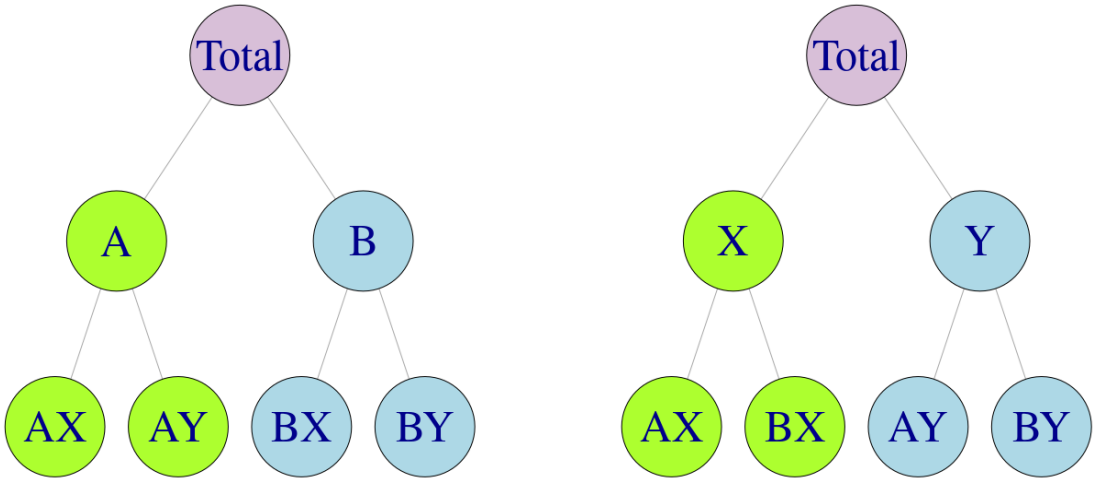


Figure 3.10 Two level grouped hierarchy showing alternative representations (Hyndman & Athanasopoulos, 2018, p. 10.1)

Incoherence

Aggregating data to each level creates a collection of individual forecastable nodes, 8 in Figure 3.9 and 9 in Figure 3.10. Due to aggregation each node can be assumed to present different information than nodes at higher or lower levels resulting in different forecasts. Meaning that more information from the data is used if all the different nodes are forecast in contrast to regarding a single level like the bottom-up or top-down approach do. The downside of this approach is that the forecasts are not *coherent* over the different levels, i.e. the sum of forecasts per level are not equal to each other, which they are in a top-down or bottom-up approach. To illustrate on the structure of Figure 3.9 and with $F(x) = forecast\ of\ x$:

- In a bottom up approach only the bottom level is forecast summing to higher level forecasts:
  $F(AA) + F(AB) + F(AC) = F(A),\ F(BA) + F(BB) = F(B), F(A) + F(B) = F(Total)$
  Resulting in:
  $\sum(F(AA), F(AB), F(AC), F(BA), F(BB)) = \sum(F(A), F(B)) = F(Total)$
- When producing separate forecasts each node has its own forecast:
  $F(A) \neq F(AA) + F(AB) + F(AC),\ \ F(B) \neq F(BA) + F(BB),\ \ F(Total) \neq F(A) + F(B)$

This can cause issues in decisions over tactical and operational levels, e.g. the forecast for the total level is used to make a budget but the lower level forecasts are more practical for operational decisions and planning. By providing individual level forecasts these different levels do not align. This issue is not present in the traditional approaches as these are naturally coherent, but do not use all of the available information. *Reconciliation* offers a method for combining the information of all the individual forecasts, leveraging the benefits mentioned in Section 3.6.1. An optimal combination exists for combining these forecasts that will always be better than traditional bottom-up or top-down forecasts (Hyndman, Ahmed, Athanasopoulos, & Shang, 2011). The main advantage of this approach is that forecasts from all levels of interest can be combined extracting more information from the data while providing a coherent forecast. Theoretically producing coherent more accurate forecasts.

Reconciliation

Reconciling exploits the fact that the hierarchical structure can be stated in matrix form. The lowest level series are aggregated through a summing matrix, as defined by Hyndman & Athanasopoulos (2018) as: $y_t = S * b_t$

With b_t as an m-dimensional vector representing all the bottom level observations and y_t as an n-dimensional vector off all observations/levels in the hierarchy. The summing matrix S defines how the bottom series aggregate to the different levels of the hierarchy. Figure 3.11 shows y_t, S and b_t for the hierarchy presented in Figure 3.9. We can observe that multiplying the bottom time series with the summing matrix results in the different nodes. While slightly different the mathematical approach does not change for the grouped example in Figure 3.10 and it is still expressed with a summing matrix, which we show in Appendix F.

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

Figure 3.11 Summing formula example (Hyndman & Athanasopoulos, 2018, p. 10.1)

Reconciliation between forecasts is done through a mapping matrix that adjusts the individual forecasts to align with each other. We can use a general notation for the coherent forecast as defined by Hyndman & Athanasopoulos (2018): $\tilde{y}_h = S * G * \hat{y}_h$

With $\hat{y}_h$ a vector of base forecasts for each element of y_t, S the summing matrix, and G the mapping matrix that translates the base forecasts to the reconciled forecast $\tilde{y}_h$. As a result the bottom level time series b_t have reconciled forecasts in $\tilde{y}_h$, then we can construct any forecast desired for an element of y_t by using the corresponding row of S. In the case of the bottom-up approach for the example in Figure 3.9 and Figure 3.11 G is an identity matrix where the first columns are empty, see Figure 3.12.

$$\boldsymbol{G} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 3.12 Bottom-up mapping matrix (Hyndman & Athanasopoulos, 2018, p. 10.6)

We see that the first three columns are empty which makes sense as only the bottom level forecast are used in a bottom-up approach. Then the last 5 columns are an identity matrix, mapping the base forecast to the bottom level series with which they correspond. This approach requires no reconciliation but illustrates how G can be used to map base forecasts for all different levels to the bottom series. An optimal approach for determining G was developed (Wickramasuriya, Athanasopoulos, & Hyndman, 2018) that uses all information from all levels. "This is important, as particular aggregation levels or groupings may reveal features of the data that are of interest to the user and are important to be modelled" (Hyndman & Athanasopoulos, 2018). In order to determine G the covariance matrix for the h step ahead forecast error W_h needs to be estimated. This can be difficult but Wickramasuriya et al. (2018) lists four simplifying assumptions. One of the estimations uses a weighted least squares regression (WLS), with weights based on the grouping structure of the data. Referred to as structural scaling and "is particularly useful in cases where forecast errors are

not available; for example, in cases where the base forecasts are generated by judgemental forecasting" (Wickramasuriya, Athanasopoulos, & Hyndman, 2018). It assigns the number of "forecast errors variances that contribute to that aggregation level" (Wickramasuriya, Athanasopoulos, & Hyndman, 2018) and applies an inversely related weight. For a structure in Figure 3.10 W_h can be imagined as a matrix with a value for each node on the diagonal as follows:

- The bottom level nodes are only affected by its own error variance; weight: 1/1=1
- Middle level nodes are affected by two bottom level error variances; weight: 1/2=0,5
- The total node is affected by the 4 middle level error variances; weight: 1/4=0,25

With W_h estimated G can be determined for which we refer to the article of Wickramasuriya et al. (2018).

3.6.3 Prediction intervals for reconciled forecasts

Section 3.5.3 mentioned the uncertainty in forecasts and how it requires wider confidence intervals called prediction intervals. Forecasts combinations are less affected by model selection caused uncertainty and might therefore also have narrower intervals. However, the combination of different methods invalidates any predefined intervals. Reconciliation experiences the same issue, no prediction interval is easily made due to the different models used. An approximation of prediction intervals was proposed by Armstrong (2001), by taking the extreme values produced for the intervals from different forecasts. The downside of this approach is that when a bad model is included the resulting prediction intervals will be too wide implying a larger than realistic uncertainty. Proper estimation of prediction intervals would require full covariance matrixes between series (Hyndman, Ahmed, Athanasopoulos, & Shang, 2011) of which computation can be difficult.

3.6.4 Conclusion

In this section we looked at ways to reduce the need for model selection. Section 3.5 provided us with performance measures to compare different models potentially pointing to the best fit, but there is always uncertainty in choosing a single model.

Forecast combinations enables the use of different models, exploiting the fact that different models extract different information from data. As a result combinations are able to outperform singular models while decreasing uncertainty involved with selecting a god model. It has been found that a simple average between the forecasts leads to results hard to beat with other weighting schemes.

Hierarchical forecasting leverages the different information extracted by model combination and data aggregation. Each relevant aggregate of demand is forecasted separately with any suitable method (combination), resulting in a structured collection of forecasts. The differences between the sums of the aggregation level forecasts can then be reconciled. Resulting in a single coherent forecast that can be used and summed to any aggregation of interest. This aligns with the research goal from Chapter 1 where we defined that more accurate budgeting and capacity decisions are desired, this corresponds with tactical and operational decisions and requires different levels of information.

Finally, we described an approximation for constructing uncertainty intervals for forecast combinations. Regular confidence intervals are too narrow and wider prediction intervals are not readily available for reconciliation. The approximation proposes to take extreme values of the prediction intervals produced by the methods included in the combination. These intervals will tend to be too wide if an unsuitable model is included in the combination and are thus not completely realistic.

3.7 Chapter conclusion

Chapter 3 focussed on identifying suitable theories from literature. In doing so we answered Research Questions 4 and 5 while providing a theoretical foundation for a proposed forecast model in Chapter 4.

Section 3.1 explained that a good forecast is always relative to the context and that accuracy can vary widely. In Section 2.3 we concluded that all demand subsets should therefore be included. Resulting in different data series of which we do not know the specific factors that contribute to it. So we conclude that our ability to forecast accurately depends on the demand subset under consideration.

Section 3.2 presented a forecasting process where a statistical forecast is preferred, adjusted judgementally when necessary and is evaluated. The current forecasting approach does not adhere to most of the presented steps and thus leads us to conclude that a more elaborate forecasting model is worthwhile and can improve on the current system.

Section 3.3 set out different methods and transformations required to effectively forecast. Additionally, we discussed the effects of aggregation. The transformation methods can be automatically applied by estimating their parameters. This possibility for automation is important in order to use them for the number of data-series under consideration. Furthermore, we concluded that the data aggregation choices in Chapter 2 were correct. We aggregate daily to monthly data while considering all possible subsets of demand. As a result only sub monthly information is lost, which is outside our scope.

Research question 4: "What forecasting methods are suitable according to literature?" was answered in Section 3.4. Both qualitative and quantitative models were discussed, ranging from simple to complex. We learned that each method extracts different information from the data. We concluded, based on our diverse dataset, that we could not pre-emptively select our reject methods. Their suitability depends on the behaviour of data which is variable, thus we need to consider every method that is practical to apply. This partially excludes forecasting with external variables, it requires additional work and further forecasting of the relevant variables.

Research question 5: "How can forecasting performance and validity be measured?" was addressed in Section 3.5 where different measures of accuracy where discussed as well as uncertainty in forecasts. We concluded that the Mean Absolute Scaled Error (MASE) is the most suitable, because it provides a scale free and model independent measure of accuracy. Cross-validation (CV) is a strong tool in determining model performance but the computational cost makes it impractical. The AIC is useful as an approximation and helps to select good models. Finally, we considered bootstrapping in order to reduce uncertainty in the forecasts, however, like CV bootstrapping is infeasible to us due to the computational costs.

Section 3.6 concluded the chapter with methods to both increase accuracy and reduce the need for model selection. We found that combining different models is a good approach to reduce uncertainty and improve accuracy. This is in line with our decision to consider all different subsets of demand and multiple methods to forecast them. Furthermore, leveraging the hierarchical and group structure of the demand allows us to use all information and reconcile the forecasts over all levels considered, i.e. ensure that the sum of demand subset forecasts equals the forecast of its aggregations.

4 Forecasting model

This chapter brings the theory from Chapter 3 and the organizational context together in order to answer research question 6. We briefly present our proposed model in Section 4.1 and the following sections provide details and the reasoning. Section 4.2 elaborates on the available data and the steps taken to ensure a usable dataset. Section 4.3 defines the hierarchical structure in our data while Section 4.4 presents our approach to forecast each node. Section 4.5 proposes a way to judgementally adjust the forecasts. Section 4.6 explains how the forecasts are reconciled to align the all level of the hierarchy. Section 4.7 details our approach to measuring our performance compared to the current method and individual models. Chapter 5 presents the results of this approach.

4.1 The forecasting model

To construct the final model we adapted the general forecasting process set out by Silver, Pike and Thomas (2017) as shown in Section 3.2. Our context was applied to their general process and adapted with additional steps, seen in Figure 4.1.

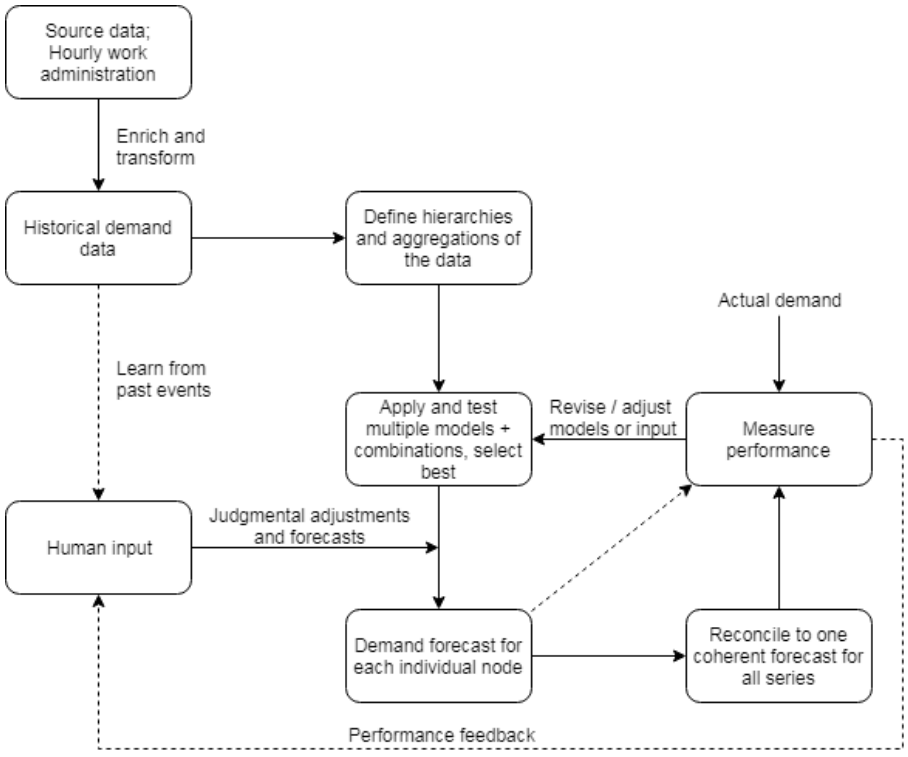


Figure 4.1 The contextual forecasting process (Authors own figure adapted from Silver, Pyke, & Thomas (2017, p. 74))

We concluded in Chapter 2 that all different demand subset should be considered because each presents different, potentially useful, information. Chapter 3 concluded that different models are able to extract different information, additionally it described methods to combine the different resulting forecasts. Both conclusions steer us toward a model where automation is necessary due to the numerous possibilities between the different subsets and models making manual tuning infeasible. Based on the model from Silver et al (2017) but extended to our context, our proposed model intends to produce consistently good forecasts for all the demand subsets with minimal manual intervention.

A summary of the steps described in Figure 4.1 and where to find details:

1. The source data is transformed and enriched to remove errors, include additional details and consolidate the information, providing historical demand data that is able to represent the demand characteristics as defined in Section 2.1. Section 4.2 provides further details.
2. From the data we determine the hierarchy/grouping structure necessary for reconciliation (see Section 3.6.2). Based on the demand characteristics we construct 1716 different demand series (nodes) divided in 16 groups/levels of aggregation, see Section 4.3 for more details.
3. For each of the nodes all suitable models are used to produce forecasts in order to exploit their ability to extract different information. Then all possible combinations between these models are produced by averaging the results, see Section 4.4.3.
4. From these results, we determine the most suitable model (combination(s)) in terms of average accuracy per node over the forecast horizon. Additionally, series with disproportionate performance can be identified leading to outlier detection, see Section 4.4.4.
5. The forecasts can then adjusted based on judgement, feedback from statistical performance and lessons learned from historical data. See Section 4.5 for details on judgemental adjustment.
6. After adjustments each node has a final forecast as good as our proposed approach can provide. However, they are not yet coherent with forecasts from other levels of demand aggregation. We reconcile the differences between aggregation levels in order to obtain a single forecast coherent for all demand, see Section 4.6.
7. Finally, we evaluate the forecasting performance on different groups and nodes and initiate a feedback loop where we can change the approach and adjust inputs, see Section 4.7.

Steps 3, 4 and 6 produce quantitative forecasts providing a statistical baseline without requiring manual input or judgemental adjustments. Based on the forecast errors we can identify problematic subsets that either require judgmental input or might need to be changed or removed in the dataset. Including this knowledge allows us to create forecasts that primarily rely on data but can be adjusted for unforeseen events. This approach answers research question 6: 'How should the forecasting be applied for engineering demand?' Forecasting should be applied with a quantitative framework as a baseline. Those results indicate series on which the model did not perform well, pointing to a need for additional information. Then, judgmental input can be used to improve the model, e.g. change the applied models, and remove/adjust outlier data and apply knowledge of external factors. These steps achieve the goal of a more fact based forecast over the current approach. The rest of the chapter will elaborate on the different steps.

4.2 Source data

Sections 2.2 and 2.3 gave some insight in the available demand data and some of the choices to make the information more accessible. This section will elaborate on how those choices were made and what steps were taken to make the quality of data uniform over the years.

4.2.1 Update to current standards

The data ranges from 2012 the end of 2017 and during that time different rules have been used in the administration of hours leading to discrepancies. Additionally, data was transferred to a new system in 2012. This resulted in a contaminated dataset and multiple errors that need to be minimized to produce accurate forecasts.

The most apparent contamination was detected by evaluating the unique combinations of the rec.order and product line as described in Section 2.2.1. The rec.order is a unique identifier for a job and the product line is a dedicated number for a specific task of that job. As a result, together they should present a unique combination linked with one task type and one customer. In practice, this was often not the case with combinations having more than one occurrence with different factors. There are three main reasons for the discrepancies.

- Human error: The administration system did not have any methods stopping people from entering wrong codes. As a result, mistakes took place and unintended contamination occurred.
- Unassigned non-routine task: When a customer asks for non-standard support this should be done under an appropriate sales order (SO). In some cases the SO is not yet available and other codes are used. Later the sales order are correctly used but the previous errors were not corrected.
- Changes in standard: During the 5 year period several administrative changes occurred in what codes to use for certain customers and tasks. As a result, the data shows a structural change without correcting the old entries.

In order to decontaminate the data, each error could be manually checked and adjusted but this is time consuming. We assume that most recent entries of a unique combination are most likely to be correct based on the observed errors and their causes, random sampling and checks have confirmed this as valid. Thus, older differing entries are overwritten with the information from the latest entries. Table 4.1 contains an example of different entries, in 2012 customer TY was referred to as SPL/TY, in 2013 errors were made in entering the hours and the customer was switched to TZ and task code to XH. In 2017, the last entries, the mistakes were corrected and the customer is labelled as TY and task as KW. Using the last entries as a guide for correcting the Customer and task codes is therefore an acceptable manner to decontaminate the data.

Table 4.1 Contaminated entries

Rec. Order	OpAc	Description	Work ctr	Pers.No.	Number	IVS code	Date	Prd cod	Month
3003341	10	Alle werkzaamheden voor H10	3022	xxxxx	2	SPL/TY	18-1-2012	KW	1-1-2012
3003341	10	werkzaamheden tbv SPL/TZ	3022	xxxxx	2	TZ	16-1-2013	XH	1-1-2013
3003341	10	TY Lab Werkzaamheden 737	3020	xxxxx	3	TY	27-10-2017	KW	1-10-2017

4.2.2 Enriching the data

Apart from cleaning the data from wrong entries we also attempted to extract additional information for our analysis. Two subjects were addressed:

1. Customer code instead of sales order (SO): In some cases a customer code was recorded in the data file while it concerned a non-routine task that should have a SO number. This information was extracted from the description field where SO identifiers were used. We cross referenced a database of SO's to any identifiers in the description field. Doing so we were able to overwrite the current IVS-code with the sales order where suitable. Table 2.4 gives an example on its row 2 where a SO overwrote the code because of the TR.xxx.xxx identifier in the description.
2. Aircraft type: In most of the data, no aircraft type is specified. However, the description often mentions a specific type for which the work was performed. As we are interested in how types might drive demand, this information was extracted from the description and

added to the data. This was done by matching the type numbers of interest (e.g. 737) to the description field. An example can be seen in Table 2.3 and Table 2.4, the first row in the first Table 2.3 has a sales order code and no aircraft type linked to it. In the description, we can match the type number 737 and the first row in Table 2.4 has this as additional information.

In this manner, we end up with a data set cleaned and enriched with additional useful information. We accept a margin of error in this approach and a more in depth data research would provide a better dataset but this is outside of our scope.

4.2.3 Preparing for analysis

In order to make the data more manageable some further steps were taken to clean the data of unnecessary information which we show with a running example. Table 4.2 shows the source data as previously explored in Section 2.2.

Table 4.2 Source data

Rec. Order	OpAc	Description	Work ctr	Pers.No.	hours	IVS code	Date	Prd cod
3005655	10	Mod and design 787	3033	xxxxxx	2	TL/787	20-12-2017	MO

After the data was analysed and enriched with further information, we are able to remove information unnecessary for the analysis. Description, work centre personnel number can all be removed in Table 4.3.

Table 4.3 Cleaned and enriched

rec.order	op.ac	hours	Date	Prd cod	IVS code	Plane type	Routine	Cust code	KLM?
3005655	10	2	20-12-2017	MO	TL	787	Yes	VOH	Yes

While the data is already easier to manage we apply further standardization and can remove the IVS code as it is translated to its customer code and the identifiers that are not relevant to our forecasts. Table 4.4 shows the resulting dataset where we only have the number of hours worked the date and identifiers for the different characteristics we want to analyse.

Table 4.4 Tidier with structured notations

hours	Date	KLM	Cust	Type	Routine?	Prd code
2	20-12-2017	KLM	VOH	787	ROU	MO

As a final step, we can combine the identifiers to create a single code that contains all the specific information, this is shown in Table 4.5. Additionally, we remove the day of the date as we are interested in monthly aggregation.

Table 4.5 Final format

hours	Date	Code
2	12-2017	KLMVOH787ROUMO

From this data we can extract the characteristics as needed through the identifying codes which are of equal length and structure. This gives us a more manageable dataset that still represents all the information we desire. But this only provides us with data series of the lowest available granularity Section 4.3 explains how we aggregate the data in order to leverage more of the available information.

4.3 Data aggregation / defining the hierarchies

Chapter 2 taught us that the engineering demand is made up of various characteristics, all with their own effects on demand behaviour. Section 2.3 subsequently showed that different combinations of these characteristics, a subset, contain different information leading to the conclusion that all different subsets should be considered. We can aggregate demand over the different characteristics and found that aggregation over time is possible as well. In this section, we look at the aggregation levels of interest in Section 4.3.1 and define the demand structure/hierarchy in Section 4.3.2.

4.3.1 Levels of aggregation

The historical data available is of a daily, individual, task specific granularity as explored in Section 2.2. In Section 3.3, we discussed how aggregation of data changes its behaviour. Noise in the data tends to even out and long term effects become clearer. In this section, we motivate our choice of aggregating to a monthly level while regarding all possible combinations of demand characteristics.

Temporal aggregation

We choose to aggregate data of daily granularity to a monthly level. This causes a loss of information on a sub-monthly level, however this choice is justified on several accounts.

- The research context asks for an approach that provides forecasts for managing capacity. Capacity is not adjustable in days or weeks and as such that level of granularity would not be beneficial. Weekly planning could benefit but falls outside our scope.
- Daily granularity shows a sparse spread of activities and would lead to many zeros in the demand series making it difficult for most time series model to forecast.
- Engineering activities do not show the volatility or sub-monthly patterns that demand like call centre calls might. Additional information from more disaggregate levels is therefore not likely to contribute to our goal.
- Further aggregation to bi-monthly, quarterly or yearly level would discard information useful for timing capacity changes on team levels.

Appendix G shows the effects of temporal aggregation on demand specific to the 777 type. We observe that variability decreases with higher level aggregation. Monthly aggregation smooths the most volatile behaviour, resulting in information more suitable for our goal. It provides an overview of demand better interpretable by the forecasting models and is in line with monthly capacity planning.

Demand characteristics aggregation

In Chapter 2 we discussed how each different combination of the demand characteristics presents different information. We concluded that each specific subset could be of interest and therefore needs to be considered. Given the five characteristics from section 2.1 this results in monthly demand for every (combination of customer (internal and external), type, routine and task. By not excluding any characteristic or combination thereof we retain all information. This is of importance as different decisions might require different levels of information. This becomes even clearer when we add time as a 'characteristic', we illustrate the benefits of retaining the information with a few examples, with and without time:

- Demand per characteristic provides insights in the necessary general capacity. It indicates the requirements for specific high level demand. E.g. total demand for the 777 type provides information on the capacity necessary with applicable skills.
- When any of the characteristics are combined more specific and detailed knowledge can be obtained providing increasingly detailed information. This enables informed decision for any level of capacity of interest. E.g. demand for repair development specific for 777 type aircraft provides information on what proportion of 777 demand is necessary for a specific task an might require other skills. On the most granular level we can combine all five characteristics.

If we add time we can assess demand from monthly level upward

- 777 Demand per month for a year provides a general overview and helps to define the budget.
- Repair development demand specific for the 777 on a monthly level provides information on when that budget should be used to manage capacity.

Through our choice of aggregating data to at a monthly, characteristics defined granularity we can offer information on demand useful for both high level tactical and lower level operational decisions on capacity. Retaining all demand subsets allows us to aggregate them in several meaningful ways leading to the demand structure in Section 4.3.2.

4.3.2 Hierarchical structure

In Section 2.3 each different combination of the demand characteristics was concluded to contain potentially useful information. Retaining demand at this detail on monthly basis is in line with the research goal as it provides information for decisions on both tactical and operational levels, see Section 4.3.1. The downside of considering all the different subsets is the complications it presents to forecasting. The number of forecasts to produce is higher than if fewer characteristics were regarded. Additionally, the forecasts between different aggregation levels will not be coherent. That is, the separate forecasts for disaggregate demand will not equal the forecast for a higher level, e.g. the total 777 demand forecast will not equal the sum of forecasts for each individual task for the 777.

A solution to this was presented in Section 3.6.2, hierarchical or grouped forecasting offers a way to use all different subsets of demand while reconciling the difference in forecasts for all these levels. In order to leverage this technique a hierarchy for the data needs to be defined. We use the characteristics from Section 2.1.1 as these are the relevant groups used in the different subsets. The lowest level of the demand series will consist of the following characteristics:

- Internal (KLM) or external customer
- Specific division/customer
- Aircraft type
- Routine or non-routine task
- Task code

Each of these characteristic then provides a possibility to aggregate demand, in line with theory of Section 3.6.2. The entire structure is too large to visualize but we will illustrate our structure by looking at two examples that show a limited part of the structure. First, the only hierarchy in our set and then a grouped example.

Hierarchical example

The customer classification of in- or external and its subdivisions are hierarchical in nature. With Figure 3.9 as an example, we have overlaid our example structure in Figure 4.2. The top level corresponds to total demand, the second level corresponds with demand subsets for internal (In), KLM related, or external (Ex) customers. Finally, the bottom level represents demand for specific subdivisions (In1, In2 and In3) or customer (Ex1 and Ex2). Figure 4.2 Only shows 5 bottom level nodes, 3 and 2 for internal and external respectively, while in reality there are as many nodes as there are distinct customers but the general structure is the same. The bottom level nodes are uniquely part of either In or Ex node and no other order of aggregation is possible.

