Essays on Electricity Market Design and the Regulation of Distribution Networks

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2021

vorgelegt
von

M.Sc. Lisa Just

aus

Sindelfingen

Referent:	Prof. Dr. Marc Oliver Bettzüge
Koreferent:	Prof. Dr. Christoph Schottmüller
Tag der Promotion:	22.10.2021

Acknowledgements

Firstly, I would like to thank Marc Oliver Bettzüge for supervising my thesis. His valuable comments and encouragement have played an important role completing my research. Furthermore, I would like to extend my heartfelt gratitude to Felix Höffler, who passed away during the writing of this thesis. Our constructive and insightful discussions will continue to leave a deep and lasting impression on me. I am also grateful to Christoph Schottmüller for co-refereeing my thesis and Oliver Gürtler for chairing the examination committee.

Many thanks go to the Institute of Energy Economics for offering such a positive working environment and providing financial support. This thesis would not have been possible without the inspiring, motivating, and of course joyful teamwork with my former colleagues and co-authors, Joachim Bertsch, Simeon Hagspiel, and Heike Wetzel. I would also like to thank Tom Brown for the pleasant cooperation during our joint research project. Furthermore, I am grateful for the helpful comments on my research from Broghan Helgeson, Helena Schweter, and Christian Tode.

Part of work in this thesis was performed within research projects. I gratefully acknowledge financial support by the Energy Storage Initiative of the German Federal Ministry for Economic Affairs and Energy and the German Federal Ministry of Education and Research through grant 03ESP239. Furthermore, funding by the German Federal Ministry of Education and Research within the Kopernikus Project ENSURE "New ENergy Grid StructURes for the German Energiewende" is gratefully acknowledged.

Finally, special thanks go to my family, especially to my parents as well as Eva and Christian, who provided me with constant support and encouragement. Foremost, my thanks go to Olli and our two little boys for giving me the joy and the confidence to complete this journey.

Lisa Just June 2021

Contents

List of Figures

List of Tables

List of Abbreviations

AC	Alternating Current
ACER	Agency for the Cooperation of Energy Regulators
AIC	Akaike Information Criterion
ARegV	Incentive Regulation Ordinance
BeNeLux	BelgiumNetherlandsLuxembourg
BIC	Bayesian Information Criterion
Bn	Billion
CCGT	Combined Cycle Gas Turbine
CEAS	Compressed Air Energy Storages
CHP	Combined Heat and Power
CO_2	Carbon Dioxide
CWE	Central Western European
DE	Germany
DEA	Data Envelopment Analysis
DC	Direct Current
EEG	Renewable Energy Sources Act
EnWG	Energy Industry Act
ENTSO-E	European Network of Transmission System Operators for Electricity
ES	Economies of Scale
ETS	Emissions Trading System
EU	European Union
EUR	Euro
FB	Flow-Based
FBMC	Flow-Based Market Coupling
FOM	Fixed Operating and Maintenance
FR	France
G-component	Generator-component

GHG	Greenhouse Gases
GTRE	Generalized True Random Effects
GW	Gigawatt
GWh	Gigawatt hours
HV	High Voltage
HVDC	High Voltage Direct Current
IGCC	Integrated Gasification Combined Cycle
km	Kilometer
kV	Kilovolt
kW	Kilowatt
kWh	Kilowatt hours
Mio	Million
MVA	Megavolt Ampere
MW	Megawatt
MWh	Megawatt hours
MWh_{th}	Megawatt hours thermal
NTC	Net Transfer Capacities
OCGT	Open Cycle Gas Turbine
PTDF	Power Transfer Distribution Factor
PV	Photovoltaic
RE	Random Effects
RC	Revenue Cap
SFA	Stochastic Frontier Analysis
Std. dev.	Standard deviation
TRE	True Random Effects
TSO	Transmission System Operator
TW	Terawatt
TWh	Terawatt hours
TWkm	Terawatt kilometer
TYNDP	Ten Year Network Development Plan
UK	United Kingdom
U.S.	United States of America
VOLL	Value of Lost Load
WEPP	World Electric Power Plants

1 Introduction

Over the last decades, electricity markets have been exposed to fundamental changes. In 1996, the single market directive adopted by the European Parliament and the Council of the European Union, initiated the liberalization process to increase competition and efficiency. Before the liberalization, vertically-integrated utilities in Germany were organized as regional monopolies responsible for the generation, retail and distribution within their respective regional service areas. The liberalization changed this paradigm by unbundling the firms' activities and introducing competition in the market segments of retail and generation – with distribution and transmission services remaining natural monopolies. High fixed costs, in particular, associated with electricity networks justified the operation by only one firm as opposed to multiple competitive firms. As such, regulation of network operators was introduced to impede the abuse of monopoly power. In Germany, the use of incentive regulation should prevent excessively high network access charges and increase the efficiency of network operators.

Apart from liberalization, further developments in electricity markets have had significant consequences for electricity networks. Most notably, climate protection and the abatement of greenhouse gas emissions have led to a global increase in the share of renewable energy technologies in electricity generation. In Germany, ambitious expansion targets and incentive mechanisms induced an increase in the installed capacity of renewable energy technologies by nearly 200 percent within only a ten-year period between 2009 and 2019 (Bundesnetzagentur, 2020a). This shift in the electricity generation structure poses key challenges for network operators, who are required by German law to connect and preferentially dispatch distributed renewable generation. Firstly, the stochastic nature of renewable resources, such as wind and solar, causes difficulties in balancing short-term electricity demand and supply. In addition, the decision as to where these renewable systems are placed are made irrespective of the network condition. Both stochastic generation and inefficient allocation may lead to increased grid congestion, which in turn may jeopardize network stability. As such, congestion management is crucial to ensure efficient coordination of generation and transmission services in the short- and long-term (Chao et al., 2000).

As the energy landscape continues to transform, both the regulation of network operators and the market design of congestion management faces new challenges. In this thesis, four papers are presented to shed light on the regulation of network operators and their interaction with other market participants. Thereby two chapters focus on regulating electricity distribution network operators, and two chapters concentrate on congestion management. In doing so, this research offers valuable insights into the efficiency of different market and regulatory designs, which may aid regulators in future decision-making processes.

1.1 Outline of the thesis

The thesis consists of four chapters based on single papers, of which three are written in co-authorship. Papers written in co-authorship were done so with equal contribution.

Chapter 2: Distributed generation and cost efficiency of German electricity distribution network operators (with Heike Wetzel). EWI Working Paper, No. 20/09, 2020.

Chapter 3: Unobserved technological heterogeneity among German electricity distribution network operators - a latent class analysis. EWI Working Paper, No. 21/05, 2021.

Chapter 4: Congestion management in power systems - Long-term modeling framework and large-scale application (with Joachim Bertsch and Simeon Hagspiel).[1]

Chapter 5: The relevance of grid expansion under zonal markets (with Joachim Bertsch, Tom Brown and Simeon Hagspiel). Published in The Energy Journal, Vol. 38, No. 5, 2017.

Chapter 2 focuses on the regulation of German distribution network operators by estimating their cost efficiency based on various model specifications. Electricity networks are considered natural monopolies with limited or no competition and are thus regulated. Over the last few decades, incentive regulation has been the most commonly chosen regulatory framework for gas and electricity network operators worldwide. Efficiency benchmarking provides a key element of incentive regulation to determine firm-specific efficiency targets. Efficiency targets could be determined based on various econometric models and model specifications, for example, the Stochastic Frontier Analysis (SFA). Especially SFA panel data models have been continuously developed to improve the estimation of firm-specific efficiency while considering firms' heterogeneity. Thereby,

[1]An updated version of the paper presented in this thesis is published in Journal of Regulatory Economics, Vol. 50, pp. 290-327, 2016.

model specifications differ according to their consideration and definition of efficiency and unobserved heterogeneity: Recent models can distinguish between persistent and transient efficiency while conventional models focus on one source of efficiency. However, the distinction may be relevant as regulators have to address different aspects to resolve inefficiency depending on whether inefficiency is persistent or transient.

An extensive and comprehensive data set of financial, technical, and structural characteristics of German distribution network operators from the years 2011 and 2017 is used to estimate the transient and persistent cost efficiency of German electricity distribution network operators. Thereby, the impact of different model specifications and the effect of an increased capacity of distributed generation from renewable energy sources are analyzed based on state-of-the-art stochastic frontier panel data models. The study of German distribution network operators is interesting for various reasons: First, with around 900 electricity distribution network operators, Germany has many heterogeneous network operators. Second, over the last two decades, there has been a significant and dynamic increase in distributed generation from renewable energy sources.

The results indicate an average cost reduction potential of about 12 percent in the short term and about 18 percent in the long-term when both sources of inefficiency are accounted for in a single model. Significant differences in efficiency estimates are identified across econometric models that account for only transient or persistent efficiency. Furthermore, distributed generation is an important cost driver in the production process of German electricity distribution network operators.

The results of Chapter 2 rely on the assumption of a common production process represented by a joint cost function among distribution operators. This assumption is relaxed in **Chapter 3** by applying a latent class model that explicitly allows for heterogeneous cost function parameters, i.e., technological heterogeneity among distribution network operators. As the benchmarking procedure's premise is that network operators are comparable, accounting for network operators' heterogeneity is crucial for regulators. In contrast to observed heterogeneity, the consideration of unobserved differences is far more challenging. Unobserved heterogeneity may impact network operators' performance as well as the production process. However, regulatory practice often neglects the possibility of heterogeneous technologies among distribution network operators and assumes a common technology. Considering a common technology and thus cost function across network operators implies, e.g., identical marginal costs and economies of scale for all network operators. It is questionable whether this assumption holds in practice. Some of the network operators' heterogeneity may influence the production process, leading to technological differences among network operators. The resulting technological differences may even increase due to the increasing distributed generation capacity from renewable energy sources, digital transformation, and electric vehicles. If technological heterogeneity is not accounted for, efficiency estimates would be biased, having direct financial consequences for network operators.

To address technological heterogeneity among network operators, a latent class stochastic frontier model is applied to the database of German electricity distribution network operators between the years 2011 and 2017, also used in Chapter 2. Latent class models account for technological heterogeneity by allowing for parameter heterogeneity in the cost function. The model assumes a sorting of network operators in different classes that share a common cost function. In particular, the capacity of distributed generation is analyzed as a possible source of technological heterogeneity.

The results indicate that German distribution network operators use different technologies and can be unambiguously classified into three statistically different classes. The findings show significant differences in the size, capacity of distributed generation and identify distributed generation capacity as an important driver of the network operators' technology.

While the previous chapters focus on the regulation of distribution network operators, the following two chapters address the market design of congestion management. **Chapter 4** presents a modeling framework for different regulatory designs regarding congestion management including both, the operation as well as the investment perspective in the generation and transmission sector. In liberalized power systems, generation and transmission services are unbundled, but remain tightly interlinked. Congestion management in the transmission network is crucial for the efficiency of these inter-linkages and thus an efficient coordination of short (i.e., operational) and long-term (i.e., investment) activities in the generation and grid sectors. Different regulatory designs and options are available to manage congestion. The efficiency of the market outcome crucially depends on the implemented congestion management and the exchanged information between the transmission and generation sector. Thereby, the design of the congestion management may vary in the definition of market areas (i.e., zonal vs. nodal markets), the regulation and organization of TSOs, i.e., one vs. multiple TSOs, the way of managing congestion besides grid expansion (e.g., redispatch and g-component), and cross-border capacity allocation routines, i.e., NTC vs. flow-based market coupling.

To compare and analyze the short and long-term efficiency of different congestion management designs, a generalized and flexible economic modeling framework is developed. Using the possibility to separate an integrated optimization problem into multiple levels, the decomposed model includes both the operation and the investment perspective in the generation and transmission sector and their inter-linkages. With this, the short and long-term effects of different congestion management designs can be analyzed and benchmarked against the welfare-optimal result. Thereby, implicit frictions and sources of inefficiencies in the various regulatory designs can be isolated and identified.

Subsequently, an algorithm to numerically solve the model is calibrated and used in a large-scale application of a detailed representation of the Central Western European region. The analysis of six different congestion management designs yields that compared to the first-best benchmark, i.e., nodal pricing, inefficiencies of up to 4.6 percent arise until 2030. Thereby, inefficiencies are mainly driven by the approach of determining cross-border capacities and the coordination of transmission system operators' activities.

Chapter 5 deepens an aspect of the analysis in Chapter 4 by focusing on the impact of restricted grid expansions under an imperfect market design, i.e., zonal markets. The European electricity market design is based on zonal markets with uniform prices, implying that scarce transmission capacities within these zones are implicitly neglected. Hence, no differentiated locational price signals are provided within these zones. In reality, this simplification is often inconsistent with physical grid properties, leading to congestion in the grid. In the short term, congestion is relieved by adjusting the original dispatch, i.e., redispatch. In the long-term, the functionality of zonal markets depends on grid expansions. However, grid expansion is often insufficient and delayed. In addition, fundamental changes in the supply and demand structure, climate protection efforts, and the increase of renewable generation increase the importance of sufficient grid infrastructure.

To study the relevance of grid expansion under zonal markets, the long-term model, developed in Chapter 4, is applied and calibrated to represent European zonal markets with redispatch. Thereby, expansions of the transmission grid are restricted per decade.

The results show that the incomplete market design, i.e., zonal markets, in combination with restricted grid expansion leads to a misallocation of generation capacities and the inability to transport electricity to where it is needed. Thus, intra-zonal congestion occurs due to missing grid expansion. The market design reveals its inherent incompleteness and leads to severe short and long-term distortions. Energy imbalances in some regions of up to 2-3 percent and the difficulty to reach envisaged political targets in the power sector are identified.

1.2 Discussion of applied methodology

The applied methodology was chosen to answer each chapter's specific research questions in the best suitable way. Chapter 2 and Chapter 3 apply empirical benchmarking techniques to estimate the network operators' cost efficiency. Efficiency benchmarking techniques are an essential element of German distribution network operators' regulation. In general, efficiency benchmarking methods can be divided into parametric and deterministic methods. While deterministic models assume that any deviation from the efficiency frontier is related to inefficiency, parametric models account for stochastic noise implying that deviations from the efficiency frontier may be driven either by inefficiency or stochastic noise. This thesis focuses solely on parametric methods and in particular on Stochastic Frontier Analysis (SFA) panel data models. SFA models rely on strong assumptions regarding the definition of a functional form of the underlying cost function and the error term distributions. In Chapter 2 we compare models that differ in the functional form and the consideration of heterogeneity. With this, a consistent framework to compare different models and their specific assumptions is provided. While Chapter 2 accounts for unobserved heterogeneity that impacts the performance of network operators but assumes that the underlying cost function is the same for all network operators, this assumption is relaxed in Chapter 3. Therefore, a latent class model is applied in Chapter 3 that provides the possibility of parameter heterogeneity in the cost function of distribution network operators. A general caveat of SFA panel data models is their relatively high data requirements and their complexity. Especially the data requirements are challenging as the German regulator does not publish data on the network operators. In particular, cost-related data must be collected from the annual reports of individual companies. Network operators do not provide information on their input prices (e.g., wages), which implies the abstraction from differences in input prices across German distribution network operators and the inability to account for allocative (in)efficiency.

The applied SFA models are solved using maximum likelihood or maximum simulated likelihood estimation. The complexity of the models often leads to difficulties in the estimation procedure and convergence problems of the models. Especially, this is the case for maximum simulated likelihood estimation and latent class models which are solved with maximum likelihood estimation. To ensure the convergence of the model, changes in the composition or specification of variables, the functional form or the distribution of the error term may be required. Thus, it is a balancing act to define a proper model and ensure its convergence.

In Chapter 4 and Chapter 5 the focus of the research questions shifts from the network operators' regulation to the analysis of congestion management. To depict various congestion management measures and thus the interaction between transmission and generation services an economic modeling framework based on a decomposed inter-temporal equilibrium model is used. The decomposed model with a generation and transmission level is able to represent different market designs that vary in the information exchange between the transmission and generation sector. In the nodal system, information is perfectly exchanged between the generation and transmission level as nodal prices represent all available information. To represent incomplete information, the information that is exchanged between transmission and generation firms is exogenously restricted. Thus, the analysis exclusively focuses on the impact of the implemented frictions and the resulting inefficiencies and thus neglects timing issues, such as uncertainty or sequential moving. As such, the approach substantially differs from classical fundamental equilibrium models that assume perfect information of all market participants and allows for imperfect information. However, the decomposition of the model comes at the costs of an iterative solution algorithm which requires convergence by reaching a pre-defined convergence criterion. Even though the behavior of the numerical solution algorithm suggests the model's convergence, the analytical proof of convergence and the existence of a global optimum is missing.

To ensure the tractability of the analytical and numerical analysis, we rely on further assumptions. First, we assume perfect foresight and perfect competition on the generation side and perfectly regulated (i.e., cost-minimizing) transmission system operators. Thus, we abstract from strategic behavior or timing issues as the generation and grid problems are solved simultaneously. Second, we consider the demand side as perfectly inelastic. For sure, it is a strong assumption that also impacts the magnitude of the measured inefficiencies of the various market designs. As it may be valid to assume an inelastic demand today, at least in the short term, it may be less suitable in the future (e.g., due to time-dependent tariffs for consumers). Third, we abstract from transaction costs that may result from a change in the market design and any socio-economic factors influencing the market outcome, for example, due to acceptance problems of grid expansion projects.

