1 INTRODUCTION

1.1 Motivation and Problem Description

Even in the early 2020s we are still in an era, where product and service portfolios increase, as customers demand more complex and higher individualized products and services. Supply networks grow in their depth, their width and their inter-dependencies (*Timmer* et al. 2015). *Rosenzweig* and *Easton* (2010) identified ‘quality’, ‘delivery’, ‘flexibility’ and ‘costs’ as basic competitive goals. *Forrester* (1958) was one of the first researchers, who described the tasks of Supply Chain Management (SCM) as the ‘interactions between the flow of information, materials, money, manpower and capital equipment’ (p. 38).

The ‘supply chain planning matrix’, as shown in figure 1.1, arranges planning tasks along a horizontal and a vertical axis (*Rohde* et al. 2000). The horizontal axis follows the flow of materials; whereas the vertical axis defines the hierarchy of the planning tasks.

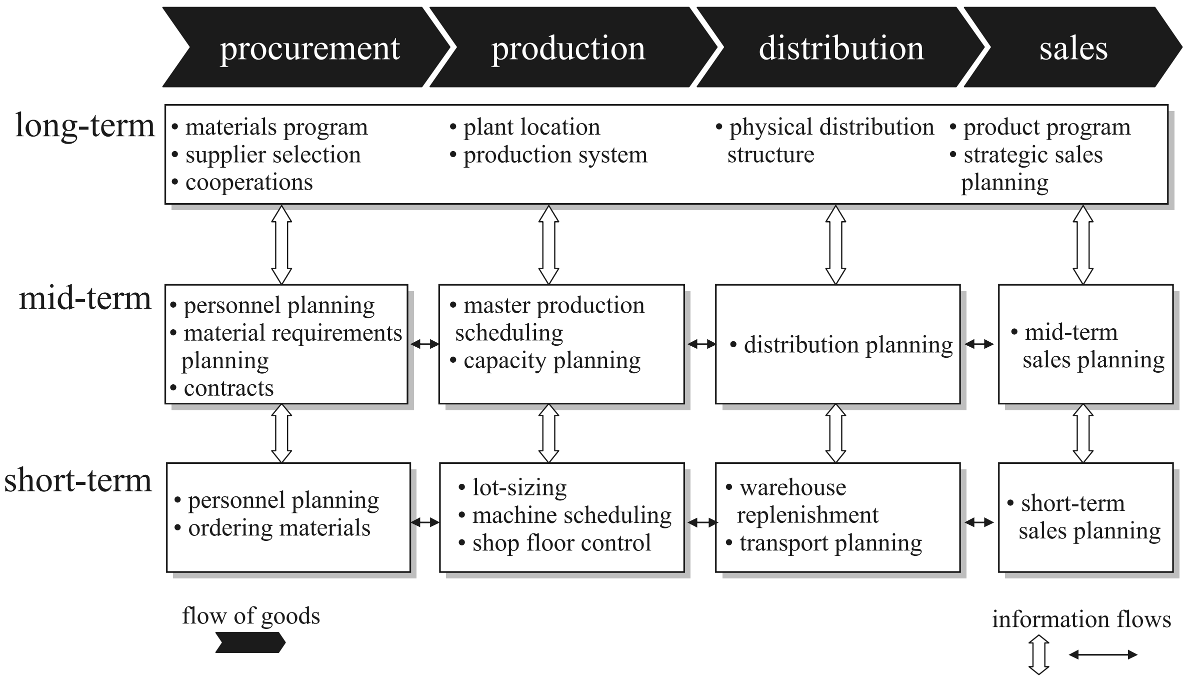

Fig. 1.1: Supply chain planning matrix (*Fleischmann* et al. 2015, p. 77)

Reliable plans for each task must be generated, often before real customer demands are known. According to *Kilger* and *Wagner* (2015) ‘replenishment decisions in a retail store are taken before a customer enters the store. Production quantities, for Make-to-Stock (MTS) products are determined prior to the point in time when the customer places orders. Decisions about procurement of raw materials and components with long lead times, have to be taken before customer orders for finished goods using these raw materials or components become known' (p. 125). Also see *Silver* et al. (2001).

Feasible plans are usually based on forecasts or demand planning. Forecasts are results of statistical methods or expert judgments 'predicting the future as accurately as possible, given all of the information available, including historical data and knowledge of any future events' (*Hyndman* and *Athanasopoulos* 2018, p. 14). Accurate forecasts play an important role in many domains (*Hyndman* and *Athanasopoulos* 2018; *Armstrong* 2001). Especially the SCM in the automotive industry is dependent on reliable forecasts for a variety of different product and process combinations.