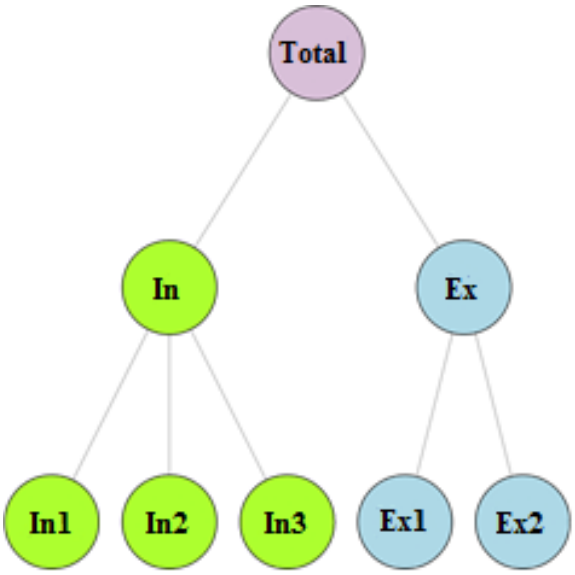


Figure 4.2 Customer hierarchy (Authors own figure adapted from Hyndman & Athanasopoulos (2018, p. 10.1)

Grouped example

We highlight a grouped structure between in- or external customers and demand for 777 and 787. Based on Figure 3.10 we apply relevant characteristics in Figure 4.3 to see how different bottom level demand produce different midlevel series with their own behaviour. To keep the example concise we restrict the figure to just two types and the in- external characteristic. The bottom level in Figure 4.3 has four separate nodes for in- and external customers combined with either 777 or 787. From there we can either aggregate to customer level or to type level producing different series. These then sum to the next higher level. There is no unique order of aggregation and thus all these nodes are valid.

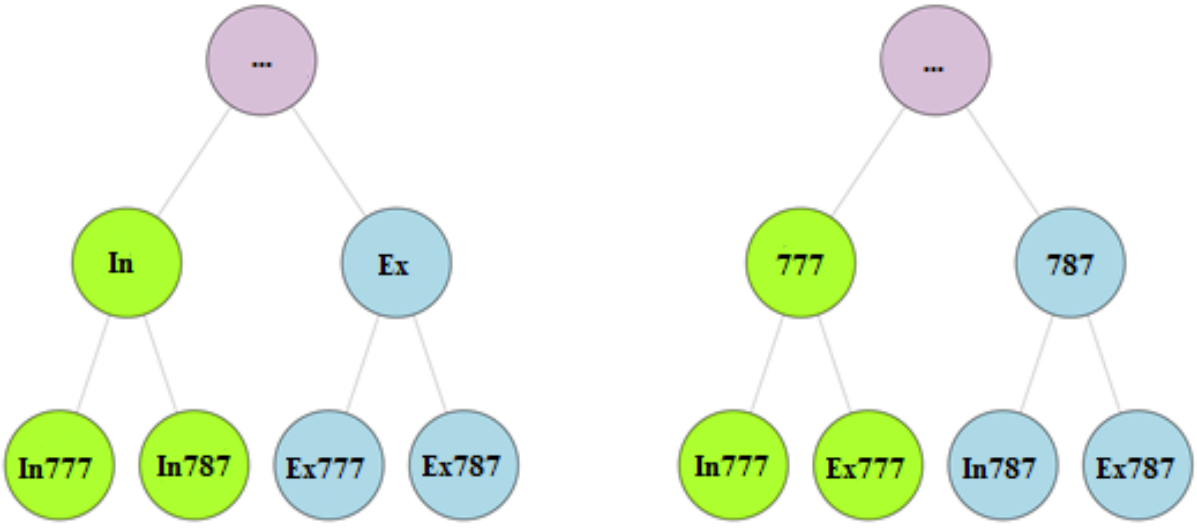


Figure 4.3 Grouped demand structure example (Authors own figure adapted from Hyndman & Athanasopoulos (2018, p. 10.1)

Total structure

The structure in Figure 4.3 was incomplete as only 2 of the several types were used to illustrate the behaviour. When included the structure becomes a complicated web of nodes and connections. Despite this, the structure always has two constants, the top level "total" node and the lowest level nodes with the highest granularity. Between these two levels we can take any possible route of aggregation over the available characteristics. Our data is defined at the most granular bottom level where all characteristics are present and no aggregation has taken place. For instance, the demand example from Section 4.2.3 is identified with KLMVOH787ROUMO which can be translated to the following

- KLM: The task was done for an internal (KLM) customer
- VOH: is the specific subdivision of KLM for which it was performed
- 787: The type linked to the task
- ROU: it was on a **rou**tine basis
- MO: It was a task related to **mo**dification

Every entry has the same structure and we can chose in what order we want to aggregate to the total demand. We can imagine that the top and bottom level nodes are constant in a figure like Figure 4.3 with the middle levels changing depending on the aggregation over characteristics of interest. If we take all different routes we can count the number of different groups and nodes present in the data, resulting in Table 4.6. From the total demand to the most disaggregate we find 16 distinct groups based on different aggregations. From 713 bottom level series, we come to a total of 1716 nodes representing distinct subsections of demand. The description in Table 4.6 either describes a singular characteristic or refers to the combination of characteristics from previous groups.

We arrive at these groups by aggregating demand over each characteristic and the possible combinations, produced in R (see Appendix H). Following the definition from Section 3.6.2 our structure is predominantly grouped. Only G1 and G2 are a hierarchy as they are uniquely structured in their ordering. A specific customer or subdivision is always exclusively an internal, KLM, or external customer. The other characteristics are independent of each other and no unique aggregation is defined. With all the different nodes defined we end up with 1716 monthly time series available to forecast, each with 72 observations for every month in 2012-2017.

Table 4.6 Defined groups and resulting nodes

Group	Description	# Nodes
Total	Total demand	1
G1	KLM or customer	2
G2	Specific division or customer	15
G3	Aircraft type	9
G4	Routine or non routine	2
G5	Specific task	62
G6	G1+G3	15
G7	G2+G3	83
G8	G1+G4	4
G9	G2+G4	26
G10	G1+G5	98
G11	G2+G5	271
G12	G3+G4	18
G13	G3+G5	302
G14	G4+G5	95
Bottom level	G2+G3+G4+G5	713

4.3.3 Outlier series

Lower level of aggregation lead to higher levels of variation in the data, as shown in 3.3.5. As a result series in lower groups of the demand structure experience more sporadic demand. This translates to months with no observed demand, this becomes prevalent at the lowest level series. Averaged over the groups 19% of the monthly observations experience no demand, ranging from 0% at the higher levels to 63% of the observations for the lowest level. Appendix K shows the characteristics of each group regarding observations with zero demand. Given the large number of zero observations in the lower levels we desire a way to indicate whether that demand is relevant for future demand. The following statements attempt to differentiate between relevant and non-relevant demand with many 0 observations:

- No observation in the past 24 months. If no demand has occurred in the past 2 years we can reasonably assume that it will not suddenly occur forecastable future and therefore a 0 forecast over the desired horizon should be adequate.
- 12 or less periods experienced demand of which non in the past 9 months. If 12 observations occurred in the last three years only 1/3 of the periods experienced demand. Combined with not occurring in the last 3 quarters we can assume them to be irrelevant for future demand.

Applying this rule over all nodes we find that an average of 12% per group is identified as an outlier. In Table 4.7 we see that it is more prevalent in lower groups, as expected with the BTS finding 44% of its nodes as outliers. A total of 540 nodes do not experience any currently relevant demand. However, they cannot be removed from the data, the nodes represent past work and affect the other groups in the demand structure. For instance, if repair development (RD) for the 777 is not relevant anymore, current and future demand might be 0. Yet, total demand of both RD and 777 are separately influenced by its past occurrence. If removed all groups are affected and past information that influences model performance would be lost. We mark the relevant nodes as outliers and will forecast them regardless. Given that our assumptions are correct the best forecasts will be 0 for the entire horizon and we should see the effects reflected in the performance.

Table 4.7 Outlier series

Group	# Nodes	# Outlier	% outlier
Total	1	0	0%
G1	2	0	0%
G2	15	0	0%
G3	9	0	0%
G4	2	0	0%
G5	62	1	2%
G6	15	2	13%
G7	83	16	19%
G8	4	0	0%
G9	26	2	8%
G10	98	15	15%
G11	271	80	30%
G12	18	1	6%
G13	302	90	30%
G14	95	20	21%
bts	713	313	44%

4.4 Forecasting the different demand subsets

Section 4.3 provides us with all the different demand subsets that are available. We saw that each of the groups, let alone the individual series, have different characteristics. In order to effectively forecast these different characteristics, we need to fit suitable models. But manually evaluating, adjusting and fitting a model to 1716 different series is impractical and an automated approach is required. In Chapter 3, we introduced different methods and concluded that each uses different ways to extract information from data. If we combine this insight with theory from Section 3.6.1 we find a way to apply multiple methods, their combinations and select the best fitting alternative per subset or group. This approach decreases the need for preliminary model selection and should produce consistently good forecasts while reducing model uncertainty caused by selecting a singular model. In Section 3.5.2, we found that the MASE was most suitable for our dataset and we will use it to compare model accuracy. We propose the following forecast approach for each demand subset:

1. Fit and forecast with a base collection of 10 methods
    - (Seasonal) naive (S, N) (discussed in Section 3.4.2.1)
    - ETS (E) (discussed in Section 3.4.2.2)
    - Theta method (Tf) (discussed in Section 3.4.2.3)
    - Arima (A) (discussed in Section 3.4.2.4)
    - TBATS (B) (discussed in Section 3.4.2.5)
    - Thief (H) (discussed in Sections 3.4.2.1 and 3.6)
    - Current practice (M) (discussed in Section 2.4)
    - Intermittent specific models (discussed in Section 3.4.2.6)
        - Imapa (I)
        - Croston (C)
2. Compute all combinations by averaging forecasts (discussed in Section 3.6.1)
3. For each compute accuracy with MASE (discussed in Section 3.5.2)

This method avoids manual model selection and uses the benefits of forecast combination from Section 3.6.1 to both limit the necessity for model selection and improve accuracy where possible. Sections 4.4.1 through 4.4.4 describe the steps in more detail.

4.4.1 Considered models

In order to leverage the power of forecast combination applying several methods is necessary before we can combine them. Several different types of models are applied of which the forecasts are combined. All implementation was performed in R using existing implementations where possible, mostly from the forecast package in R (Hyndman R. J., et al., 2018).

(Seasonal-) naive

The naive method is a simple method from Section 3.4.2.1, it assumes that the last observed value will continue in the future, seasonal naive does the same with the previous season's observation. The method is easy to grasp and explain but simplistic considering the different time series for review. The method is included for two reasons: First, it serves as a benchmark for the more complicated methods. Second, some series contain data that can offset models and make them overreact (ETS for example is sensitive to outliers), including a simple and stable forecast might balance this reaction when combined. Applied with the '(s)naive' functions in R (Hyndman R. J., et al., 2018).

Exponential smoothing

Exponential smoothing is a versatile method and widely used in forecasting. Using the state space framework (Hyndman, Koehler, Snyder, & Grose, 2002) the ETS fitting process can be automated by choosing parameters that result in the best in-sample fit. Practically this can be applied by fitting all combinations of parameters within a desired range (see Section 3.4.2.2) and choosing the combination with the lowest AIC. This approach works, because the AIC provides a balance between information captured and overfitting (see Section 3.5.2) pointing to the model. The flexibility of ETS in its capabilities of handling errors, trend and seasonality in both additive and, be it limited, multiplicative fashion explain why it is so widely used in practice. Additionally, it makes ETS suitable for the diverse dataset under consideration and requiring forecasts. The model is fit to the data with the 'ets()' function, which applies the AIC methodology, and passed to the 'forecast()' function in R (Hyndman R. J., et al., 2018).

Theta method

The theta method is a special case of exponential smoothing, SES with a trend of a fixed expression. It is less flexible but also more intuitive than the ETS framework. We include this method, because of its good performance in practice (Makridakis & Hibon, 2000) and it allows insights in the accuracy more complex ETS models offers. The Theta method is applied through the 'thetaf()' function in R (Hyndman R. J., et al., 2018).

ARIMA

Arima is one of the most widely used techniques used in practical forecasting. The ability to handle time series characterises and stationarity makes it very versatile and is therefore included. In order to fit an Arima model the parameters p, d and q need to be estimated for the autoregressive, differencing and moving average parts of the model. Like ETS the choice of different parameters for both p and q can be determined by minimizing the AIC, see Section 3.5.2. The order of differencing changes the AIC and can therefore not be estimated through that approach, a unit root test determines a suitable number of differences. We fit the Arima model with the 'auto.arima()' function in R (Hyndman R. J., et al., 2018), which iterates through different parameter combinations finding the lowest AIC. It uses a reasonable default approach determined by Hyndman and Khandakar (2008) to balance computation time with the found model.

Dynamic regression

The Arima framework can extend regular regression with time series capabilities by applying an Ar(i)ma model on the regression errors, see Section 3.4.2.7. This method can potentially include any number of external variables to explain part of the variability of the data. Preliminary testing showed mixed results with external variables and defining what series they should be applied on. To illustrate, the total demand for each type was tested against their number in the fleet on the assumption that more craft might cause more maintenance, leading to more engineering tasks. The results of our testing are presented Table 4.8.

*Table 4.8 Types regressed and forecast against number of craft, with **Best** and 2nd best*

Type	Correlation	R-squared	adjusted r-squared	P-value	MASE type regression	MASE no regression	MASE regression all types
787	0,768	0,59	0,58	0,00	0,44	0,46	**0,40**
777	-0,228	0,05	0,04	0,05	**0,31**	**0,31**	0,32
330	-0,330	0,11	0,10	0,00	1,24	**1,23**	1,54
744	0,426	0,18	0,17	0,00	1,90	**1,08**	2,70
73N	-0,100	0,01	0,00	0,40	1,56	**1,38**	1,59

What we learn from Table 4.8 is that the results are mixed, only the 787 and 744 experience positive correlation between the number of aircraft and engineering demand. Evaluating the R-squares we observe that this translates to 60% of variation explained for the 787 but nothing comparable for the other types. Furthermore, R-square is not a measure of forecast power and three different forecasts with a 12 month horizon are compared on the MASE.

- Dynamic regression forecast on the number of a specific type
- Regular Arima forecast without external variables
- Dynamic regression with all the types as external variables

We observe from the resulting MASE that no forecasts, except that of the 787, benefit from including the number of aircraft as external variables. This implies that the number of aircraft has no real predictive power on type related demand. In some cases forecast accuracy is not greatly affected while in others it is significantly reduced, especially the 744.

The conflicting 787 result is partly explained by an upward trend that both the demand and external variables experienced. The 787 was phased in during the period and as such demand has increased as did the number of aircraft, creating a trend for both. As a result, regression identifies correlating behaviour with no apparent causation if the other results are realistic. Regressing total 787 demand on a simple time trend substantiates this as it results in an R-square of 89%. This implies that the number of aircraft in the fleet has no direct relation to the size of engineering work. Given that a lot of engineering work is done for a type in general and not per individual aircraft, e.g. a repair manual applies to each craft, this makes sense but some relation was expected. On the other hand the forecasts of the 787 became accurate by including all variables implying some predictive power after all.

This leaves us with mixed results and no clear conclusion on the usability of dynamic regression. The discussion is further complicated when regarding the demand structure defined in 4.3.2. If an external variable is found to correlate with and explain the variation of a node, how should this be applied to the rest of the structure? Assuming that the number of aircraft are actually beneficial for total 787 demand we have to decide whether that is also the case for any other related aggregation node. For instance, is the variable still useful when just repair development tasks for the 787 are considered or when only non-routine tasks are of interest? Given that our demand structure has 1716 different demand series we have several choices for including external variables:

- Define useful variables for each individual node; This requires a substantial amount of work to assess each of the 1716 different series
- Define useful variables per group; this reduces the different groups to 16 but there is no clear expectation on how that would affect each of the underlying series. The MASE results from

regressing on all different types in Table 4.8 show that nearly all forecasts actually became worse by including irrelevant information.
- Apply all variables to all nodes; the least amount of work as the variables would only need to be collected. Yet, like defining variables for the groups will probably not result in better forecasts.
- Define useful variables for high level nodes and apply them only to the related lower level nodes, i.e. let the variables 'trickle down' the demand structure.

A proper investigation on the effects of external variables on the entire demand structure would incur a high cost in time and necessary domain knowledge. Furthermore, forecasting with external variables requires forecasted values of the variables, introducing additional uncertainty and requiring further investment in time and knowledge. We believe that dynamic regression could produce very interesting results but it requires a dedicated research to properly execute. Due to the high complexity and required time to properly apply we exclude dynamic regression and focus on applying time series models.

TBATS

TBATS is capable of handling, multiple, complicated patterns and seasonality unlike Arima or ETS naturally can. This is mainly of interest with high interval frequencies like sub daily observations. But, as we do not know what kind of patterns all our subsets exhibit we include the method to observe its performance. If it does well it can potentially indicate interesting subsets to further investigate on the detected patterns. We implement TBATS like described in Livera et al. (2011) and implemented with the 'tbats()' function in R (Hyndman R. J., et al., 2018).

Temporal hierarchy forecasting

As described in Section 3.6.2, temporal characteristic of data can be considered to be a hierarchical as well. We have opted to implement a form of temporal forecasting. It requires aggregating the demand to higher time levels (e.g. month →quarter→year) forecasting for each of these and then reconciling as described in 3.6.2. The approach was proposed by Athanasopoulos et al. (2017) and then implemented in the R package thief (Hyndman & Kourentzes, 2018). We opt for the implementing it with forecasts from the ETS framework. With this choice, we can benchmark the effects of temporal forecasting against regular ETS to see its merits.

Intermittent demand models

In Section 3.4.2.6, we described methods to handle series with a majority of zero values. These are important to include as our low level demand subsets frequently contain zero values and show sporadic demand. Croston's method and iMapa are both applied, Croston is the classic method for intermittent series and is used as implemented in the forecasting package in R with the 'croston()' function (Hyndman R. J., et al., 2018). iMapa combines concepts of intermittent demand methods and temporal hierarchical forecasting. Implemented based on the work of Kourentzes (2014) and Petropoulos & Kourentzes (2015) in the 'tsintermittent' package in R (Kourentzes & Petropoulos, 2016). It uses concepts of forecast combination, hierarchical forecasting and intermittent models to produce a final forecast. We therefore expect it to perform well on some of the demand subsets.

4.4.2 Fitting models and forecasting

Section 4.4.1 defined the methods to be considered for forecasting demand. We can now fit the models on historical data and forecast future values. Section 4.3.3 highlighted that some series might be considered as outliers and that 0 forecasts might be the most appropriate, however this might not

always be the case. Some might represent low frequency events and a 0 forecast would underestimate its impact. Applying the methods from Section 4.4.1 allows us to see what forecast would be best according to the MASE. After producing the forecasts we can confirm whether we were correct in defining the series as outliers. For the regular series all models are applied, while performance of different models is expected to vary we have no reason to exclude any pre-emptively. Following the fact that different models extract different information and contributes this to the forecast combination all have potential. In this manner no information is lost and all different model combinations can be examined.

For each model a forecast is produced for every month of the required forecast horizon (12 for a monthly forecast over a year). If all models proposed in Section 4.4.1 are applied we end up with 10 individual forecasts that can be combined before any accuracy measures are applied. The forecasts are produced in R, for additional information see Appendix H and Appendix I. Figure 4.4 and Figure 4.5 show the 10 resulting forecasts for total demand over 2017, with full historical data and zoomed in respectively. Table 4.9 Provides abbreviations for the methods which will be used from hereon in figures and text. It is apparent that each model extracts different information and thus produces different forecasts. Most of the models are able to detect patterns and appear to be within reasonable range of the actual values.

Table 4.9 abbreviations

N	Naive
S	Seasonal naive
M	Mean
E	ETS
A	Arima
H	Thief
B	TBATS
C	Croston
I	iMapa
Tf	Theta

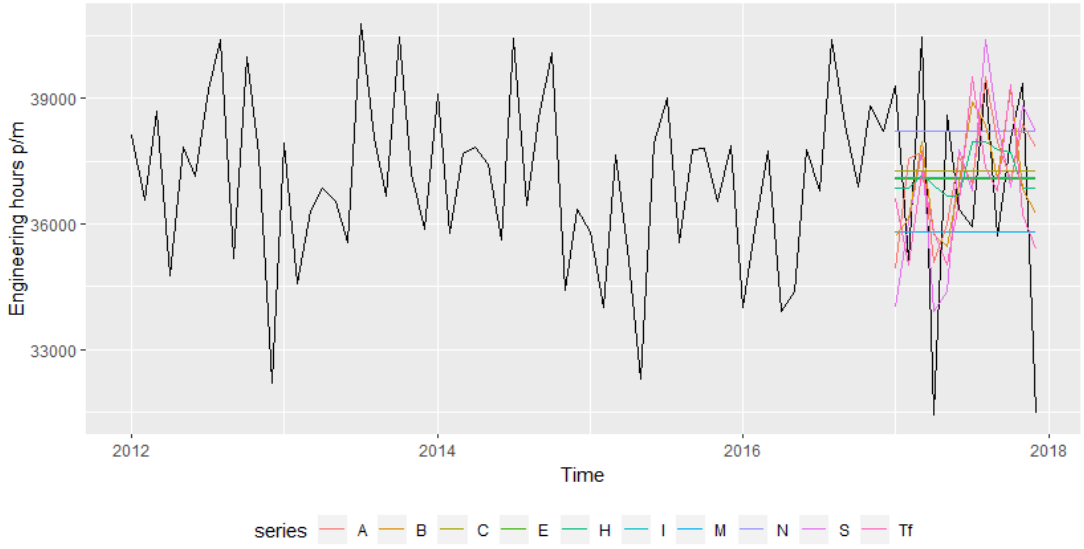


Figure 4.4 10 Base model forecasts of Total demand

Figure 4.6 and Figure 4.7 show the same for 777 demand. Observe the larger deviations from the actual values. Demand of the 777 has changed over time and as a result models experience difficulty in extracting patterns and forecasting them. Some methods are able to extract the correct information of what looks like a downward, but levelling off, trend. From these 10 models we could determine the best performer but Section 3.6 has taught us that combining the information from different models will generally increase performance.

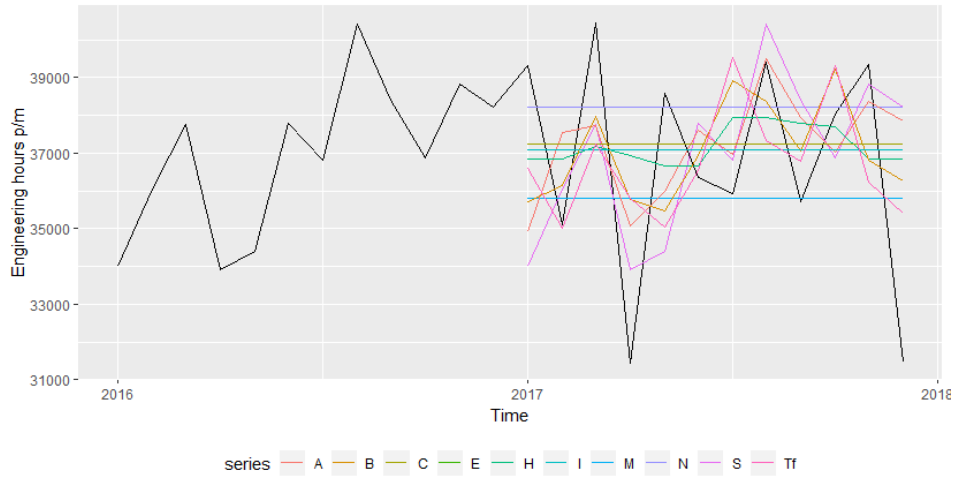


Figure 4.5 10 Base model forecasts of Total demand (2016-2017)

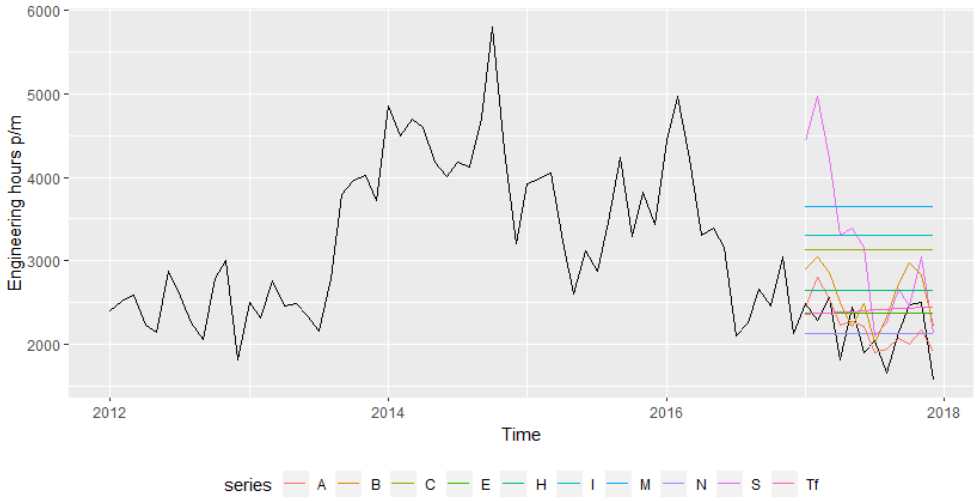


Figure 4.6 Base forecasts 777 total

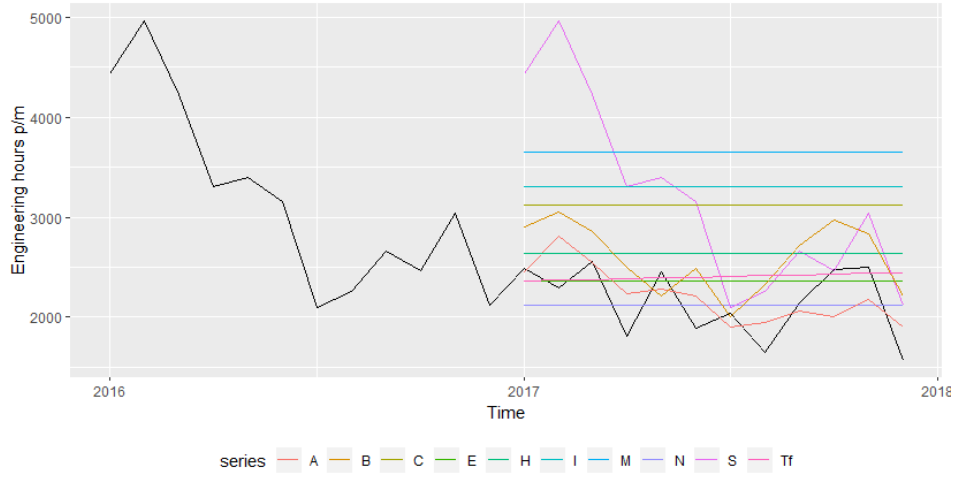


Figure 4.7 Base forecasts 777 demand (2016-2017)

4.4.3 Combining the forecasts

With each of the, maximum of, 10 different forecasts produced per node in Section 4.4.2 we have captured different patterns from the demand. For each period of the forecast horizon multiple values are available that al describe different possibilities of the future value. We found in Section 3.6.1 that the combination of different forecasts regularly outperform single model forecasts due to reducing model uncertainty and including more of the information available in the data. Different weighting schemes can be used to combine the different forecasts but the arithmetic mean is often the best choice due to the forecast combination puzzle (see Section 3.6.1). All average combinations of the forecasts are calculated resulting in a large number of different forecasts for each node. With n the number of models considered for that subset, we get:

$$Number\ of\ forecast\ combinations: \sum_{r=1}^{n} \frac{n!}{r!\,(n-r)!}$$

If 10 methods are applied this results in 1023 unique combinations which provides a large sample of forecasting results we can use for assessing model accuracy. By evaluating all combinations we are able to pick those that work well on the data or a group and perhaps omit models that do not. Due to the different capabilities between models we expect to observe differences in performance between them. Figure 4.8 shows some selected combinations of forecasts for total demand, the series shown are the best and worst performer (determined by MASE, see Section 4.4.3), mean of the base forecasts and the min and max values, indicating the spread over the different forecasts. Figure 4.9 does the same for 777 demand and we can clearly see the much wider spread and distance from the actual values. This clearly illustrates the spread results of each combinations ant thus also of their accuracy. In order to determine the best performing methods and combinations we will assess the accuracy of the different methods.