The assumptions made are necessary to ensure the numerical and analytical tractability of the model but leave room for extensions and improvements by future research. Nevertheless, the presented analytical and numerical approach provides a valuable tool to assess several further relevant questions, such as different forms of congestion management in other European regions or the valuation of grid expansion projects.

In general, it is important to remember the assumptions and restrictions of the empirical, analytical or numerical models when interpreting the results, especially when general conclusions should be drawn. This thesis provides a consistent framework to compare the impact of different regulatory designs of congestion management on the one hand and the efficiency benchmarking of distribution network operators on the other hand.

2 Distributed generation and cost efficiency of German electricity distribution network operators

2.1 Introduction

Electricity networks, which are considered natural monopolies with limited or no competition, are regulated. Over the last few decades, incentive regulation has been the most commonly chosen regulatory framework for gas and electricity network operators worldwide. In 2009, Germany introduced incentive regulation for electricity networks with the aim of preventing excessively high network access charges and increasing the efficiency of network operators. A key element of incentive regulation is the use of efficiency benchmarking techniques that determine individual efficiency targets for each network operator. Regulators use a variety of different econometric methods, such as data envelopment analysis (DEA), stochastic frontier analysis (SFA) and semi-parametric methods, to identify these targets (Kumbhakar and Lien, 2017). Furthermore, within the econometric methodologies, model specifications differ widely with respect to the selection of variables, the assumptions concerning the underlying functional form and the sources of inefficiency considered.

In this paper, we employ state-of-the-art stochastic frontier panel data models to investigate the influence of different model specifications on the estimated individual cost efficiency targets of a large number of German electricity distribution network operators. In particular, we focus on the performance of recently developed SFA models that account for both transient and persistent inefficiency (Colombi et al., 2014, Filippini and Greene, 2016, Tsionas and Kumbhakar, 2014), as opposed to the widely used conventional SFA models that focus only on one source of inefficiency.

A distinction between transient and persistent inefficiency is important for regulatory purposes, as policy implications for improving persistent and transient efficiency are different. Transient inefficiency, for example induced by short-term managerial misbehavior, could be addressed relatively easily by implementing appropriate incentives in the existing regulatory framework. In contrast, persistent inefficiencies indicate structural problems that may require a general adjustment of the regulatory approach (Filippini et al., 2018, Kumbhakar and Lien, 2017).

There are only a few empirical applications that consider different types of efficiency of electricity distribution network operators. Kumbhakar and Lien (2017) use a panel of Norwegian electricity distribution companies between 2000 and 2013. They find differences among models that account for different outcomes in terms of short-term, long-term and overall efficiency and conclude that a proper regulatory design is crucial to obtain correct efficiency measures. Kumbhakar et al. (2020) quantify the cost of input misallocation, differentiating between the persistent and transient technical efficiency of Norwegian electricity distribution firms between 2000 and 2016. The results show evidence of persistent inefficiency and non-negligible costs of input misallocation. Using the data of 28 New Zealand electricity distribution companies from 2000 to 2011, Filippini et al. (2018) analyze the impact of the distinction between transient and persistent efficiency components in terms of price cap regulation. Based on the identification of differences, the authors conclude that the regulator should apply differentiated efficiency measures.

For Germany, there is only one study that distinguishes between transient and persistent efficiency. Using a panel data set with 1,370 observations between 2006 and 2012, Badunenko et al. (2021) investigate the effect of restructuring electricity distribution systems following the German reunification and find that Eastern and Western distribution network operators exhibit the same transient efficiency but vary in their persistent efficiency. However, as in the few earlier studies for Germany that dealt with only one source of efficiency (Cullmann, 2012, Hess and Cullmann, 2007, von Hirschhausen et al., 2006), Badunenko et al. (2021) analyze only technical efficiency and do not consider cost efficiency.

The lack of studies on the cost efficiency of German distribution network operators is largely due to the fact that, unlike in other countries, such as Norway and New Zealand, the German regulator does not publish data on the network operators included in its efficiency benchmarking. In particular, cost-related data must be collected from the annual reports of individual companies, which is a cumbersome process. In this paper, we use a unique and comprehensive panel data set of the financial, technical and structural characteristics of German distribution network operators from 2011 to 2017 that allows us to estimate both persistent and transient cost efficiency for a large segment of German electricity distribution network operators.

Such an analysis is particularly interesting for two reasons: First, with around 900 electricity distribution network operators, Germany has a high number of heterogeneous network operators. Second, over the last two decades, there has been a significant and dynamic increase in distributed generation from renewable energy sources. As the major share of distributed generation is connected to the distribution network and given that network operators are legally obliged to connect and preferably dispatch distributed generation, it is likely that they will be financially affected by an increase of distributed generation. In the short term, the stochastic nature of decentralized generation makes it difficult to ensure safe and reliable grid operation, which increases the need for active intervention by the operator. In the long term, network expansion, modernization and innovation are essential.

The development of distributed generation from renewable energy sources has also been taken into account by the German regulator. While accompanying studies to the German regulatory benchmarking do not find a significant influence of distributed generation on the costs of electricity distribution network operators in the first and second regulation period, they do find one in the third regulation period (Sumicsid and EE2, 2008, Swiss Economics and Sumicsid, 2014, Swiss Economics et al., 2019). For the regulatory benchmarking, however, only cross-sectional data from a single year and a small number of network operators are used.

In this context, our study is the first to use a large panel data set to analyze the impact of a large amount of distributed generation from renewable energy sources on the total costs of German electricity distribution network operators. Second, we analyze how different model specifications in terms of assumptions regarding the underlying functional form and the sources of inefficiency influence the estimated cost function parameters and both transient and persistent cost efficiency estimates. As electricity generation from renewable energy sources is increasing worldwide and electricity network regulation is becoming increasingly complex, our results are of high interest for not only German policy makers but also electricity network regulators globally.

The remainder of the paper is structured as follows: Section 2.2 offers a brief description of the German incentive regulation and the development of distributed electricity generation, while Section 2.3 presents the methodology. Section 2.4 describes the data, and Section 2.5 reports the empirical results. Finally, Section 2.6 summarizes the main results and concludes.

2.2 The German incentive regulation and distributed electricity generation

In 2009, Germany introduced incentive regulation with the aim of simulating competition among electricity distribution network operators and providing incentives to increase their cost efficiency. To prevent network operators from setting excessively high network tariffs and earning monopoly rents, an annual revenue cap is assigned to each operator.[2]

The starting point of this revenue cap is a regulatory period of five years. The first regulatory period was from 2009 to 2013, the second was from 2014 to 2019 and the third started in 2019 and will end in 2023. The third year of a regulatory period is also the base year t = 0 for the following period. This means that the regulator takes the costs of this year as a starting value for determining the annual revenue caps in the following regulatory period and in a first step breaks them down into two elements: permanently non-controllable costs and controllable costs. Examples of permanently non-controllable costs include concession fees and taxes neither of which can be influenced by a network operator.

In a second step, the regulator carries out an efficiency benchmarking of the controllable costs among the network operators and on this basis further breaks down operator-specific controllable costs into temporarily non-controllable costs and controllable costs. From an efficiency benchmarking perspective, the temporarily uncontrollable costs represent efficient costs, whereas the controllable costs represent inefficient costs. If, for example, the efficiency score for a network operator obtained from the benchmarking is 0.8, the regulator considers 80 percent of the operator's controllable costs efficient and 20 percent inefficient.

Finally, on the basis of the cost evaluation and the efficiency benchmarking, the regulator sets an annual revenue cap RC_{it} for network operator i in year t according to the following formula:

$$RC_{i,t} = C_{pnc,i,t} + (C_{tnc,i,0} + (1 - DF_{i,t})C_{c,i,0} + \frac{B_0}{T})(\frac{CPI_t}{CPI_0} - PF_t) + X_{i,t}, \tag{2.1}$$

where $C_{pnc,i,t}$ denotes the permanently non-controllable costs in year t and $C_{tnc,i,0}$ and $C_{c,i,0}$ denote the temporarily non-controllable costs and the controllable costs, respectively, in the base year $t = 0$. DF represents a distribution factor that increases from 0.2 up to 1 within a regulatory period and thus defines a reduction path for the controllable (i.e., the inefficient) costs. Furthermore, the formula includes an efficiency bonus

[2]The following description in based on the German Incentive Regulation Ordinance (ARegV). More details can also be found in Swiss Economics et al. (2019).

B_0 for very efficient network operators that is distributed over the length of the regulation period T, an inflation correction via the consumer price index $\frac{CPI_t}{CPI_0}$, a general sectoral productivity factor PF_t and some additional factors summarized in $X_{i,t}$, which are not discussed in detail here for simplicity.[3] Overall, the revenue cap formula shows that the German incentive regulation and thus the network operator-specific revenues strongly depend on a correct determination of the individual efficiency scores of network operators.

To determine individual efficiency scores, electricity network regulators worldwide use non-parametric, parametric and semi-parametric benchmarking techniques. These techniques differ in terms of their flexibility and consideration of stochastic noise. Non-parametric techniques such as DEA are extremely flexible and do not require the assumption of a functional form. However, non-parametric techniques do not take stochastic noise into account and therefore carry the risk of data errors. In contrast, parametric techniques such as SFA take stochastic noise into account but have the disadvantage of requiring a functional form, which can lead to specification errors. The German regulator, the Bundesnetzagentur, uses a mixture of both methods to determine individual efficiency scores. Each network operator is assigned the best efficiency score out of four model specifications that result from two DEA and two SFA specifications ("best of four principle").

Furthermore, the model specifications differ considerably in terms of their consideration of the heterogeneity of network operators. Since both supply tasks and regional characteristics can vary considerably from one network operator to another, performing only a simple comparison of the costs would lead to misleading results. For example, network operators operate in different landscapes, have different network sizes, different numbers of customers and so forth. Therefore, it is crucial to consider both the individual and regional characteristics of network operators in the efficiency benchmarking. Accordingly, structural variables that are expected to have an impact on costs must be included in the benchmarking process. In Germany, the Incentive Regulation Ordinance (ARegV) defines the variables that have to be considered (§13 AregV). These variables include the number of exit and metering points, the length of underground cables and overhead lines, the annual peak load, the supply area and the capacity of distributed generation.

In Germany, the consideration of distributed generation is of particular interest, as there has been a significant and dynamic increase in distributed generation in recent years. This increase is mainly due to the support scheme established by the Renewable Energy

[3] The additional factors include a cost of capital premium for investments after the base year, a quality factor to ensure quality of supply, volatile cost shares and surcharges or discounts on the regulatory account. For a detailed description of all factors included in the revenue cap formula and the calculation of the efficiency bonus, see Incentive Regulation Ordinance (ARegV).

Sources Act (EEG). A guaranteed feed-in tariff for 20 years, a connection obligation and a preferential feed-in have led to a rapid increase in distributed generation. From 2008 to 2017, the number of renewable power plants rose from about 0.5 million to more than 1.7 million. As shown in Figure 2.1, this increase in facilities is related to an increase in installed capacity from about 35 GW in 2008 to about 108 GW in 2017 (Bundesnetzagentur, 2019a). Renewable energies also play an important role in the consumption mix: In 2017, 36 percent of gross electricity consumption came from renewable sources (Federal Ministry for Economic Affairs and Energy, 2019).

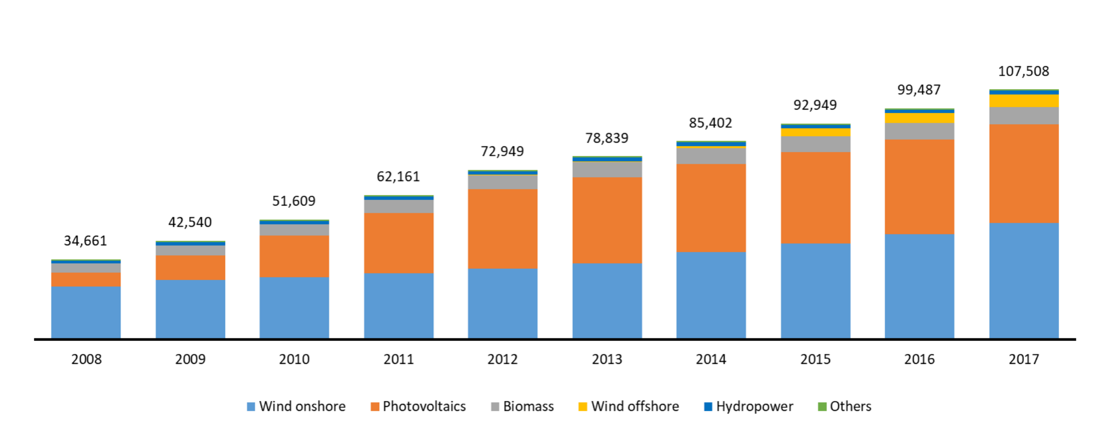


FIGURE 2.1: Installed capacity of distributed generation eligible for EEG payments in megawatt [MW] Source: Bundesnetzagentur (2019a)

A substantial majority of 98 percent of installed renewable power plants are connected to the distribution network (E-Bridge et al., 2014). The German distribution network is operated by almost 900 network operators, which are extremely heterogeneous in terms of size and structure. Due to the increase in distributed generation, there has been an intensive discussion concerning the associated cost consequences for network operators. Challenges arise not only from the pure number of renewable power plants and their capacity but also from the associated increase in volatility in power generation and the changing structure of consumers who simultaneously become generators. Stabilizing the system therefore requires more active intervention by network operators (BDEW, 2016), which could lead to increasing operating costs. Furthermore, increasing connections of renewable power plants could result in a need for network expansions. Due to different network structures (e.g., in urban vs. rural areas), the extent of the necessary network expansions and thus the investment costs may differ considerably (E-Bridge et al., 2014). The heterogeneity of network operators and, moreover, the unequal distribution of distributed generation among network operators may lead to these operators not being equally affected by the increase in distributed generation. Hence, in the following, we analyze whether distributed generation has a cost effect and whether it causes differences in the efficiency of network operators.

2.3 Methodology

In line with previous studies on the cost efficiency of electricity distribution network operators, we define the total costs of an electricity distribution network operator as a function of input prices, outputs and a number of network characteristics that account for the heterogeneity of an operator's network environment. Using two outputs and five network characteristics, the total cost function can be written as

$$TC = C(QE, QC, ND, SC, DG, I, East, DT) \tag{2.2}$$

where TC denotes the total costs, QE is the amount of electricity supplied, QC is the number of connection points, ND denotes the network density, SC is the share of underground cable in the total network and DG denotes the installed capacity of distributed generation. Furthermore, we include two dummy variables, I and $East$. I refers to structural differences in terms of whether or not operators are integrated (i.e., operators that operate an electricity and a gas distribution network or only operate an electricity distribution network). In addition, $East$ captures structural differences of operators located in East or West Germany. Finally, the dummy variables DT capture changes over time.

Due to a lack of data, we do not include input prices in our analysis. Thus, we assume that there exist no significant differences in input prices across distribution network operators in Germany, an assumption that has been used in other electricity network studies and is also used by the German regulator (Filippini and Wetzel, 2014, Swiss Economics et al., 2019). A detailed description of the variables and their expected impact on total costs is provided in the following section (see Section 2.4).

In the next step of the analysis, we define a functional form for the total cost function. In empirical studies on electricity network operators, the Cobb-Douglas and translog functional forms are most commonly used. The Cobb-Douglas functional form is relatively simple and easy to estimate. However, it has the drawback that it imposes a number of a priori restrictions on the structure of the underlying technology. A translog functional form is more flexible but more difficult to estimate. Particularly when it comes to highly correlated variables, the translog function form can easily suffer from multicollinearity problems (Filippini et al., 2018). For reasons of comparison and to investigate the impact of different functional forms on the inefficiency estimates, we use both the Cobb-Douglas and the translog functional forms in our analysis. Based on Equation 3.5, the translog cost function can be written as

$$
\begin{aligned}
lnTC_{it} =& \beta_0 + \beta_{QE} lnQE_{it} + \beta_{QC} lnQC_{it} + \beta_{ND} lnND_{it} + \beta_{SC} lnSC_{it} + \beta_{DG} lnDG_{it} \\
& + 0.5\beta_{QEQE}(lnQE_{it})^2 + 0.5\beta_{QCQC}(lnQC_{it})^2 + 0.5\beta_{NDND}(lnND_{it})^2 \\
& + 0.5\beta_{SCSC}(lnSC_{it})^2 + 0.5\beta_{DGDG}(lnDG_{it})^2 + \beta_{QEQC} lnQE_{it} lnQC_{it} \\
& + \beta_{QEND} lnQE_{it} lnND_{it} + \beta_{QESC} lnQE_{it} lnSC_{it} + \beta_{QEDG} lnQE_{it} lnDG_{it} \\
& + \beta_{QCND} lnQC_{it} lnND_{it} + \beta_{QCSC} lnQC_{it} lnSC_{it} + \beta_{QCDG} lnQC_{it} lnDG_{it} \\
& + \beta_{NDSC} lnND_{it} lnSC_{it} + \beta_{NDDG} lnND_{it} lnDG_{it} + \beta_{SCDG} lnSC_{it} lnDG_{it} \\
& + \beta_I I_i + \beta_{East} East_i + \beta_t T_t + \epsilon_{it}
\end{aligned}
\tag{2.3}
$$

where i indicates the firm and t the time period, the βs are the unknown parameters to be estimated and ϵ_{it} is the error term. The Cobb-Douglas functional form is nested in the translog functional form with all second-order and interaction terms dropped from Equation 2.3. Later, we use postestimation likelihood ratio tests to determine which functional form represents a better fit to the data.