To stress the practical relevance, this work addresses a real industial problem at the car manufacturers AUDI AG and Volkswagen Group. To use forecasts for an end-to-end planning, e.g., from n-tier suppliers over Original Equipment Manufacturer (OEM) to final customers, a large number of forecasts for different levels of hierarchies and aggregation levels must be generated. Suppliers, for example, need exact individual forecasts for hundreds of thousands of parts. Or the sales department of an OEM needs to know how many private and business customers will order how many cars in which region, ideally including the exact take rates of sales options. In this context two major problems arise:

1. The number of forecasts needed to be generated is high. For example, the theoretically possible sales options of the 2012 model AUDI A3 lead to approximately $1.1 \cdot 10^{38}$ combinations. This number only includes sales options. Other characteristics, such as the sales region and the type of the customer increase the number of theoretical forecasts.

2. Forecasts need to be *coherent*. This means, they need to be consistent and summed up correctly over different levels of a hierarchy. For example, the sum of all forecasts for *business* and *private* orders must be equal to the forecast of *total* orders. This relation is called 'hierarchical forecasting' (*Hyndman* and *Athanasopoulos* 2018).

Figure 1.2 shows an example. The top row (a) represents the number of total orders. The second row (b) shows the quantities for specific customer segments; in this example *private* and *business* customers. The straight lines represent the number of daily orders of cars of a specific type in a specific region. The dashed lines represent the forecasts. The time-series on the second row (b) add up correctly for each day to the *total* orders in row (a).

At most of the brands at the Volkswagen Group, a car is technically specified by setting the characteristics for about 160-200 options.[1] Options (also known as 'PR-Familie' or 'PR-family') can be seen as attributes. They represent superior groups or modules such as wheels, seats or multi-functional displays. They are encoded by three capital letters, e.g., 'RAD' for the module wheels. To specify an order, an exact characteristic must be assigned to every

[1] To completely specify a car, additional information must be defined. The model key includes information about the vehicle category, the type of body, standard packages of options, the type of engine, and the type of gearbox. Also the model year, the sales region and colors must be defined upfront. This work neglects these additional information.