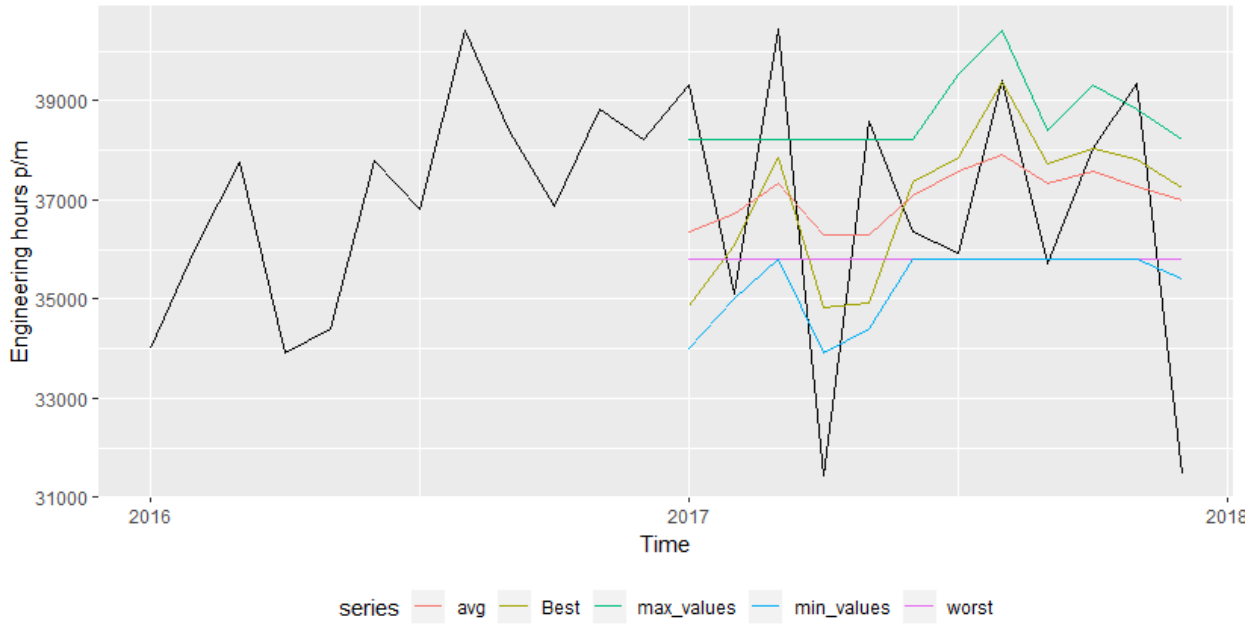


Figure 4.8 Selection of forecast combinations for total demand

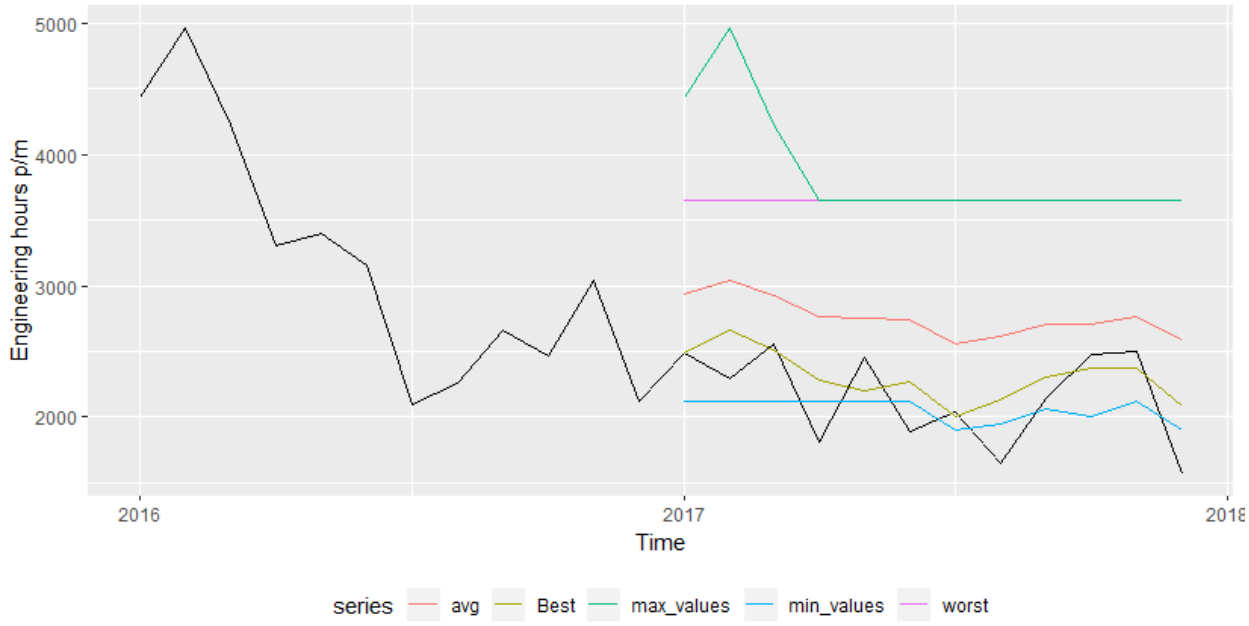


Figure 4.9 Selection of forecast combinations for 777 demand

4.4.4 Measuring accuracy

Each combination produced in Section 4.4.3 uses information from the different included models. Their respective forecast accuracy is therefore expected to differ and needs to be measured. In Section 3.5.4 we found the MASE to be suitable because it is scale free, provides comparison between models and provides a benchmark to the (seasonal) naive forecast. We calculate the MASE for each period in the forecast horizon and average the performance over the entire forecast horizon. This provides us with average forecast accuracy over the entire desired forecast horizon.

To fairly estimate how well a combination of methods performs, we require an additional validation step in addition to the training and test set. Our approach will be to split the available data in three parts a training, test, and validation set allowing us to verify results. By fitting the models on the training set, forecasting over the test set horizon and measuring that accuracy, we can determine the best performer over that period. However, this is no guarantee for future accuracy and we require verification of its performance. Fitting the models again and the training and the test set increases the available data for the models, then a forecast over the validation period provides a second independent measure of accuracy. Average performance of methods over both horizons will help indicate which combinations of models perform well. We take the following approach:

- Determine suitable training, test and validation sets
    - E.g. Training: 2012-2015, Test: 2016, Validation: 2017
- Fit the models on a training set and forecast over the test set period
    - E.g. Fit on 2012-2015 and forecast for 2016
- Measure accuracy of all different forecast combinations on a test set
    - E.g. Determine MASE of 2016 forecast
- Fit models on the training and test set and forecast over validation period

    - E.g. Fit on 2012-2016 and forecast 2017
- Measure accuracy on a validation set
    - E.g. Determine MASE of 2017 forecast
- Determine the average MASE per each model combination over the 2 forecast periods
- Determine overall best performing models
    - Average MASE over each method for each group as defined Section 4.3
    - Average MASE for each method for every subset

Following these steps, we can determine the method (combination) that has performed most consistently on historical data. This provides a reasonable way to select a combination of model to be used to forecast future periods for which no test data is available. We can make 3 different selections of different models all decreasing overall accuracy:

- Preferable combination selected for each of the 1716 nodes, the most accurate but also lead to the most diverse set of models to consider.
- Combinations that performs best for each of the 16 groups, averaging the performance over all the nodes in a group provides a best overall performer presenting 16 different combinations to forecast with. One can expect overall forecasting accuracy to drop as more generalized methods are applied.
- Combinations that performed best over the entire collection of nodes. Further generalizing the results decreases the models to consider but will also further decrease overall accuracy.

In this manner, we provide a method that forecasts without a need for manual input. It provides reasonable indications of the most suitable models based on the MASE. As the results come from test and validation sets their performance has been verified over two different parts of data and is therefore stable overtime. Additionally, because forecast combination is applied the chance of model misspecification is lowered and should therefore have lower risk of suddenly performing badly. We can therefore assume, that this combination should also perform well in forecasting the actual future.

On the other hand bad performance in either the test or validation period might lead to immature discarding of certain methods. Additionally, if a method performs well on either set or the data is assumed to retain its behaviour it should also perform well on an adjacent time period. When this is not the case we can examine whether the underlying behaviour has actually stayed the same or if the data exhibits patterns unforeseeable from historical data. If this is the case then in order to produce accurate forecasts these effects might need to be adjusted for which we need judgmental input.

4.5 Judgemental adjustments

Section 3.4.1 described the merits and dangers of judgmental forecasting and adjustments, summarized as, "trust the statistical forecast unless your reasoning is justifiable and according to clear rules". From the context and observed from data we know that external factors can heavily influence demand introducing behaviour that statistical methods cannot predict. As such, the proposed statistical method is expected to produce bad results on certain series. This is where trust in the statistical forecast should be replaced with judgmental forecasting. Currently the organization has no structured way to apply this and we identify two situations in which judgmental input is required and then a method to apply it.

Bad in-sample performance

As seen in Figure 4.1 a feedback loop is initiated after measuring accuracy and performance on both the individual and the reconciled forecasts. From these evaluations we can identify series on which the statistical framework has performed badly. The MASE scores per series over both the test and validation set serve as outlier detection for our method, scores that are significantly higher than expected indicate problematic series. The MASE benchmarks against the (seasonal) naive method and returns < 1 when performance is better and > 1 otherwise. Results significantly larger than 2 would merit further evaluation. Judgmental input and evaluation of these series can then try to determine what caused this and what would provide a fit to the data and potentially the forecasts.

Statistically unforeseeable effects

Experts in the organization have knowledge of effects that cannot be foreseen from historical demand. This knowledge is essential for producing reasonable out of sample forecasts. An example of demand that would not be correctly predicted by the model can be seen in Figure 4.10.

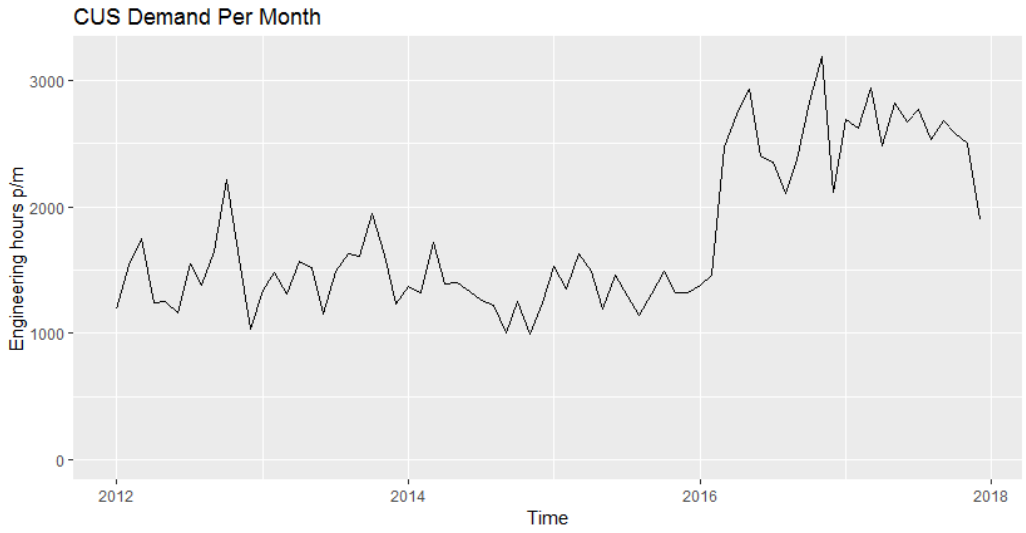


Figure 4.10 Customer demand per month

Here demand clearly experiences a shift, seemingly jumping to a new level just after the start of 2016. The demand nearly doubles in size and the statistical methods cannot foresee this when forecasting with data up to 2016. It would predict that the fairly level behaviour would continue into the future and therefore be off by a large margin. This would produce a bad in sample accuracy as address in the previous paragraph. In the case of Figure 4.10 a customer contracted engineering for additional

work. This contract was set to start from 2016 and a rough amount of hours was determined. In such a situation judgemental input is necessary to provide an accurate forecast. Experts should therefore always be consulted on specifying events that are out of the ordinary for past behaviour.

Applying judgmental input

The organization currently has no guidelines on judgemental forecasting apart for consulting the relevant parties and having them adjust to their insight. Our advice would be to adopt clear rules in the forecasting process on how to apply judgment, we propose the following steps:

- First run the statistical forecasting model
- Identify bad in-sample performance based on the MASE
- Determine whether the bad performance of the model can be improved
    - If so, apply improvements and return to the first step,
      E.g. an outlier subset exists due to erroneous administration and can be cleaned
    - Otherwise continue
- If no improvable bad performance remains, consult experts on at least the following topics
    - Major projects that cannot be accounted for (e.g. cabin overhaul projects or introduction of new aircraft type)
    - A new department is going to focus on a specific topic requiring extra capacity
    - When substantial contracts have been made with customers that were previously out of scope
- If external influences have been identified consult experts and use the Delphi or panel of expert's method (Section 3.4.1.2) to get an agreeable forecast and have them justify their decision.
- Compare the judgemental forecast and only consider it if the change is large compared to the statistical forecast
- Apply the forecast to the relevant node overriding the statistical forecast.

By setting out stricter rules on when and how to apply judgment forecast accuracy is more likely to increase. Our recommendations help to ensure that only large adjustments are made, "small adjustments to forecasts tend to hurt forecasting performance while large adjustments improve the performance. This is primarily due to the fact that large adjustments are made only when a manager has significant and relevant information" (Silver, Pyke, & Thomas, 2017, p. 119). However, what should be considered as large is difficult to estimate as it will depend on the point of view of the involved expert. This will require further specification but anything within 10-20% of the statistical forecast or within 1 standard deviation might be considered small. If a forecast is to be adjusted a thing to consider is the effect on the other levels of the demand structure defined in Section 4.3.2. Nodes are part of different aggregations at different levels and if a node's forecast is significantly changed it needs to be considered whether this needs to apply to higher/lower levels as well. This becomes more important when considering the reconciliation in Section 4.6 where different levels of the structure exchange information. An adjustment applied to a single node, and not its higher/lower level connections, might not show its effects.

4.6 Reconciling the forecasts

The goal of the research, as defined in Chapter 1 is to create a model that can make demand more predictable and exert more control over capacity on both a tactical and operational level. Section 4.4 describes which methods are used to quantitatively forecast all demand, as defined in Section 4.3, these forecasts can be adjusted according to judgement, when necessary, following Section 4.6. This provides us with accurate forecasts for all different levels of aggregation. However, accurate these forecasts may be they are not aligned with each other. Different aggregations of data and applying different models produces forecasts that do not sum to forecasts of their aggregates, this might cause misalignment between tactical and operational decisions.

To illustrate what this means; Table 4.6 defined the demand structure resulting in 16 different levels of aggregation. Each of the 16 groups represent total demand divided in different pieces, each of which was separately forecast. Aggregating data presents different information (see Section 3.3.5), resulting in different forecasts. Thus summing all forecasts in a group represents a forecast for the total demand, but these summed forecasts are not equal to each other. Different levels of aggregation present different information, leading to incoherent forecasts over the different levels. We extend our example from Section 4.3.2, specifically Figure 4.3, where we have 4 groups (Total, in- and external customers, type and the bottom level) and 9 unique nodes instead of 16 and 1716. When each of the nodes is forecast separately we can visualize how the forecasts do not add up. Figure 4.11 visualizes the simplified structure with dummy forecasts. What we observe is that several sums of forecasts, that all portray total demand, do not add up. Table 4.10 shows the discrepancies between group level forecasts and the sum of disaggregate series forecasts. G2 and its bottom level time series are consistent with each other but G1 and the total forecasts are different for each different summation.

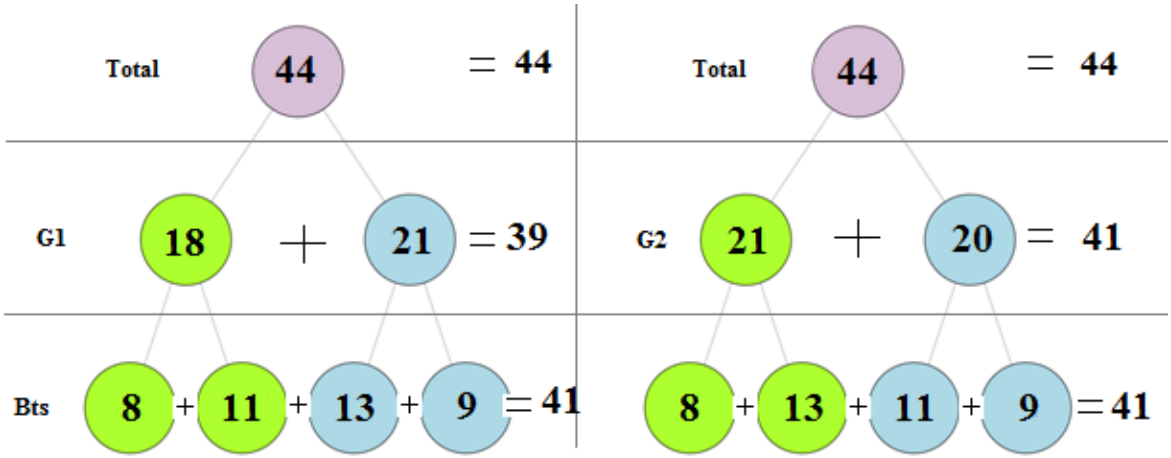


Figure 4.11 Grouped structure unaligned forecasts

The information presented in the nodes differs due to aggregation and as a result models produce forecasts that are not coherent. When the organization makes a decision it is useful that forecasts on all levels portray the same information, possibly leading to misalignments. If budgeting is based on the total forecast, 44 in our example, while planning considers the sum of the lowest level forecasts, 41 in our example, resulting expectations and decisions could be different. Reconciliation removes the differences between forecasts on different aggregations, i.e. forecasts are adjusted in such a way that the sum of lower level forecasts always adds up to forecasts of higher levels.

Table 4.10 Group forecasts versus the sum of disaggregate series forecast

Group	Nodes	Parts	Sum
Total forecast	Total	44	44
	In + Ex	18 + 21	39
	787 + 777	21 + 20	41
	In777 + In787 + Ex777 + Ex787	8 + 11 +13 + 9	41
G1 Internal	In	18	18
	In787 + In777	8 + 11	19
G1 External	Ex	21	21
	Ex777 + Ex787	13 + 9	22

We apply reconciliation as described in Section 3.6.2 by estimating the mapping matrix G. It projects information from the entire structure through the summing matrix onto the bottom level time series. As a result the forecasts are evened out between levels. In Figure 4.11 we see that the total forecast is significantly higher than the other totals, 44 vs 41, 41 and 39. We can expect reconciliation to change the values to even out the difference. Applying structural scaling in the reconciliation leads to the, rounded, values in Figure 4.12.

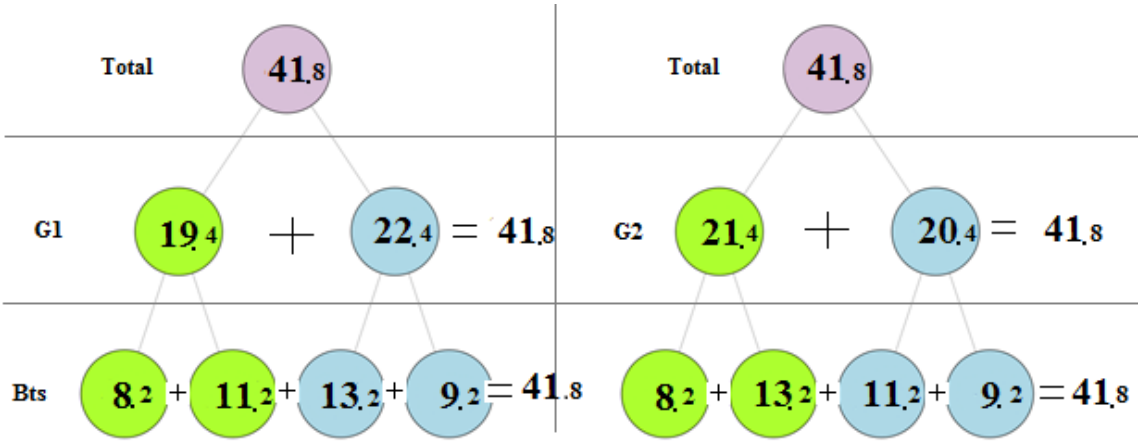


Figure 4.12 Grouped structure, rounded, reconciled forecasts

We can observe how the forecasts are changed depending on their scale. The degree of change increases when going from the lowest level of aggregation to the total forecast. This is exactly what structural scaling describes by using weights based on the number of forecast errors that influence a forecast, see Section 3.6.2 and Wickramasuriya et al. (2018) for further details. The summed forecasts of any underlying group is now equal to the forecast of its higher level, they are coherent and because of this the entire structure can be represented with the bottom level forecast.

We apply the same principle to the entire structure defined in Section 4.3.2. Forecasts for all 1716 nodes in 16 different groups are reconciled. The final forecasts for the 713 nodes at the lowest level can then be aggregated to any desired forecast by summing over the relevant characteristic. This ensures that forecasts used for decisions on different organizational level, e.g. budgeting and operational planning, align. Additionally, theory implies that using information from all different levels of aggregation could improve overall accuracy, but this depends on the individual quality of the forecasts. After reconciliation, all relevant forecasts have been created and we can move forward to evaluating performance.

4.7 Evaluating performance

After all the previous steps we have constructed a statistical forecast but have no real idea of its performance. In our proposed framework judgemental adjustments always comes after, and are based on, the statistical results. Therefore, we desire the statistical results to be as good as possible without manual input. As such, we focus on evaluating the results of the framework without applying judgmental input. In order to properly test performance, we will take several steps to see how well the proposed model works.

4.7.1 Statistical performance

Our main goal is to provide the statistical framework necessary for forecasting and thus we need to know how well the applied methods worked. We will compare accuracy according the MASE for four different forecasting results:

- The current forecasting practice
- The best forecast that resulted from the proposed model before reconciliation
- A benchmark combination forecast of ETS, Arima, Tbats and seasonal-naive
- The reconciled forecast based on the best results of the proposed model

These 4 collections of forecasts help us determine the value of the proposed method. First, we can compare the results of the proposed approach, before and after reconciliation, to that of the current approach indicating whether more accurate forecasts are actually produced. Then, by including a benchmark of just 4 models we assess the benefits of including a larger flexible selection of models. In order to fairly asses the performance we take the train, test, validate approach from 4.4.4. We fit on 2012-2015 and forecast 2016, then we fit on 2012-2016 and forecast 2017, resulting in three separate measures of accuracy, 2016, 2017 and the average over both. This will teach us whether the proposed method outperforms the current process, does not include too many models and what the effects of reconciliation are.

4.7.2 Individual forecast method performance

We evaluate the best outcomes of the proposed method in Section 4.7.1 but this provides no further insight in the performance of individual methods. We choose to apply a set of multiple models and incur a computational penalty for each we include. Determining whether the included models add accuracy to the forecasts allows us to evaluate if such a large collection of models is necessary. Perhaps some models do not contribute to high performance and could be excluded. Additionally, if only a select number of models are responsible for the best forecasts a, large, ensemble could also be unnecessary. In both cases we require information on the performance of individual methods over the different demand subsets. We propose the following comparisons for each method:

- Best average performance per group vs best average performance while excluding a method
- Mean performance per group vs mean performance excluding a method

By comparing the effects a method has on the best performance we can see if there is an immediate impact on overall accuracy. Comparing mean performance lets us see whether the inclusion of a method improves or diminishes overall performance. For instance, we expect that including Croston's method or iMapa will impact the bottom time series more than it does the higher aggregations. As a result their in- or exclusion might have significant effects on forecast accuracy. By evaluating the performance of each method we can conclude which can be omitted without drastic

influence on the results. It will indicate which methods perform significantly better or worse than others as well as provide information on the merits of including more methods.

4.7.3 Total forecast accuracy

All previous comparisons work by comparing a scaled value the MASE in order to say something about the forecast accuracy. This indicates whether improvements in accuracy have been made but not what the actual effect on this is for the organization or the actual numbers. Comparing absolute numbers and errors between groups does not work because of their differences in scale and behaviour but each is a disaggregation of total demand. We propose to sum all forecasts in a group in order to construct 16 different forecasts of total demand. Add the current approach and the reconciled forecast and we can comprehensively compare the differences in accuracy in numbers that are relevant to the organization

4.8 Chapter conclusion

Research question 6: 'How should the forecasting methods be applied to engineering demand?' was answered in this chapter by defining our forecasting approach. We combined conclusions from the research goals, the organizational context and theory to define a suitable forecast method. In Chapter 2 we concluded that all subsets contain different information and should be included. In Chapter 3 we found that different models are suited for different data. This resulted in the conclusion that no single model is suitable to forecast all subsets and multiple would be needed. Additionally, we concluded that manual application and tuning of these models would require too much time.

We propose that engineering demand should be forecast with multiple model to use as much of the available information as possible. By combining the forecasts from all models we produce forecasts that combine the different information and produce more accurate forecast. In order to select the best resulting forecast from all these combinations we measure their accuracy with the MASE. This is done for every node in the demand structure resulting in a collection of accurate but incoherent forecasts. These forecast can then be reconciled in order to align all levels of the structure with each other sharing information from the different forecasts over all the levels. With the statistical forecasts in place we can measure and compare performance of different approaches, e.g. current practice vs reconciled forecasts, and determine the most accurate results. Allowing us to conclude whether a more complex approach is useful. Finally, the proposed model then initiates a feedback loop to evaluate the applied models and improve performance where possible.

The proposed framework provides a structured approach to forecasting. There are clear defined steps and trust is placed on the statistical outcomes. An initial setup for applying judgemental forecasting is provided but good judgemental forecasting is dependent on trustworthy statistical forecasts. Therefore we omit further testing in favour of focussing on the statistical framework as our main goal. With the different statistical forecasts in hand we have defined how to compare their results. Chapter 5 is dedicated to evaluating the forecasts and determining the organizational impact of more accurate forecasts.

5 Model performance and results

Chapter 4 provides a guide on how to implement the proposed forecasting method. It focusses on producing trustworthy statistical forecasts, evaluating them which leads to, judgemental, reassessment of the model and its forecasts, see Figure 4.1 for a summarized overview. The proposed method is deemed successful if we can improve over forecasts from the current process. Defining and comparing performance also provides an answer to research question 7.

We focus on the results from the statistical framework while not testing judgmental adjustments. Section 3.2 describes how the first step of good forecast has its foundation in an accurate and trustworthy quantitative forecast which is our main focus. As such we discuss the results of the statistical forecasts, compare different methods and describe the organizational impact. We follow the approach from Section 4.7 to evaluate the outcomes. Section 5.1 discusses the results per demand group and compares our method, the current approach, a benchmark combination, and the effects of reconciliation. Section 5.2 discusses the sensitivity of our method to the inclusion or exclusion of single models. Section 5.3 attempts to discuss the organizational implications of increased forecast accuracy and Section 5.4 presents our conclusions about the forecasting results.

5.1 Group results

Through the proposed method we create several different forecasts that can be used for assessing model performance. We compare the best results of all combinations, the reconciled forecast that aligns all levels of the demand structure and two benchmark forecasts, the current system and a combination of four models. As described in Section 4.4.4 performance is to be tested and validated, therefore we evaluate performance for 2016, 2017 and the average of both. All results shown are after accounting for and removing outlier series, Appendix L elaborates and shows the effects on forecast accuracy and why they are omitted from the results here.

5.1.1 Current method vs proposed method

The most important indication that the proposed method is of value is by providing more accurate forecasts than the current method. The proposed method has a higher degree of complexity implying an ability to extract more information from the data and thus more accurate forecasts. Table 5.1 shows the averaged best MASE per group for 2016, 2017 and the best over both those periods.