A large number of SFA models for panel data can be used to estimate the defined total cost function. For reasons of comparison, we use three models in our analysis. As shown in Table 2.1, the models differ in terms of the econometric specification of the error term ϵ_{it} and thus in their measurement of both persistent and transient inefficiency.

	Model I - RE	Model II - TRE	Model III - GTRE
Error term	$\epsilon_{it} = v_{it} + u_i$	$\epsilon_{it} = \omega_i + v_{it} + u_{it}$	$\epsilon_{it} = \omega_i + h_i + v_{it} + u_{it}$
	$v_{it} \sim N[0, \sigma_v^2]$	$v_{it} \sim N[0, \sigma_v^2]$	$v_{it} \sim N[0, \sigma_v^2]$
	$u_i \sim N^+[0, \sigma_u^2]$	$u_{it} \sim N^+[0, \sigma_u^2]$	$u_{it} \sim N^+[0, \sigma_u^2]$
		$\omega_i \sim N[0, \sigma_\omega^2]$	$\omega_i \sim N[0, \sigma_\omega^2]$
			$h_i \sim N^+[0, \sigma_h^2]$
Estimator			
Persistent inefficiency	$E(u_i \mid \epsilon_{i1}, \ldots, \epsilon_{iT})$		$E(h_i \mid \epsilon_{it})$
Transient inefficiency		$E(u_{it} \mid \epsilon_{it})$	$E(u_{it} \mid \epsilon_{it})$

Table 2.1: Econometric specification

The first model is the random effects (RE) model proposed by Pitt and Lee (1981). In this model, the error term consists of two components: a half-normally distributed time-invariant component u_i which measures persistent inefficiency and a normally distributed time-varying component v_{it} which captures random noise. The model estimates are obtained by maximum likelihood estimation, and, as proposed by Jondrow et al. (1982), the individual level of inefficiency is predicted by the conditional mean of the inefficiency term u_i. The major shortcomings of this model are that it only estimates persistent inefficiency and that all time-invariant individual effects are included in the inefficiency estimates. Consequently, should any time-invariant unobserved heterogeneity exist, the RE model will tend to overestimate the level of persistent inefficiency (Greene, 2005a).

The second model is the true random effects (TRE) model developed by Greene (2005a,b). This model accounts for the shortcomings of the RE model by adding an individual random effect ω_i to the error term. As a result, the normally distributed time-invariant component ω_i captures time-invariant unobserved heterogeneity, and the half-normally distributed time-variant component u_{it} measures transient inefficiency. The model estimates are obtained by maximum simulated likelihood estimation, and, as in the RE model, the individual level of inefficiency is predicted by the conditional mean of the inefficiency term u_{it}. This specification separates time-invariant unobserved heterogeneity from time-varying inefficiency and therefore addresses one of the two limitations of the RE model. However, since all time-invariant individual effects are treated as unobserved heterogeneity, any persistent inefficiency is not included in the inefficiency term. Consequently, while the RE model tends to overestimate inefficiency, the TRE model tends to underestimate it (Farsi et al., 2006).

This problem is accounted for by the third model, the recently proposed generalized true random effects (GTRE) model (Filippini and Greene, 2016). The GTRE model accounts for both persistent and transient inefficiency by adding a fourth component h_i to the error term. Thus, h_i, captures persistent inefficiency and is assumed to be half-normally distributed. As before, v_{it} accounts for random noise, ω_i reflects unobserved time-invariant heterogeneity and u_{it} measures transient inefficiency. As for the TRE model, the model estimates are obtained by maximum simulated likelihood estimation, and the individual levels of persistent and transient inefficiency are predicted by the conditional mean of the corresponding inefficiency terms u_{it} and h_i following the approach outlined in Filippini and Greene (2016).

Individual persistent efficiency scores are calculated as $PE_i = \exp(-\hat{u}_i)$ in the RE model and as $PE_i = \exp(-\hat{h}_i)$ in the GTRE model. Individual transient efficiency scores are calculated as $TE_{it} = \exp(-\hat{u}_{it})$. A value of one indicates 100 percent efficiency, and a value lower than one indicates some degree of inefficiency.

Despite the fact that the GTRE model is the most sophisticated of the three models and addresses several of the shortcomings of the other two models, we retain the RE and the TRE models in our analysis for reasons of comparison. The RE and TRE model are still extremely popular in the field of parametric efficiency analyses, and one contribution of our analysis is showing the impact of different modeling approaches on the estimated efficiency scores in a regulatory framework.

2.4 Data

For our analysis, we use a comprehensive and unique database of German electricity distribution network operators from 2011 to 2017. Financial data are obtained from the publicly available financial reports of energy firms that operate an electricity distribution network. By law, vertically integrated energy firms in Germany have to publish separate accounts concerning their activities in electricity and gas distribution (§6b Energy Industry Act (EnWG)). These separate accounts allowed us to collect financial data that are directly connected to the operation of electricity distribution networks and are not combined with other activities that these firms engage in. In addition, we extended the database with technical data from the professional data provider ene't and with data on renewable power plants from the renewable power plant installation register maintained by the four German transmission system operators (50 Hertz Transmission GmbH et al., 2018). The register contains all renewable power plants that are subsidized by the EEG, which applies to more than 95 percent of all renewable power plants in Germany (Bundesnetzagentur, 2019a, Federal Ministry for Economic Affairs and Energy, 2019).

Year	2011	2012	2013	2014	2015	2016	2017
Network operators [number]							
Sample	299	318	326	348	354	351	113
Germany	869	883	883	884	880	875	878
Percent	34	36	37	39	40	40	13
Network length [million km]							
Sample	0.7	0.8	0.8	0.8	1.1	1.1	0.4
Germany	1.9	1.8	1.8	1.8	1.8	1.8	1.8
Percent	37	44	44	44	61	61	22
Connection points [million]							
Sample	21.8	22.6	22.9	25.4	31.7	32.8	13.0
Germany	47.7	48.8	49.9	50.1	50.3	50.7	50.5
Percent	44	46	46	51	63	65	26
Distributed generation [GW]							
Sample	33.8	39.2	38.2	44.3	58.8	66.2	14.9
Germany	62.2	72.9	78.8	85.4	92.9	99.5	107.5
Percent	37	54	48	52	63	67	14

Sources: Bundesnetzagentur (2013c, 2014a,b, 2016a,b, 2017, 2019a,b)

Table 2.2: Sample overview

Table 2.2 shows the number of observed electricity distribution network operators, the total network length, the total number of connection points and the total installed capacity of distributed generation included in our sample by year. In addition, we also present the total numbers for Germany for reasons of comparison. Our sample contains 2,109 observations from 453 distribution network operators. Notably, there are much fewer observations in 2017 than in the other years, as not all firms had published their annual reports for 2017 at the time of data collection. In consequence, the volumes of the other variables are also lower in 2017 than in 2016. Overall, the table shows that our sample covers a large proportion of the German electricity distribution sector. In particular, in 2015 and 2016, our sample represents more than 60 percent of the sector in terms of network length, connection points and distributed generation. The scope of our dataset thus renders it quite unique, since, in contrast to other countries, no detailed publicly available data on electricity network operators are provided by the German regulator.

Table 2.3 shows descriptive statistics of the variables used in the total cost function defined in Equation 2.3. Total costs (TC) are defined as the sum of variable and capital costs. The variable costs consist of personnel expenses, material expenses and other operating expenses, while the capital costs consist of depreciation and opportunity cost of capital. The opportunity cost of capital is calculated by multiplying fixed assets by the interest rate paid on long-term debt. All monetary values are adjusted for inflation using the consumer price index for Germany and are stated in year 2010 Euros. To define the cost function, we consider two outputs: the number of connection points (QC) and the electricity transferred in gigawatthours (QE). These two outputs reflect the two elements of the joint service of electricity distribution: network connection and electricity supply (Neuberg, 1977). Numerous empirical studies on electricity distribution networks have used this output combination (Cullmann, 2012, Filippini and Wetzel, 2014, Growitsch et al., 2012, Hess and Cullmann, 2007).

	Mean	Median	Std. dev.	Minimum	Maximum
Total costs (million 2010€)	61.54	11.53	263.60	0.25	5,235.34
Electricity transferred (GWh)	1,993.72	244.93	10,191.12	4.76	247,549.60
Connection points (thousand)	80.69	18.58	273.13	0.45	4,965.61
Distributed generation (MW)	140.10	14.70	908.11	0.35	16,120.73
Network density (connection points/network km)	37.12	32.97	24.67	6.49	268.97
Share of cable (cable km/network km)	0.93	0.97	0.09	0.14	1.00
Integrated	0.80			0	1
East	0.19			0	1

Table 2.3: Descriptive statistics

Furthermore, to account for the rapidly increasing number of renewable power plants connected to German electricity distribution networks, we include the variable distributed generation (DG). DG is measured in megawatts and comprises the installed capacity of wind energy (on and offshore), solar power, biomass, hydropower, deep geothermal energy, mine gas, landfill gas and sewage gas. Since connecting fluctuating renewable energy sources to the network incurs both connection and system stability costs, we expect a positive sign for the distributed generation coefficient, indicating higher costs for distribution operators with a higher amount of distributed generation connected to the network.

Figure 2.2 presents frequency distributions of the two outputs and the distributed generation. For clarity, the graphs are limited to the 90 percent percentiles and thereby show that approximately 90 percent of network operators transfer less than 3,000 GWh electricity, have fewer than 165,000 connection points and less than 80 MW installed capacity of distributed generation connected to their network in all years. The right-skewed distributions shown in Figure 2.2 illustrate the high number of small network operators in the sample. The median values in Table 2.3 show that about 50 percent of the distribution network operators transferred less than 245 GWh electricity, have fewer than 19,000 connection points and less than 15 MW installed capacity of distributed generation connected to their network in all years. These numbers show the highly fragmented structure of the German electricity distribution sector. Nevertheless, our data set also includes a number of very large distribution network operators. The largest network operator in the sample is Westnetz, which transferred 247 TWh of electricity and had almost 5 million connection points in 2015. With more than 16 GW installed capacity of distributed generation connected to its network, Avacon AG is the distribution network operator with the highest capacity of distributed generation in the sample.

In addition to the large differences in the levels of the two outputs and the installed capacity of distributed generation connected to the networks, we also see large differences in the other structural variables included in our analysis. Network density varies from very dense networks with up to 270 connection points per network km to very low density networks with only six connection points per network km. A similar picture emerges for the share of underground cables in the total network, which ranges between 14 and 100 percent. We expect a negative sign for the network density coefficient, indicating that networks with higher density can benefit from density effects and can therefore operate at lower costs than other networks (see, e.g., Filippini and Wetzel (2014)). In contrast, we expect a positive sign for the cable share coefficient, indicating higher construction and operating costs for networks with a higher share of underground cables (see, e.g., von Hirschhausen et al. (2006); Hess and Cullmann (2007)).

The variable 'integrated' equals one if a network operator operates both an electricity and a gas distribution network and zero otherwise. As indicated in Table 2.3, about 80 percent of the electricity distribution operators in our sample are integrated (i.e., operate an electricity and gas distribution network). Assuming that integrated operators can benefit from scope economies, we expect a negative sign for the integrated coefficient,

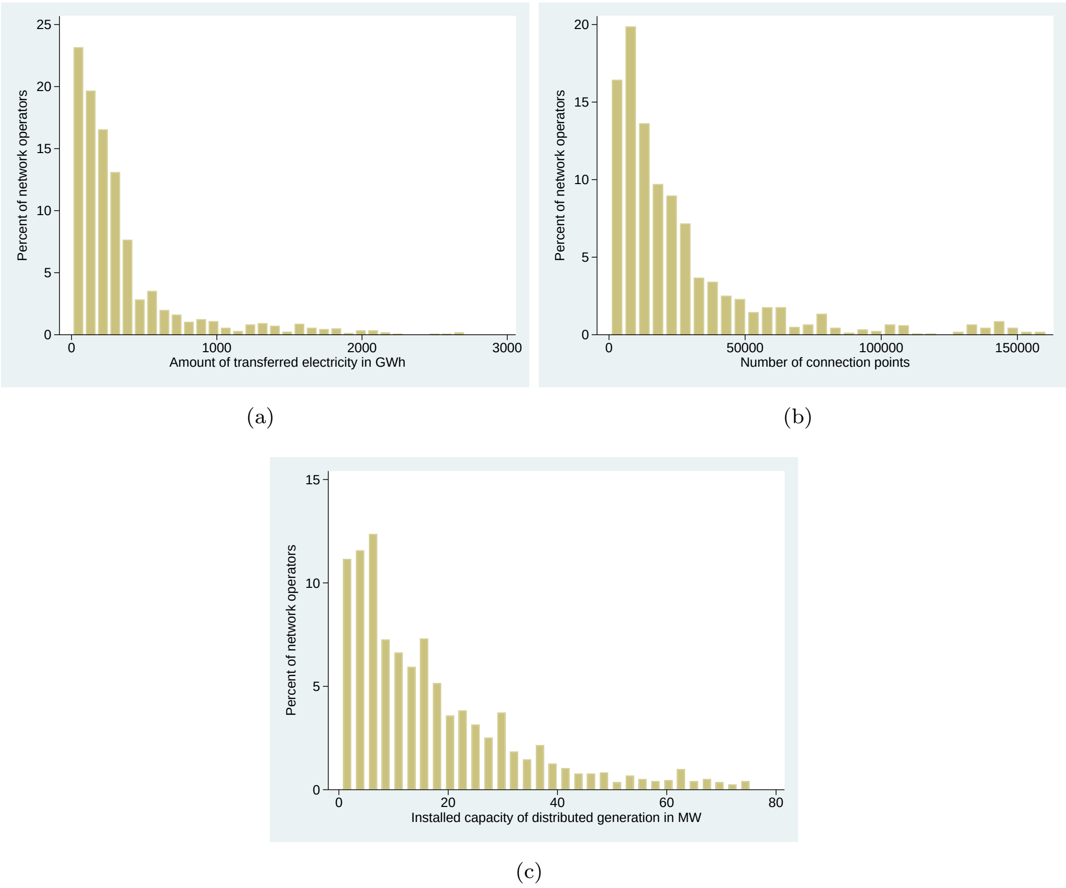


FIGURE 2.2: Frequency distributions of transferred electricity [GWh] (a), connection points (b) and distributed generation [MW] (c)

indicating lower costs for integrated operators than for non-integrated operators (see, e.g., Fetz and Filippini, 2010, Triebs et al., 2016).

Finally, the variable ‘east’ equals one if the network is located in East Germany (including Berlin) and zero otherwise. By including this variable, we account for differences that might stem from the history of socialism and the related modernization investments in East Germany after the German reunification. Thus, we expect a positive sign for the east coefficient, indicating higher costs for networks in East Germany than for networks in West Germany (see, e.g., von Hirschhausen et al. (2006); Hess and Cullmann (2007)). About 20 percent of the electricity networks in our sample are located in East Germany. Overall, our data show that the German electricity distribution sector is not only highly fragmented but also characterized by a high degree of heterogeneity.

2.5 Results

2.5.1 Cost function estimation

The estimated results for the Cobb-Douglas and the translog specifications of the three econometric models are presented in Table 2.4. Since total costs and the regressors are in logarithmic form and the regressors are normalized at their sample median, the coefficients can be interpreted as cost elasticities evaluated at the sample median. The first-order coefficients for the two outputs have the expected signs and are highly statistically significant in all estimations. The estimated coefficients for the electricity supplied vary between 0.142 and 0.156 in the Cobb-Douglas specification, indicating that a 10 percent increase in the number of kilowatt-hours supplied increases total costs by about 1.4 to 1.6 percent. The corresponding coefficients in the translog specification are slightly lower, indicating a total cost increase of about 0.8 percent as a result of a 10 percent increase in electricity supplied. Similar to many other studies (e.g., Badunenko et al., 2021, Filippini and Wetzel, 2014), the estimated coefficients for the second output, the number of connection points, are much higher. Across both cost function specifications, the estimated coefficients vary between 0.563 and 0.712. Hence, a 10 percent increase in the number of connection points is estimated to increase total costs by about 5.6 to 7.1 percent.

The estimated coefficients for the network characteristics (network density, share of cable and integrated network operator) also have the expected signs, and most of them are statistically significant at least on the 10 percent level. Only in the Cobb-Douglas specification the coefficients for the share of cable are not statistically significant. Overall, our estimation results indicate that network operators with a higher network density and integrated network operators that provide both electricity and gas supply services have lower costs, while network operators with a higher share of cable face higher costs. In particular, for integrated network operators, the estimated coefficients suggest costs that are about 4 to 9 percent lower.