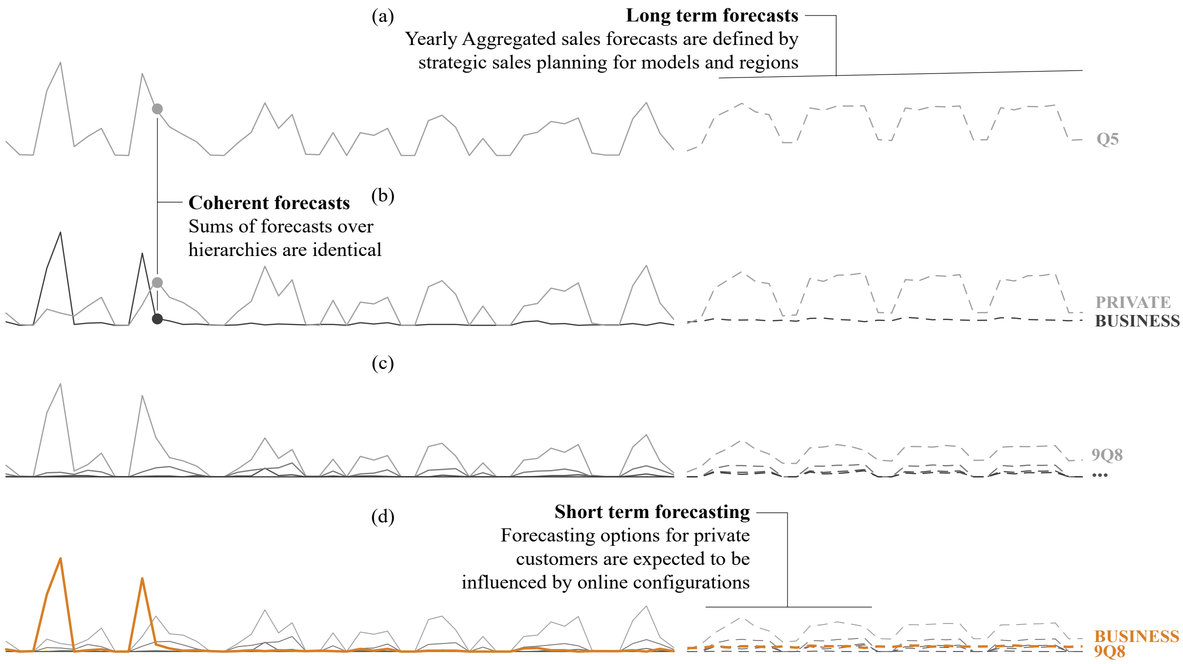

Fig. 1.2: Example of coherent hierarchical forecasts
Absolute numbers of orders (continuous lines) and forecasts (dashed lines) per day: The top row (a) shows the total quantities per model Q5, row (b) shows quantities for each customer segment, row (c) shows quantities of all 'Primäre Eigenschaft' (primary characteristic) (PR)-numbers of a single PR-family, and (d) the quantities of all combinations.

PR-family. For example, for the option RAD a specific type of wheel must be chosen. The possible choices for each PR-family are called 'PR-number' (or 'PR-Nummer') and they are also encoded by three letters, e.g., 40R, 40S, or 40T. If a customer chooses a specific wheel (e.g., 40S), the PR-family RAD assigns the PR-number 40S. From the assignments of the 160-200 PR-families the demand for each part can be derived through the Bill of Materials (BOM). Some parts depend on multiple PR-families. For example, a front ends' specific drilling pattern depends on the car model, the number and the type of sensors, headlights etc. Some combinations are forbidden by technical rules. For example, 19" disc brakes can only be assembled with 19" wheels which prohibits the specification of all other wheels.

Examples of sales quantities and forecasts for different PR-numbers of a single PR-family are shown in row (c) of figure 1.2. To be coherent, the sum of all PR-numbers must add up to the total sales forecasts. Row (d) shows the combination of the PR-family (from row c) and the customer type (from row b) as a Cartesian product. Each option of customers can be combined with each PR-number of the PR-family. As highlighted in orange, a sales manager could be interested in the sales forecast of a specific option and for a specific customer segment (e.g., the number of business customers, ordering cars with PR-number 9Q8).

Besides the model, the type of customer and the technical options a variety of additional attributes are relevant. An exemplified overview of possible combinations and their relationships is shown in figure 1.3. The attributes are arranged as a tree. The horizontal axis represents different levels of the hierarchy, e.g., the attributes brand, model, region, or the