General performance

What is immediately clear from Table 5.1 is that applying more sophisticated methods results in more accurate forecasts. All different groups experience higher forecast accuracy than the current method would produce with an average decrease in the MASE of 25% if the years are regarded separately and 20% for the average performance. A 20% MASE improvement implies 20% more accurate forecasts in comparison to a one step ahead, in-sample, naive forecast. The decrease in performance when regarding the average over both years is due to inconsistency in performance. Choosing the most consistently performing model reduces overall accuracy by we sacrifice performance in one year for long term performance. For instance in case of forecasting Total demand:

- A naive, seasonal naive and Arima (NSA) combination is most accurate over 2016 with 0,78
- A seasonal naive + tbats (SB) combination was most accurate over 2017 with a MASE of 1,45
- An average would imply a MASE of (0,78+1,45)/2= 1,12 but consists of 2 different models
- Lowest average MASE: NSA with 0,78 in 2016, 1.52 in 2017 and an average of 1,15

Table 5.1 Forecasting accuracy per group of current and proposed method

Group	2016 Best	2016 Current	2016 Best - current	% diff	2017 Best	2017 Current	2017 Best - current	% diff	Avg Best	Avg Current	Avg Best - current	% diff
Total	0,78	1,02	-0,24	23%	1,45	1,56	-0,11	7%	1,15	1,29	-0,14	11%
G1	2,63	3,09	-0,46	15%	0,96	1,03	-0,07	7%	1,83	2,06	-0,23	11%
G2	3,65	4,22	-0,57	14%	0,78	1,10	-0,32	29%	2,29	2,64	-0,35	13%
G3	0,45	0,64	-0,20	30%	0,53	0,61	-0,08	13%	0,53	0,62	-0,09	15%
G4	0,55	0,82	-0,27	33%	1,05	1,78	-0,73	41%	0,89	1,30	-0,41	32%
G5	0,75	1,03	-0,28	27%	0,77	1,05	-0,28	27%	0,83	1,04	-0,21	20%
G6	0,99	1,25	-0,26	21%	0,46	0,54	-0,08	15%	0,76	0,90	-0,14	16%
G7	1,21	1,67	-0,46	28%	0,80	1,08	-0,29	26%	1,05	1,28	-0,23	18%
G8	1,58	2,22	-0,64	29%	1,12	1,81	-0,69	38%	1,60	2,02	-0,42	21%
G9	0,75	1,17	-0,42	36%	0,94	1,43	-0,49	34%	2,36	2,70	-0,34	13%
G10	0,98	1,26	-0,28	22%	0,91	1,30	-0,39	30%	1,27	1,49	-0,22	15%
G11	0,85	1,10	-0,25	23%	1,64	1,97	-0,33	17%	0,87	1,10	-0,23	21%
G12	0,77	1,11	-0,34	30%	0,50	0,68	-0,18	26%	0,70	0,90	-0,19	22%
G13	1,10	1,56	-0,46	29%	0,61	0,90	-0,29	32%	1,04	1,41	-0,37	26%
G14	0,94	1,32	-0,38	29%	0,68	1,10	-0,42	38%	0,87	1,12	-0,26	23%
BTS	0,96	1,18	-0,22	19%	0,88	1,33	-0,45	34%	0,98	1,28	-0,30	23%

The same model is rarely most accurate in both years and so accuracy in either of the tests needs to decrease in order to pick the method with the highest overall performance. To illustrate the difference in forecasts Figure 5.1 shows the current approach forecast, the best for 2016, 2017 and average over both. The current approach captures none of the patterns and seems consistently too high. The difference in accuracy between the individual and average best occurs in 2017 where the most accurate forecast is consistently lower than the forecasts that results from the on average better method. By averaging performance over the years we sacrifice forecast accuracy in one year for more consistent performance.

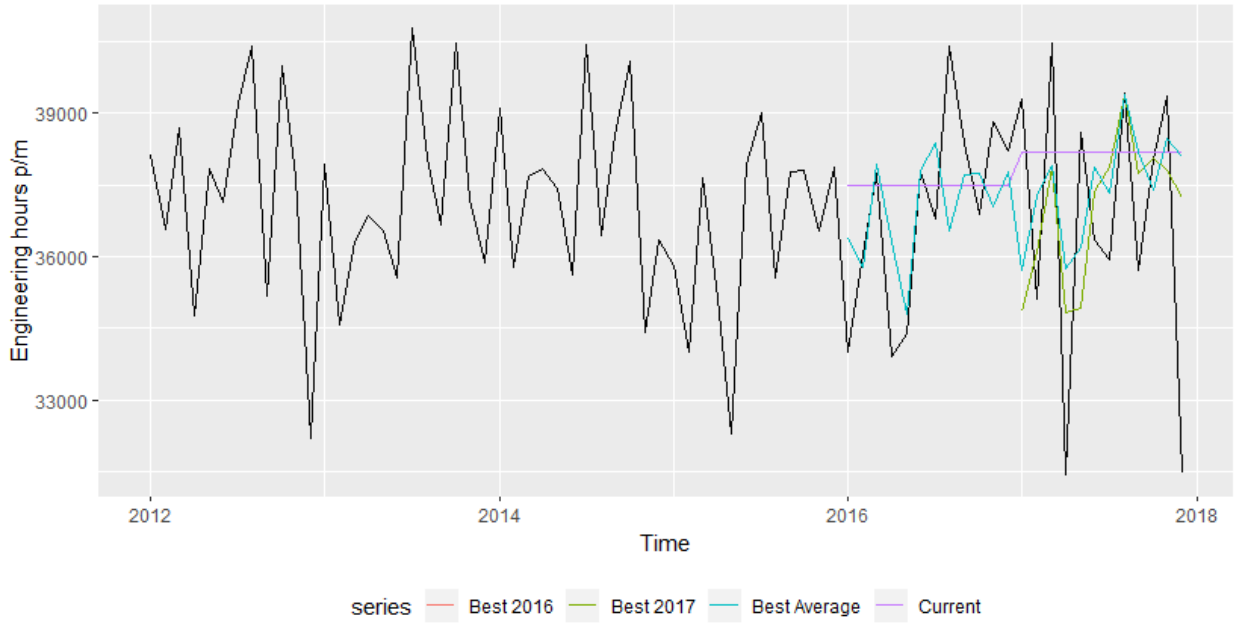


Figure 5.1 Forecasts of Total demand over 2016, 2017, the average best and the current method

We can see similar effects in Figure 5.2 which shows the demand of a task from group G5 and Figure 5.3 which show the demand of KLMVOH787ROUMO a node from the bottom level time series and the example used in defining the demand structure in Section 4.3.2. The best combinations for these series were Tbats + Croston + ETS + Theta (BCETf) and seasonal naive + mean + theta (NMTf) respectively.

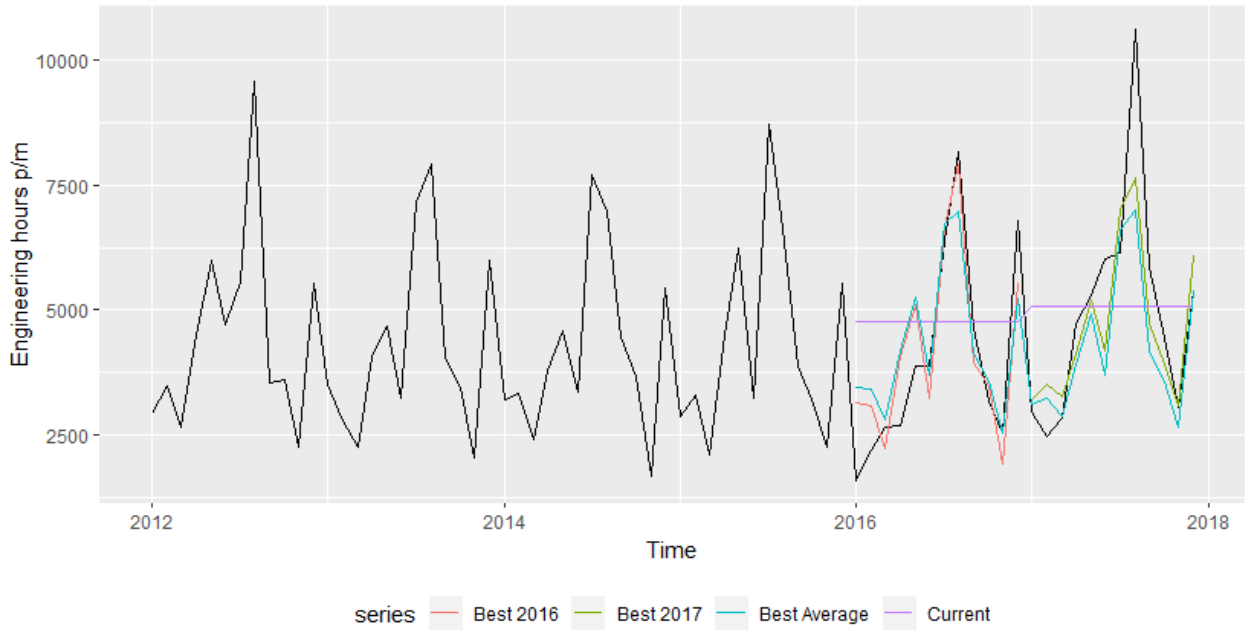


Figure 5.2 Forecasts of VA task demand over 2016, 2017, the average best and the current method

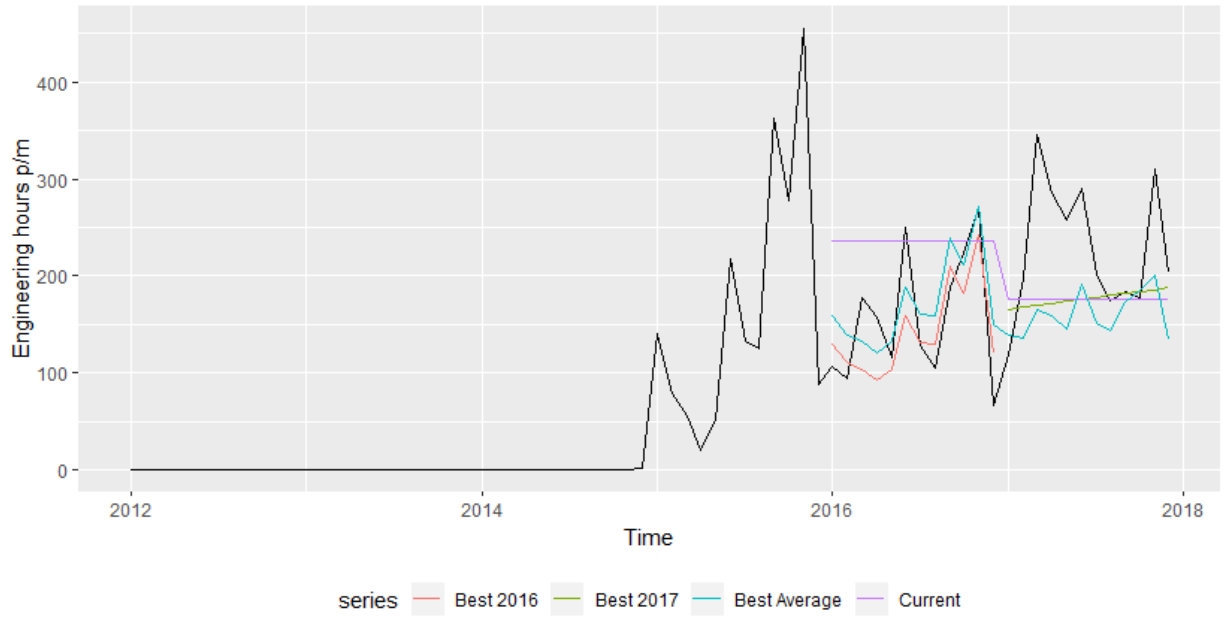


Figure 5.3 Forecasts of KLMVOH787ROUMO demand over 2016, 2017, the average best and the current method

Identifying series that require judgemental input

Taking a closer look at the average MASE results we see that they range from 0,53 to 2,63 for the proposed method. Following the definition of the MASE a score of 1 would mean comparable accuracy as the one step ahead in sample naive forecast. Given that our forecasts span 12 months we can expect to deviate from this benchmark, forecasts get more inaccurate with further horizons. But the range implies that performance varies for certain aggregations of the data with G2 and G9 experiencing the highest values. Closer inspection of the groups point to the same cause, one of the underlying nodes experiences such bad forecasts that it skews the entire average. The node involved in G9 is actually a part of the demand in G2 and responsible for the out of the ordinary behaviour, Figure 5.4 shows the demand and the forecast. What we can see in the figure is that demand suddenly explodes at the start of 2016, behaves somewhat erratically before falling again in 2017. Time series models learn from past behaviour but nearly all the behaviour of this demand occurs in the years we were forecasting. No long term patterns could therefore be detected resulting in very high MASE values. This could be regarded as an outlier not suitable to forecast with the proposed methods. We discuss the topic of outlier further in Appendix L and why we have chosen not to include them in the presented averages. This example still fell within the defined thresholds and even though it skews the overall performance it also serves as an example. In this case the change in demand behaviour occurred due to a new customer contracting work to engineering, not forecastable by data but possibly so by experts.

What the example shows is how bad performance of the quantitative methods (resulting in a high MASE) serve as an indicator for out of the ordinary series. Any MASE upward of single numbers is overtly suspicious and should trigger the feedback loop proposed in Section 4.1, prompting necessary judgmental input to explain the behaviour, adjusting the model if possible or adjust the resulting forecast. In the case of Figure 5.4 the forecast for 2016 could be adjusted to reflect the expected hours of work for that customer, resulting in more accurate results. See Appendix L for additional examples and the number of nodes removed per group.

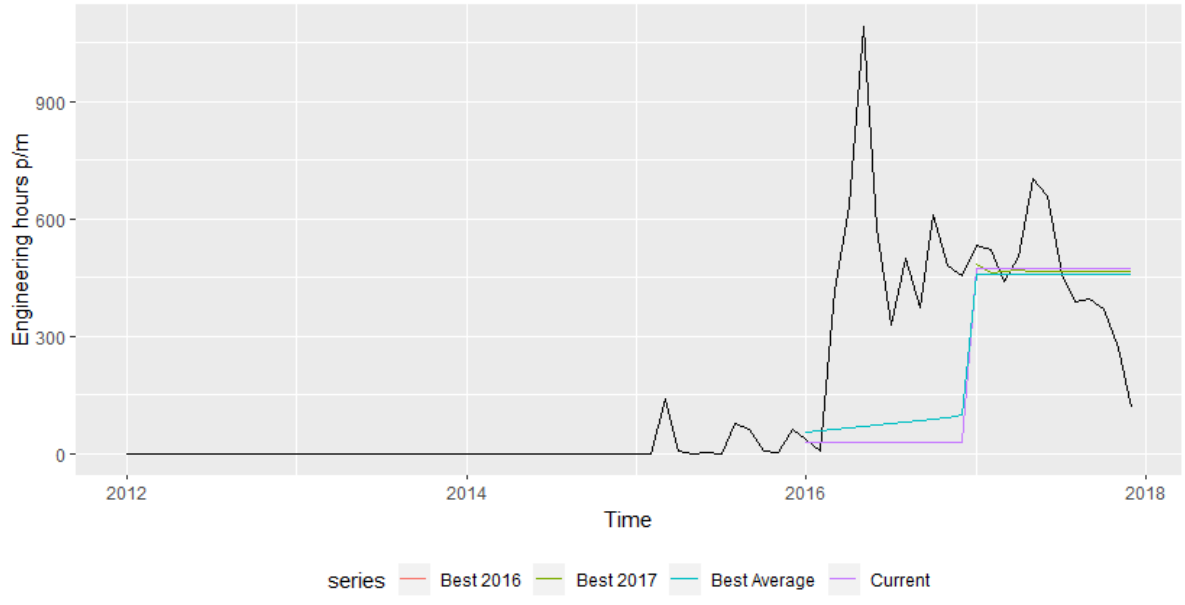


Figure 5.4 Bad forecasts in G9 for the best forecasts of 2016, and 2017, the average best and the current method

Conclusion

Regardless of including or excluding the outlier nodes, performance of the current method was consistently worse. We can conclude from Table 5.1 that the current method is not suitable for forecasting demand which clearly has more intricate behaviour than a simple average can capture. This leads us to an early conclusion that the proposed method is preferable over the current one if forecast accuracy is desired. We cannot yet conclude whether the extensive selection of models we included are actually of use, perhaps we could do with a smaller selection while producing similar results. Section 5.1.2 compares the results against a fixed combination of 4 models to see the effect of a smaller selection.

5.1.2 Benchmark combination

To assess whether the inclusion of ten models and all their possible combinations is actually beneficial we compare the results against a benchmark combination of four different models, Arima, ETS, Tbats and seasonal naive. We have chosen them to represent different capabilities of the forecasting models. Versatile and commonly used in Arima and ETS, simple and intuitive through seasonal naive and complex with Tbats. We have discussed the reason for their inclusion in more detail in Section 4.4.1. We compare the results achieved over forecasting 2017 with the best forecast and the current method. This puts the emphasis on performance over the most current results. Additionally, because the model is fixed in its combination there is no need to see whether results stay consistent like we did in the previous section. Table 5.2 shows the resulting average MASE per group compared with the best forecast of 2017 and that of the current method.

Table 5.2 Benchmark MASE comparison

Group	2017 best	2017 benchmark	2017 current
Total	1,45	1,47	1,56
G1	0,96	1,05	1,03
G2	0,78	1,06	1,1
G3	0,53	0,68	0,606
G4	1,05	1,69	1,78
G5	0,77	1,13	1,05
G6	0,46	0,61	0,541
G7	0,80	1,18	1,08
G8	1,12	1,72	1,81
G9	0,94	1,47	1,43
G10	0,91	1,30	1,3
G11	1,64	2,08	1,97
G12	0,50	0,72	0,68
G13	0,61	0,96	0,896
G14	0,68	1,05	1,1
BTS	0,88	1,40	1,33

What we conclude from the results is that the combination of the four models is not able to accurately capture the behaviour of demand over the different groups. On many occasions it is outperformed by the current approach which is a simple method that does not model any patterns. We illustrate the results of the benchmark methods compared to the current approach and the best models of 2017 with three figures. What we can observe is that the benchmark combination does extract relevant patterns from the data but the combination does not reach either the correct intensity of the pattern or applies it at a wrong level. Figure 5.5 shows the total demand with the best, current and benchmark forecast. The benchmark forecast shows patterns similar to that of the best forecast but at a lower intensity. Figure 5.6 shows the demand of 777 and again we can see suitable behaviour of the forecast but this time at an overall too high level. Finally Figure 5.7 shows demand of KLMVOH787ROUMO where we see that the benchmark forecast has picked up a pattern but misses the level of the data. Additionally, the pattern might be wrong as the best forecast from all combinations was a trended straight-line.

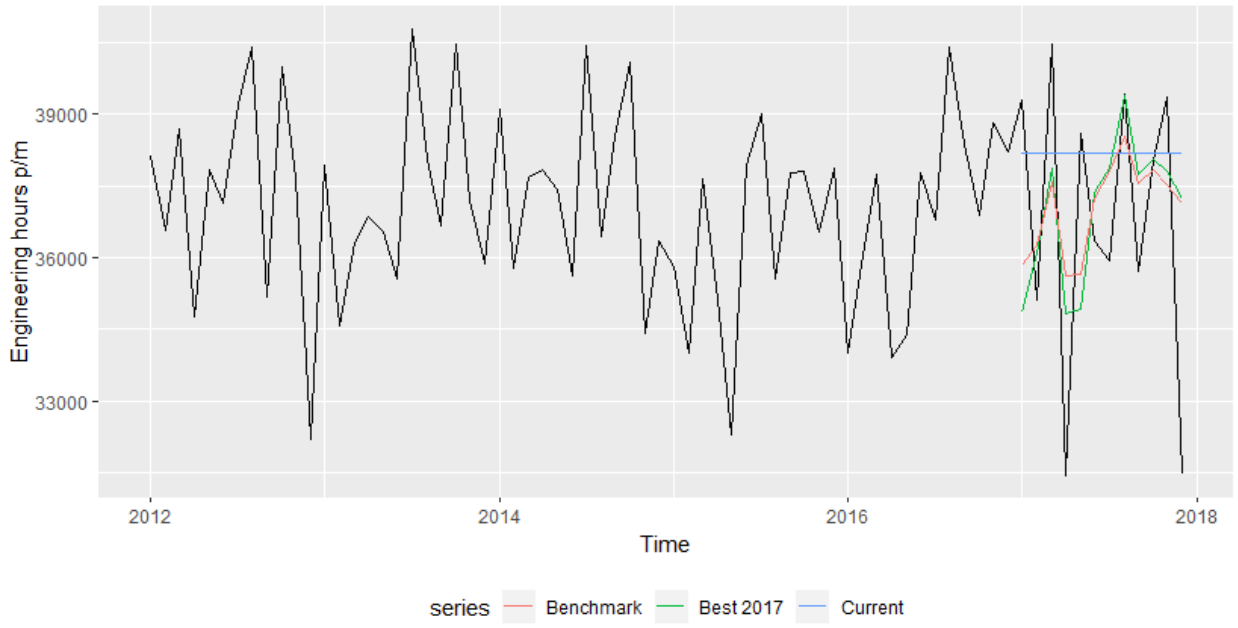


Figure 5.5 Forecasts for total demand over 2017 from the best, current method and benchmark combination

This indicates that the data, and its behaviour, are indeed as diverse as we first saw in Section 2.3 and must require different models to accurately capture the diversity. In order to determine which models add the most predictive power we can evaluate the effect of excluding a certain model on the average performance, Section 5.2 discusses this topic further. For now we can conclude that not pre-emptively excluding models in either Section 3.4 and 4.4.1 was a valid choice as the overall results are better because of the variety of models included.

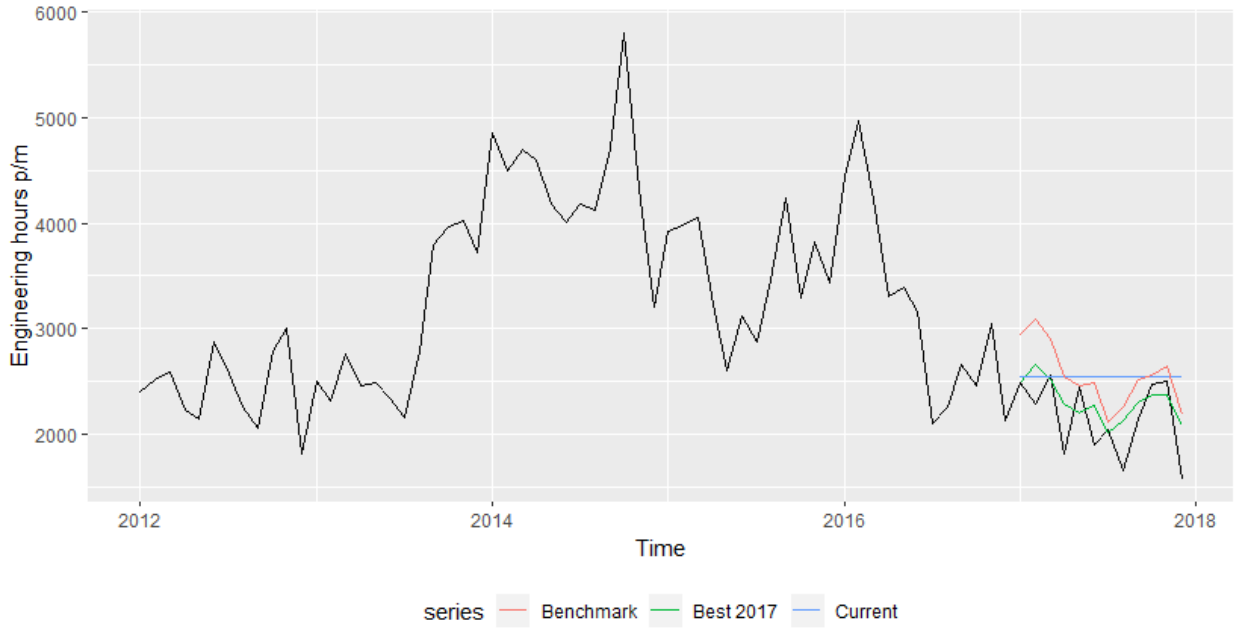


Figure 5.6 777 forecasts benchmark vs best over 2017 and the current method

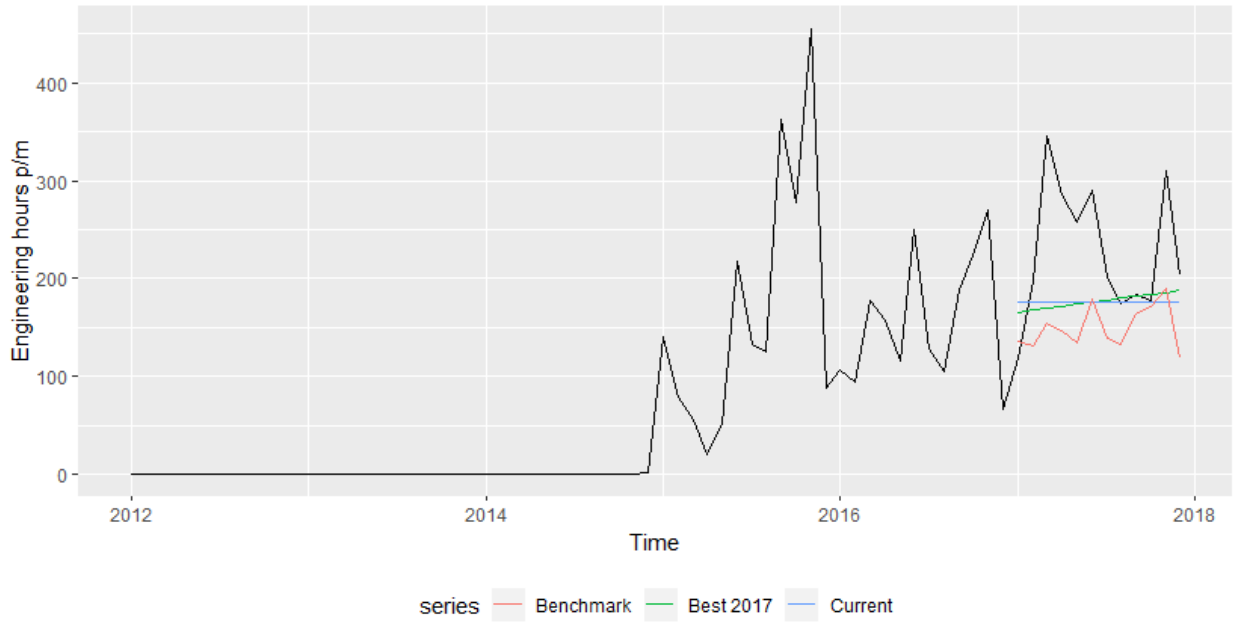


Figure 5.7 KLMVOH787ROUMO forecast benchmark vs best over 2017 and the current method

5.1.3 Reconciliation

Until now we have found that our produced forecasts are significantly better than those of the current method and that this can attributed to the inclusion of multiple models. Each node was forecast as accurately as possible with a combination of the applied base methods. By producing individual forecasts for the nodes we have introduced incoherence between the levels of the demand structure. Sections 3.6.2, 4.3.2 and 4.6 address the issue of forecast reconciliation, i.e. aligning all the individuals' nodes of the demand structure so they aggregate to, and coincide with higher levels and their forecasts in the demand structure. Applying reconciliation to all nodes provides us with a new forecast which we can evaluate on its accuracy. Table 5.3 and Table 5.4 show comparisons for the average MASE per group. Table 5.3 provides a comparison between MASE forecast accuracy of 2017, that of the best forecast, the reconciled forecast and of the current method. Table 5.4 provides a comparison over both years and of the average.