As expected, the estimated coefficients for the dummy variable "east" are positive in all models. However, within the translog specification, only one of three coefficients is statistically significant, indicating costs about 5 percent higher for operators located in East Germany compared to operators located in West Germany. Within the Cobb-Douglas specification, all coefficients are statistically significant and indicate a cost difference of about 5 to 7 percent. The estimated time dummy coefficients are very similar across all models and suggest a significant cost increase over time. In 2017, the network operators

	Cobb-Douglas			Translog		
Variable	RE	TRE	GTRE	RE	TRE	GTRE
Constant	15.789***	16.288***	16.303***	15.728***	16.172***	16.134***
	(0.024)	(0.017)	(0.016)	(0.029)	(0.031)	(0.016)
$ln\,QE$	0.142***	0.144***	0.156***	0.080***	0.080***	0.084***
	(0.014)	(0.007)	(0.007)	(0.021)	(0.008)	(0.008)
$ln\,QC$	0.657***	0.581***	0.563***	0.712***	0.629***	0.635***
	(0.018)	(0.009)	(0.009)	(0.031)	(0.011)	(0.011)
$ln\,ND$	-0.265***	-0.267***	-0.264***	-0.236***	-0.282***	-0.299***
	(0.020)	(0.010)	(0.010)	(0.036)	(0.014)	(0.013)
$ln\,SC$	0.068	0.014	0.010	0.318**	0.114*	0.193***
	(0.067)	(0.038)	(0.037)	(0.153)	(0.060)	(0.053)
I	-0.066***	-0.090***	-0.064***	-0.055*	-0.068***	-0.038***
	(0.024)	(0.010)	(0.010)	(0.029)	(0.011)	(0.010)
$ln\,DG$	0.158***	0.224***	0.225***	0.173***	0.221***	0.219***
	(0.012)	(0.005)	(0.005)	(0.017)	(0.006)	(0.006)
East	0.060**	0.052***	0.073***	0.004	0.008	0.050***
	(0.025)	(0.010)	(0.010)	(0.030)	(0.011)	(0.010)
2012	0.014	-0.001	0.006	0.016	0.010	0.011
	(0.023)	(0.021)	(0.020)	(0.024)	(0.021)	(0.019)
2013	0.064***	0.046***	0.051***	0.065***	0.059***	0.060***
	(0.018)	(0.018)	(0.017)	(0.020)	(0.018)	(0.016)
2014	0.086***	0.062***	0.068***	0.087***	0.076***	0.076***
	(0.018)	(0.017)	(0.016)	(0.019)	(0.017)	(0.016)
2015	0.098***	0.069***	0.077***	0.099***	0.082***	0.085***
	(0.020)	(0.017)	(0.017)	(0.022)	(0.018)	(0.017)
2016	0.162***	0.133***	0.139***	0.163***	0.148***	0.146***
	(0.021)	(0.018)	(0.018)	(0.022)	(0.019)	(0.017)
2017	0.181***	0.142***	0.155***	0.179***	0.161***	0.162***
	(0.028)	(0.024)	(0.023)	(0.034)	(0.025)	(0.023)
Log likelihood	6.221	100.948	91.792	66.793	167.132	139.139

Notes: To conserve space the first-order coefficients are presented only. The second-order and interaction coefficients are available from the authors upon request. Standard errors in parentheses. ***, ** and *: Significant at the 1%-, 5%-, and 10%-level. The estimations have been performed in NLOGIT 6.

Table 2.4: Estimation results

are estimated to have had costs about 14 to 18 percent higher when compared to the costs in 2011.

Finally, with regard to distributed generation, all coefficients are statistically significant at the 1 percent level and show a positive sign. The coefficients vary between 0.158 and 0.225, which indicates that a 10 percent increase in the installed capacity of distributed generation results in a total costs increase by about 1.6 to 2.3 percent. These results show that distributed generation is a significant cost driver in the production process of German electricity distribution network operators and thus should be included in the benchmarking procedure.

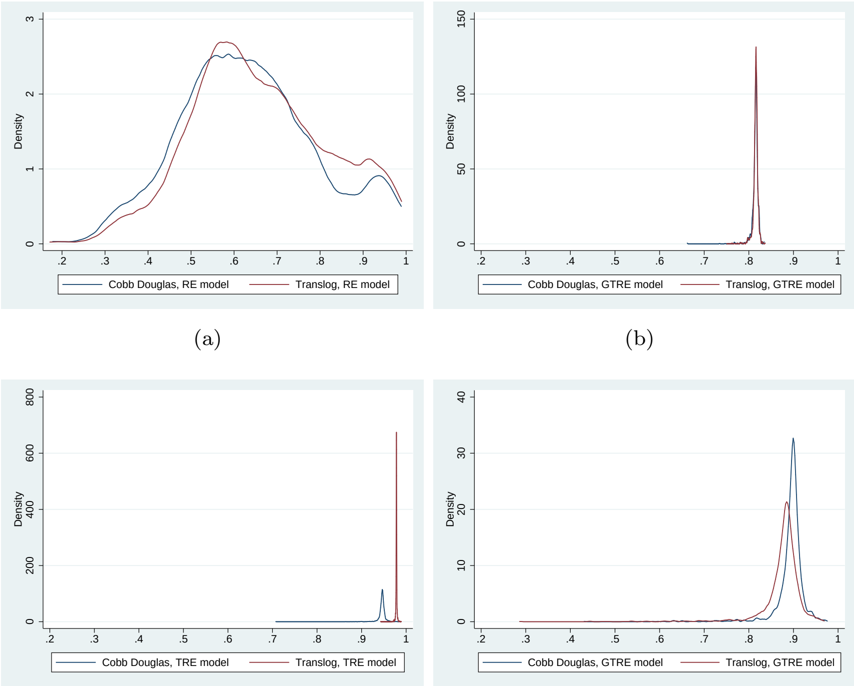
2.5.2 Cost efficiency

Table 2.5 shows descriptive statistics for both the transient and persistent efficiency scores, which are estimated with the three econometric and the two functional form specifications. First, it can be seen that the estimates from the Cobb-Douglas and the translog specifications are very similar in all models. The highest difference in the average efficiency level of about 3.25 percentage points is shown for the TRE model. This result indicates that with respect to the estimated efficiency levels, the choice between a Cobb-Douglas and a translog specification when it comes to functional form is of minor importance in the case of German electricity distribution network operators. This result is also supported by the Kernel density estimates shown in Figure 2.3. For all models, the distribution obtained from the Cobb-Douglas specification is very similar to the distribution obtained from the translog specification. Nevertheless, post-estimation likelihood-ratio tests indicate that, in all models, the translog specification is preferred over the Cobb-Douglas specification.

	Mean	Std. dev.	Minimum	Maximum
Persistent efficiency				
Cobb-Douglas, RE model	0.6343	0.1597	0.1782	0.9888
Translog, RE model	0.6601	0.1583	0.1716	0.9895
Cobb-Douglas, GTRE model	0.8146	0.0082	0.6619	0.8346
Translog, GTRE model	0.8153	0.0059	0.7505	0.8369
Transient efficiency				
Cobb-Douglas, TRE model	0.9458	0.0096	0.7073	0.9844
Translog, TRE model	0.9783	0.0014	0.9427	0.9892
Cobb-Douglas, GTRE model	0.8933	0.0350	0.4309	0.9760
Translog, GTRE model	0.8750	0.0478	0.2861	0.9693

Table 2.5: Cost efficiency scores

A completely different picture emerges when comparing the estimated efficiency values across the three econometric models. Focusing on the mean values obtained with the translog specification, our results show that the mean values of persistent efficiency vary from 66.01 percent in the RE model to 81.53 percent in the GTRE model. This difference of about 15.52 percentage points is due to the fact that, in the RE model, all unobserved time-invariant heterogeneity is included in the efficiency scores, while this is not the case in the GTRE model. Hence, our results indicate that the RE model significantly underestimates the persistent efficiency of German electricity distribution operators.



(a) (b) (c) (d)

FIGURE 2.3: Kernel density estimates of persistent efficiency (a,b) and transient efficiency (c,d)

A similar picture emerges for the estimates of transient efficiency obtained from the TRE and GTRE models. For the translog specification, the mean value in the TRE model is about 10.33 percentage points higher than in the GTRE model. As the GTRE model is the more sophisticated model in that it accounts for both persistent and transient efficiency, this result suggests that the TRE model overestimates the transient efficiency of German electricity distribution operators.

In addition to the differences in the efficiency levels, possible differences in the efficiency rankings are also relevant (see, e.g., Filippini et al., 2018). Table 2.6 reports the correlations among the estimated efficiency levels estimated by the three econometric and the two functional form specifications. The correlation coefficients for the Cobb-Douglas and the translog specifications vary between 0.88 and 0.97 in the three econometric specifications and thus indicate only minor differences in the efficiency rankings obtained from the two functional forms. Only slightly lower correlation coefficients are shown for transient efficiency as estimated by the TRE and the GTRE models, namely 0.88 for the Cobb-Douglas and 0.80 for the translog specification. However, the lowest correlation coefficients among the estimated efficiency levels and hence the largest differences in the efficiency ranking are observed for persistent efficiency. The correlation coefficient between the RE and GTRE model is 0.61 for the Cobb-Douglas specification and 0.65 for the translog specification.

Persistent efficiency	Cobb-Douglas, RE model	Translog, RE model	Cobb-Douglas, GTRE model	Translog, GTRE model
Cobb-Douglas, RE model	1	0.918	0.614	0.624
Translog, RE model	0.918	1	0.558	0.654
Cobb-Douglas, GTRE model	0.614	0.558	1	0.882
Translog, GTRE model	0.624	0.654	0.882	1
Transient efficiency	**Cobb-Douglas, TRE model**	**Translog, TRE model**	**Cobb-Douglas, GTRE model**	**Translog, GTRE model**
Cobb-Douglas, TRE model	1	0.965	0.878	0.815
Translog, TRE model	0.965	1	0.834	0.802
Cobb-Douglas, GTRE model	0.878	0.834	1	0.931
Translog, GTRE model	0.815	0.802	0.931	1

Table 2.6: Correlation coefficients

Overall, our results show that, in particular, the econometric specification of the model significantly influences the estimated efficiency scores for German electricity distribution operators. In comparison, the choice of the functional form is of minor importance. Nevertheless, with regard to the estimated coefficients, some differences can also be identified in this context.

In addition, if we consider the results of the translog GTRE model as being most reliable, we find an average transient efficiency of about 88 percent, while the average persistent efficiency is lower, at about 82 percent. This result suggests that German electricity distribution network operators could reduce their total costs by an average of about 12 percent by improving their short-term management performance and by about 18 percent by long-term restructuring efforts. One reason for the higher level of persistent inefficiency could be the highly fragmented structure of the sector. Indeed, the estimated results in Table 2.4 indicate that the sector is characterized by economies of scale.[4] Therefore, one way to increase the level of persistent efficiency could be to merge small operators into larger units.

2.5.3 Distributed generation

With regard to distributed generation, the results presented in Table 2.4 show that the installed capacity of renewable power plants is a significant cost driver. In addition, we are interested in whether there are significant differences in the transient efficiency

[4]Economies of scale (ES) are defined as the proportional increase in total costs brought about by a proportional increase in outputs, holding all other explanatory variables fixed: $ES = 1/(\frac{\partial lnTC}{\partial lnQE} + \frac{\partial lnTC}{\partial lnQC})$. Economies of scale are present if ES is greater than 1 (Filippini and Wetzel, 2014).

levels among network operators with a high and a low installed capacity of distributed generation. Table 2.7 presents the estimated mean values of transient efficiency for distribution networks with a very low (below 2.69 MW), a low (between 2.69 and 5.82 MW), a high (between 30.43 and 75.55 MW) and a very high capacity of distributed generation (above 75.55 MW).[5] It can be seen that, in all model specifications, the mean efficiency values are very similar for the considered four groups of distribution network operators. Furthermore, non-parametric Wilcoxon rank-sum tests indicate that at least at the 5 percent level of significance, the hypothesis that the mean transient efficiency level is the same for the four considered capacity levels of distributed generation cannot be rejected. Hence, we do not observe any significant differences in the efficiency levels of distribution network operators with high and low installed capacity of distributed generation, at least as long as the installed capacity of distributed generation is included as a cost driver in the total cost function.

	Installed capacity of distributed generation				Wilcoxon very low vs. very high	Wilcoxon low vs. high
	Very low (N=210)	Low (N=210)	High (N=317)	Very high (N=318)		
Transient efficiency	Mean	Mean	Mean	Mean	p-value	p-value
Cobb-Douglas, TRE model	0.9460	0.9460	0.9454	0.9452	0.132	0.743
Translog, TRE model	0.9783	0.9783	0.9783	0.9782	0.077	0.803
Cobb-Douglas, GTRE model	0.8933	0.8931	0.8932	0.8903	0.640	0.331
Translog, GTRE model	0.8759	0.8729	0.8767	0.8705	0.750	0.285

Table 2.7: Comparison of mean transient efficiency by installed capacity of distributed generation

2.5.4 Locational differences

Finally, we consider efficiency differences between distribution network operators located in East and West Germany. Badunenko et al. (2021) find that East German distribution network operators on average perform better than West German distribution network operators in terms of persistent efficiency but not in terms of transient efficiency. Figure 2.4 shows the Kernel density estimates of the persistent and transient efficiency scores obtained from the translog GTRE model with the estimates being displayed separately for East and West German distribution network operators. The distributions are very similar for both kinds of efficiency. This finding indicates that there are hardly any significant differences in terms of both transient and persistent efficiency between

[5] The thresholds are chosen based on the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentiles of the variable "distributed generation".

Eastern and Western electricity distribution operators. This finding is also supported by complementary Wilcoxon rank-sum tests for all other model specifications.

Regarding persistent efficiency, this result is contrary to that obtained by Badunenko et al. (2021). This inconsistency may be due to the fact that our observation period starts in 2011 and runs until 2017, while the observation period of Badunenko et al. (2021) starts in 2006 and ends in 2011. Thus, the data of Badunenko et al. (2021) are much closer to the date of the German reunification and thus to the beginning of the restructuring of the electricity sector in East Germany than ours. Furthermore, the empirical models differ: While Badunenko et al. (2021) use an input distance function approach and hence focus on technical efficiency measures, we apply a cost function approach and consider cost efficiency measures.

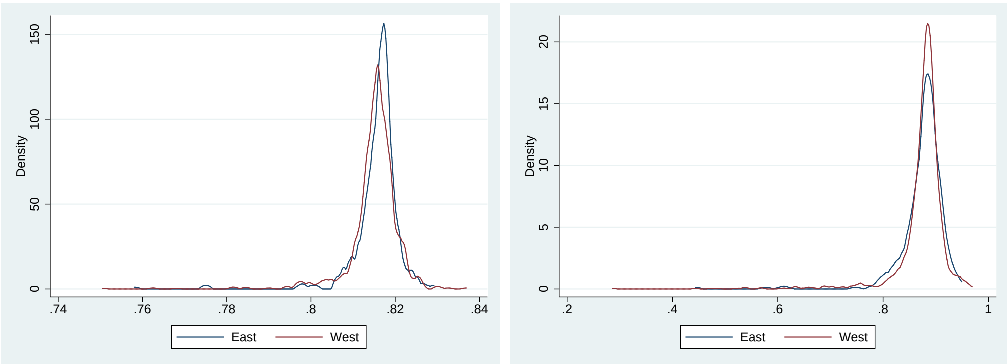


FIGURE 2.4: Comparison of distribution network operators' persistent (left) and transient (right) efficiency in East and West Germany

2.6 Conclusions

Efficiency benchmarking techniques are an essential component of incentive regulation in many network industries, such as telecommunications, water, electricity and gas. Regulators around the world use a variety of econometric methods and model specifications, and academic research is continuously contributing to the development of these techniques. Particularly for regulated sectors, recent developments in stochastic frontier panel data models are of special interest, as such advancements make it possible to distinguish between transient and persistent inefficiency in a single model.

Using a comprehensive and unique data set of financial, technical and structural characteristics of German distribution network operators from 2011 to 2017, we estimated both the transient and persistent cost efficiency of German distribution network operators. Our results indicate an average cost reduction potential of about 12 percent in the short term and about 18 percent in the long term when both sources of inefficiency are accounted for in a single model. These results are robust to the choice of a Cobb-Douglas or translog functional form. However, the comparison with econometric models that account for either only transient or only persistent efficiency shows significant differences in terms of efficiency estimates and suggests that ignoring one or the other source of inefficiency can lead to false efficiency targets in incentive regulation. In accordance with Filippini et al. (2018) and Kumbhakar et al. (2020), we therefore conclude that the German regulator should consider both transient and persistent cost efficiency in its regulatory approach and use different measures to address inefficiencies resulting from both short-term management mistakes and from long-term structural problems.

Due to the significant and dynamic increase in decentralized generation from renewable energy sources in Germany over the last two decades, we were also interested in the effects of an increasing capacity of distributed generation on the total costs of distribution network operators. As expected, we find that distributed generation is a significant cost driver in the production process of German electricity distribution network operators. Our results indicate that a 10 percent increase in distributed generation capacity leads to a total costs increase of about 1.6 to 2.3 percent. In terms of cost efficiency, however, we did not find any significant differences among network operators with high and low installed distributed generation capacity in transient efficiency. These results indicate that distributed generation has no impact on cost efficiency, at least in Germany. However, it is important, particularly in contexts with a high share of distributed generation, to take distributed generation into account in the cost function and thus in the regulatory approach.