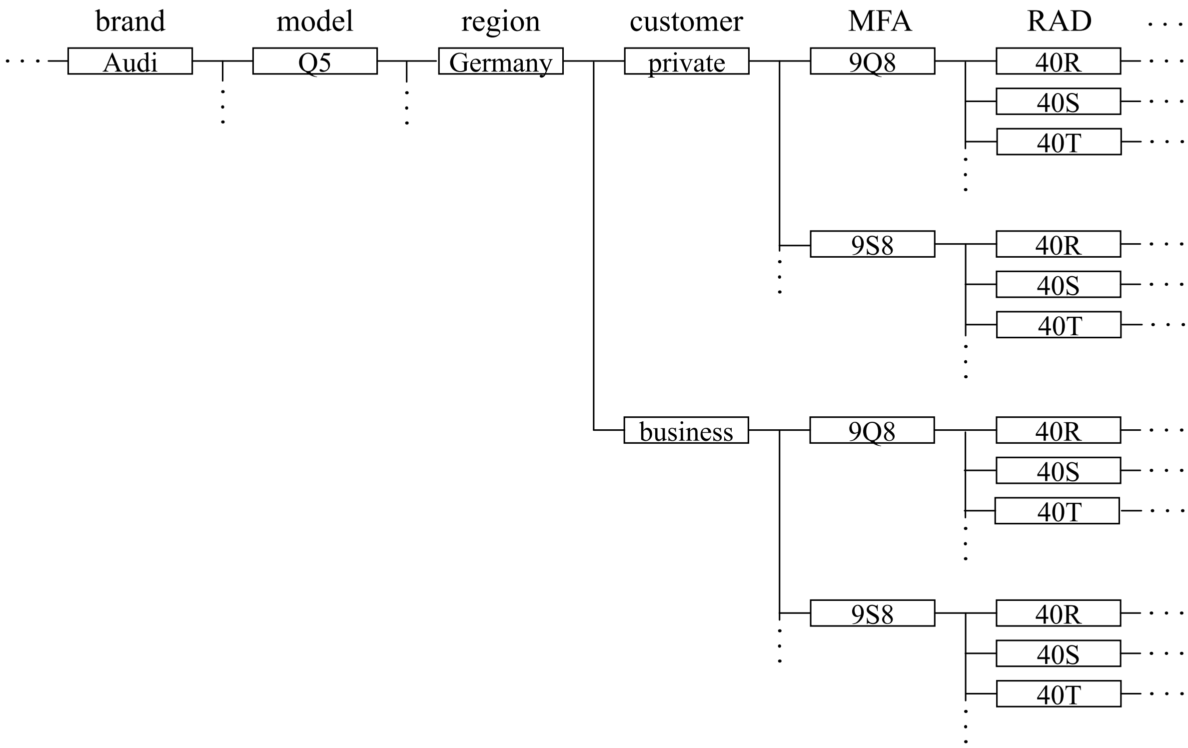

Fig. 1.3: Exemplified combinations of attributes as a tree
Each column represents a level in the hierarchy, where all variants of an attribute, e.g. the PR-numbers of the PR-family MFA are added to all previous combinations as additional branches. The leaves consist of a set containing all possible combinations of all attribute values. The number of leaves grows exponentially with every additional level.

PR-families. For each level additional branches are added for all possible characteristics of the attribute. For example, at the customer level, branches for *private* and *business* are added. For each following hierarchy branches are added to the existing branches. For example, 9Q8 and 9S8 are added to all previous branches. For the complete tree each leaf represents a combination with a defined characteristic for each level.

Besides the increasing complexity of products and supply chains additional data sources become available. At the research partner AUDI AG data, generated by potential customers using the online car configurator, are stored in databases since 2017. The stored data includes all information which are needed to specify a complete car.

1.2 Research Question and Contribution

Based on the context described, three central research questions arise:

1. **Which models and processes allow the integration of high-dimensional web configurations as predictors to forecast sales options in the automotive industry?**
   There exist a variety of publications which propose statistical models, including the usage of different predictors for the generation of sales forecasts. However, the majority of research, especially in the domain of the automotive industry focus on three aspects: (1) The number of global product sales, (2) the number of sales for specific regions or countries and (3) the number of sales for specific brands (see Section 3.6.1). To the knowledge of the author, there exist neither publications on the integrated forecasting of sales options using time-series, nor on using online web configurations as predictors. This work proposes a process which enables the usage of high-dimensional online configurations for the evaluation of different models. The process further allows the tuning of hyper parameters as well as the training, testing and evaluation of different statistical learning methods.

2. **Does the integration of different predictors, especially the integration of online configurations increase the forecasting accuracy for sales options?**
   An important aspect is the evaluation of the proposed process and models. To evaluate the statistical performance and show the practical relevance, the models are trained, tuned, and tested on multiple real industrial data sets. The prediction accuracy is compared to forecasts which are generated without the usage of online configurations as predictors. This allows one to answer the question, of whether the integration of online predictors leads to an improvement or not.

3. **How can coherent forecasts be generated for high-dimensional hierarchies?**
   Rob Hyndman proposed the so-called 'reconciliation' approach, which allows one to manipulate a-priori time-series in a way, they become coherent. The resulting a-posteriori time-series sum up correctly over all hierarchy levels (*Hyndman* et al. 2010; *Hyndman* et al. 2016; *Wickramasuriya* et al. 2018). However, there exist domains where the number of forecasts are to high to use this approach, as it generates time-series for all possible combinations of all levels in the hierarchy. This work proposes a different reconciliation approach which systematically neglects some parts in the hierarchy. This allows the efficient computation of coherent a-posteriori time-series for large hierarchies.

The contribution to the scientific body can be summarized as:

- An in-depth **description of automotive production processes and its planning tasks** as the practical process which needs accurate high-dimensional hierarchical forecasts
- A **theoretical data pipeline** which allows one to train, test, and evaluate different model classes, with a variety of hyper parameters

- A **theoretical reconciliation model** which allows the generation of coherent time-series of large hierarchies
- The **practical implementation** of the theoretical models and processes
- Examples of **well-parametrized regression models** of important classes
- The **practical evaluation** using real industry data for all models and processes

This dissertation is addressed to academics and practitioners, who are interested in applying forecasting methods to complex hierarchical problems and using high-dimensional predictors to improve forecasts:

- **Business forecasters and mangers** can use the reference model to evaluate their current processes and forecasting models. Furthermore, they can use it as a recommendation to improve their methods, data flow, Information Technology (IT) systems and their organization. Also, the model can trigger the integration of further complex predictors, state-of-the-art metrics and measurements.
- **Project managers for forecasting projects** receive information about how well different integrated methods and data flows might perform. This can be used to further investigate particular combinations and neglect others.
- **ERP software companies** can use the proposed processes and models to improve their products and to develop new products for forecasting complex hierarchical structures.
- **Business and IT consultants** can use the statistical and practical results of the industry use case as a reference.
- **Forecasting and Machine Learning researchers** receive a contribution to the scientific body, especially about the integration of online configurations and the associated processes and methods.