Table 5.3 MASE comparison of reconciled vs best and current in 2017

Group	Best17	Rec17	Current17
Total	1,45	1,44	1,56
G1	0,96	1,09	1,03
G2	0,78	0,86	1,10
G3	0,53	0,54	0,61
G4	1,05	0,98	1,78
G5	0,77	1,02	1,05
G6	0,46	0,66	0,54
G7	0,80	2,24	1,08
G8	1,12	1,00	1,81
G9	0,94	2,06	1,43
G10	0,91	1,37	1,30
G11	1,64	4,65	1,97
G12	0,50	0,56	0,68
G13	0,61	2,39	0,90
G14	0,68	2,40	1,10
BTS	0,88	5,24	1,33

The first thing that becomes apparent is that overall accuracy drastically decreases after reconciliation. Nearly every group loses accuracy in the process and the error increases with lower level series. Some decrease in accuracy was expected but this result is overly negative, the current forecasting approach appears much more accurate and because it forecasts averages it is also reconciled. This appears to demerit reconciliation, but why does it perform so poorly?

Table 5.4 MASE comparison reconciled forecast

Group	Rec16	Current16	*Diff*	Mase17	Current17	*Diff*	Avg	CurrentAvg	*diff*
Total	0,79	1,02	-0,23	1,44	1,56	-0,12	1,12	1,29	-0,17
G1	2,31	3,09	-0,78	1,09	1,03	0,06	1,70	2,06	-0,36
G2	3,45	4,22	-0,77	0,86	1,10	-0,24	2,15	2,66	-0,51
G3	0,50	0,64	-0,15	0,54	0,61	-0,06	0,52	0,62	-0,11
G4	0,65	0,82	-0,17	0,98	1,78	-0,81	0,81	1,30	-0,49
G5	1,20	1,03	0,17	1,02	1,05	-0,03	1,11	1,04	0,07
G6	1,15	1,25	-0,10	0,66	0,54	0,12	0,91	0,90	0,01
G7	2,82	1,67	1,15	2,24	1,08	1,16	2,53	1,37	1,16
G8	1,60	2,22	-0,62	1,00	1,81	-0,81	1,30	2,02	-0,72
G9	2,65	1,17	1,48	2,06	1,43	0,63	2,36	1,30	1,06
G10	1,64	1,26	0,38	1,37	1,30	0,07	1,50	1,28	0,22
G11	3,44	1,10	2,34	4,65	1,97	2,68	4,05	1,54	2,51
G12	0,80	1,11	-0,31	0,56	0,68	-0,12	0,68	0,90	-0,22
G13	3,13	1,56	1,57	2,39	0,90	1,49	2,76	1,23	1,53
G14	2,64	1,32	1,32	2,40	1,10	1,30	2,52	1,21	1,31
BTS	4,94	1,18	3,76	5,24	1,33	3,91	5,09	1,26	3,83

Inspecting the lower groups makes clear why reconciliation decreases overall performance. In theory borrowing information from all levels is beneficial, but this advantage breaks down in series that have many 0 value observations. Reconciliation 'projects' patterns prevalent at higher levels to series that do not exhibit such patterns normally, mainly the series discussed in Section 4.3.3. As a result forecasts that were near faultless are adjusted to be more errors prone. Figure 5.8 shows an example of a 'perfect' forecast wrongly adjusted to reflect patterns from higher levels. Both the current method and best forecasts correctly predict that there is no further demand to be expected. Yet reconciliation detects the patterns in the higher levels and adjusts the zero forecasts accordingly, this introduces two big flaws. First, the accuracy as measured by MASE spikes, the average in-sample naive error is low and suddenly large errors appear causing large deviations. Secondly, reconciliation does not care that our demand is represented in hours and therefore has a strict border at 0.

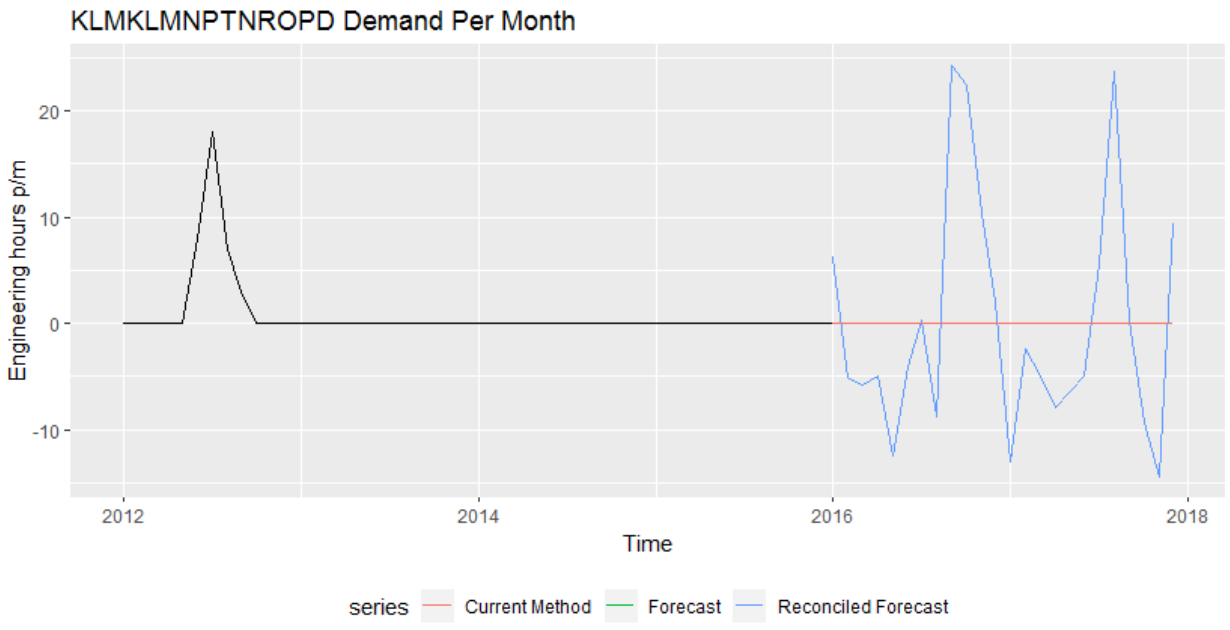


Figure 5.8 Reconciled zero forecast

Reconciliation performance is thus clearly biased by series that do not need adjustments. In order to get a better picture of the actual effect of using information from the different levels we can try to improve reconciliation by iteratively coercing the method to set problematic series to 0. Appendix M further discusses the approach where we iterate to increase performance. This introduces a bias to the forecast but is justifiable given the underlying data behaviour. Table 5.5 presents the results which are much more in line with what we would expect, overall there is a decline compared to the best forecast but the current method is generally outperformed. Figure 5.9 shows the effect on the example presented in Figure 5.8 for 2017. The most notable changes in the forecast is its scale and adherence to the positivity constraint.

Table 5.5 Improved MASE comparison

Group	Best Mase	Rec mase	Current
Total	1,45	1,44	1,56
G1	0,96	1,12	1,03
G2	0,78	0,85	1,10
G3	0,53	0,50	0,61
G4	1,05	0,92	1,78
G5	0,77	1,01	1,05
G6	0,46	0,46	0,54
G7	0,80	1,20	1,08
G8	1,12	0,88	1,81
G9	0,94	1,16	1,43
G10	0,91	1,20	1,30
G11	1,64	2,47	1,97
G12	0,50	0,50	0,68
G13	0,61	0,96	0,90
G14	0,68	0,98	1,10
BTS	0,88	1,81	1,33

The errors introduced on the lower levels imply that reducing the number of demand characteristics, and thus the number of groups, might provide higher accuracy by avoiding part of the outlier series. This assumption was tested in Appendix N where we find that reconciliation performance is slightly worse when a demand characteristic is removed. Thus we can conclude that forecasts of lower demand groups suffer in quality and do not benefit from reconciliation but they still add value by increasing reconciliation accuracy for higher level demand groups.

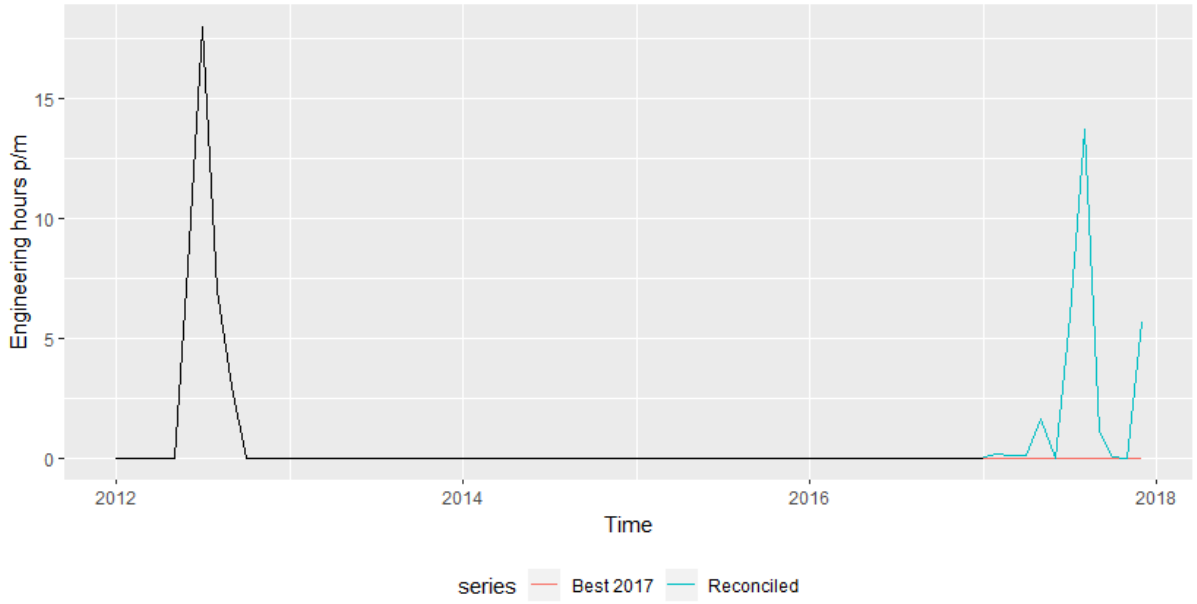


Figure 5.9 Iterated reconciliation forecast

With the accuracy results at an acceptable level we illustrate the effects of reconciliation with the same examples as in Section 5.1.2. Figure 5.10 and Figure 5.11 show the forecast for total and 777 demand respectively. No large changes to the forecast are apparent. Figure 5.12 present the lower scale forecasts for KLMVOH787ROUMO demand, in line with the precious observations the changes are more apparent on lower levels of demand where they have bigger impact on accuracy as well.

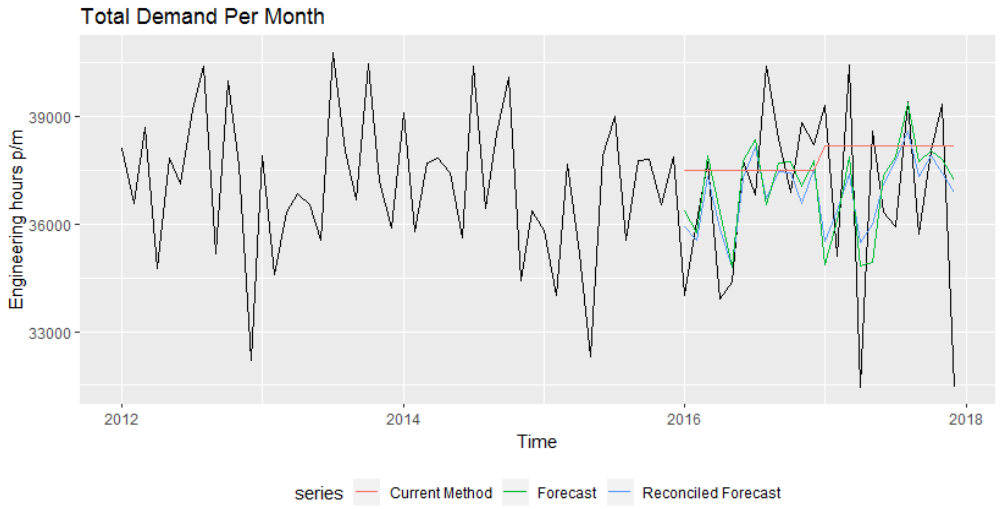


Figure 5.10 Total demand forecasts reconciled vs best and current method

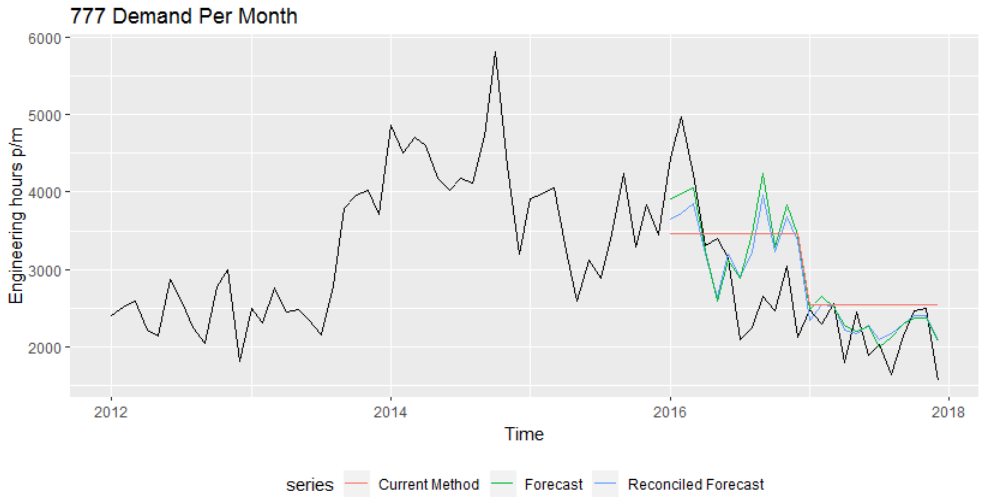


Figure 5.11 777 demand forecasts reconciled vs best and current method

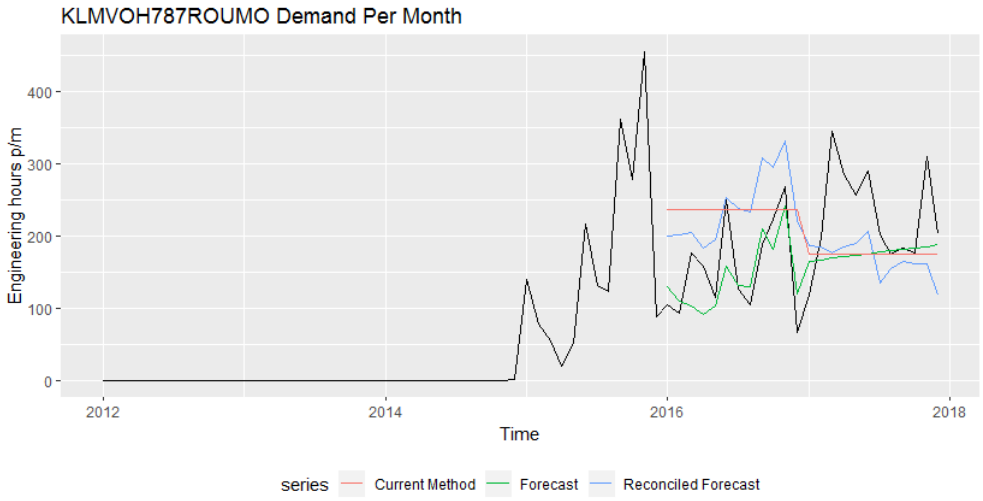


Figure 5.12 KLMVOH787ROUMO demand forecasts reconciled vs best and current method

From the results we conclude that reconciliation is useful when applied diligently but it might have more potential applied to continuous demand with fewer 0 observations. We find that some forecasts, mainly the higher levels, are improved while others, especially the lower groups, deteriorate. Nevertheless, the coherency introduced by reconciliation makes it a useful tool for making aligned business decisions, be it budgeting for the coming year or planning for the coming months. Finally, the current method is generally outperformed which further merits the use of more complex methods.

5.1.4 Conclusion

For each of the comparisons we are able to come to clear conclusion about the effectiveness of forecasting with multiple models. In each comparison the best results were always provided by the proposed approach. This makes sense as our forecast is able to choose the best method from a large selection of forecast combinations allowing for a better fit. Comparing the results to the current method we find that on average our approach reduces the MASE with 25% or 20%, respectively when regarding the years separately or averaged.

A similar conclusion holds when assessing the comparison against the benchmark combination. We chose four models and compared their average forecast against the best results and the current method. The benchmark performed poorly at a level comparable to that of the current approach. The fact that a combination of four different models is outperformed by the current approach confirms that different data needs different models to produce accurate results. By choosing a fixed combination, we force all four models to participate and supply information even though they might be poorly matched to the data.

Reconciliation proved to be a complicated affair to produce accurate results. Its capability of taking information from all levels and project it throughout the demand structure is beneficial in a theoretical sense. However, the abundant 0 observations at the lower demand levels made the approach less effective. By iterating the method we were able to improve the results to a point where the current method is generally outperformed. What we can safely conclude is that the current approach to forecasting demand is insufficient for capturing the underlying patterns. The best forecasts drastically improve accuracy over the groups and if reconciled we can still achieve an increase over current accuracy while aligning the forecasts.

5.2 Sensitivity to inclusion of methods

Comparing the results of the best forecast against that of the benchmark combination confirmed that a larger selection of models is beneficial for forecast accuracy. Including ten different models and all of their combinations creates questions about which method adds the most predictive power to the approach. In order to assess their individual influence we produce the minimal and mean MASE for each group given that a single method is removed from the pool.

5.2.1 Effect on minimal MASE

Table 5.6 shows the effects of excluding a method on the average minimal MASE per group. For each node we disregard all forecast combinations that include the relevant model, determine the minimal MASE and determine the average per groups. If a model is suited for nodes in a certain group we expect the best performance to drop, i.e. result in a higher MASE. Table 5.6 highlights the biggest increases in MASE to see the effects per method. What becomes clear is that the naive method

consistently adds the most predictive power to combinations for the lower level groups which makes sense. These are characterised by demand that is increasingly difficult to model due to higher variation and 0 value observations. In those situations the best forecast is often simple and naïve predicts optimally for a random walk, so effective under large variation. The higher level groups where demand is more consistent benefit from the more complex models. From Table 5.6 we conclude that ETS, temporal hierarchical forecasting, Croston's methods and the theta approach do not significantly add to overall best forecast performance and might be excluded from the model pool without impacting the results. But comparing minimal values only indicates the effect on best performance and does not tell us whether a model, on average, adds predictive power to the forecasts, this is discussed in Section 5.2.2.

*Table 5.6 Method exclusion effect on minimal MASE, the highest increase in **bold**, 2nd highest increase underlined*

Group	N	S	M	E	A	H	B	C	I	Tf	Best
Total	1,45	**1,46**	1,45	1,45	1,45	1,45	1,45	1,45	1,45	1,45	1,45
G1	0,96	0,96	0,98	0,96	0,97	0,97	**0,99**	0,96	0,96	0,96	0,96
G2	0,79	0,79	0,78	0,78	0,80	0,78	0,78	0,78	**0,83**	0,78	0,78
G3	0,53	0,53	**0,55**	0,53	0,53	0,53	0,53	0,53	0,53	0,53	0,53
G4	**1,54**	1,05	1,05	1,05	1,05	1,05	1,09	1,05	1,05	1,05	1,05
G5	**0,81**	0,79	0,78	0,77	0,78	0,77	0,77	0,77	0,79	0,77	0,77
G6	0,46	0,47	**0,48**	0,46	0,46	0,46	0,47	0,46	0,46	0,46	0,46
G7	0,78	0,78	0,77	0,76	0,76	0,76	0,77	0,76	**0,78**	0,76	0,76
G8	1,12	1,12	1,12	1,13	1,23	1,12	1,13	1,12	**1,33**	1,12	1,12
G9	**1,08**	1,03	1,02	1,02	1,03	1,02	1,04	1,02	1,04	1,02	1,02
G10	**0,95**	0,92	0,92	0,91	0,91	0,91	0,92	0,91	0,92	0,91	0,91
G11	**1,94**	1,90	1,90	1,89	1,90	1,89	1,89	1,89	1,89	1,89	1,88
G12	**0,53**	0,50	0,51	0,50	0,52	0,50	0,51	0,50	0,51	0,50	0,50
G13	**0,71**	0,69	0,68	0,67	0,67	0,68	0,67	0,67	0,68	0,67	0,67
G14	**0,74**	0,72	0,71	0,70	0,71	0,70	0,70	0,71	0,71	0,70	0,70
BTS	**1,09**	1,06	1,06	1,05	1,05	1,05	1,05	1,05	1,05	1,05	1,05
Avg	**0,97**	0,92	0,92	0,92	0,93	0,91	0,92	0,91	0,94	0,91	0,91

5.2.2 Effect on mean MASE

Table 5.7 shows the effect of excluding a model on the average MASE per group. To determine the effects we disregard any forecasts made with the relevant model, determine the average MASE over the remaining forecasts and average over the nodes in the groups. In the case of a model that mostly adds predictive power to the forecasts we expect the mean to increase, conversely removing a bad model is expected to see the mean MASE decrease. If the mean MASE is unchanged the model contributes evenly to good and bad forecasts. Table 5.7 highlights the **increased** and decreased values, to see the general effects per model. As in Section 5.2.1 we can observe the naive method to consistently add predictive power and we can see that all models add predictive power to at least one group. The mean approach seems to reduce performance consistently, this is a further indication the current approach to forecasting is unsuitable as a mean forecast is currently used (see Section 2.4). Apart from the naïve method we can also observe that the more complex methods (ETS, Arima and TBATS) have a more positive effect on the result than the simple approaches.

*Table 5.7 Method exclusion effect on mean MASE, the highest value in **bold**, lowest underlined*

Group	N	S	M	E	A	H	B	C	I	Tf	Best
Total	1,51	**1,52**	1,51	1,51	1,51	1,51	**1,52**	1,51	1,51	**1,52**	1,51
G1	1,17	**1,19**	**1,19**	**1,19**	**1,22**	**1,20**	**1,23**	1,17	1,12	**1,20**	1,18
G2	**1,08**	1,04	1,05	**1,08**	**1,09**	1,07	**1,09**	1,06	1,07	**1,08**	1,07
G3	**0,72**	**0,72**	0,71	**0,73**	**0,73**	**0,73**	**0,73**	**0,72**	0,69	**0,73**	0,72
G4	**1,79**	1,68	1,67	**1,72**	**1,72**	1,70	**1,73**	1,69	**1,72**	1,69	1,71
G5	**1,11**	1,10	1,10	1,09	1,10	1,10	1,09	1,10	**1,11**	1,10	1,10
G6	**0,69**	**0,68**	0,67	**0,69**	**0,68**	**0,69**	**0,69**	0,66	0,65	**0,69**	0,68
G7	**1,10**	1,09	1,09	1,08	**1,10**	1,09	1,09	1,09	1,09	**1,10**	1,09
G8	**1,83**	1,79	1,78	1,79	**1,84**	1,78	1,80	1,79	1,80	**1,81**	1,80
G9	**1,53**	1,51	**1,53**	1,51	1,47	1,48	**1,53**	**1,55**	**1,55**	1,50	1,51
G10	**1,29**	1,27	1,27	**1,28**	1,27	**1,28**	**1,28**	**1,28**	**1,28**	1,27	1,27
G11	**2,39**	2,36	2,36	**2,37**	**2,37**	**2,37**	2,36	**2,37**	2,36	**2,38**	2,36
G12	**0,76**	0,74	0,73	**0,76**	**0,76**	**0,75**	**0,76**	0,74	0,74	**0,76**	0,75
G13	**1,05**	1,04	1,03	1,04	1,04	1,04	1,03	1,04	1,04	1,04	1,04
G14	**1,07**	1,06	1,05	1,05	1,06	1,06	1,05	1,06	1,06	1,06	1,06
BTS	**1,56**	1,53	1,53	**1,54**	**1,54**	**1,54**	1,53	**1,54**	**1,54**	**1,54**	1,53
Avg	**1,29**	1,27	1,27	**1,28**	**1,28**	1,27	**1,28**	1,27	1,27	**1,28**	1,27

5.2.3 Conclusion

Based on the results in this section we conclude that the inclusion of different models is beneficial. Best and average results vary depending on the inclusion of a suitable method. Confirming the principles of forecast combination that make it a more accurate approach. Selecting one model to fit on data introduces a level of uncertainty, applying multiple and combining the results mitigates this.

We can further state that the choice to include all models was valid, since each has added at least some predictive power. Naive shows to influence nearly all of the results positively, implying that some part of the variation is best explained by a simple process and therefore benefits from including a naive forecast. Its function might also be to act a normalizer in the forecast combination, it forces more complex forecasts to average with the last observed value tethering it to the most recent outcome. A more detailed exploration of the individual model effects is necessary to infer causation of model performance.

Based on the results some models could probably be excluded, for instance Croston's method does not or only marginally improves its forecasts. But in order to truly claim that a model adds no predictive power requires a deeper analysis. The order of removing models will also affect the accuracy of the remaining combinations, thus to fully see the effect of a model every possible order of removal should be analysed. As this step only serves to optimize the runtime of the model (no predictive power can be gained by removing a model) we defer the deeper analysis to a future research/implementation.

5.3 Total forecast accuracy and organizational impact

Each node in the demand structure is part of a group that represents total demand divided into smaller parts. This means that the sum of all forecasts in one group is also a forecast for total demand. As different levels of aggregation present different behaviour it is meaningful to see whether group forecasts can provide accurate predictions of total demand. Additionally, this allows us to compare performance in absolute numbers and explore the organizational impact of producing more accurate forecasts.

5.3.1 Accuracy of total forecast

To determine forecasting accuracy for total demand all node forecasts per group were added together. Depending on the group this ranges from summing 2 to 713 forecasts, see Section 4.3.2 for details on the demand structure. Table 5.8 shows the resulting MASE for forecasting total demand in 2017. We can observe differences in accuracy ranging from 1,43 to 2,25. Total demand accuracy appears to decrease with groups that consist of more nodes. This makes sense as more nodes imply a higher granularity of data, leading to more difficult to forecast series and thus higher errors. On the other hand, the highest overall accuracy is achieved by groups other than the Total node. This appears to be a confirmation of the benefits implied by using different levels of data aggregation. Splitting the data, until a certain level, separates causes of variation making them easier to detect and as a result higher overall accuracy can be achieved. While we can observe the differences in accuracy between the groups it is hard to conclude on the actual effects that has on the forecast. We cannot conclude what impact a 10% increase in MASE means without looking at actual values, which we do in Section 5.3.2.

Table 5.8 Total demand forecast accuracy 2017

Group	MASE
Total	1,45
G1	1,43
G2	1,43
G3	1,52
G4	1,44
G5	1,61
G6	1,51
G7	1,53
G8	1,45
G9	1,47
G10	1,6
G11	1,94
G12	1,57
G13	1,78
G14	1,55
BTS	2,25
Current	1,56
Reconciled	1,44

5.3.2 Organizational impact

Applying the proposed forecasting method improves forecast accuracy as demonstrated throughout Chapter 5. In order to compare the different forecasts we chose to apply the MASE as this allowed us to compare forecasts from different scales. The downside of the MASE is that it does not directly translate values relatable for the organization, apart from indicating that more accurate forecasts are possible. By summing all the forecasts to the total level the differences in scale have been removed, enabling comparisons on the actual values. Table 5.9 shows a comparison between group, current and reconciled forecasts. The sum of the forecast of 2017, the mean and the difference with the actual values are also included.

Table 5.9 Accuracy comparison with absolute values

Group	MASE	Sum	Mean	sum - actual	mean- actual
Total	1,45	444035	37003	2895	241
G1	1,43	441933	36828	793	66
G2	1,43	438855	36571	-2286	-190
G3	1,52	448532	37378	7391	616
G4	1,44	449175	37431	8035	670
G5	1,61	433066	36089	-8074	-673
G6	1,51	444209	37017	3068	256
G7	1,53	432227	36019	-8913	-743
G8	1,45	440286	36691	-854	-71
G9	1,47	432392	36033	-8748	-729
G10	1,6	429823	35819	-11317	-943
G11	1,94	419944	34995	-21197	-1766
G12	1,57	439867	36656	-1273	-106
G13	1,78	429918	35826	-11223	-935
G14	1,55	433474	36123	-7667	-639
BTS	2,25	412168	34347	-28972	-2414
Current	1,56	458255	38188	17115	1426
Reconciliation	1,44	442829	36902	1689	141
Actual	0	441140	36762	0	0

The first thing we learn from Table 5.9 is that a higher MASE coincides with worse absolute performance, Figure 5.13 clearly shows the relation. This validates that the MASE was a good fit for measuring the forecast accuracy and optimized forecasts to minimize overall errors. Results for 2016 are similar to that over 2017 with absolute performance decreasing with higher MASE. The 2016 results are presented in Appendix O.