We conclude that the distinction between transient and persistent efficiency is highly relevant for German electricity distribution companies. Transient and persistent inefficiency have different causes and different policy implications in terms of improving efficiency. Furthermore, generally speaking, regulatory approaches globally and not just in Germany should be constantly adapted to novel circumstances, such as the increasing expansion of renewable energies. Electricity distribution network operators worldwide are facing major challenges in transforming power supply systems towards a sustainable energy supply. Not only on the generation side, the increase in the use of new technologies will necessarily lead to adjustments and thus to different cost structures. Increased electric mobility or the use of heat pumps and electricity storage systems will also have an impact. Such future developments must be taken into account in further research and regulation and thus also in the efficiency comparison of electricity distribution network operators.

3 Unobserved technological heterogeneity among German electricity distribution network operators - a latent class analysis

3.1 Introduction

Electricity distribution is a common example of a market segment that exhibits characteristics of a natural monopoly. In order to avoid monopoly rents and to ensure a cost-efficient operation of the network, distribution network operators are regulated. Thereby, incentive regulation is often used in regulatory practice.

Yardstick competition was one of the first concepts of incentive regulation introduced by Shleifer (1985). Based on the idea of yardstick competition, efficiency benchmarking was implemented in Germany in the year 2009 (Swiss Economics et al., 2019). Efficiency benchmarking aims at determining the individual cost efficiency based on a comparison to the best practice, i.e., firms with the lowest costs. With this simulated competition, benchmarking should lower costs and thus increase efficiency. The efficiency of benchmarking is based on the relatively strong assumption that firms are identical or that heterogeneity across firms is accounted for correctly (Shleifer, 1985). However, as identical firms' premise is rarely fulfilled in practice, regulators are especially concerned about addressing heterogeneity across network operators to achieve reliable efficiency estimates.

In this paper, we estimate the cost efficiency of German network operators, explicitly accounting for technological heterogeneity[6] across network operators. In particular, we investigate whether technological differences across network operators are related to

[6]Throughout the paper, the technology of network operators is described by the network operators' cost function. Thus, technological heterogeneity refers to the fact that different cost function parameters may be suitable to describe heterogeneous production processes of network operators.

the capacity of distributed generation from renewable energy sources that is installed in the network operators' area. This is particularly challenging, as network operators differ according to many characteristics, some of which are beyond and some within their control. Differences may be, for example, due to technical (e.g., network length, condition of the networks), structural (e.g., number and type of customers, network density, capacity of renewable energy sources), environmental (e.g., weather conditions, landscape), and other characteristics. Some of these differences can be observed by the regulator and can thus be accounted for in the benchmarking procedure. For example, technical data on the network infrastructure must be submitted to the regulator by the network operators, and data on renewable energy sources are publicly available. In contrast, there may be factors that cause differences across network operators but are unobservable, too complex, and too expensive to account for – or no data exists (Cullmann, 2012). Amongst others, only network operators are aware of their network condition. Furthermore, environmental differences may to some extent be observable, albeit expensive (e.g., rain hours, days of snow), but a majority may be unobservable or hard to measure (e.g., ground condition, flat or mountainous landscape). Thus, a residual of heterogeneity remains, that the regulator is not able to account for, i.e., unobserved heterogeneity (Agrell et al., 2014).

As the benchmarking results have a direct financial impact, addressing unobserved heterogeneity is a crucial and challenging task for regulators. Recently developed panel data models provide several approaches to account for unobserved heterogeneity (e.g., Greene (2005a,b) and Filippini and Greene (2016)). These models assume a common production process of network operators. However, unobserved heterogeneity may also impact the production process of network operators (Cullmann, 2012). Thereby, the production process, and thus the technology of network operators, is described by the production or cost function of network operators. The aforementioned models assume a common technology, i.e., efficiency frontier, and thus neglect to account for heterogeneous cost and production functions among network operators. In doing so, homogeneous assets, services, and operations of firms (Agrell and Brea-Solis, 2017) as well as identical technological characteristics (e.g., marginal costs and economies of scale) are assumed (Llorca et al., 2014); however, it is questionable whether this holds in practice. For example, an increase in the capacity of distributed generation from renewable energy sources may only have a low cost impact for some network operators while others have a significantly higher cost increase. This difference in marginal costs may, for example, be driven by different plant types (e.g., wind vs. photovoltaic (PV)) or environmental differences (e.g., a mountainous landscape in contrast to a flat landscape). Thus, recent developments such as the dynamic and vast increase in renewable energies and the development of smart and active grids make the assumption of homogeneous conditions for network operators even more difficult (Agrell and Brea-Solis, 2017) and may increase technological heterogeneity. If any technological heterogeneity exists, current regulatory approaches either overlook the differences or misinterpret them as network operators' inefficiency.

To account for technological differences, i.e., heterogeneous cost and production functions, across network operators, so-called latent class models can be applied. Latent class models allow for firms to be sorted into different classes with a common cost or production function. The idea of combining latent class modeling with stochastic frontier analysis models is based on the work of Orea and Kumbhakar (2004) and Greene (2005b). Using data on the Spanish banking system, Orea and Kumbhakar (2004) identify four different classes to account for technological differences across banks between the years 1992 and 2002. They conclude that bank heterogeneity can be fully controlled for, which would otherwise bias efficiency results if technological differences remain uncontrolled. Greene (2005b) applies a latent class model to the U.S. banking industry. He compares different empirical approaches to consider heterogeneity across firms and concludes that a latent class model with two classes can address unobserved heterogeneity in the U.S. banking industry. Assuming that unobserved characteristics may impact German network operators' production process, Cullmann (2012) estimates a latent class model for a panel data set of German distribution network operators from the years 2001 to 2005 to disentangle inefficiency from heterogeneity. She concludes that there are indicators that small and large German distribution operators use different technologies. Llorca et al. (2014) apply a latent class analysis to account for unobserved differences in U.S. electricity distribution operators' technologies between the years 2001 and 2009. They conclude that a latent class analysis outperforms other ex-ante clustering methods and traditional panel data models to address unobserved heterogeneity. The latent class model used can account for environmental differences across network operators without explicitly including them in the cost function. Agrell and Brea-Solis (2017) estimate various latent class models of Swedish electricity distributors between the years 2000 and 2006 and contrast the results with the results of deterministic outlier detection models. In contrast to the one-step approach of, e.g., Agrell and Brea-Solis (2017), Agrell et al. (2014) propose a two-step approach. In the first step, Norwegian power distribution operators are clustered using a latent class model. In the second step, network operators' efficiency is evaluated with the group-specific frontiers using deterministic and stochastic frontier analysis. They find that efficiency estimates for the analyzed Norwegian power distributors are higher and more realistic than a conventional one-step approach. Orea and Jamasb (2017) develop a nested latent class model that can differentiate between fully efficient network operators and those that are, to some part, inefficient. At the same time, they also account for differences in the underlying technology. They estimate the model for Norwegian distribution network operators between the years 2004 and 2011. The possible importance of distributed generation from renewable energy sources was only considered by Agrell and Brea-Solis (2017). They detect a large heterogeneity of injections of power from distributed generation but find no clear pattern regarding their cost impact across classes.

With this paper, we contribute to the existing literature in various ways. First, we analyze whether German electricity distribution network operators use heterogeneous technologies represented by heterogeneous cost functions. In particular, we focus on the capacity of distributed generation from renewable energy sources as a possible source of technological heterogeneity. The analysis of German network operators provides an interesting and relevant case study, as a diverse and large number of electricity distribution network operators have been confronted with changing market environments. Amongst others, this includes the increase in renewable power plants but also the digital transformation over the previous years. This aspect has been neglected by Cullmann (2012), who also accounts for heterogeneous technologies across German network operators. Thus, we are to the best of our knowledge the first to estimate whether structural differences in the technology across German distribution network operators are related to the capacity of distributed generation from renewable energy sources. To address the research question, we apply a latent class stochastic frontier analysis to a large and unique panel of German distribution network operators between the years 2011 and 2017. Second, we compare the results of the latent class analysis with those of two alternative clusters of network operators to gain insights into the importance of accounting for heterogeneous cost function parameters across network operators. With this, we provide valuable insights for the regulation of German distribution network operators and the relevance of considering unobserved heterogeneity.

The paper is organized as follows. Section 3.2 gives a brief overview of the regulatory framework in Germany. Section 3.3 describes the methodology applied, while Section 3.4 introduces the data used in the empirical analysis. Section 3.5 presents the estimation results, and Section 3.6 concludes.

3.2 Regulatory framework in Germany

In 2009, incentive regulation for German electricity distribution operators was introduced by the German regulator, the Bundesnetzagentur, via the Incentive Regulation Ordinance (ARegV). Incentive regulation is based on the economic idea that competition between network operators should be ensured to increase efficiency. Therefore, a revenue cap is assigned to each network operator for one regulation period. By introducing a revenue cap, network operators have the incentive to reduce their costs within the regulation period to increase profits or decrease losses.[7] The calculation of the individual revenue cap is based on a cost assessment of controllable and non-controllable costs as well as an individual efficiency value. In the cost assessment, the regulator reviews the cost situation of each network operator. Subsequently, the individual efficiency value is determined by an efficiency benchmarking of the controllable costs. In practice, regulators apply a variety of parametric and non-parametric approaches. Non-parametric methods such as the data envelopment analysis (DEA) have the advantage of high flexibility as no functional form has to be specified. However, these models do not account for statistical noise, meaning any deviation from the efficiency frontier is considered as inefficiency. In contrast, parametric models such as the stochastic frontier analysis (SFA) require the definition of a functional form but can account for statistical noise. In the case of German benchmarking, a "best-of-four" approach is used: Network operators are assigned the highest efficiency value resulting from four different model specifications, two DEA and two SFA models.

As mentioned above, the efficiency of the benchmarking procedure and the reliability of efficiency estimates are based on the assumption that network operators are comparable. In this context, the German Energy Industry Act (EnWG) states that network operators' financial obligations that result from the benchmarking require their structural comparability (§21a EnWG). Germany is characterized by a high number and high heterogeneity of network operators, who differ according to individual and regional characteristics. Amongst others, the number and structure of customers, the size and condition of networks, the number and installed capacity of renewable power plants connected the grid, and the landscape and its characteristics may vary considerably across network operators. Thus, regulators cannot compare network operators based on their costs only. Some of the differences across network operators can be observed by the regulator, while others cannot. The regulator accounts for observable differences in the benchmarking by including so-called structural variables. These may consist of the area served, length of lines, and the capacity of renewable power plants connected to the grid (§13 ARegV).

Even though regulators account for observed differences across network operators, the issue of possible unobserved heterogeneity remains. The use of panel data models that can account, at least to some extent, for unobserved heterogeneity across network operators is not feasible for Germany, as data restrictions result from only two regulation periods having been completed. In consequence, the Bundesnetzagentur applies a cross-sectional

[7] The description of the regulation procedure in Germany is based on the Incentive Regulation Ordinance (ARegV) as well as Swiss Economics et al. (2019).

model, which means that network operators are not observed over time but treated as independent observations. As such, the observation and identification of unobserved heterogeneity are impossible. Furthermore, the regulatory approach does not explicitly consider the possibility that network operators differ regarding their technologies and thus their cost and production functions. Nevertheless, regulators try to account for at least some technological heterogeneity. In doing so, the Bundesnetzagentur specifies two groups of network operators according to their size. As such, a group of network operators of the same size should be more homogeneous, and thus better to compare in the benchmarking (Agrell and Brea-Solis, 2017). Network operators with more than 30,000 connected customers are obliged to participate in the regular benchmarking procedure. In contrast, network operators with less than 30,000 customers can participate in the standard benchmarking process or be part of a simplified procedure. In the simplified procedure, network operators are assigned an average efficiency value without the requirement of a cost assessment.[8] This proves that the regulator has the implicit assumption that unobserved technological differences between small and large network operators may exist (Cullmann, 2012). In the past regulation period, only a minority of 20 percent of total network operators participated in the regular benchmarking.

Such an a priori clustering may be easy to implement but is inefficient for various reasons: For one, the classification may be based on the wrong criterion and thus results in misleading groups. In the case of Germany it is questionable whether the network operators' size is correct or sufficient to describe possible technological differences across network operators. Especially over the past years, the increase in distributed generation from renewable energy sources imposes new heterogeneity across network operators: Between the years 2011 and 2017, the installed capacity of renewable power plants increased by 45 GW (Bundesnetzagentur, 2019a). Around 98 percent of all renewable power plants are connected to the distribution network (E-Bridge et al., 2014). Moreover, the capacity of renewable power plants is not evenly distributed across network operators, which could lead to some network operators being more affected than others. In the year 2016, 75 percent of the capacity of renewable power plants was installed in network areas of only 15 network operators (IAEW, 2016). The question remains as to whether such a situation may induce structural differences across network operators which are not taken into account under the current regulatory regime and by the size clustering of network operators. Furthermore, efficiency estimates for different groups may be estimated without considering the information of other groups. In this case, the regulator neglects important information as network operators still belong to the

[8] The simplified procedure should also protect smaller network operators from excessive effort to be benchmarked. This aspect is, however, not considered in this paper.

same industry and share common characteristics (Kumbhakar et al., 2015). This loss of information may induce inefficiency in the regulatory procedure.

As the benchmarking procedure has a direct financial impact on network operators, an adequate consideration of technological heterogeneity is of crucial importance. In the following, we investigate whether technological heterogeneity represented by different cost functions is present across German network operators using the latent class modeling approach. Furthermore, we analyze whether possible groups of network operators differ significantly regarding observable characteristics (e.g., the capacity of renewable power plants).

3.3 Methodology

This section first determines the cost function and describes conventional SFA panel data models that assume a common technology, i.e., cost function, and homogeneous technological characteristics for network operators. In a second step, we focus on latent class models that exhibit the possibility of differences in the cost function parameters across network operators.

3.3.1 Definition of cost function

We estimate a cost function of electricity distribution network operators that can be described by:

$$TC = TC(QE, QC, ND, DG, DT) \tag{3.1}$$

where TC denotes the total costs, QE the amount of electricity supplied, QC the number of customers connected to the network, ND the network density, and DG the capacity of distributed generation from renewable energy sources in the specific network area. DT comprises dummy variables that capture changes over time. As frequently done in the literature, QE and QC are defined as outputs (Filippini and Orea, 2014). Due to data restrictions, we abstract from input prices and assume no substantial differences in input prices across electricity distribution network operators. The cost function estimation requires the specification of a functional form. In this paper, a Cobb-Douglas function of the following form is estimated:

$$ln\,TC_{it} = \beta_0 + \beta_{QE}\,ln\,QE_{it} + \beta_{QC}\,ln\,QC_{it} + \beta_{ND}\,ln\,ND_{it} + \beta_{DG}\,ln\,DG_{it} + \sum_{t=2}^{T} \alpha_t\,DT_t + \epsilon_{it} \tag{3.2}$$

where the subscript i denotes the firm, t the time period, the β's are the unknown parameters to be estimated, and ϵ_{it} is the error term. There exists a broad range of SFA models that can estimate the cost function. In the following, we concentrate on panel data models. In contrast to cross-sectional models, panel data models observe an individual i at different time periods t and are thus able to account for unobserved heterogeneity, at least to some extent. In general, SFA panel data models can be written as:

$$y_{it} = \beta' x_{it} + \epsilon_{it} \tag{3.3}$$

where y_{it} denotes the costs, x_{it} is a vector of inputs, outputs and firm-specific characteristics and ϵ_{it} is the error term. Thereby, ϵ_{it} comprises unobserved heterogeneity, inefficiency, and random noise. The major concern of panel data models is how to disentangle inefficiency from heterogeneity and random noise. Empirical models differ in their econometric specification of the error term and thus in the way heterogeneity is accounted for (Just and Wetzel, 2020). In earlier panel data models (e.g., Pitt and Lee (1981)), the error term is defined as: $\epsilon_{it} = u_i + v_{it}$. Thereby, inefficiency, u_i, is assumed to be half-normally distributed and constant over time, while stochastic noise, v_{it}, is normally distributed and varies over time. This implies that all time-invariant firm-specific unobserved heterogeneity is considered as inefficiency, and all time-varying effects are absorbed in the stochastic noise term. In order to separate inefficiency and firm-specific heterogeneity, Greene (2005a) and Greene (2005b) introduced individual time-invariant effects with the so-called True Random Effects (TRE) model where the error term now consists of three parts: $\epsilon_{it} = v_{it} + u_{it} + \omega_i$. v_{it} denotes the normally distributed, time-varying stochastic noise term, u_{it} the half-normally distributed, now time-varying inefficiency term and the additional normally distributed term ω_i captures firm-specific, time constant random variable effects that account for firm-specific heterogeneity. The TRE model is described by:

$$y_{it} = \beta' x_{it} + \omega_i + v_{it} + u_{it} \tag{3.4}$$

Thus, all effects that are constant over time are considered as firm-specific unobserved heterogeneity. It follows, that time-invariant structural inefficiencies are also regarded as unobserved heterogeneity rather than as inefficiency.[9]

The presented models assume a common production process of network operators represented by a joint cost function, i.e., efficiency frontier. With this they neglect the possibility that unobserved heterogeneity may influence firms' technology. On the one hand, the TRE model assumes that individual efficiency frontiers may differ according to firm-specific intercepts that capture unobserved heterogeneity. On the other hand, firms are assumed to share the same technological characteristics, e.g., marginal effects, economies of scale, and scopes (Llorca et al., 2014). However, especially in an environment where observed entities are very heterogeneous, the assumption of a common cost function for all network operators may not be valid. For example, the connection of renewable power plants or new customers may induce different costs for network operators depending on, e.g., their specific network conditions or environmental characteristics (e.g., mountains, urban area). In Germany, it may be that, due to the further increase in the capacity of renewable energies, the ongoing digital transformation and also the increasing use of electric vehicles, technological heterogeneity across network operators will even increase. If common technology characteristics are assumed, even though heterogeneous technology characteristics are present, efficiency estimates may be misleading. As technological differences may be considered as inefficiency if a common cost function is wrongly assumed, models will underestimate efficiency. In the following, we discuss a modeling approach that considers differences in firms' technology and its characteristics by allowing for heterogeneous cost function parameters.