1.3 Research Methodology

The design of an appropriate research methodology guarantees the fulfillment of multiple requirements: (1) The identified research questions are sufficiently handled and the proposed artifacts are provided. (2) At the same time the recent state-of-the-art is considered. (3) The results of the research are published in an appropriate way, and (4) the research process is transparent and comprehensible (*Yin* 2009). The methodology of a scientific work has to be chosen and adapted according to the object of investigation (*Creswell* and *Creswell* 2018).

This dissertation combines elements from three streams of research theory: (1) The basic concept is based on the design-oriented paradigms of Design Science Research (DSR) (*Peffers* et al. 2007). (2) To guarantee the practical relevance, process steps of the Action Research (AR) method are included (*Davison* et al. 2004). (3) As the object of investigation is the generation of high-dimensional hierarchical time-series forecasts, the steps are adapted to this specific task proposed by *Armstrong* (2001). In the following the applied methods are described.

DSR and AR

To follow the tradition of 'design research' whose foundations were laid by Herbert A. Simon's publication 'The Sciences of Artificial' (*Simon* 1968), this work contributes to the scientific body by providing domain independent results which can be implemented to a wide range of applications (*Hevner* et al. 2004; *Denyer* et al. 2008). According to *Peffers* et al. (2007) design is 'the act of creating an explicitly applicable solution to a problem' (p. 49). The resulting artifact include the process of gaining the scientific insights, as well as the results themselves. To guarantee the practical relevance, while fulfilling the scientific requirements the DSR approach was extended by including practical companies into the research process (*Davison* et al. 2004). The extended method is the AR method. According to *Peffers* et al. (2007), the research process can be divided into four steps:

1. **Analysis**
   During the Analysis phase, the problem and the motivation of research is identified and research questions as well as the research gap are derived. The authors highlight the importance of combining knowledge from both, the scientific research body and the practitioner community.

2. **Design**
   In the Design phase, artifacts are constructed, using systematic methods (*Peffers* et al. 2007), such as AR. Artifacts include 'constructs, theories, models, methods, and instantiations' (*Österle* and *Otto* 2010, p. 288).

3. **Evaluation**
   In the Evaluation phase, artifacts are evaluated against the previously defined research goals, e.g., against defined metrics and tested with practical pilot implementations.

4. **Diffusion**
   During the Diffusion phase, the results are published, students are taught and the results are transferred to participating entities.

Forecasting Process
The AR process can be further specified to account for the nature of the object of investi-

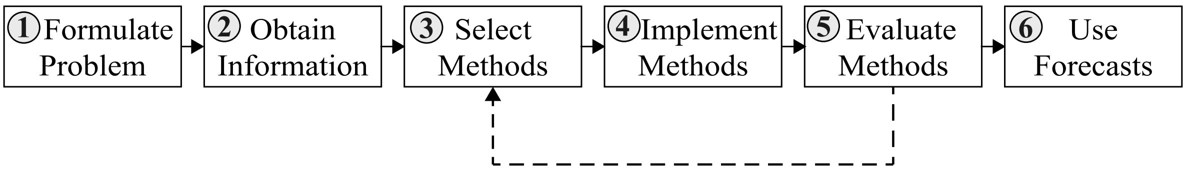

Fig. 1.4: Process of developing and implementing forecasting models (*Armstrong* 2001, p. 8)

gation: The high-dimensional hierarchical time-series forecasts. *Armstrong* (2001) proposes a reference process for the development and implementation of forecasting. Its six steps are shown in figure 1.4. One can see similarities to the AR process: The steps of 'Formulate Problem' and 'Obtain Information' can be compared to the AR's step of 'Analysis'. The steps 'Select Methods' and 'Implement Methods' relate to the 'Design' phase, 'Evaluate Methods' to the 'Evaluation Phase', and 'Use Forecasts' to the 'Diffusion' phase. In the following the six forecasting steps are described and related to the specific object of investigation:

1. **Formulate Problem**
   The first step of a forecasting process includes the analysis and formulation of the underlying problem. A brief introduction on this work's problem is given in Section 1.1. In Chapter 2 the physical process and the related planning tasks are described in detail. Furthermore, available data sources are statistically explored and described (Chapter 4).