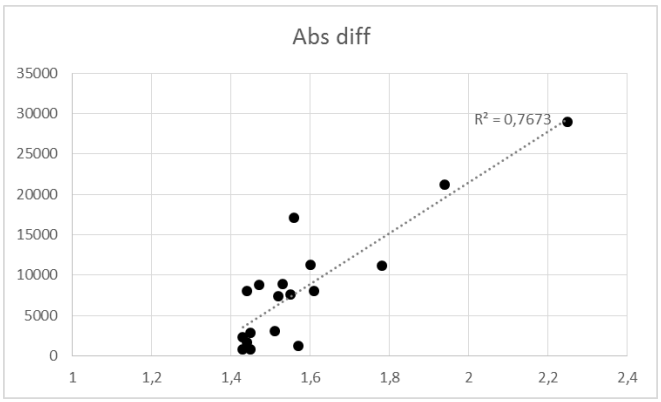


Figure 5.13 MASE vs absolute difference with actual values

The current method overestimates the total amount of work in 2017 with 4% (17115/458255≈4%) roughly equal to 9,3 FTE (assuming 1840 working hours p/y, 17115/1840=9,3 FTE). The most accurate forecast for total demand is the G1 forecast where an overestimation of 793 hours occurs, or 0,4 FTE. The difference in accuracy leads to nearly 9 FTE allocated to demand where it does not appear necessary from the data. The downside of using the most accurate forecast is that division of hours over the different groups is not possible due to incoherence between the levels. This is where reconciliation has a big advantage, using information from all the levels it produced a more accurate forecast than most of the groups and it overestimates demand with only 0,9 FTE (less than 0,5% difference with the actual value) over 2017, a difference with the current method of 8,4 FTE. Over 2016 the best and reconciled forecast produce errors of the same magnitude within 1% of the actual yearly outcome verifying the consistency of the approach.

Another conclusion is that the granularity of a group negatively influences its overall accuracy. The BTS forecast is the most granular and it underestimates the required FTEs by nearly 16 in 2017 and 23 in 2016. This is a further confirms our conclusion in Section 5.1.3 that granular series are difficult to accurately forecast for their own level but should nevertheless be included for the information they provide to the higher groups through reconciliation.

Using more accurate means of forecasting demand would allow more flexible and accurate decision making regarding capacity and lead to less loss of opportunity. The proposed forecast implies that in 2017 8,4 FTE was be redundantly forecast by the current approach. In a knowledge heavy and project oriented department like engineering this could lead to opportunities not taken as capacity might have been deemed insufficient. A more accurate forecasts allows the identification of flexibility in capacity and empowers decisions on where to apply it to its full potential.

5.3.3 Conclusion

The organization would do well to adopt more accurate forecasting techniques. Comparing more complex forecasting and the current approach on absolute deviation from the actual values has shown that a significant difference in performance exists. The equivalent of 8,4 fulltime employees in 2017 is overestimated by the current method in comparison to a more complex approach which consistently forecasts within <1% error margins of yearly totals. Knowing accurately beforehand how capacity might change is an asset in exerting more control over both the long and short term. Potentially saving 8,4 FTE for the coming year creates possibilities in cost saving but more importantly in more flexibly utilizing the available knowledge and skills. Projects or other business opportunities previously deemed unrealistic due to capacity restrictions might become viable as no budget increases are necessary to free up space for such endeavours.

5.4 Conclusion

Throughout Chapter 5 we have focussed on assessing the performance of the forecasting approach proposed in Chapter 4. In doing so we answered research question 7: 'How does the proposed method perform?' Through several comparisons we have found that the proposed forecasting method consistently produces forecasts significantly more accurate than the current approach.

In Section 5.1 we performed several comparisons to assess overall performance. Comparing the results of the most accurate forecasts for 2016, 2017, and the average best over both against the current method clearly illustrated the potential for improvement. Over the individual years an improvement, in terms of MASE, of 25% was achieved and 20% over the combined period. A benchmark forecast of a four model combination failed to outperform either the proposed method or the current approach, indicating that a large part of the predictive power is caused by including sufficiently different models and allowing for a flexible selection of their combinations.

Reconciliation was shown to decrease overall accuracy compared to the best possible forecasts. Accounting for demand with mostly 0 values increased overall performance to a point where the current method is outperformed on a majority of the groups. While the forecasts are not as accurate as the individual best, the major advantage is alignment over the different groups. The reconciled forecast provides consistent information regardless of the group, be it the BTS, G9 or Total demand. Regardless of the bad performance on low level groups we found that overall accuracy of reconciliation decreases if lower level groups are removed. Confirming theory that all aggregations contain information useful for increasing forecast accuracy.

Section 5.2 looked at the individual contribution of a model to the predictive power of combinations. We learned that the naive approach consistently improves the forecasts in which it is included, especially for lower level groups. Its assumption that the most recent observation will persist, positively influences its combinations. This could imply that part of the variation in demand is random for which a naïve forecast is suited. Some models do not seem to actively contribute to the best results, while others work consistently well on higher level groups. We conclude that a pool of models is beneficial as each introduces additional information to the forecasts potentially increasing the predictive power. However, models can be redundant or miss-specified for certain groups and might be removed without consequences.

Section 5.3 aggregated the different group forecasts to forecasts of total demand. This allowed us to see on a high level which group forecasts are closest to actual demand. We found that over the entirety of 2017 the current approach overestimates demand with 4% equivalent to 9,3 full time employees. The best performing and reconciled forecasts overestimate demand with 0,2% and 0,4%, or 0,43 and 0,92 FTE, respectively. Applying a more accurate forecasting approach would free that capacity in a planning phase and enable its use for more pressing matters. In general confirming that a more accurate forecast allows for better and more flexible control over capacity.

We can conclude that the proposed forecasting method is significantly better than the current approach and that forecast reconciliation, while resulting in lower accuracy, has benefits for managing capacity. The potential impact on the organization is increased flexibility in deploying capacity or potential costs reductions. This aligns with our goal of producing more accurate forecasts for both tactical (budgeting) and operational (planning) purposes as determined in Chapter 1. We have focussed on confirming and validating that the approach works, further organizational impacts are expected after implementation.

6 Conclusions, discussion and recommendations

In this final chapter we conclude the research. Section 6.1 presents our conclusions and answers out main research question. Section 6.2 contains a discussion of our research and results. Finally, Section 6.3 presents recommendations for implementation within KLM and for further research.

6.1 Conclusions

This research focussed on answering the question:

"How to accurately forecast uncertain demand for KLM engineering with quantitative and qualitative methods?"

After testing and validating the approach in Chapter 5 we conclude that the proposed forecasting method in Chapter 4 is a, more than, sufficient framework for forecasting demand for KLM engineering. We present key points to consider when forecasting engineering demand:

1. Engineering demand consists of several distinct characteristics that subset demand into parts with unique behaviour, all of which present different information.
2. No single method is able to extract all the information presented by the different demand subsets. Multiple methods are therefore necessary to forecast different subsets.
3. Combining forecasts from different methods combines their information and is able to more accurately forecast than any single method. A simple average of different forecasts is good enough to produce combinations that outperform their parts.
4. To achieve accuracy in forecasting for all subsets a level of automation is required as the number of different subsets makes manual forecasting infeasible. Otherwise a critical selection of which subsets to forecast needs to take place while identifying their most appropriate methods.
5. Reconciliation improves performance for higher level demand but is less effective on more granular groups.
6. Judgemental forecasts should only be applied when there is good reason not to trust the statistical forecast. In such a case clear rules and guidelines should be followed. Bad performance of the statistical forecast can serve as a reliable indicator for human input.
7. Periodic evaluation of the method and the results is necessary to ensure continued performance of forecasting.

These points follow from our findings in literature, their application, and validation. By applying ten different models and testing all their combinations we were able to significantly increase forecast accuracy. On average the current method is outperformed by 25% over a single year and with 20% when the best forecast performance over two years is regarded. We can therefore conclude that the application of multiple models and their combinations is an effective strategy.

The choice for a large and flexible pool of forecast combinations was validated by comparison against a static four model combination. Not only did the static combination fail to perform close to the best combinations it was often outperformed by the current method. Confirming that a flexible selection of models, while requiring automation, drastically decreases the risk of selecting unfit models for a demand subset. This was further validated by assessing the individual impact of the applied models where we found that each adds predictive power to at least one of the demand groups.

Reconciliation had mixed results. It was negatively affected by series with many 0 observations leading to bad performance. Iterating the method lead to improved results and the conclusion that it decreases overall performance but still outperforms the current method while using more information from the data while providing the opportunity for aligned decisions on both tactical, budgeting and operational, planning, decisions.

The increase in forecast accuracy by the proposed method was validated by evaluating the potential impact on the organization. We found that the current method over estimates demand with 9,3 FTE in 2017 where the best forecast and the reconciled forecast overestimate demand with 0,4 and 0,9 FTE respectively. A difference of at least 8,4 FTE over one year potentially negatively impacting decisions where additional capacity was required. Over both 2016 and 2017 the proposed forecasting method was able to forecast the yearly total to within 1% of the actual values.

6.2 Discussion

Throughout researching and testing the approach we have found our results to be valid by outperforming the current approach. Our success in this seemed self-explanatory beforehand but turned out to be harder than expected on some occasions. This is illustrated by the benchmark combination that actually performs wore than the current approach. In hindsight this is explained by the all-round capability of a mean forecast. Regardless of patterns or behaviour the mean is a fairly stable characteristic in most demand series, it is not prone to radically change except for some outlier occurrences. As such it will usually provide a reasonable forecast, often better than miss specified complex models as was probably the case in the benchmark combination.

The same holds for the first results of reconciliation, based on literature the expectation was to see an equal or increase of accuracy over the demand structure. This turned out not to be the case to due demand with a high degree of 0 observations. While we had expected these series to have some influence we had not foreseen that reconciliation could not properly handle them at all.

Forecasts can only get as accurate as the available data. In forecasting engineering demand we were dependent on using a proxy for demand through booked man-hours, which are prone to different kinds of error. A follow up research/project that looks into properly cleaning, maintaining and future proofing this data is an interesting topic for increasing accuracy and leveraging more information from the data. Additionally, the assumptions that validate it as a proxy for demand should be tested to see whether it is in fact suitable.

One of the largest limitations was the available knowledge and hardware. Our initial knowledge about forecasting was limited and the department had no expertise on the topic triggering a search for suitable methods and resulting in an over eager framework. Applying a large collection of models on a large number of series resulted in a lot of data. Requiring suitable hardware and data knowledge to properly manage. The hardware met its limitations on certain calculations and data, disqualifying potentially interesting improvements to the forecasting framework. Additionally, our limited knowledge prevented us from accurately gauging the required effort for certain extensions of the model. Implementing external variables looked very promising but proved to taxing to apply.

The complexity also presents difficulties for implementation, the proposed framework is not so straightforward to directly implement in the organization. KLM needs to define what kind of accuracy they require and should accordingly invest in implementing a fitting forecasting approach.

6.3 Recommendations and further research

6.3.1 Recommendations

Based on our results we recommend the following:

- KLM needs to decide what kind of forecasting accuracy they require for what subsets of demand and implement a suitable forecasting method accordingly.
- Implement the presented framework for a comprehensive approach capable of handling all different subsets.
- Clean the source data and analyse its function as a proxy for demand.
- Define rules and guidelines on how to apply judgemental forecasting and investigate its effectiveness.
- Integrate forecasting within the available IT-systems instead of the responsibility of employees that need to use Excel

We recommend to put the highest priority on implementing a more advanced statistical forecast combined with the need for clean and up to date data. Then the way judgmental forecasting is applied should be formalized in order to leverage as much information from both the statistical forecasts and expert knowledge. Finally, KLM should work toward an integrated forecasting method that can automatically predict future demand while using relevant information from external sources to improve accuracy but this should be a goal on the horizon.

6.3.2 Suggestions for future research

Due to restriction in expertise, time and computing power several potentially interesting topics have not been pursued. Here are our recommendations for further research:

Bootstrap forecasting and intervals

As describes in Section 3.5.3 bootstrapping provides a way to simulate slightly different forecast outcomes by randomly changing the remainders of the data under consideration. This might provide more average forecast on average and also allows for prediction intervals based on actual data instead of a theoretical model. The costs incurred are a high computation time as each demand subset would need to be simulated at least a 100 times to provide accurate results.

Regression with external variables

Due to the complexity and time consuming effort of collecting testing and predicting suitable external variables this has been omitted from this research. It would be very interesting to see how much of the variant in demand could be explained by such a model. Some preliminary ideas for external variables:

- GDP, general economic prosperity has effect on the intensity of commercial aircraft operation. This could influence maintenance and from there engineering
- Utilization of an aircraft, both flight hours per day and seats: potentially correlated with GDP but perhaps also easier and more accurate to come by for KLM.
- Amount of certain aircraft types and their age
- Dummy variables for major projects and customer work. If it is known that certain jobs are only for a certain period dummy variables can help filter out this variation from the demand data.

Working day correction

As booked working hours are used as a proxy for demand the amount of working days per month can have a significant effect on the demand per month. This realization cam to late ion our research to implement but initial findings, presented in Appendix P, are very promising. It appears we could be able to significantly increase forecast accuracy through a straightforward data transformation that should not be too hard to implement in the suggested framework. From all suggested further research this appears to provide the best balance between required effort and probable results.

Forecast reconciliation with non-negativity restriction

The method we used to reconcile the forecast leads to negative demand on series that experience low level demand. These results are not realistic, setting the negative values to zero creates a bias in the forecast which is not advisable. A more general approach using a constrained least square regression to produce the reconciled summing matrix for the grouped time series could force non-negativity. This will result in non-biased coherent forecasts but will incur a cost in computation time and complexity as it does not optimally use the unique structure of a hierarchical time-series.

Apply machine learning algorithms

In order to reduce our scope, computational time and increase the interpretability of our method we opted not to use machine learning methods. These models are becoming increasingly powerful and could add predictive power to our approach. A good extension of the research would be applying neural networks and genetic algorithms to see if they can outperform the more classical time-series approach.

7 References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 716-723.

Armstrong, J. S. (2001). *Principles of Forecasting.* Bostan, MA: Kluwer Academic.

Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting, 16*(4), 521-530.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*.

Bates, J. M., & Granger, C. W. (1969). The Combination of Forecasts. *Operational Research Society Vol 20, No. 4*, 451-468.

Bergmeir, C., Hyndman, R. J., & M.Benítez, J. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting, 32*, 303-312.

Box, G. E., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Box, G. E., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* Wiley.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time series and Forecasting, Third edition.* Springer International Publishing. doi:10.1007/978-3-319-29854-2

Carson, R. T., Cenesizoglu, T., & Parker, R. (2011). Forecasting (aggregate) demand for US commercial air travel. *International Journal of Forecasting*, 923-941.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wange, W. (2016). The forecast combination puzzle: A simple theoretical. *International Journal of Forecasting Vol 32*, 754-762.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 3-33.

Croston, J. D. (1972). Forecasting and Stock Control for Intermittent Demands. *Operational Research Quarterly, 23*(3), 289-303.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of business & economic statistics*, 253-263.

Duffuaa, S. O., & Raouf, A. (2015). Maintenance Strategic and Capacity Planning. In *Planning and Control of Maintenance Systems.* Springer, Cham.

Fildes, R., & Goodwin, P. (2007). Against Your Better Judgment? Can Improve Their Use of Management Judgment in. *Interfaces 37(6)*, 570-576. doi:https://doi.org/10.1287/inte.1070.0309

Franses, P. H. (2016). A note on the Mean Absolute Scaled Error. *International Journal of Forecasting, 32*, 20-22.

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research Volume: 86, issue: 8*, 1678-1685.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting, 12*, 37-48. doi:10.1002/for.3980120104

Hyndman, R. J. (2014). *Measuring forecast accuracy.* Retrieved from https://robjhyndman.com/papers/forecast-accuracy.pdf

Hyndman, R. J., & Athanasopoulos, G. (2018, 06). *Forecasting: Principles and Practice* (2nd ed.). Monash University, Australia: OTexts. Retrieved June 2018, from https://otexts.org/fpp2/

Hyndman, R. J., & Billah, B. (2003). Unmasking the Theta method. *International Journal of Forecasting, 19*(2), 287-290.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software, 3*, 1-22.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679-688.

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

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., . . . Yasmeen, F. (2018). *Forecasting functions for time series and linear models*, R package version 8.4. Retrieved from http://pkg.robjhyndman.com/forecast

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting Volume 18, Issue 13*, 439-454.

Hyndman, R. J., Lee, A., Wang, E., & Wickramasuriya, S. (2018). hts: Hierarchical and Grouped Time Series. Retrieved from https://CRAN.R-project.org/package=hts

Hyndman, R., & Kourentzes, N. (2018). thief: Temporal HIErarchical Forecasting. Retrieved from http://pkg.robjhyndman.com/thief

Hyndman, R., Lee, A., Wang, E., & Shanika. (2018). hts: Hierarchical and Grouped Time. Retrieved from https://CRAN.R-project.org/package=hts

Kolassa, S. (2008). Can we obtain valid benchmarks from published surveys of forecast accuracy? *Foresight: The International Journal of Applied Forecasting*(11), 6-14.

Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* New York: Springer-Verlag.

Kourentzes, N. (2014). On intermittent demand model optimisation and selection. *International Journal of Production economics*, 180-190.

Kourentzes, N., & Petropoulos, F. (2016). tsintermittent: Intermittent Time Series Forecasting. Retrieved from https://CRAN.R-project.org/package=tsintermittent

Kourentzes, N., Barrow, D., & Petropoulos, F. (2018). Another look at forecast selection and combination: Evidence from from forecast pooling. *International Journal of Production Economics*, 1-10.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*(1-3), 159-178.

Lawrence, M., Edmundson, R., & O'Connor, M. (1986). Accuracy of Combining Judgemental and Statistical Forecasts. . *Management Science,*, 1521-1532.

Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 1513-1527.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International Journal of Forecasting, 16*(4), 451-476.

Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, 914-924.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research, 268*, 545-554.

Silver, E. A., Pyke, D. F., & Thomas, D. J. (2017). *Inventory and Production management in Supply chains.* Boca Raton: Taylor & Francis Group.

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting, 21*(2), 303-314.

Teunter, R. H., & Duncan, L. (2009). Forecasting Intermittent Demand: A Comparative Study. *The Journal of the Operational Research Society*, 321-329.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2018). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*.

Appendix A. **Product codes engineering tasks**

ID	Category	Code	Task
1	AMP Management	MR	AMP continuous evaluation and amendment
1	AMP Management	MP	Aircraft phase-in / phase-out in AMP
1	AMP Management	MS	AMP special programs and other special operator requests
1	AMP Management	RP	Repair assesment
1	AMP Management	VS	Preparation of work instructions and jobcards.
1	AMP Management	VM	Managing of packaging optimization
2	Fleet Performance Management	RB	Reliability monitoring and reporting
2	Fleet Performance Management	FM	Reliability Engineering
2	Fleet Performance Management	SB	Evaluation of OEM originated documents (SB's etc)
2	Fleet Performance Management	AD	Evaluation of authority originated documents (AD's etc)
3	Data Management	AU	Certificates of airworthiness / other ship documents (BvL, AD status, etc)
3	Data Management	OM	Acquisition and update of Maintenance Documentation (Operator)
3	Data Management	DB	Engineering and documentation and configuration control
4	Operator Support	SP	Specification of aircraft, engines and components
4	Operator Support	DA	Aircraft Acceptance
4	Operator Support	FI	Aircraft phase-in support
4	Operator Support	FO	Aircraft phase-out support
4	Operator Support	FS	Feasibility Studies
4	Operator Support	ME	ETOPS support
4	Operator Support	OC	Consulting
4	Operator Support	RO	Operator Representation
4	Operator Support	RT	Emergency / recovery assistance
4	Operator Support	WM	Aircraft weight reporting & monitoring
4	Operator Support	WC	Warranty / claims support
5	Production Support	PO	Support during maintenance execution (FAR 145)
5	Production Support	TF	Flight test support (troubleshooting, maintenance check)
5	Production Support	NT	NDT Inspection Planning
5	Production Support	XH	NDT Inspection During day time
5	Production Support	WW	Aircraft Weighing During daytime
6	Maintenance Package	VT	Adaption of packages to maintenance unit requirement
6	Maintenance Package	MX	One-time approval for extension of maintenance interval

ID	Category	Code	Task
7	Maintenance Consulting	MC	Maintenance consulting
7	Maintenance Consulting	ES	Parts & material standardization
7	Maintenance Consulting	ID	Suggestion box items
7	Maintenance Consulting	TQ	Acquisition & purchasing support
8	Data Management	TM	Preparation / update of manuals requested by Maintenance Unit
8	Data Management	MD	Distribution of maintenance documentation
9	Design Engineering	MO	Modifications and design certification
9	Design Engineering	RD	Repair development and certification
9	Design Engineering	LD	Livery Drawings
10	Transaction Services	KW	Laboratory Services Quality control and sample testing
10	Transaction Services	RH	Laboratory Services Research related to materials in a/c
10	Transaction Services	PE	Process Engineering
10	Transaction Services	PH	Technical Photography
12	Internal	AR	Business planning
12	Internal	LE	Management
12	Internal	FA	Financial Management
12	Internal	PP	Project Management
12	Internal	PM	Performance Management
12	Internal	PD	Process / product development / business analysis
12	Internal	DO	Document Management intern
12	Internal	QM	Quality System Management
12	Internal	AO	KEI / DOA recognition
12	Internal	CU	Knowledge Management
12	Internal	KT	Department meeting, etc
12	Internal	OR	Work council (OR, GC, etc.)
12	Internal	SE	Secretarial Activities
12	Internal	IT	IT Management
12	Internal	ET	Equipment and tooling
12	Internal	EQ	Acquisition Engineering customers
13	Absenteeism	VA	Vacation (holidays, ATV)
13	Absenteeism	AF	Illness / special leave

Appendix B. **Forecasting process cycle**

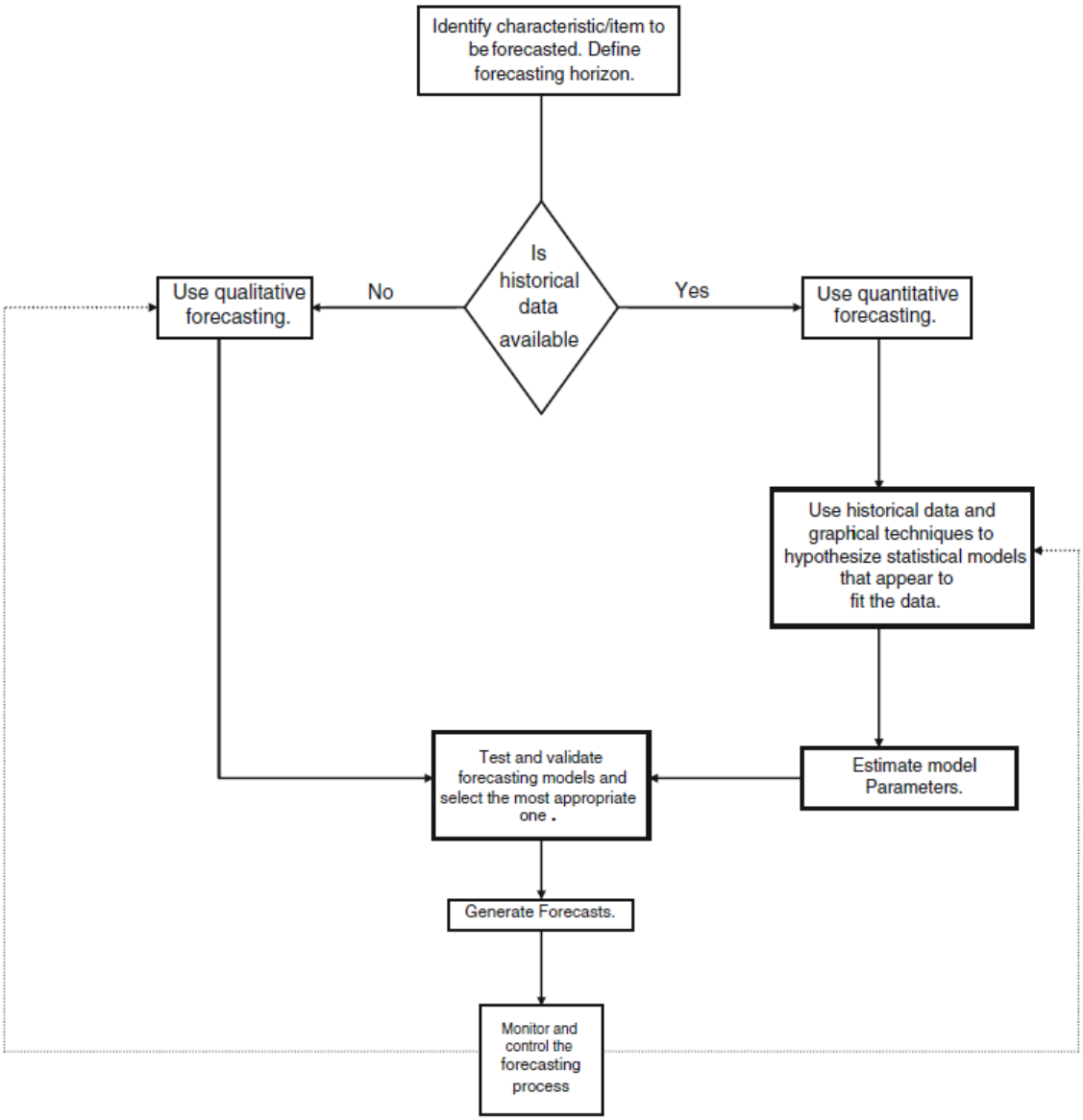


Figure 7.1 Forecasting process cycle (Duffuaa & Raouf, 2015, p. 21)

Appendix C. **Data transformation example**

Year	Apr	Aug	Dec	Feb	Jan	Jul	Jun	Mei	Mrt	Nov	Okt	Sep
2000	379.	544.	355.	286.	300.	606.	474.	468.	334.	398.	411.	467.
2001	357.	539.	370.	317.	313.	622.	523.	472.	347.	390.	402.	532.
2002	422.	575.	381.	269.	280.	584.	555.	467.	351.	354.	426.	567.
2003	361.	611.	386.	299.	295.	625.	504.	482.	373.	381.	485.	515.
2004	418.	568.	402.	291.	296.	647.	510.	453.	351.	446.	416.	602.
2005	394.	631.	408.	351.	296.	692.	566.	478.	358.	458.	457.	514.
2006	401.	642.	387.	349.	367.	706.	617.	484.	367.	467.	465.	633.
2007	483.	675.	393.	334.	314.	730.	645.	543.	379.	476.	508.	616.
2008	479.	716.	399.	363.	353.	736.	654.	591.	444.	481.	498.	610.
2009	465.	729.	397.	379.	363.	725.	651.	622.	381.	447.	542.	701.