3.3.2 Latent class modeling

Latent class models allow for parameter heterogeneity in the cost function and by this account for different technologies across firms. It is assumed that a latent sorting of network operators in various classes according to differences in cost function parameters exists (Greene, 2007). Network operators that belong to the same class share a common cost function and efficiency frontier. Thus, latent class models control for heterogeneity between groups rather than for individual heterogeneity (Llorca et al., 2014). Each network operator belongs to one class in which he remains over time. Compared to an arbitrarily a priori clustering of firms, latent class models use all firms' information to determine the cost function of the different classes, irrespective to which class they are assigned.

[9] The most recent panel data models are also able to differentiate between persistent and transient efficiency (Filippini and Greene, 2016).

Under the latent class framework the overall cost function becomes:[10]

$$\begin{aligned} ln\,TC_{it} = \beta_{0j} + \beta_{QEj}\,ln\,QE_{it} + \beta_{QCj}\,ln\,QC_{it} + \beta_{NDj}\,ln\,ND_{it} + \beta_{DGj}\,ln\,DG_{it} \\ + \sum_{t=2}^{T} \alpha_{tj}\,DT_t + \epsilon_{it}|j \end{aligned} \tag{3.5}$$

where in addition to the introduced notation subscript j denotes the class. The $\beta's$ are estimated for each class j, and the error term is conditioned on class j. For the SFA model, $\epsilon_{it}|j$, is again divided into a random error term, $v_{it}|j$, which is assumed to be normally distributed and independent of the half-normally distributed inefficiency component, u_{it}.

The probability of observing individual i given that it is part of class j is described by $P(i|j)$, which is the conditional likelihood function of individual i.[11]

$$P(i|j) = \prod_{t=1}^{T} P(i,t|j) \tag{3.6}$$

As a network operator can be a member of every class, the unconditional likelihood function is the sum of the J class likelihood functions multiplied with the prior class probabilities, π_{ij}. Thus, the unconditional likelihood function for individual i is denoted as:

$$P(i) = \sum_{j=1}^{J} \pi_{ij} P(i|j) = \sum_{j=1}^{J} \pi_{ij} \prod_{t=1}^{T} P(i,t|j). \tag{3.7}$$

Prior class probabilities, π_{ij}, are estimated simultaneously with the parameters of the cost frontiers and are assumed to be constant over time (e.g., Agrell and Brea-Solis (2017)). They are parametrized as multinominal logit model:

$$\pi_{ij} = \frac{exp(\theta_j)}{\sum_{j=1}^{J} exp(\theta_j)}, \theta_J = 0, \sum_{j=1}^{J} \pi_{ij} = 1. \tag{3.8}$$

[10] Alternative model specifications differing in their functional form (Cobb Douglas vs. translog) as well as in the variable definition (e.g., density or number of distributed generation from renewable energy sources) were tested but not able to achieve convergence. Convergence problems of latent class models are reported frequently in the literature (e.g., Llorca et al. (2014), Orea and Kumbhakar (2004), Cullmann (2012), Agrell and Brea-Solis (2017)).

[11] The following model description is based on Greene (2005b) and Greene (2016).

θ_j captures all class specific parameters. The restriction, $\theta_J = 0$, is imposed as only J-1 probabilities have to be calculated to specify J probabilities. Furthermore, prior class probabilities have to sum up to one. It is important to note again that a network operator can be a member of only one class in which he remains over time. As the respective class is, however, unknown to the researcher, the prior class probability can be interpreted as the uncertainty of the researcher (Greene, 2005b).

From this, the overall likelihood function results:

$$logLF = \sum_{i=1}^{N} logP(i) = \sum_{i=1}^{N} log(\sum_{j=1}^{J} \pi_{ij} \prod_{t=1}^{T} P(i,t|j)). \tag{3.9}$$

The model is estimated using maximum likelihood. In contrast to the traditional models (e.g., TRE model), the model estimates J different cost frontiers each with an individual probability, i.e., the prior class probability. Thereby, technological heterogeneity is caused by differences in the technology parameters of each frontier, i.e., the β's, and the variance of the inefficiency, σ_u, and error term, σ_v (Agrell et al., 2014).

After the estimation, posterior class probabilities can be computed using Bayes theorem:

$$w(j|i) = \frac{P(i|j)\pi_{ij}}{\sum_{j=1}^{J} P(i|j)\pi_{ij}}. \tag{3.10}$$

Posterior class probabilities denote the probability of class j given that we have observed individual i. With the posterior class probabilities, it is possible to assign individuals to one class, i.e., the class with the highest posterior probability.

A shortcoming of latent class models is that the number of classes J can not be estimated but has to be defined by the researcher. In the empirical literature the number of classes is most commonly defined using the Akaike (AIC) or Bayesian information criterion (BIC). Both information criteria are suitable for comparing models with different classes as they both favor the model's goodness of fit while they penalize the number of included parameters in the model (Orea and Kumbhakar, 2004). The preferred number of classes is obtained by estimating models with a different number of classes in the first step and comparing them regarding their specific information criterion in the second step. The model with the lowest information criterion is preferable.

3.4 Data

To answer our research question, we construct a comprehensive database of electricity distribution network operators between the years 2011 and 2017. The database contains financial and technical data for each network operator. We collect the financial data from the network operators' balance sheets and profit and loss accounts. An amendment to the Energy Industry Act (EnWG) in 2011 strengthens the reporting provisions by enforcing network operators to make separate activity reports for gas and electricity distribution. These reporting provisions enable the direct allocation of financial data to the activities associated with electricity and gas, respectively. The professional data provider ene't provides the technical data. The installed capacity of renewable power plants is obtained from the installation register of the Renewable Energy Sources Act (EEG) published by the four transmission system operators (TSO) (50 Hertz Transmission GmbH et al., 2018). The register contains all renewable power plants that are subsidized by the EEG. The data are merged with the financial and technical data through the name of the distribution operator.

According to the ARegV, the regulator should apply outlier detection procedures based on statistical tests (e.g., Cook's distance) and delete observations above a predefined threshold. A standard threshold for Cook's Distance is a value greater than four divided by n (Bollen and Jackman, 1990). After deleting outliers that were detected based on Cook's distance and dropping observations with odd information, our unbalanced panel comprises 1,911 observations. We observe a maximum of 330 electricity distribution operators, which corresponds to more than one-third of the number of German electricity distribution operators.[12]

Table 3.1 shows the descriptive statistics of the variables included in the total cost function. Total costs, TC, the dependent variable, are the sum of labor costs (personnel expenses), capital costs (sum of material and depreciation expenses), opportunity costs of capital, and costs of other inputs. Thereby, opportunity costs are defined as fixed assets multiplied by the interest rate paid on long-term debt. Monetary values are expressed in year 2010 Euros by deflating them with the consumer price index. We define two outputs: the amount of transferred electricity, QE, measured in megawatt-hours (MWh), and the number of connected customers, QC. As can be seen in Table 3.1, there is a considerable variation in the size of the variables: While Elektra-Genossenschaft Effeltrich eG, the smallest network operator, transferred 5 GWh electricity in the year 2015

[12] In the year 2017, the Bundesnetzagentur listed 878 electricity network operators (Bundesnetzagentur, 2020b). In our sample, the number of network operators in the year 2017 deviates strongly from the other years as not all network operators have published their annual report for the year 2017 at the time of data collection.

and has approximately 1,000 connection points, the largest network operator, Westnetz, transferred almost 250,000 GWh of electricity in the year 2015 and has with nearly 5,000,000 connection points the highest value in the sample. Even though there are very large network operators, most of the sample consists of smaller network operators: 90 percent of network operators transfer less than 2,537 GWh electricity which corresponds to only 1 percent of the maximum value, and have less than 160,785 connection points. We expect a positive cost impact if the amount of supplied electricity or the number of connected customers increases (e.g., Just and Wetzel (2020)).

We also include structural variables that may influence the costs of network operators but are beyond their control. The network density, ND, is defined as the number of connection points per network kilometer. The sample includes network operators with very dense networks and network operators with a low network density: Netz Leipzig GmbH has 129 connection points per network kilometers, the densest network. In comparison, nvb Nordhorner Versorgungsbetriebe GmbH has only 8 connection points per network kilometers and thus the lowest network density. Due to density effects, we expect network operators with a higher network density to have lower costs, i.e., a negative coefficient (e.g., Filippini and Wetzel (2014), Just and Wetzel (2020)).

	Mean	Median	Std. dev.	Minimum	Maximum
Total costs (million 2010€)	52.45	11.64	232.36	0.35	5,235.34
Electricity transferred (GWh)	1,911.42	250.03	10,299.59	4.76	247,549.60
Connection points (thousand)	76.68	18.86	269.78	1.04	4,965.61
Network density (connection points/network km)	35.72	33.21	16.86	7.68	129.20
Distributed generation (MW)	96.58	14.64	690.60	0.35	14,827.64

Table 3.1: Descriptive statistics

In particular, we are interested in the impact of the installed capacity of distributed generation from renewable energy sources, DG. DG comprises the installed capacities of PV, wind energy, biomass, hydropower, deep geothermal energy, sewage gas, mine gas and landfill gas that are subsidized by the EEG. The installed capacity of renewable power plants is unequally distributed across network operators, as shown in Table 3.1. While network operators have, on average, an installed capacity of renewable power plants of 97 MW, 50 percent of the network operators have less than 15 MW installed. On the other hand, Avacon AG has an installed capacity of renewable power plants of 14.8 GW in the year 2015, which is the maximum in the sample. Due to a preferential dispatch and the obligation to connect renewable power plants to the grid, the capacity of distributed generation from renewable energy sources is expected to induce connection costs and challenges in guaranteeing a stable network. Thus, we expect a positive cost impact (e.g., Just and Wetzel (2020)).

3.5 Results

3.5.1 Latent classes' cost function estimation

To test whether the assumption of different technologies, i.e., cost functions, is suitable for German network distribution operators, we estimate the latent class model from one to six classes. We determine the number of classes using the BIC criterion (e.g., Agrell and Brea-Solis (2017), Llorca et al. (2014)). Indicated by the lowest BIC, three classes are preferred over more or fewer classes.

The cost function estimation results are obtained using maximum likelihood estimation and are shown in Table 3.2. The prior class probabilities are simultaneously estimated with the cost function estimates and indicate that 41 percent of the sample belong to class 1, 37 percent to class 3, and 21 percent to class 2 which is the smallest class. The parameters are expressed in natural logarithms and are normalized at their sample median and can thus be interpreted as estimates of cost elasticities at the sample median.

Variable	Class 1	Class 2	Class 3
Constant	16.267***	15.589***	16.107***
	(0.017)	(0.027)	(0.018)
$ln\, QE$	0.130***	0.044**	0.097***
	(0.010)	(0.021)	(0.008)
$ln\, QC$	0.600***	0.918***	0.674***
	(0.012)	(0.029)	(0.011)
$ln\, ND$	-0.400***	-0.374***	-0.447***
	(0.016)	(0.038)	(0.015)
$ln\, DG$	0.250***	0.052***	0.220***
	(0.008)	(0.014)	(0.007)
2012	0.063***	0.036***	0.038***
	(0.017)	(0.034)	(0.013)
2013	0.076***	0.140***	0.074***
	(0.017)	(0.034)	(0.014)
2014	0.088***	0.115***	0.077***
	(0.017)	(0.033)	(0.013)
2015	0.104***	0.079***	0.073***
	(0.017)	(0.032)	(0.013)
2016	0.148***	0.170***	0.134***
	(0.017)	(0.033)	(0.013)
2017	0.172***	0.354***	0.162***
	(0.023)	(0.051)	(0.019)
Prior Class Probabilities	0.410***	0.218***	0.373***
	(0.026)	(0.022)	(0.026)

Notes: Standard errors in parentheses. ***, ** and *: Significant at the 1%-, 5%-, and 10%-level. The estimations have been performed in NLOGIT 6.

TABLE 3.2: Estimation results

All coefficients have the expected sign and are highly statistically significant. We observe differences in the coefficients across the three classes, which suggests differences in marginal costs and thus the technology. In all classes, increasing the outputs leads to an increase in costs. An increase of 10 percent in the supplied electricity induces a cost increase between 0.4 and 1.3 percent. Increasing the number of connected customers by 10 percent yields a very similar cost increase of 6.0 percent for class 1 and 6.7 percent for class 3. Network operators in class 2 face higher connection costs (9.2 percent). According to Cullmann (2012) differences in connection costs may be due to differences in the customer structure (e.g., industrial vs. household customers). As expected, there are density effects across all three classes, i.e., costs decrease between 3.7 and 4.5 percent when the network density increases by 10 percent. Interestingly, the coefficient of the capacity of renewable power plants varies significantly across classes. While for network operators in class 1 and 3, an increase in the capacity of renewable power plants results in a comparatively high cost increase of 2.5 and 2.3 percent, respectively, it is with 0.5 percent lower for class 2. Environmental conditions (e.g., flat vs. mountainous landscape) and differences in the structure of the capacities of renewable energy plants across classes may be an explanation. For example, wind capacities are predominantly connected to the medium- and high-voltage network while PV is usually connected to the low-voltage network. Thereby, networks of the lower voltage levels have a lower capacity to integrate renewable power plants and are more often confronted with voltage problems than medium- and high-voltage networks. Thus, the same capacity of renewable power plants causes higher costs in the low-voltage network than in the medium- and high-voltage network (Swiss Economics et al., 2019). Positive and significant coefficients of the time dummies indicate a cost increase over time compared to 2011 for all classes.

Besides the differences in the marginal costs of network operators, economies of scale are another technology parameter that indicates whether technological heterogeneity across classes is present. Economies of scale measure the percentage cost increase that results from a simultaneous increase in all outputs by 1 percent.[13] If economies of scale are present ($ES > 1$), network operators would benefit from increasing their size, e.g., by expansion or mergers. In contrast, diseconomies of scale ($ES < 1$) indicate that a reduction of size would be beneficial for network operators. Consequently, $ES = 1$ reveals that network operators operate at their optimal size given the underlying technology (Badunenko and Kumbhakar, 2017). We observe positive economies of scale for all classes, which nevertheless differ. While classes 1 and 3 have similar economies of scale, 1.4 and 1.3, respectively, class 2 has quite lower economies of scale near unity. Thus, efficiency increases are possible for classes 1 and 3 by increasing their size, while network operators in class 2 operate near their optimal size.

[13] $ES = 1/(\frac{\partial lnTC}{\partial lnQE} + \frac{\partial lnTC}{\partial lnQC})$ (Filippini and Wetzel, 2014).

The results indicate that the latent class approach exhibits technological heterogeneity among German network operators, captured by three different classes that would have been ignored, assuming a common cost function.[14] Notably, while cost function parameters are similar in classes 1 and 3, they differ for class 2. Thereby, the main difference results from higher marginal costs to connect customers to the network and lower marginal costs of distributed generation capacity from renewable energy sources in class 2. Furthermore, economies of scale are similar in classes 1 and 3 but vary in class 2. Thus, the cost functions, i.e., technologies, of classes 1 and 3 are more alike than the cost function of class 2. In the following, we provide possible explanations for the differences between the technologies of class 2 compared to classes 1 and 3, focusing primarily on the impact of the installed capacity of renewable power plants.

3.5.2 Characteristics of Classes

Using the estimated posterior class probabilities, we assign network operators to the class with the highest individual probability. The average posterior class probability varies between 95.8 percent for class 3, 96.54 percent for class 2, and 96.92 percent for class 1. Thereby, only five observations of two network operators have a posterior class probability of less than 50 percent. Thus, we conclude that the vast majority of observations can be clearly assigned to one class (Agrell et al., 2014). This results in the following class sizes: Class 1 is with 779 observations the largest class, followed by class 3 which contains 752 observations. Class 2 comprises 380 observations and is thus the smallest class.

Table 3.3 shows the descriptive statistics of the variables included in the total cost function differentiated by class. The two outputs, electricity supplied and the number of connected customers, basically reflect network operators' size. For both variables, we observe significant differences between classes. Class 1 comprises at least some larger operators indicated by the highest average supplied electricity and the highest number of connected customers. Furthermore, the standard deviation of the outputs is highest in class 1 which shows that it consist of large and small network operators. Whereas the picture is clear for class 1, the differences between classes 2 and 3 are not apparent. The average amount of supplied electricity is similar in both classes, while the maximum amount differs significantly: The largest network operator in class 2 delivers more than twice as much electricity than the largest network operator in class 3 (68 TWh vs. 31 TWh). Regarding the number of connected customers, class 3 comprises the smallest network operators: the mean, the standard deviation, and the maximum amount of connected customers are the lowest across classes. The relatively low and similar median

[14] The impact of considering a common cost function is analyzed in Section 3.5.4.