2. **Obtain Information**
   Traditionally two types of information are used to generate forecasts: '(a) Statistical data, and (b) the accumulated expertise of the people who collect the data and use the forecasts' (*Hyndman* and *Athanasopoulos* 2018, p. 21). This work uses statistical data. However, the proposed models and processes are designed in a way they can be adapted manually by so-called 'judgmental forecasts'.

   The available statistical data in this work consists of two main sources: *Historical orders* and *online web configurations*. Historical orders are derived from previous customer orders. They include the information shown in figure 1.3, such as the model, the sales region, the type of customer, all PR-families, and the order date.[2] Online configurations are generated, whenever a potential customer finishes a configuration using a web-based car configurator. They include all technical information which are relevant to define a car, such as the model and all PR-families. Furthermore, they include information,

[2]For Build-to-Stock (BTS) orders, where the final customer is not known in advance, the order date is the date when a retailer places an order.

about the time and date when a configuration was completed. Personal information, such as the IP address of a customer or the region are usually not stored. The statistical characteristics and correlations of both datasets are evaluated in Chapter 4.

Some forecasting models, especially regression models, use one or more *input variables* X to calculate one or more *output variables* Y. Input variables are also called predictors, independent variables, features, or just variables. Output variables are sometimes called response or dependent variables (*James* et al. 2013). Customers in some markets use web-based car configurators during their decision-making process, before purchasing a real car. This work evaluates the hypothesis that the data from online configurations can be used as input variables X, to explain some behavior of car sales time-series and to increase the prediction accuracy of sales forecasts as output variables Y.

3. **Select Methods**

   An important step is the selection of methods. There exist a large number of different forecasting methods for different tasks. An overview over time-series forecasting methods is given by *Hyndman* and *Athanasopoulos* (2018). Supervised learning methods are reviewed by *Dey* (2016). However, this work does not focus on a structured or comprehensive selection of possible models. Instead, it provides a data pipeline which allows one to train, test, and evaluate different types of models. To demonstrate the practical significance, the data pipeline is evaluated using a combination of the following methods:

   - Autoregressive Integrated Moving Average (ARIMA)
   - Artificial Neural Network (ANN)
   - Random Forests Regression (RFR)

   Furthermore, the concept of coherent ‘Hierarchical models’ is extended, and a theoretical approach of integrating probabilistic graphical ‘Markov models’ is described. An overview over the methods is given in Chapter 3.

4. **Implement Methods**

   An important element of the implementation is the tuning of the models. Especially ANN and RFR models lack of scientific publications, which describe structured methods and processes for the effective tuning of hyper-parameters regarding this work or a similar context. Chapter 5 introduces a process which implements all models and allows a structured parametrization of hyper parameters.

   To allow the generation of forecasts in large hierarchies, an effective reconciliation approach based on mathematical optimization is proposed in Chapter 6. It is extended by integrating correlations, as given by Markov Networks.

5. **Evaluate Methods**

   A key aspect of AR is the integration of practitioners into the scientific process. Besides the practical problem formulation, the results of the process and the models are evaluated using real industry data. In this context different statistical performance metrics are used:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- symetric Mean Absolute Percentage Error (sMAPE)

The metrics are constantly applied to all results and intermediate steps. The final forecasts are compared to ARIMA based forecasts. For further information see Section 5.2.

6. **Use Forecasts**
   Besides the evaluation using a large number of real industry data, the practical usage at the industrial partner AUDI AG is not described in this work.

The phases of AR and the steps of forecasting projects now can be joined as shown in figure 1.5. The steps are summarized in the following section.