Figure 7.2 Dataset

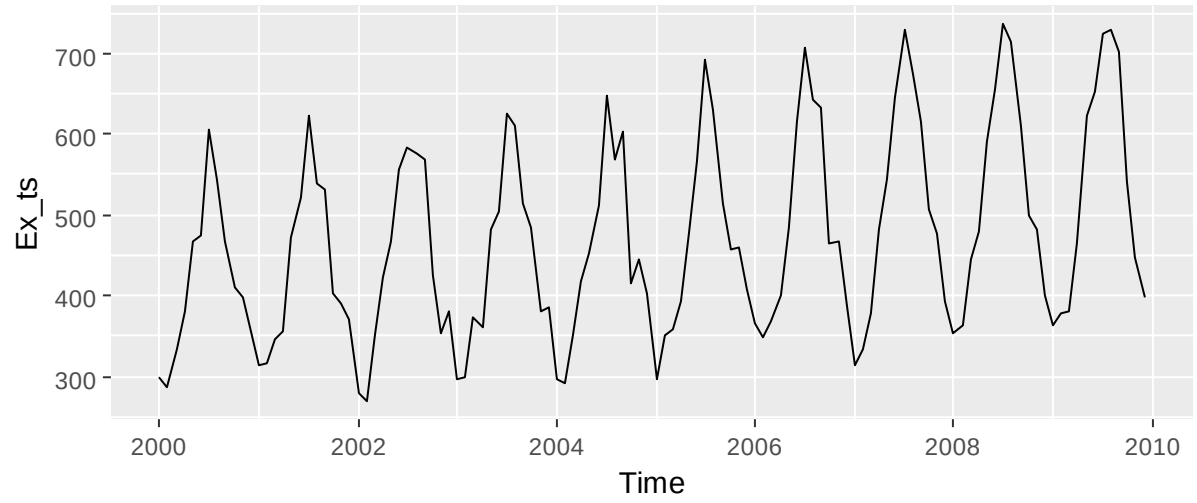


Figure 7.3 Data plotted

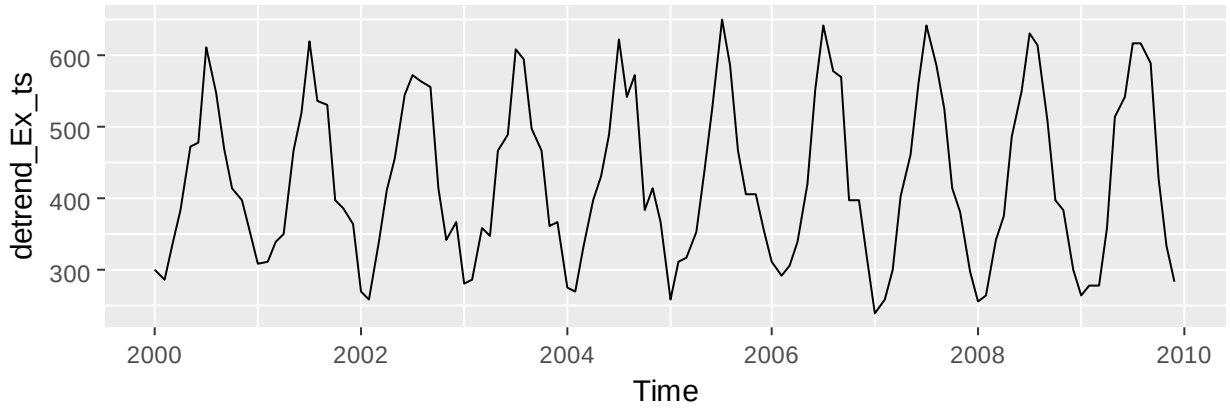


Figure 7.4 Detrended (season + remainder)

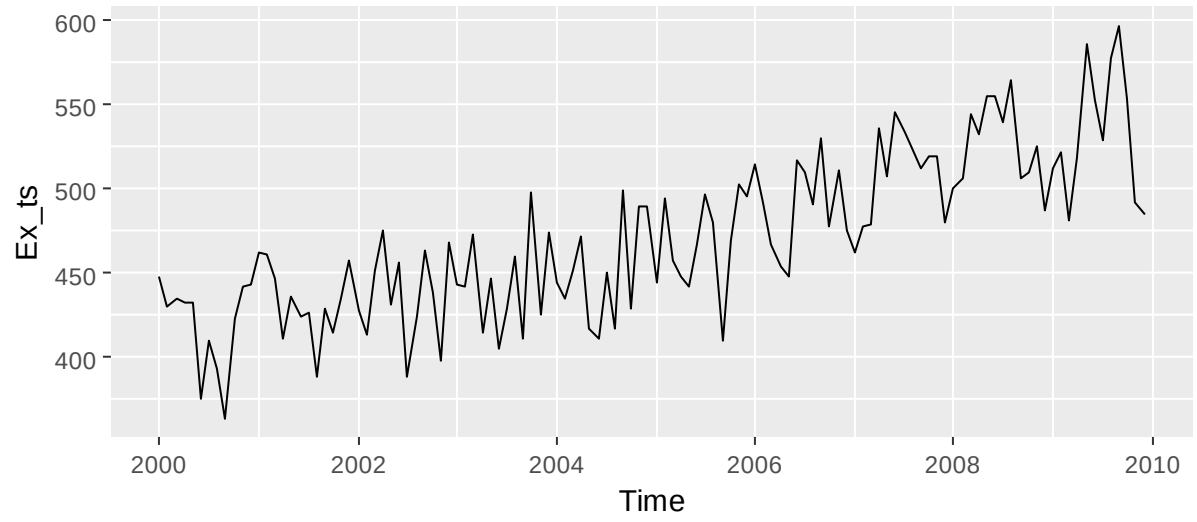


Figure 7.5 Deseasonalized (trend + remainder)

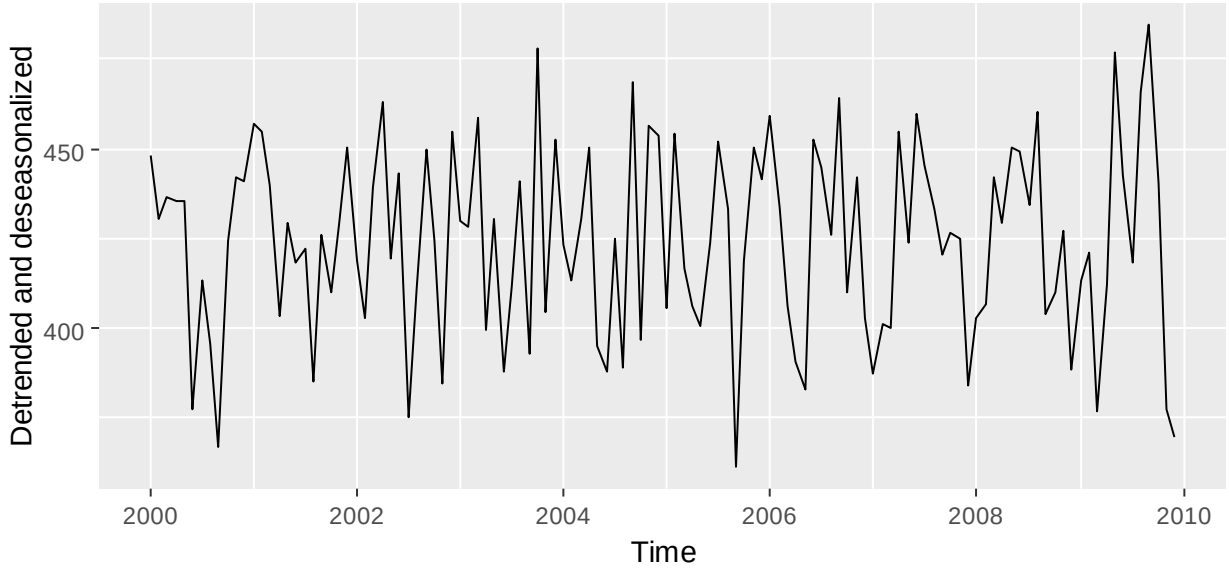


Figure 7.6 Detrended + deseasonalized

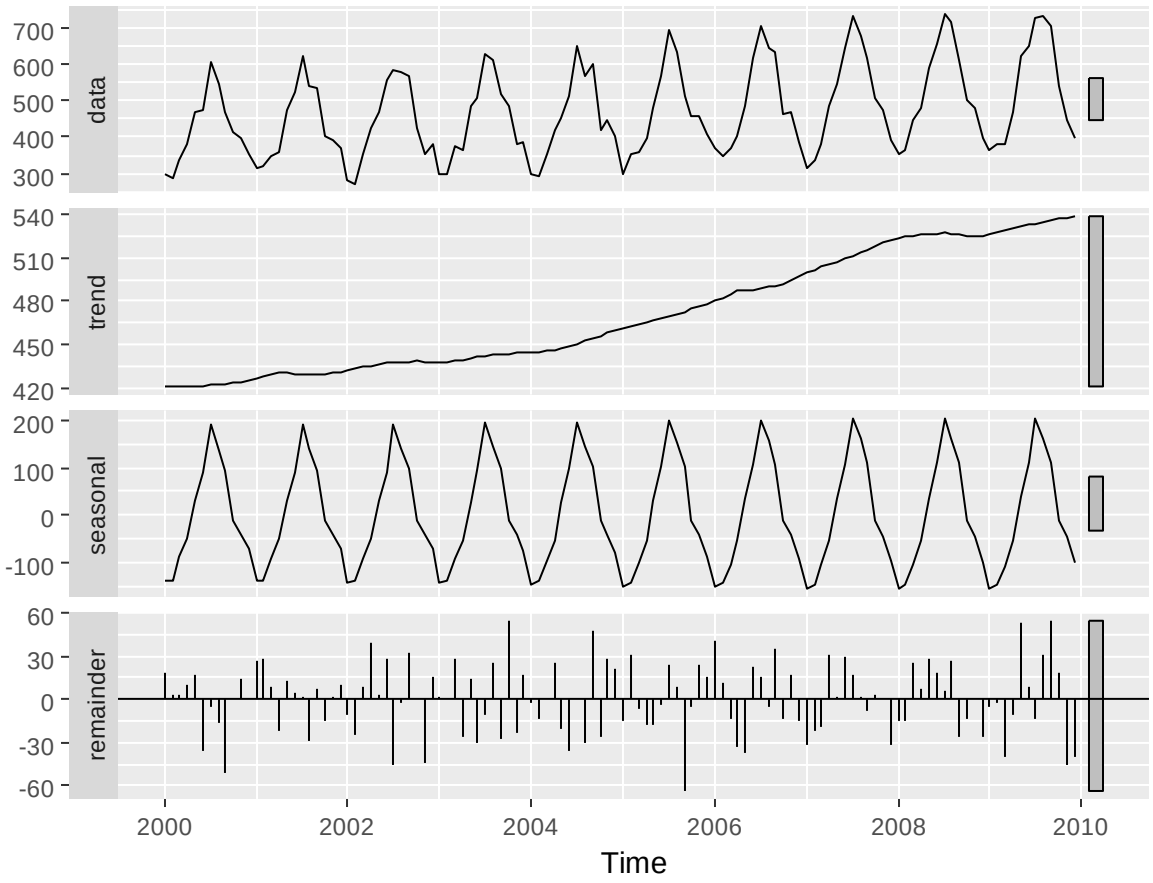


Figure 7.7 Decomposition

Appendix D. **Key principles of judgemental forecasting**

Fildes and Goodwin (2007) designed a set of principles to follow when applying judgemental forecasts. These were based on the *Principles of Forecasting Handbook* by Armstrong (2001) and subsequent research. A total of 11 principles were designed and sub divided into 3 categories. Each of the principles have been shown in research to be best practice and improve forecasting accuracy when applied diligently. Additionally to the improved accuracy the principles make sure that the process is repeatable and testable, reducing uncertainty and bias in the forecasts.

The principles

***When to Use Judgment*:**

Principle 1. Use quantitative rather than qualitative methods.

Principle 2. Limit subjective adjustments of quantitative forecasts.

Principle 3. Adjust for events expected in the future.

How to Apply Judgment

Principle 4. Ask experts to justify their forecasts in writing.

Principle 5. Use structured procedures to integrate judgmental and quantitative methods.

Principle 6. Combine forecasts from approaches that differ.

Principle 7. If combining forecasts, begin with equal weights.

How to Assess the Effectiveness of Judgment

Principle 8. Compare past performance of various forecasting methods.

Principle 9. Seek feedback about forecasts.

Principle 10. Use error measures that adjust for scale in the data.

Principle 11. Use multiple measures of forecast accuracy.

Appendix E. **State space ETS models**

State space equations for each additive error model in the classification. Multiplicative error models are obtained by replacing ε_t by $\mu_t\varepsilon_t$ in the equations

Trend component	Seasonal component: N (none)	A (additive)	M (multiplicative)
N (none)	$\mu_t = l_{t-1}$ $l_t = l_{t-1} + \alpha\varepsilon_t$	$\mu_t = l_{t-1} + s_{t-m}$ $l_t = l_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = l_{t-1}s_{t-m}$ $l_t = l_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/l_{t-1}$
A (additive)	$\mu_t = l_{t-1} + b_{t-1}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$	$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (l_{t-1} + b_{t-1})s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(l_{t-1} + b_{t-1})$
M (multiplicative)	$\mu_t = l_{t-1}b_{t-1}$ $l_t = l_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/l_{t-1}$	$\mu_t = l_{t-1}b_{t-1} + s_{t-m}$ $l_t = l_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/l_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (l_{t-1}b_{t-1})s_{t-m}$ $l_t = l_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \alpha\beta\varepsilon_t/(s_{t-m}l_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(l_{t-1}b_{t-1})$
D (damped)	$\mu_t = l_{t-1} + b_{t-1}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$	$\mu_t = l_{t-1} + b_{t-1} + s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (l_{t-1} + b_{t-1})s_{t-m}$ $l_t = l_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \alpha\beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(l_{t-1} + b_{t-1})$

Figure 7.8 State space ETS models (Hyndman, Koehler, Snyder, & Grose, 2002, p. 443)

Appendix F. **Grouped time series summing matrix**

$$\begin{bmatrix} y_t \\ y_{\mathrm{A},t} \\ y_{\mathrm{B},t} \\ y_{\mathrm{X},t} \\ y_{\mathrm{Y},t} \\ y_{\mathrm{AX},t} \\ y_{\mathrm{AY},t} \\ y_{\mathrm{BX},t} \\ y_{\mathrm{BY},t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{\mathrm{AX},t} \\ y_{\mathrm{AY},t} \\ y_{\mathrm{BX},t} \\ y_{\mathrm{BY},t} \end{bmatrix}$$

Figure 7.9 Summing formula example (Hyndman & Athanasopoulos, 2018, p. 10.2)

Appendix G. **Temporal aggregation effect on 777 MO demand**

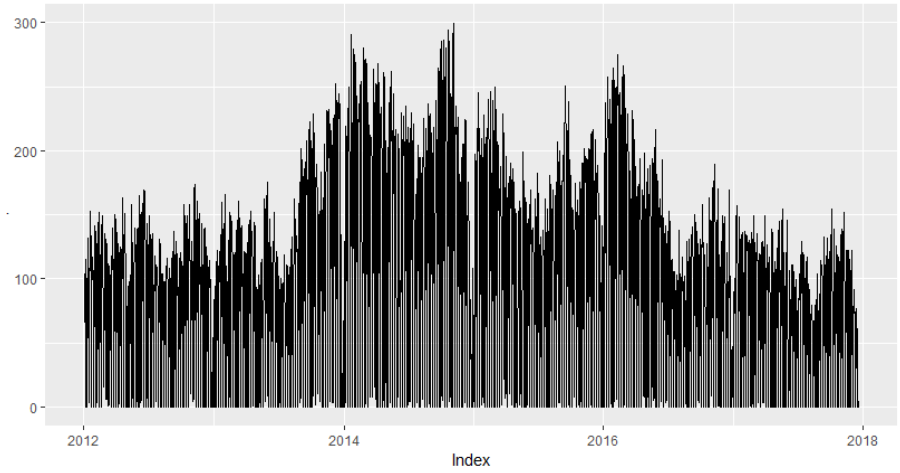


Figure 7.10 777 daily demand

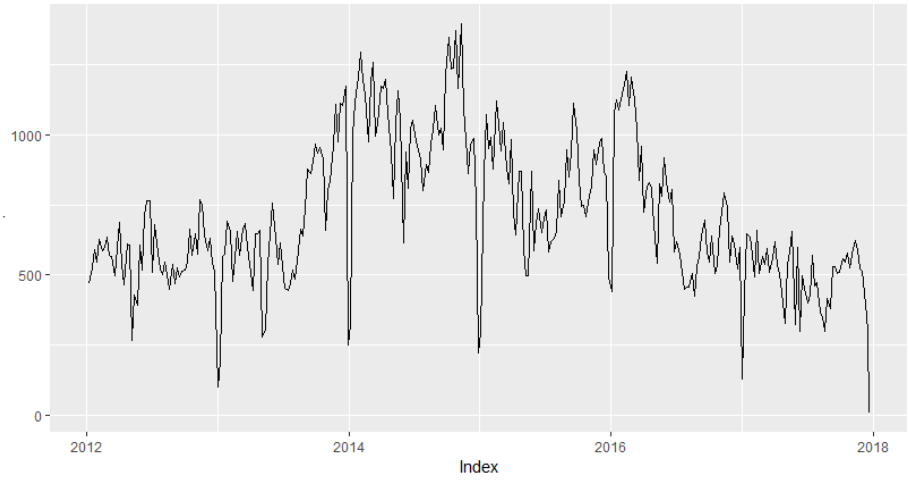


Figure 7.11 777 weekly demand

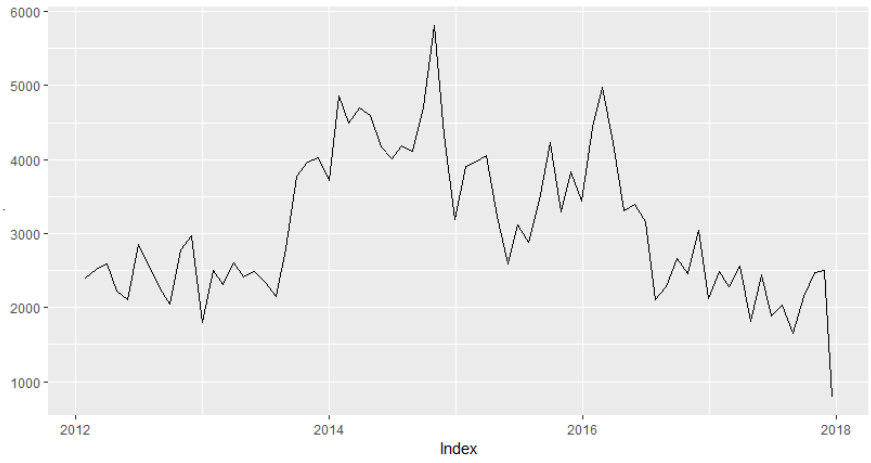


Figure 7.12 777 monthly demand

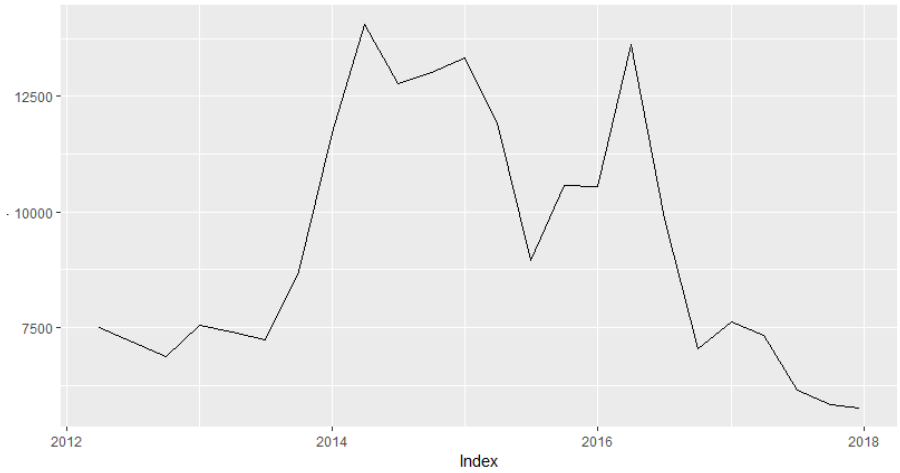


Figure 7.13 777 quarterly demand

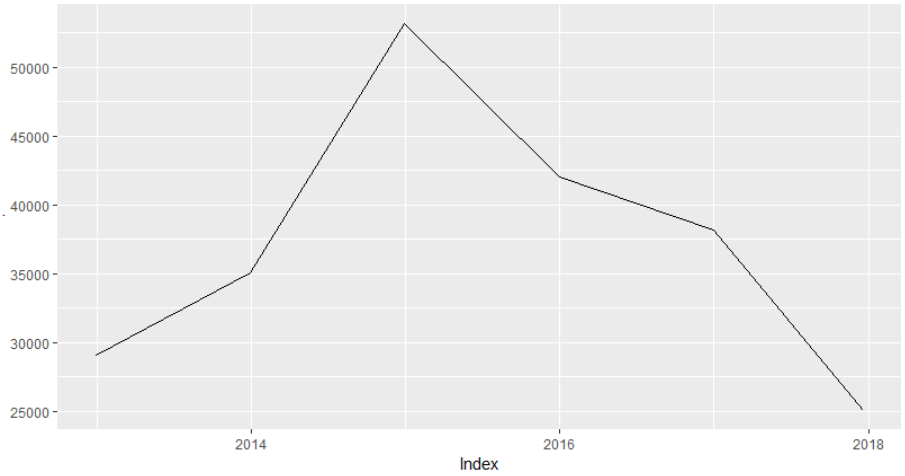


Figure 7.14 777 yearly demand

Appendix H. **Implementation in R**

As mentioned throughout Chapter 4 the implementation of the proposed model was done in R. Here we present some of the steps taken where necessary supported with the used (pseudo-)code.

A. Define all demand levels

In order to create all demand subsets of interest we need to apply the hierarchy as defined in Section 4.3. By using standard length identifiers for the characteristics of interest defined in Section 2.3 and 4.3.1 we are able to group the data from a spreadsheet layout to a grouped time series using the hts package in R (Hyndman R. J., Lee, Wang, & Wickramasuriya, 2018). To illustrate an identifier code, we refer to the example of KLMVOH787ROUMO presented in Section 4.2.3 and its in Section 4.3.2. It translation:

- KLM: The task was done for an internal (KLM) customer
- VOH: is the specific subdivision of KLM for which it was performed
- 787: The type linked to the task
- ROU: it was on a **rou**tine basis
- MO: It was a task related to **mo**dification

We can define what the groups and hierarchies by defining the number of characters per group or hierarchy as follows:

- R call: gts(y, characters=c(c(3,3),3,3,2)) where y is the spreadsheet data.

The information given to the character argument, c(c(3,3),3,3,2), coincides with the number of characters in the identifier code. First c(3,3) refers to KLMVOH which is hierarchical, it was a task for KLM specifically for VOH. Then the 3, 3 and 2 correspond to the other groups, 3 for the type: 787, 3 for (non-)routineness: ROU and 2 for the task: MO.

B. Applying the methods

With the grouped time series in R we have all the different demand subsets available. From this we construct a 'tibble' in R that contains a subset per row and the following information:

- NodeName: Which node is on this row
- Testset: the relevant test set
- Trainingset: the original data minus the test set
- Forecasts: a matrix containing all forecasts and their combinations for all steps in the forecast horizon
- Errors: the accuracy per method per forecast horizon
- Accuracy: the average accuracy over all steps of the forecast horizon
- MASE: The resulting lowest MASE for this demand subset
- Best forecast: The forecast (combination) that produced the best MASE

The code used to produce this can be seen in 0 and requires the gts object created in Appendix H.A.

C. Reconciling the results

In order to reconcile the forecast we collect all the resulting forecasts and turn them back into a grouped time series. We collect all the best forecasts from Appendix H.B and call combinef() from the hts package (Hyndman R. J., Lee, Wang, & Wickramasuriya, 2018) in order to reconcile the series and retrieve the bottom level reconciled forecasts. Appendix J presents the applied R code. A limitation of this method is that bottom level forecasts that have values near zero can be reconciled to be negative, which is not allowed in our setting.

Appendix I. **R code for applying the forecasts**

```
#Build tibble, forecast and add errors and measures
Comp_tibbler <- function(gts){
  f    <- Fore_tibbler(gts)
  ft   <- Err_tibbler(f)
  return(ft)
}
```

Figure 7.15 high level function call

```
#Build tibble and forecast
Fore_tibbler <- function(gts,h=12){
  f <- tibbler(gts)
  for (i in 1:length(f$Node)) {
    f$FNode <- lapply(f$Train,benchmarks,h)
  }
  return(f)
}
```

Figure 7.16 calls function that produces the tibble (tibbler) and applies the forecasts (benchmarks)

```
#build Forecast tibble
tibbler <- function(gts, end=c(2016,12), start=c(2017,1)){
  f<-tibble(Node=colnames(allts(gts)[,]))
  train=  window(allts(gts),end=end)
  test=   window(allts(gts),start=start)
  for (i in 1: dim(allts(gts))[2]) {
    f$Train[[i]]  <- train[,i]
    f$Test[[i]]  <- test[,i]
  }
  return(f)
```

Figure 7.17Builds the base tibble

```
#Add errors measures and Mase
Err_tibbler <- function(f){
  for (i in 1:length(f$FNode)) {
    f$Errors[[i]]   <- errors(f$FNode[[i]], train = f$Train[[i]], test = f$Test[[i]])
    f$Acc[[i]]      <- f$Errors[[i]] %>% group_by(Method) %>%
      summarise(MASE=mean(ASE),sMAPE=mean(sAPE)) %>%
      arrange(MASE)
    f$best[[i]]   <- f$FNode[[i]][f$Acc[[i]]$Method[1],]
    f$MASE[i]       <- f$Acc11[[i]]$MASE[1]
  }
  return(f)
}
```

Figure 7.18 Adds the accuracy measures and selects the best performer

```r
# calculate the required forecasts and compute all combinations (average)
benchmarks <- function(y, h=12) {
  require(forecast)
  # Compute four benchmark methods
  if (sum(tail(y,24)!=0)==0) {
    fcasts <- rbind(
      S = snaive(y, h)$mean,
      N  = naive(y,h)$mean,
      M = meanf(window(y,start=end(y)[1]),h)$mean
      )
  } else  {
  fcasts <- rbind(
    N  = naive(y,h)$mean,
    S  = snaive(y, h)$mean,
    E  = forecast(ets(y), h,biasadj = T)$mean,
    A  = forecast(auto.arima(y,stepwise = FALSE), h,biasadj = T)$mean,
    Th = thief(y,h=h)$mean,
    #Tc = thief_combine(y, h),
    Tb = forecast(tbats(y),h,biasadj = T)$mean,
    Tf = thetaf(y, h)$mean,
    #H = hybridForecast(y,h)$mean,
    M = meanf(window(y,start=end(y)[1]),h)$mean

  )
  }
```

Figure 7.19 Method that produces the relevant forecasts part 1/2

```r
  if (sum(y!=0)>2)
    fcasts <- rbind(
      fcasts,
      C  = croston(y,h)$mean,
      I = imapa(y,h)$frc.out,
      Tf = thetaf(y, h)$mean
      )

  colnames(fcasts) <- seq(h)
  method_names <- rownames(fcasts)
  # Compute all possible combinations
  method_choice <- rep(list(0:1), length(method_names))
  names(method_choice) <- method_names
  combinations <- expand.grid(method_choice) %>% tail(-1) %>% as.matrix()
  # Construct names for all combinations
  for (i in seq(NROW(combinations))) {
    rownames(combinations)[i] <- paste0(method_names[which(combinations[i, ] > 0)],
                                        collapse = "")
  }
  # Compute combination weights
  combinations <- sweep(combinations, 1, rowSums(combinations), FUN = "/")
  # Compute combinations of forecasts
  return(combinations %*% fcasts)
}
```

Figure 7.20 Method that produces the forecasts part 2/2

Appendix J. **Reconciling the forecast**

```
#Reconcile Forecasts to gts
recon_tibble <- function(f,gts,year=2017){
  allf <- matrix(NA,nrow = 12 ,ncol = nrow(f))
  for(i in 1:nrow(f)){
     allf[,i] <- f$best[[i]]
  }
  allf <- ts(allf, start = year,frequency = 12)
  y <- combinef(allf, groups = get_groups(gts), keep ="gts",
                algorithms = "slm")
  y$labels <- gts$labels
  y$groups <- gts$groups
  dimnames(y$bts) <- dimnames(gts$bts)
  return(y)
}
```

Figure 7.21 Method that takes the relevant tibble produced in 0

Appendix K. **Zero value observations per group**

Table 7.1 Characteristics of observation per group

Group	# Nodes	Hours per node	Total observations	Total zeros observations	avg # zeros per node	% zero	Avg hours per observation	avg hours non-zero observation
Total	1	2666359	72	0	0	0%	37033	37033
G1	2	1333180	144	0	0	0%	18516	18516
G2	15	177757	1080	88	6	8%	2469	2688
G3	9	296262	648	7	1	1%	4115	4160
G4	2	1333180	144	0	0	0%	18516	18516
G5	62	43006	4464	444	7	10%	597	663
G6	15	177757	1080	112	7	10%	2469	2755
G7	83	32125	5976	1891	23	32%	446	653
G8	4	666590	288	0	0	0%	9258	9258
G9	26	102552	1872	538	21	29%	1424	1999
G10	98	27208	7056	1759	18	25%	378	503
G11	271	9839	19512	9859	36	51%	137	276
G12	18	148131	1296	78	4	6%	2057	2189
G13	302	8829	21744	9636	32	44%	123	220
G14	95	28067	6840	2057	22	30%	390	557
BTS	713	3740	51336	32101	45	63%	52	139

Table 7.1 shows the characteristics related to zero value observations per group an explanation for each column:

- Group: the group as defined in Section 4.3.2
- # nodes: the number of different demand series, or nodes, in the demand structure
- Hours per node: the number of demand hours per each node
- Total observations: Each node runs for 72 months (2012-2017) and each provides an observation of demand
- Total zero observations: total number of zero value observation per group
- Avg # of zeros per node: the total number of zero value observations per node (total zeros per group/ # of nods in group)
- % zero: percentage of zero value observations per group (avg # of zero per node/72, or total zero/total observations)
- Avg hours per observation: hours per node/72
- Avg hours non-zero observations: hours per node/(72-# of zero observations)

Appendix L. **Outlier handling**

Section 4.3.3 shortly handled the subject of outlier demand series by highlighting that a significant portion of the lower level nodes experience a lot of 0 values. In order fairly asses the different series these 540 series were marked as outlier series, either for removal of analysis. But other outliers were present in the data. Somewhat unexpectedly our proposed method serves as a decent outlier detection for series that need judgmental input.