	Mean	Median	Std. dev.	Minimum	Maximum
Electricity supplied [GWh]					
Class 1	2,637.13	250.16	15,186.99	24.68	247,549.60
Class 2	1,402.76	260.54	5,091.56	4.76	68,301.84
Class 3	1,416.69	249.05	4,105.17	9.88	31,492.64
Connection points [thousand]					
Class 1	101.85	19.78	387.94	1.63	4,965.61
Class 2	66.92	17.72	175.53	1.04	1,936.85
Class 3	55.55	18.98	111.76	1.54	754.67
Network density [Connection points/network km]					
Class 1	36.56	34.36	18.94	7.68	127.52
Class 2	36.30	32.57	18.81	11.01	129.20
Class 3	34.57	33.24	13.06	11.78	84.60
Distributed generation [MW]					
Class 1	102.89	15.33	581.70	0.47	9,781.43
Class 2	165.50	15.72	1,278.85	0.58	14,827.64
Class 3	55.21	12.34	181.42	0.35	1,635.70

Table 3.3: Descriptive statistics per class

values of the outputs across classes indicate that all classes mainly consist of smaller network operators. However, classes 1 and 2 also include some larger network operators. Referring to the estimated economies of scale per class, it seems reasonable that the smallest network operators assigned to class 3 may benefit from a size increase. The positive economies of scale of class 1 may also be driven by the large share of smaller network operators as it seems rather surprising that increasing their size would also benefit the largest network operators.

The mean and median network density show only slight variation across classes. However, the distribution of the network density is narrower in class 3 than in class 1 and 2: The standard deviation is lowest, and also the maximum value is with 84 connection points per network kilometer considerably smaller than those in class 2 and 1 (129 and 128 connection points per network kilometer, respectively). Thus, most of the network operators in class 3 have a relatively low network density.

We see significant differences in the cost impact of the installed capacity of renewable power plants across classes and are now interested, whether classes also differ according to the amount of renewable power plants' installed capacity. Class 3 is with an average installed capacity of 55 MW and a median capacity of 12 MW clearly characterized as the class with the lowest installed capacity. With 166 MW, the average installed capacity in class 2 is three-time higher, and with 103 MW in class 1 almost twice as high as in class 3. Besides, class 2 also contains network operators with the highest distributed generation capacity from renewable energy sources having a maximum of 14.8 GW compared to 9.8 GW in class 1 and only 1.6 GW in class 3. All classes contain network operators with very low distributed generation capacities from renewable energy sources, resulting in comparatively low median values and large standard deviations.

In sum, class 1 contains the largest network operators with high, but not the highest, installed capacity of renewable power plants. Network operators allocated to class 2 have the highest installed capacity of renewable power plants and are of relatively medium size. The picture is evident for class 3, consisting of small network operators operating networks with low density and low distributed generation capacities from renewable energy sources.

Even though we observe differences in class characteristics across classes, we are interested in whether the deviations are statistically significant different from zero. Hypothesis tests (t-tests) on mean equality across classes show that classes statistically differ, at least in some parts. As shown in Table 3.4, the minor differences are found between classes 1 and 2 that only differ significantly in the average number of connected customers (QC). In contrast, classes 1 and 3 statistically differ in the mean of all variables. Class 2 and 3 do not vary statistically in their size but their network density and installed capacities of distributed generation from renewable energy sources. Thus, a classification of German network operators according to their size only may be misleading.

	Class 1 & Class 2		Class 1 & Class 3		Class 2 & Class 3	
Variable	mean difference	p-value	mean difference	p-value	mean difference	p-value
QE	1,234.37	0.1233	1,220.43	0.0333 **	-13.94	0.9604
QC	34.93	0.0946 *	46.30	0.0017 ***	11.37	0.1860
ND	0.26	0.8234	1.99	0.0171 *	1.73	0.0717 *
DG	-62.62	0.2523	47.68	0.0318 **	110.29	0.0205**

Notes: ***, ** and *: Significant at the 1%-, 5%-, and 10%-level.

Table 3.4: Differences in means of variables across classes

We conclude that there is technological heterogeneity between German network operators related to differences in the size, the network density, and the capacity of distributed generation from renewable energy sources identified by differences in the cost function parameters. Furthermore, differences in class characteristics are present. As we are particularly interested in the impact of distributed generation capacity from renewable energy sources, we try to find even more related differences across classes. Therefore, we analyze whether classes differ according to the installed capacities of PV and wind power.[15] Especially wind and PV plants have very heterogeneous demands on network operators as they differ in the voltage level of the network they are usually connected to and in further characteristics, e.g., plant-specific size, locational requirements, and so forth. The summary statistics are shown in Table 3.5.

[15]Convergence problems and characteristics of the data impede the inclusion of these variables in the total cost function.

	Mean	Median	Std. dev.	Minimum	Maximum
Installed PV capacity [MW]					
Class 1	53.24	10.05	264.86	0.47	3,339.93
Class 2	34.04	9.49	183.56	0.39	2,222.87
Class 3	28.27	8.85	77.29	0.35	755.90
Installed wind capacity [MW]					
Class 1	35.11	0	289.07	0	5,579.58
Class 2	118.20	0.55	1,015.48	0	11,753.48
Class 3	19.86	0	84.96	0	824.36
Wind share					
Class 1	0.10	0	0.16	0	0.69
Class 2	0.17	0.02	0.23	0	0.91
Class 3	0.13	0	0.20	0	0.89

Table 3.5: Descriptive statistics of non-included variables per class

As for the total installed capacity of renewable power plants, wind and PV capacities are lowest for class 3. However, the differences between class 1 and 2 become clearer now: In class 2, the high installed capacity is driven by wind capacities that are, on average, almost five times higher than in class 3 and more than twice as high as those in class 1. In contrast, class 1 comprises the highest average and median installed capacity of PV. The differences in the installed wind capacities across classes clarify the heterogeneous cost function parameters. The high wind capacities lead to lower marginal costs of distributed generation capacities from renewable energy sources in class 2 (see Table 3.2). Due to voltage problems that occur more often in the low-voltage level and the limited capacity to integrate renewable power plants, the same capacity of renewable power plants causes higher costs in the low- than in the medium- and high voltage network (Swiss Economics et al., 2019). These findings are supported by the wind share per class, defined as the installed wind capacity divided by the total installed capacity of renewable power plants. In class 2, 17 percent of the installed capacity of distributed generation from renewable energy sources comes from wind power compared to 10 and 13 percent in class 1 and 3, respectively. Thus, network operators of class 2 may have lower marginal costs of the capacity of distributed generation from renewable energy sources due to a high capacity and share of wind power connected to the medium- and high-voltage level. In contrast, classes 1 and 3 have a lower capacity of wind power and a higher share of capacities connected to the low-voltage level, resulting in a higher cost impact.

The previous findings indicate that heterogeneity among German network operators exists regarding their cost function parameters. Especially class 2 seems to differ significantly from the other classes. These differences may be driven by the high installed wind capacities in class 2. To validate the results and demonstrate the general impact of wind capacities on network operators' costs, we estimate a standard random effects model abstracting from any efficiency effects. We split the sample into network operators with and without wind capacities and estimate Equation 3.2 separately for both groups and compare the results.[16] The "wind group", i.e., network operators with installed wind capacities larger than zero, comprises 880 observations, and the "no-wind group", i.e., network operators with no installed wind capacity, contains 1,031 observations.[17] The results support the previous findings: Network operators without any installed wind capacities face higher marginal costs from increasing distributed generation capacity from renewable energy sources than network operators with installed wind capacities. In the case of network operators having no wind, the effect measures mainly the impact of an increase in PV capacities which induce higher costs than wind capacities. As previously mentioned, these differences result from the wind capacities connected to the medium- and high-voltage network, while PV is usually connected to the low-voltage network.

To sum up again: Network operators with the highest capacity of renewable power plants, mainly driven by wind power, are medium-sized and attributed to class 2. Thereby, the high capacity of wind power explains the lower marginal costs of the capacity of renewable power plants in class 2. Class 1 comprises the largest network operators with a high capacity of distributed generation from renewable energy sources in general and PV in particular. In contrast, class 3 includes operators combining the lowest installed capacity of renewable power plants, small size, and the lowest network density. The higher marginal costs of the capacity of renewable power plants in classes 1 and 3 are driven by the comparatively higher share of capacities from renewable energy sources connected to the low voltage level.

3.5.3 Efficiency estimates

Besides the differences in parameter estimates and class-specific characteristics, we are interested in differences in the efficiency estimates. In a latent class framework there are as many frontiers as classes. Thus, the estimation of (in)efficiency is not as straightforward as in the standard SFA models. Inefficiency can be estimated based on two approaches: First, network operators are benchmarked against the frontier of the class with the highest posterior class probability, i.e., the "most-likely" frontier. Second, the network operator is benchmarked against a "weighted-average" frontier consisting of the probability-weighted average of all classes (Greene, 2002). The higher the posterior class probabilities, the lower are the differences between the two approaches (Orea and

[16] The distribution of wind capacities across network operators and the low variation of network operators' wind capacities over time prevent the estimation of a single model, including the capacities of wind and PV.

[17] The detailed estimation results are depicted in Table 3.9 in the Appendix.

Kumbhakar, 2004). In our case, the mean posterior class probability varies between 95.8 and 96.9 percent per class. To be consistent with network operators' assignment to a particular class, we chose the approach with the "most-likely" frontier. Due to the high posterior probabilities, differences between the two methods are only of minor importance. The individual inefficiency is estimated using the approach of Jondrow et al. (1982), where individual inefficiency is predicted by the conditional mean of the estimated inefficiency term, $u_{i,t}$.

	Mean	Std. dev.	Minimum	Maximum
Total sample	0.9033	0.0977	0.3169	0.9891
Class 1	0.8962	0.0700	0.6254	0.9891
Class 2	0.7959	0.1358	0.3169	0.9833
Class 3	0.9649	0.0154	0.8907	0.9891

TABLE 3.6: Cost efficiency estimates

Table 3.6 shows the summary statistics of efficiency estimates per class and of the whole sample. Comparing the efficiency estimates across classes, it is remarkable that class 2 has a relatively low efficiency (79.6 percent). In contrast, class 3 contains very efficient network operators with efficiency estimates that only vary between 89.1 and 98.9 percent. The average efficiency of network operators in class 1 is 89.6 percent and thus close to the overall sample average of 90.3 percent. The differences across classes are illustrated by the distribution of efficiency estimates shown in Figure 3.1, which differ substantially. Network operators in class 3 show a very narrow distribution around very high efficiency values, while the distribution of class 2 is much flatter and at a lower level. Hypotheses tests (t-tests) with the hypothesis of equal average efficiency estimates across classes confirm the results: The null-hypothesis of equal efficiency has to be rejected for all class comparisons (class 1 with 2, class 2 with 3, and class 1 with 3) at the 1 percent significance level. Thus, efficiency estimates differ significantly between classes.

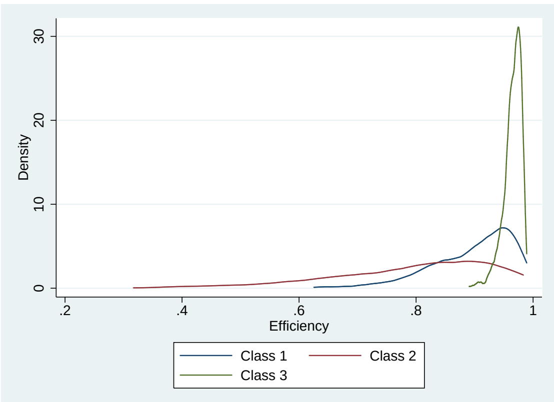


FIGURE 3.1: Kernel density estimates of cost efficiency

As we observe differences in the efficiency and installed capacity of distributed generation of renewable energy sources across classes, we are interested in whether these are related. Analogous to Just and Wetzel (2020), we compare the efficiency of network operators with different shares of distributed generation from renewable energy sources but cannot find significant differences in the efficiency. Thus, the results indicate that differences in network operators' efficiency are not driven by the installed capacity of renewable power plants.

3.5.4 Comparison with two- and one-class model

To validate our results, we compare them with those of two different model specifications. First, we estimate a latent class model with only two classes. In Germany, network operators are clustered in two groups according to their size, revealing the indirect assumption of structural differences between small and large network operators. Thereby, the regulator defines the clustering criterion as 30,000 connected customers. As we have pointed out, such an exogenous clustering is inefficient per definition. Thus, we are rather interested in whether a latent class model with two classes would reflect the regulator's assumption that network operators' size is the relevant clustering criterion. Second, we estimate a True Random Effects (TRE) model where network operators are considered one group, i.e., a one-class model. Even though the TRE model allows individual frontiers due to a random shift of the constant, the cost function, i.e., technology, is assumed to be equal for all network operators. The detailed estimation results of the two-class and the TRE model can be found in the Appendix.

We first analyze the results obtained by the two-class model. The estimation of a latent class model with two classes yields the following class sizes: Class A consists of 1,248 network operators and class B of 663.[18] The summary statistics of both classes are shown in Table 3.7. It is striking that the average differences between the two classes are smaller compared to the differences across classes in the three-class model. However, class A contains small network operators with lower installed capacities of distributed generation from renewable energy sources, while class B consists of large network operators with comparably higher installed capacities of distributed generation from renewable energy sources. Again, the installed capacity of wind power is not included in the cost function but shows significant variation across the two classes. The high capacities of distributed generation from renewable energy sources in class B are driven by high wind capacities being more than twice as high as in class A. Analogous to the results of the three-class model, network operators in class B face lower marginal costs of the installed capacity of

[18] We name the classes A and B rather than 1 and 2 to avoid confusion with the names of the three-class model.

distributed generation from renewable energy sources than network operators in class A (see Table 3.10 in the Appendix). The heterogeneous impact of distributed generation capacity from renewable energy sources is neglected in the TRE model, assuming a joint cost function, i.e., technology, for all network operators. The results show that the estimated marginal costs of increasing distributed generation capacity from renewable energy sources are smaller than in class A and larger than in class B. Thus, the TRE model overestimates the cost impact of distributed generation from renewable energy sources for network operators with relatively high wind capacities and underestimates it for those with lower wind capacities (see Table 3.11 in the Appendix).[19]

Referring to our question of whether an endogenous clustering reflects the German regulator's assumption, we have to state that this is rather not the case. Classes differ in the number of connected customers, but the specific sampling criterion of 30,000 connection points seems not to be relevant here. Even though class A consists of relatively small operators, nearly one-third has more than 30,000 connection points, and even 66 percent of the rather large network operators in class B have less than 30,000 connection points. Thus, we cannot confirm the German regulator's clustering criterion if we assume the existence of two classes.

	Mean	Median	Std. dev.	Minimum	Maximum
Electricity supplied [GWh]					
Class A	1,895.16	239.14	11,649.46	9.88	247,549.60
Class B	1,942.03	261.46	7,101.38	4.76	68,301.84
Connection points [thousand]					
Class A	67.59	18.41	257.71	1.60	4,965.61
Class B	93.80	20.20	290.56	1.04	2,382.71
Network density [Connection points/network km]					
Class A	34.60	31.56	17.16	7.68	127.52
Class B	37.84	35.66	16.08	11.01	129.20
Distributed generation [MW]					
Class A	89.78	14.69	479.75	0.35	9,781.43
Class B	109.36	14.45	907.70	0.58	14,827.64
Installed wind capacity [MW]					
Class A	31.44	0	237.21	0	5,579.58
Class B	72.34	0.01	770.44	0	11,753.48

Table 3.7: Descriptive statistics per class of the two-class model

As efficiency estimates have direct financial consequences for network operators, we are furthermore interested in the impact of different clustering approaches on efficiency estimates. Therefore, we compare the three-class model's efficiency estimates with those of the two-class model and the one-class model, i.e., the TRE model, which assumes equal technology characteristics and thus a common cost function for network operators. Thus,

[19]The estimation results of the TRE model are obtained by maximum simulated likelihood estimation.

technological heterogeneity, of which our previous results indicate that it is present, is considered as inefficiency. In consequence, we expect a lower efficiency in the TRE model than in the latent class models. Average efficiency estimates across the three models are shown in Table 3.8. The average efficiency estimates differ only slightly but are, as expected, highest for the three-class model. The two-class model yields, on average, slightly higher efficiency values than the TRE model. This may be because the two-class model considers technological heterogeneity among network operators, which is attributed to inefficiency in the TRE model. Applying t-tests to test the equality of efficiency estimates across the different models, we find that efficiency between the three- and the two-class model and between the three-class model and the TRE model differ significantly at the 1 percent level. Differences between the two-class model and the TRE model are only slightly significant at the 10 percent level.