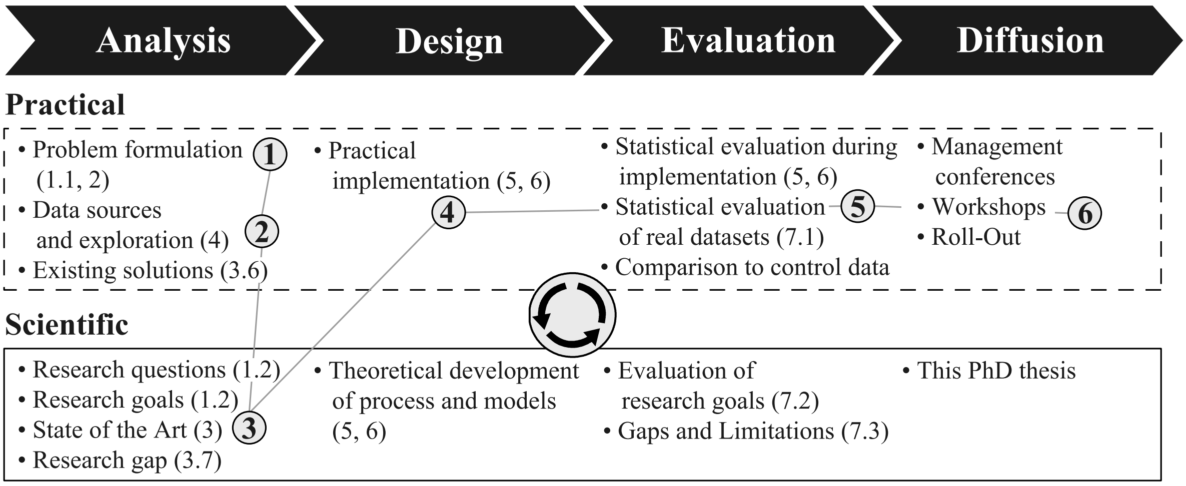

Fig. 1.5: Research methodology of this dissertation
The black arrows on the top divide the method into four phases in accordance with the AR approach (*Peffers* et al. 2007). Each phase is further divided into practical and scientific aspects. The gray bubbles represent the major tasks of forecasting projects, as shown in figure 1.4. The circle in the middle represents regular iteration workshops between the practical users and the scientific researchers. The numbers in the brackets link to corresponding chapters of this dissertation.

1.4 Structure of the Work

As shown in figure 1.6 this dissertation is structured as follows: Chapter 1 gives an introduction into the work. In Section 1.1 the general business problem is described. Section 1.2 introduced the research question and explains the contribution to the scientific research and to practitioners. To guarantee scientific rigor results which are at the same time beneficial to the practitioner community, an appropriate research methodology is introduced in Section 1.3. This section (1.4) gives an overview over the work's structure.

The problem formulation is further specified in Chapter 2. It answers the question 'What needs to be forecasted and why?' Section 2.1 gives a detailed overview over the physical production processes in automotive supply chains. The relevant planning processes are introduced in Section 2.3. Section 2.2 describes the forecasting objects, which are necessary to fulfill these planning processes.

Chapter 3 gives an overview over the state-of-the-art. It consists of seven sections. First statistical forecasting methods are introduced. Each section consists of a *literature review* followed by a *model description* and possibilities for the *implementation*. The first section (3.1) introduces 'regression models with ARIMA errors'. The second section (3.2) describes 'Artificial Neural Networks' as one of two supervised learning methods. The second method is 'Random Forests Regression' which is introduced in Section 3.3. The probabilistic graphical model of 'Markov Networks' is explained in Section 3.4. Section 3.5 introduces 'Hierarchical Models' which allow the reconciliation of coherent time-series. Section 3.6 reviews forecasting publications in the domain of automotive supply chains. It consists of three subsections: SubSection 3.6.1 gives an overview over publications on forecasting of specific individual time-series, e.g., for specific brands or regions. In subSection 3.6.2 integrated forecasting frameworks are reviewed. Finally, subSection 3.6.3 shows how probabilistic graphical models are used to generate forecasts of sales options. In Section 3.7 the research gap is derived.

Chapter 4 focuses on the analysis of available data sources. Section 4.1 explains the data sources. In Section 4.2 statistical explorations and descriptive tests are performed. These include basic descriptive statistics, Lorenz curves, histograms, the detection of outliers and high leverage points, time-series decomposition, testing for ACF and PACF and unit root tests. In Section 4.3 a method for detecting correlations is introduced and clustering tests are performed.

In chapters 5 and 6 the data sources are used to manipulate time-series using supervised learning methods and generate hierarchical forecasts using an optimization model. Chapter 5 is divided into five sections. The first section (5.1) describes why time-series should be manipulated with supervised machine learning techniques. In Section 5.2 a generic data pipeline for shifting, splitting, training, predicting and testing is developed. This generic pipeline is used to generate forecasts with feed-forward ANN models in Section 5.3 and multivariate RFR models in Section 5.4. The last section (5.6) uses these short-term results to manipulate mid- to long-term time-series.