Series requiring judgmental input

The results presented in Chapter 5 have most of the problematic series excluded. This was necessary to give a realistic picture of the forecasting accuracy, resulting for instance in Table 5.1 and Table 5.4. Were the outliers not removed we find Table 7.2 and Table 7.3. The majority of the MASE scores still appear reasonable but we experience some serious outliers in their average performance.

Table 7.2 MASE before outlier removal

Group	Mase16	Current16	*Diff*	Mase17	Current17	*Diff*	Avg	CurrentAv	*Diff*
Total	0,78	1,02	-0,24	1,45	1,56	-0,11	1,12	1,29	-0,17
G1	2,63	3,09	-0,46	0,96	1,03	-0,07	1,80	2,06	-0,26
G2	3,65	4,22	-0,57	0,78	1,10	-0,32	2,22	2,66	-0,44
G3	0,45	0,64	-0,20	0,53	0,61	-0,08	0,49	0,62	-0,14
G4	0,55	0,82	-0,27	1,05	1,78	-0,73	0,80	1,30	-0,50
G5	0,75	1,03	-0,28	0,77	1,05	-0,28	0,76	1,04	-0,28
G6	0,99	1,25	-0,26	0,46	0,54	-0,08	0,72	0,90	-0,17
G7	32,00	32,60	-0,60	0,80	1,08	-0,29	16,40	16,84	-0,44
G8	1,58	2,22	-0,64	1,12	1,81	-0,69	1,35	2,02	-0,67
G9	29,80	30,40	-0,60	0,94	1,43	-0,49	15,37	15,92	-0,54
G10	0,98	1,26	-0,28	1,51	1,90	-0,39	1,24	1,58	-0,34
G11	1,38	1,64	-0,26	2,99	3,59	-0,60	2,19	2,62	-0,43
G12	0,77	1,11	-0,34	0,50	0,68	-0,18	0,64	0,90	-0,26
G13	1,57	2,04	-0,47	29,90	30,20	-0,30	15,74	16,12	-0,39
G14	0,94	1,32	-0,38	0,68	1,10	-0,42	0,81	1,21	-0,40
BTS	3,62	4,17	-0,55	3,52	3,98	-0,46	3,57	4,08	-0,51

Table 7.3 Reconciled MASSE before outlier removal

Group	Mase16	Current16	*Diff*	Mase17	Current17	*Diff*	Avg	CurrentAv	*diff*
Total	0,79	1,02	-0,23	1,44	1,56	-0,12	1,12	1,29	-0,17
G1	2,31	3,09	-0,78	1,09	1,03	0,06	1,70	2,06	-0,36
G2	3,45	4,22	-0,77	0,86	1,10	-0,24	2,15	2,66	-0,51
G3	0,50	0,64	-0,15	0,54	0,61	-0,06	0,52	0,62	-0,11
G4	0,65	0,82	-0,17	0,98	1,78	-0,81	0,81	1,30	-0,49
G5	1,20	1,03	0,17	1,02	1,05	-0,03	1,11	1,04	0,07
G6	1,15	1,25	-0,10	0,66	0,54	0,12	0,91	0,90	0,01
G7	28,30	32,60	-4,30	3,05	1,08	1,97	15,70	16,80	-1,10
G8	1,60	2,22	-0,62	1,00	1,81	-0,81	1,30	2,02	-0,72
G9	29,70	30,40	-0,70	2,06	1,43	0,63	15,90	15,90	0,00
G10	2,64	1,26	1,38	5,56	1,90	3,66	4,10	1,58	2,52
G11	11,60	1,64	9,96	25,70	3,59	22,11	18,70	2,62	16,08
G12	0,80	1,11	-0,31	0,56	0,68	-0,12	0,68	0,90	-0,22
G13	6,89	2,04	4,85	32,10	30,20	1,90	19,50	16,10	3,40
G14	2,64	1,32	1,32	2,40	1,10	1,30	2,52	1,21	1,31
BTS	17,80	4,17	13,63	20,00	3,98	16,02	18,90	4,07	14,83

Several of the groups experience very bad performance which warrants a closer look. The pattern that appeared was that a select few series skews the entire average performance per group with a MASE as high as 2300. We find that it is mostly caused by outlier behaviour that the models cannot predict. Figure 7.22 shows a series responsible for the unbalancing of the results. We can see that slight demand is observed at the end of 2015, this is what the MASE uses to scale the forecast errors. Then in 2016 the demand jumps with several orders of magnitude. As a result the scaling factor is not realistic and the resulting MASE over 2016 is 2351, Over 2017 the models could correctly interpret the change in level and the MASE was 0,84.

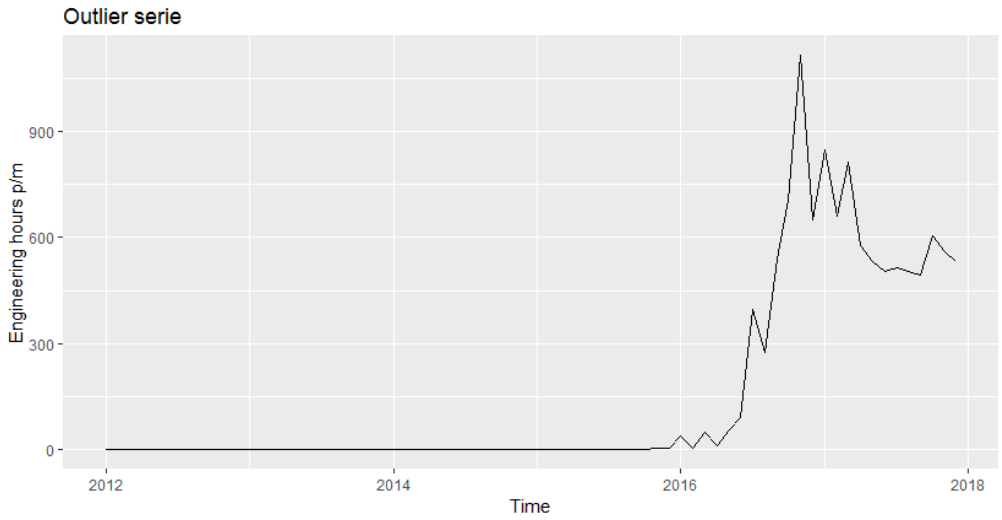


Figure 7.22 Outlier series G7 2016

This is one of the cases where judgmental forecasting as described in Section 4.5 is necessary to produce forecasts for unforeseen situations. The change incurred here is unsystematic and models that depend on patterns in data cannot handle such behaviour. Sampling the series with such high MASE's

confirmed that the cause is usually very slight demand in the training period and an explosive jump in demand in the test period causing the in-sample errors to be drastically different from the out of sample errors. As this is what the MASE scales on it is unrepresentative of the forecast accuracy for the other series. Additionally, this means that our model can serve as a detection tool for outlier data apart from forecasting and as such can help to improve the available data set or identify true outliers and their cause. The out of the ordinary MASE serve as an indicator and from these series the organization can learn when it is important to make judgemental adjustments.

In order to get a better picture of general performance we deem any MASE >50 as outliers that need judgmental assessment. By doing this we exclude 23 results from 2016 and 25 from 2017, the results of this are reflected in Table 5.1 and Table 5.4 and clearly visible as the overall performance is much more in line with expectations.

Appendix M. **Iterative Reconciliation performance**

As referred to in Section 5.1.3 forecast reconciliation introduces large errors by applying patterns to series that should experience no forecast. By iteratively setting the relevant forecasts to 0 and reconciling again we coerce the algorithm to assign values closer and closer to the expected forecast. After setting the forecast to zero and reconciling we further coerce reconciliation to set all negative value to 0. After several iterations no significant improvements in accuracy occur and we can assume that a representative result is achieved. Table 7.4 shows the MASE progression per group over 2017, the results are colour coded per row (group) to indicate the relative best results.

Table 7.4 Iterative progression of reconciliation MASE compared to the best and current

Group	OLS	iter 1	iter2	iter 3	iter 4	iter 5	iter 6	Best Mase	Current
Total	**1,44**	**1,44**	**1,44**	**1,44**	**1,44**	**1,44**	**1,44**	1,45	1,56
G1	1,09	*1,09*	1,10	*1,11*	1,12	1,12	1,12	**0,96**	1,03
G2	0,86	*0,84*	0,84	*0,85*	0,85	0,85	0,85	**0,78**	1,10
G3	0,54	*0,51*	0,51	**0,50**	**0,50**	**0,50**	**0,50**	0,53	0,61
G4	0,98	*0,96*	0,94	*0,93*	**0,92**	**0,92**	**0,92**	1,05	1,78
G5	1,02	*1,01*	1,01	*1,01*	1,01	1,01	1,01	**0,77**	1,05
G6	0,66	*0,47*	0,47	**0,46**	**0,46**	**0,46**	**0,46**	**0,46**	0,54
G7	2,24	*1,21*	1,19	*1,19*	1,20	1,20	1,20	**0,80**	1,08
G8	1,00	*0,92*	0,90	*0,89*	**0,88**	**0,88**	**0,88**	1,12	1,81
G9	2,06	*1,50*	1,35	*1,26*	1,21	1,18	1,16	**0,94**	1,43
G10	1,37	*1,60*	1,40	*1,29*	1,23	1,21	1,20	**0,91**	1,30
G11	4,65	*3,52*	2,99	*2,74*	2,55	2,49	2,47	**1,64**	1,97
G12	0,56	*0,51*	**0,50**	**0,50**	**0,50**	**0,50**	**0,50**	**0,50**	0,68
G13	2,39	*1,20*	1,05	*0,99*	0,97	0,96	0,96	**0,61**	0,90
G14	2,40	*1,33*	1,14	*1,05*	1,01	0,99	0,98	**0,68**	1,10
BTS	5,24	*2,54*	2,13	*1,95*	1,81	1,84	1,81	**0,88**	1,33

Appendix N. **Reconciling with one less characteristic**

In Section 5.1.3 we observed that errors were introduced in the lower level groups after reconciliation. We were able to reduce the errors by accounting for outlier series, see Appendix L, and iteratively adjusting them toward required values, see Appendix M. Given that the errors occur and become larger in more granular demand groups this implies that reducing the granularity, potentially increasing information density, could result in better overall accuracy. In order to test this we iterated through the forecast process for 2017 while having removed the product code from the characteristics. The product code is the most diversifying characteristic with 62 base nodes, see Table 4.6 for the original structure.

By removing the product code the granularity of the demand is drastically reduced. Table 7.5 illustrates the resulting nodes per level with a total of 285 vs 1716 for the original demand structure. This drastically reduces the number of nodes in the structure and keeps the level of demand per node higher. As shown in Table 7.6 the average outliers per series also decreasing the highest percentage of outliers was 44% in the original structure vs 19% in the new.

Forecasting the smaller demand subset, reconciling and iterating to the first step as in Appendix M leads to the MASE comparison in Table 7.7. We can observe that overall no significant accuracy is gained and while losses are marginal in some groups we can identify large decreases in accuracy in others. Based on these results we can conclude that including lower level, more granular demand adds information to higher levels demand groups even if their own performance might not be of the same quality.

Table 7.5 Demand structure without product code

New groups	Description	# Nodes
Total	Total	**1**
G1	KLM or customer	2
G2	Specific division or customer	**15**
G3	Aircraft type	**9**
G4	Routine or non routine	**2**
G5	G1 + G3	**15**
G6	G1 + G4	**4**
G7	G2 + G3	**83**
G8	G2 + G4	**26**
G9	G3 + G4	**18**
BTS	G2+G3+G4	**110**

Table 7.6 Outliers in new structure

New groups	Old equivalent	# Nodes	# Outlier	% outlier
Total	Total	1	0	0%
G1	G1	2	**0**	0%
G2	G2	15	0	0%
G3	G3	9	0	0%
G4	G4	2	0	0%
G5	G6	15	2	13%
G6	G8	4	0	0%
G7	G7	83	16	19%
G8	G9	26	2	8%
G9	G12	18	1	6%

Table 7.7 MASE comparison old vs removed

New groups	Old equivalent	New MASE	Old MASE
Total	Total	1,45	**1,44**
G1	G1	**1,08**	1,09
G2	G2	0,86	**0,84**
G3	G3	0,54	**0,51**
G4	G4	1,01	**0,96**
G5	G6	0,51	**0,47**
G6	G8	0,95	**0,92**
G7	G7	1,34	**1,21**
G8	G9	2,35	**1,5**
G9	G12	0,52	**0,51**

Appendix O. **2016 absolute forecast results**

Group	MASE	Sum	Mean	sum - actual	mean- actual
Total	0,78	444181	37015	736	61
G1	0,88	435210	36268	-8235	-686
G2	1,10	439438	36620	-4008	-334
G3	0,84	447810	37318	4365	364
G4	0,67	441467	36789	-1978	-165
G5	1,15	431403	35950	-12042	-1004
G6	0,95	434983	36249	-8463	-705
G7	1,17	433458	36121	-9988	-832
G8	0,73	438158	36513	-5288	-441
G9	1,10	439735	36645	-3711	-309
G10	1,20	428895	35741	-14551	-1213
G11	1,33	425361	35447	-18084	-1507
G12	1,17	437091	36424	-6355	-530
G13	1,53	415226	34602	-28220	-2352
G14	1,27	426676	35556	-16770	-1397
BTS	2,30	399855	33321	-43591	-3633
Current	1,02	449956	37496	6511	543
Reconciliation	0,79	440622	36718	-2824	-235
Actual	0	443446	36954	0	0

Figure 7.23 2016 Accuracy comparison with absolute values

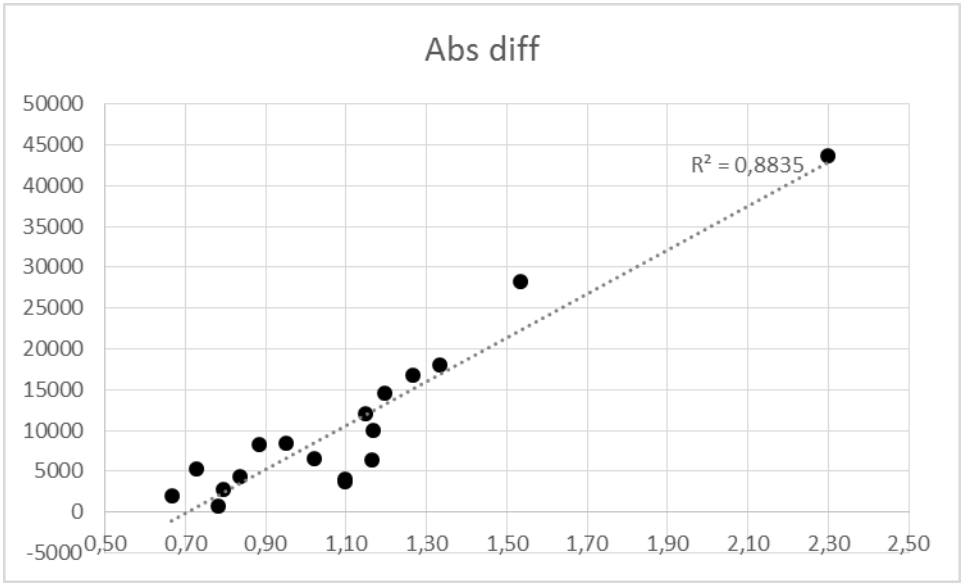


Figure 7.24 2016 MASE vs absolute difference with actual values

Appendix P. **Effect of correcting for working days per month**

A realization too late to extensively examine was correcting the working hours per month for the actual working days in those months. As this is a large driver in the total amount worked it potentially explains a part of the variation seen in the monthly demand. Figure 7.25 presents the total engineering demand divided by the working days per month. That is all days in that month that are not weekends or holidays. It appears that this somewhat stabilizes demand compared to the unadjusted numbers.

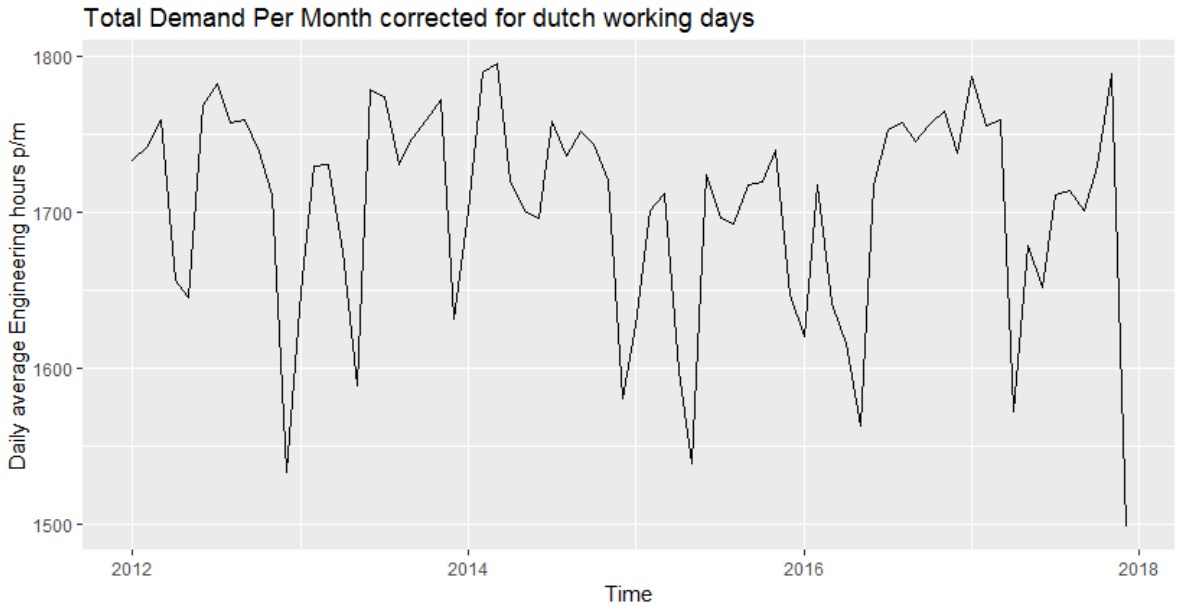


Figure 7.25 Demand corrected for dutch monthly working days

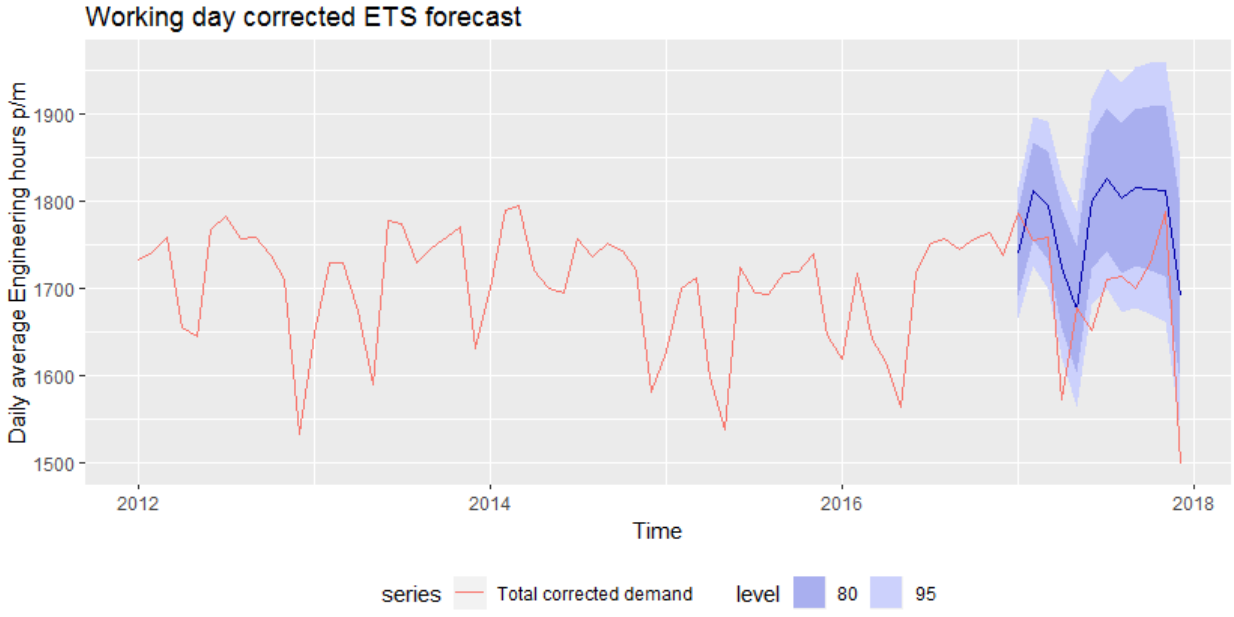


Figure 7.26 ETS forecast of corrected demand

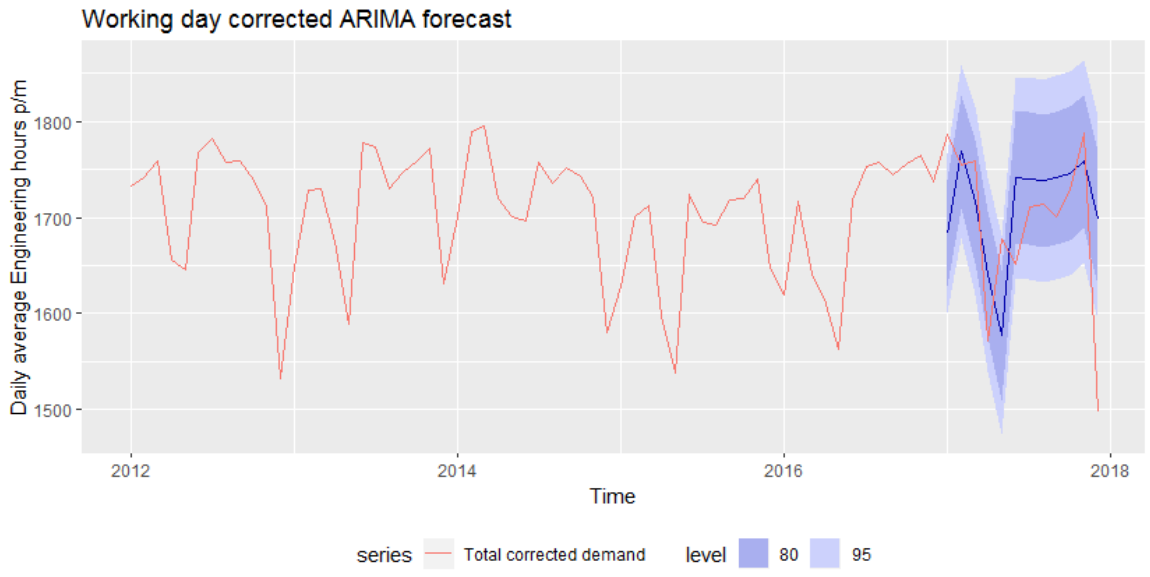


Figure 7.27 Arima forecast of corrected demand

We can see in Figure 7.26 and Figure 7.27 that forecasting performance by ETS does not appear very accurate but Arima appears to do does better. Table 7.8 Shows the resulting MASE on the corrected values where we find a 1,35 higher than accuracy reached on the uncorrected total demand.

Table 7.8 Corrected MASE

MASE	Non-corected	Corrected
Arima	1,51	1,35
ETS	1,54	1,89

Because the data was transformed the comparison is not completely fair and back transforming the forecast will provide a better indication of accuracy. The resulting back transformed Arima forecast vs a Arima forecast on the untransformed data are shown in Figure 7.28.

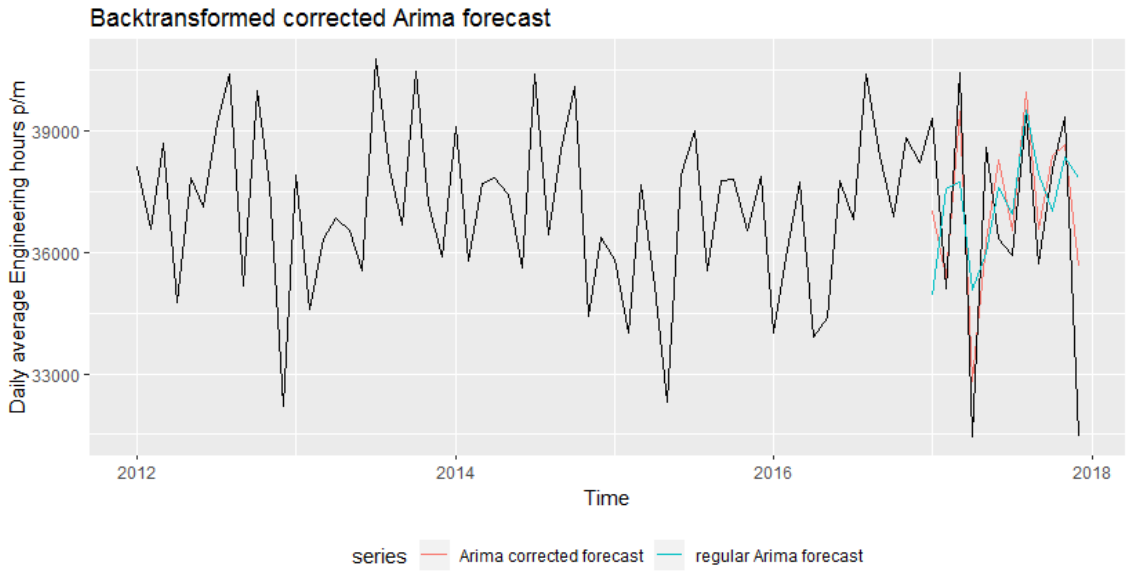


Figure 7.28 Corrected vs uncorrected Arima forecast

The corrected forecast appears to produce significantly more accurate results, especially in moving with the higher peaks and falls in data. This implies that a significant part of variation in the data can be explained by correcting for the working days per month and merits follow up research and results. Confirmed by the resulting MASE of 0,87 for the Arima forecast in Figure 7.28, far lower than any previous result for the total forecast. Given the hierarchical/grouped demand structure it would be interesting to see whether improvements can be seen over all different nodes. Implementing the working day transformation into our proposed framework is very possible and not too complicated and would thus serve as a natural next step in extending the framework.