Chapter 6 introduces an optimization model which allows the generation of coherent forecasts for large hierarchies. In Section 6.1 the mathematical model is described. Section 6.2 proves that the model can be used for high-dimensional data. The practical implementation is shown in Section 6.3. A theoretical extension, using Markov Networks, which allows the consideration of correlations between multiple time-series is described in Section 6.4.

In Chapter 7 the developed processes and models are evaluated using instances of real industry data. Section 7.1 describes the application at the company AUDI AG. Final conclusions are drawn in Section 7.2. Section 7.3 shows opportunities for further research directions.

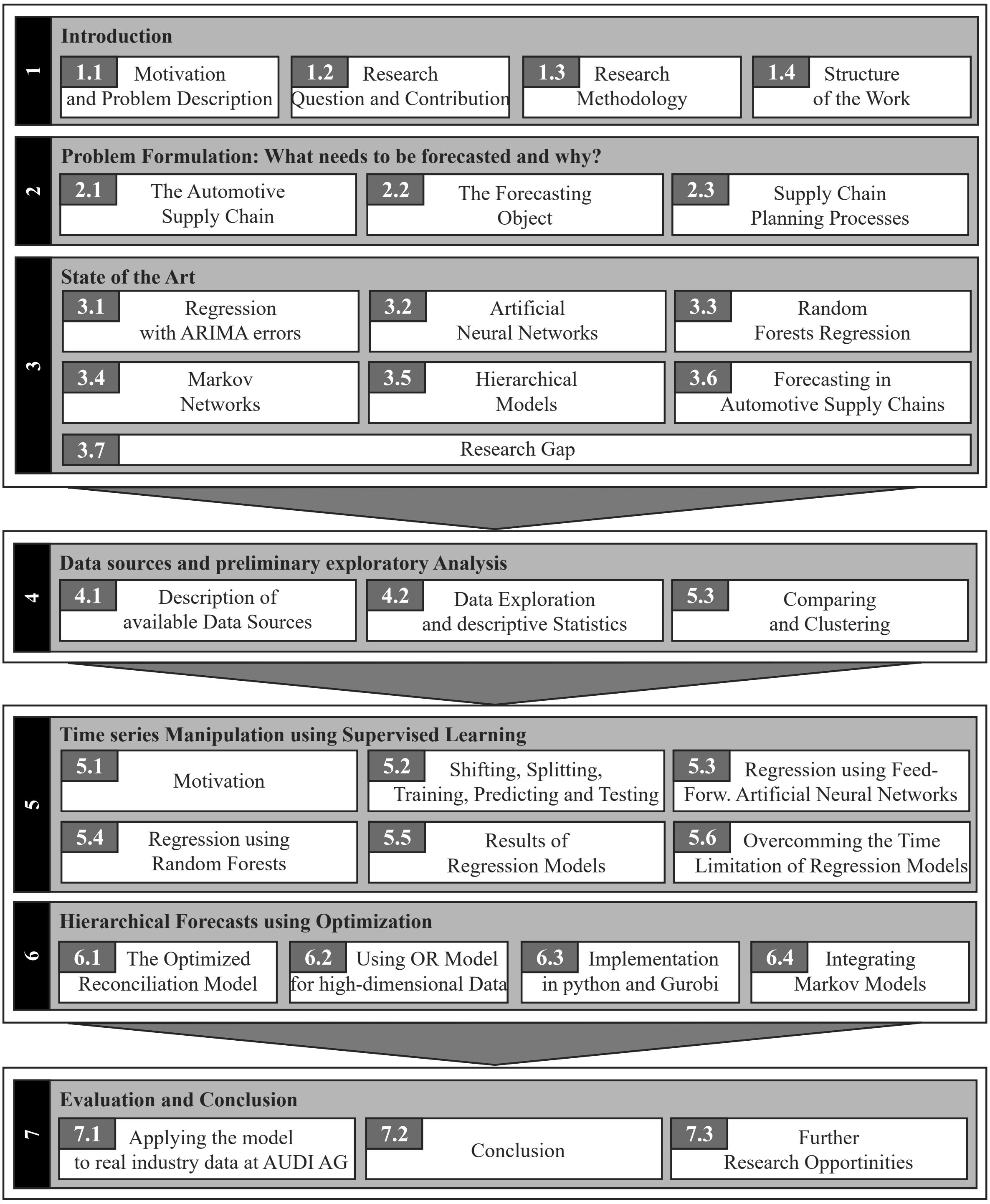

Fig. 1.6: Structure of the dissertation