	Mean	Std. dev.	Minimum	Maximum
Three-class model	0.9033	0.0977	0.3169	0.9891
Two-class model	0.8982	0.0601	0.5997	0.9857
One-class model (TRE)	0.8960	0.0611	0.3414	0.9855

Table 3.8: Comparison of cost efficiency estimates

Even though efficiency estimates may, on average, differ only slightly, differences for individual network operators may be more prominent. If we account for two technology classes represented by two different cost functions, 75 percent of the network operators will face lower efficiency estimates than with three technology classes. Suppose only one class is considered, and thus, differences in the underlying cost function are ignored at all. In that case, 69 percent of the network operators will have lower efficiency estimates than in the three-class model. Thus, individual efficiency estimates are sensitive to the model specification. The results correspond to our expectation and previous research: The consideration of technological differences reduces unobserved heterogeneity within the classes, and therefore, the heterogeneity conventional models may consider as inefficiency (Agrell et al., 2014).

3.6 Conclusions

The regulation of electricity distribution network operators is most commonly based on incentive regulation, of which benchmarking is an essential element. The use of benchmarking procedures requires the existence of a set of comparable network operators. As this assumption is seldom fulfilled in practice, addressing the heterogeneity among network operators in benchmarking is one of the major concerns and challenges for regulators. If differences among network operators are observable, the regulator accounts for them directly in the benchmarking procedure. To deal with unobserved heterogeneity is far more challenging.

Unobserved heterogeneity may impact network operators in various ways. The fact that it can affect the production process is, however, often neglected in the regulatory practice, with a common technology, i.e., production process, being assumed among network operators instead. This implies, for example, that all network operators are represented by the same cost function and thus face the same marginal costs, e.g., to connect new customers or distributed generation plants. However, network operators differ substantially in many aspects that are likely to influence the production process and thus contradict this assumption. For example, network operators in a mountainous landscape face higher connection costs than those in a rather flat landscape. Moreover, costs of distributed generation from renewable energy sources differ according to the plant type: Wind capacities at higher voltage levels induce lower costs than PV in a low-voltage network. These differences remain uncontrolled for if regulators assume a common technology, i.e., cost function.

Technological differences may even increase due to changing market conditions. Increasing distributed generation from renewable energy sources as well as the use of electric vehicles and heat pumps would require an adaption of network operators. The ability to adapt may also be influenced by unobserved differences among network operators. These developments together with the structural diversity of network operators raise doubts on the assumptions of homogeneous conditions among network operators at the moment and even more in the future. If technological heterogeneity remains uncontrolled and a joint cost function is assumed for all network operators, efficiency estimates would be biased and have direct financial consequences for network operators.

In this paper, we estimate the cost efficiency of German network operators and explicitly account for technological differences represented by heterogeneous cost functions among network operators. Based on a latent class model, our results show that German network operators can be unambiguously classified into three statistically different classes sharing a common technology, i.e., cost function, based on size and distributed generation variables. We find significant differences in the size, installed capacity of distributed generation from renewable energy sources, and efficiency estimates across classes. The results indicate that distributed generation from renewable energy sources is a relevant driver of technological heterogeneity among classes. First of all, differences in the installed capacity of renewable power plants across classes are partly driven by the total capacity but especially by differences in the installed capacities of wind power. Furthermore, we observe a heterogeneous cost impact of an increase in distributed generation from renewable energy sources across classes related to their installed wind capacities.

Thus, clustering German network operators solely according to their size, as done by the German regulator, leads to misleading results. The importance to account for technological heterogeneity among German network operators is confirmed by comparing the results of the three-class model with those of a one- and two-class model. Efficiency values are highest in the three-class model as, at least some, technological heterogeneity is overlooked and thus considered as network operators' inefficiency in the one- and two-class model.

We conclude that German network operators use heterogeneous technologies represented by different cost function parameters. The consideration of technological differences reduces unobserved heterogeneity within classes, avoiding the misspecification of heterogeneity as inefficiency in conventional models. As efficiency estimates have a direct financial impact on network operators, the importance of correctly accounting for heterogeneity can be clearly seen in the results. In this context, the application of latent class analysis may provide valuable insights for regulators to detect and address technology differences among network operators in Germany and worldwide.

3.7 Appendix

Variable	wind group	no-wind group
Constant	16.203***	16.320***
	(0.024)	(0.023)
$\ln QE$	0.127***	0.138***
	(0.021)	(0.023)
$\ln QC$	0.684***	0.599***
	(0.032)	(0.032)
$\ln ND$	-0.459***	-0.346***
	(0.041)	(0.037)
$\ln DG$	0.184***	0.247***
	(0.021)	(0.019)
2012	0.024*	0.011
	(0.013)	(0.014)
2013	0.063***	0.036***
	(0.014)	(0.014)
2014	0.071***	0.040***
	(0.014)	(0.015)
2015	0.076***	0.046***
	(0.014)	(0.015)
2016	0.143***	0.112***
	(0.015)	(0.016)
2017	0.170***	0.140***
	(0.020)	(0.023)

Notes: Standard errors in parentheses. ***, ** and *: Significant at the 1%-, 5%-, and 10%-level. The estimations have been performed in STATA 12.

Table 3.9: Estimation results of wind and no-wind group

Variable	Class A	Class B
Constant	16.184***	15.771***
	(0.016)	(0.031)
$ln\,QE$	0.135***	0.117***
	(0.010)	(0.019)
$ln\,QC$	0.568***	0.781***
	(0.017)	(0.027)
$ln\,ND$	-0.315***	-0.235***
	(0.019)	(0.033)
$ln\,DG$	0.265***	0.105***
	(0.007)	(0.013)
2012	0.047***	0.028
	(0.016)	(0.029)
2013	0.060***	0.098***
	(0.016)	(0.029)
2014	0.072***	0.103***
	(0.016)	(0.029)
2015	0.075***	0.101***
	(0.016)	(0.029)
2016	0.125***	0.191***
	(0.016)	(0.030)
2017	0.131***	0.303***
	(0.023)	(0.045)
Prior Class Probabilities	0.630***	0.370***
	(0.028)	(0.028)

Notes: Standard errors in parentheses. ***, ** and *: Significant at the 1%-, 5%-, and 10%-level. The estimations have been performed in NLOGIT 6.

Table 3.10: Estimation results of the two-class model

Variable	TRE
Constant	16.129***
	(0.006)
$ln\,QE$	0.144***
	(0.004)
$ln\,QC$	0.632***
	(0.006)
$ln\,ND$	-0.391***
	(0.007)
$ln\,DG$	0.199***
	(0.003)
2012	0.037***
	(0.009)
2013	0.078***
	(0.008)
2014	0.083***
	(0.008)
2015	0.087***
	(0.008)
2016	0.150***
	(0.008)
2017	0.176***
	(0.013)

Notes: Standard errors in parentheses. ***, ** and *: Significant at the 1%-, 5%-, and 10%-level. The estimations have been performed in NLOGIT 6.

Table 3.11: Estimation results of the TRE model

4 Congestion management in power systems – Long-term modeling framework and large-scale application

4.1 Introduction

The liberalization of power systems entails an unbundling of generation and grid services to reap efficiency gains stemming from a separate and different organization. While there is competition between generating firms, transmission grids are considered a natural monopoly and are operated by regulated transmission system operators (TSOs). However, strong inter-linkages remain between these two parts of the power system: From a transmission perspective, TSOs are responsible for non-discriminatory access of generating units to transmission services while maintaining a secure grid operation. They are thus strongly influenced by the level and locality of generation and load. Furthermore, due to Kirchhoff's laws, operation and investment decisions of one TSO may affect electricity flows in the area of another TSO. From a generation firms' perspective, activities are impacted by restrictions on exchange capacities between markets or operational interventions by the TSOs to sustain a reliable network.

An efficient regulatory design of those inter-linkages between generation and grid will positively affect the overall efficiency of the system, for instance by providing locational signals for efficient investments into new generation or transmission assets. To ensure an efficient coordination of short (i.e., operational) and long-term (i.e., investment) activities in the generation and grid sectors, congestion management has been identified to be of utmost importance (e.g., Chao et al. (2000)). Different regulatory designs and options are available to manage congestion, including the definition of price zones as well as various operational and investment measures. Because it is able to deliver undistorted and hence efficient price signals, nodal pricing is a powerful market design to bring along short-run efficiency. This was shown in the seminal work of Schweppe et al. (1988) and Hogan (1992). Nevertheless, many markets deviate from that short-run efficient congestion management and pursue alternative approaches, e.g. due to historical or political reasons. For instance, most European countries deploy national zonal market areas with uniform electricity prices. Implicitly, several challenges are thus imposed upon the system: First, in zonal markets, intra-zonal network congestion remains unconsidered by dispatch decisions. However, if a dispatch induces intra-zonal congestion (which is typically often the case), it might be necessary to reconfigure the dispatch, known as re-dispatch. Alternatively, a generator-component (g-component) may be implemented as part of the market clearing process. The latter makes generators pay an additional fee according to their impact on the zonal network (and hence, changes the dispatch). Second, cross-border capacity needs to be managed. Whereas historically, cross-border capacities have often been auctioned explicitly, many market areas are now turning to implicit market coupling based on different allocation routines, such as net-transfer capacities (NTC) or flow-based algorithms.[20]

The literature has investigated various regulatory designs to manage congestion in power systems from different perspectives. Static short term efficiency of nodal pricing – as shown by Schweppe et al. (1988) – was confirmed, e.g, by van der Weijde and Hobbs (2011) who compare nodal pricing and NTC based market coupling in a stylized modeling environment. Furthermore, several papers have quantified the increase in social welfare through a switch from zonal to nodal pricing for static real world case studies (see for example: Green (2007), Leuthold et al. (2008), Burstedde (2012), Neuhoff et al. (2013)). Similarly, Daxhelet and Smeers (2007) show that generator and load components reflecting their respective impact on congestion have a positive effect on static social welfare (as well as its distribution), while Oggioni and Smeers (2012) investigate different congestion management designs in a six node model and find that multilateral arrangements may improve efficiency. Oggioni et al. (2012) and Oggioni and Smeers (2013) show that in a zonal pricing system, the configuration of zones as well as the choice of counter-trading designs have a significant impact on efficiency.

A second line of literature deals with the dynamic long term effects of congestion management, i.e., the investment perspective. On the one hand, issues of timing (e.g., due to uncertainty or commitment) in settings consisting of multiple players (such as generation and transmission) have been addressed. Höffler and Wambach (2013) find that long term commitment of a benevolent TSO may lead to inefficient investment decisions due to the locational decisions of investments in generation. In contrast, Sauma and Oren (2006) and Rious et al. (2009) formulate the coordination problem between a generation

[20] Cross-border capacities and prices are implicitly taken into account during the joint clearing of coupled markets.

and a transmission agent as a decomposed problem, and find that a prospective coordinated planning approach as well as transparent price signals entail efficiency gains. On the other hand, imperfect simultaneous coordination (e.g., due to strategic behavior or hidden information) has been investigated by Huppmann and Egerer (2014) for the case of multiple TSOs being active in an interconnected system. They find that a frictionless coordinated approach outperforms the system outcome with strategic TSOs maximizing social welfare within their own jurisdiction.

With this paper, we contribute to the above literature with a generalized and flexible economic modeling framework for analyzing the short as well as long term effects of different congestion management designs in a decomposed inter-temporal equilibrium model including generation, transmission, as well as their inter-linkages. Specifically, with our framework we are able to represent, analyze and compare different TSO organizations, market areas (i.e., nodal or zonal pricing), grid expansion, redispatch or g-components, as well as calculation methods for cross-border capacity allocation (i.e., NTC and flow-based). A major advantage of our analytical and numerical implementation is its flexibility to represent different congestion management designs in one consistent framework. We are hence able to identify and isolate frictions and sources of inefficiencies by comparing these different regulatory designs. Moreover, we are able to benchmark the different designs against a frictionless welfare-optimal result, i.e., the "first best". In order to exclusively focus on the frictions and inefficiencies induced by the congestion management designs, we do not address issues of timing, such as uncertainty or sequential moving. Instead, we assume perfect competition, perfect information, no transaction costs, utility-maximizing agents, continuous functions, inelastic demand and an environment where generation and grid problems are solved simultaneously. As an additional contribution, we calibrate and numerically solve our model for a large-scale problem. Specifically, we investigate a detailed representation of the Central Western European (CWE) region.[21] Thereby, we offer a sound indication on how different congestion management designs perform in practice, and provide empirical evidence that nodal pricing is the efficient benchmark while alternative designs imply inefficiencies of up to 4.6 percent until 2030.

The paper proceeds as follows: In Section 2, we analytically develop our modeling framework. In Section 3, a numerical solution method to solve this framework is proposed. In Section 4, we apply the methodology to a detailed representation of the CWE region in scenarios up to the year 2030. Section 5 concludes and provides an outlook on future research.

[21] The CWE region is one of seven regional initiatives to bring forward European market integration. The countries within this area are Belgium, France, Germany, Luxemburg and the Netherlands.

4.2 Economic framework

In order to develop a consistent analytical modeling framework for different congestion management designs, we start with the well-known model for an integrated optimization problem for planning and operating a power system.[22] By design, this model does not contain any frictions and inefficiencies. Hence, its result is necessarily first best and may serve as the efficient benchmark for alternative settings. Moreover, it corresponds to the concept of nodal pricing as introduced by Schweppe et al. (1988).[23]

Conceptually similar to Sauma and Oren (2006), we then make use of the possibility to separate an integrated optimization problem into multiple levels (or, in other words, subproblems). Even though the model structure is different, it can be shown that both formulations of the problem yield the same results. However, in the economic interpretation we can take advantage of the separated model structure representing unbundled generation and transmission sectors. On the generation stage, competitive firms decide about investments in and dispatch of power plants, whereas the transmission stage consists of one or multiple TSOs that efficiently expand and operate transmission grid capacities. Lastly, with generation and transmission separated, we are able to introduce six practically relevant congestion management designs through the manipulation of the exchange of information between and among the two levels, and show how they deviate from the first best.

Even though the modeling framework would allow to study an extensive range of congestion management designs, we restrict our attention to four settings (and two additional variations) that are both, relevant in practical applications and sufficiently different from each other. Specifically, our settings vary in the definition of market areas (nodal and zonal pricing), the regulation and organization of TSOs (one or several TSOs), the way of managing congestion besides grid expansion (redispatch and g-component) and different alternatives for cross-border capacity allocation (NTC vs. flow-based market coupling). The analyzed settings are summarized in the following Table 4.1.

Noticeably, despite the separated generation and transmission levels, in all settings agents are assumed to act rationally and simultaneously while taking into account the activities of the other stage.[24] Furthermore, we assume perfect competition on the generation stage and perfect regulation of the TSOs in the sense that TSO activities are aligned with social objectives. TSOs as well as generators are price taking, with an

[22] Such a model is typically applied to represent the optimization problem of a social planner or an integrated firm optimizing the entire electricity system, including generation and transmission.

[23] Nodal pricing usually describes prices based on short-term marginal costs. In our case, however, since we consider investments and inelastic demand, the interpretation of the nodal prices differs. We discuss implications in Section 4.2.1.

[24] I.e., sequential moving and issues of timing are not considered.

	Market area and coupling	TSO scope	TSO measures
I	Nodal markets	One TSO	Grid expansion
II - NTC	Zonal markets, NTC-based coupling	One TSO	Grid expansion, zonal redispatch
II - FB	Zonal markets, Flow-based coupling	One TSO	Grid expansion, zonal redispatch
III - NTC	Zonal markets, NTC-based coupling	Zonal TSOs	Grid expansion, zonal redispatch
III - FB	Zonal markets, Flow-based coupling	Zonal TSOs	Grid expansion, zonal redispatch
IV	Zonal markets	Zonal TSOs	Grid expansion, zonal g-component

TABLE 4.1: Analyzed congestion management designs

independent institution (e.g., the power exchange) being responsible for coordinating the activities of the different participating agents and for market clearing.[25] Importantly, while in the first best design all information is available to all agents, alternative congestion management designs may induce an adverse (e.g., aggregated) availability of information. The solution of the problem is an intertemporal equilibrium which is unique under the assumption of convex functions. Hence, a unique ordering of alternative problem settings, i.e., in our case, congestion management designs, can be obtained. Noticeably, with the above assumptions, our general modeling approach can be thought of as a way to compare today's and future performances of different congestion management designs based on today's state of the system, today's information horizon, as well as rational expectations about future developments and resulting investment decisions.[26]

For developing the economic modeling framework in the following subsections, we will deploy parameters, variables and sets as depicted in Table 5.5.

4.2.1 Setting I – First Best: Nodal pricing with one TSO

By design, nodal pricing avoids any inefficiency by covering and exchanging all information present within the problem of optimizing the dispatch. It hence represents the first best setting in our analysis of different congestion management designs. According to Schweppe et al. (1988), a distinct price per grid node results from the optimal dispatch of the entire system. Nodal prices are based on efficient locational short-run marginal costs, obtained from a simultaneous market clearing that implicitly considers the physics of the electricity network (specifically, loop flows). Introducing a dynamic perspective

[25] By assuming perfect competition and an inelastic demand, we are able to treat the general problem as a cost minimization problem. This assumption is commonly applied for formulation of electricity markets in the literature. An alternative formulation with a welfare maximization approach would be possible, but wouldn't impact the general conclusions.

[26] In our numerical application, this approach is supplemented with discounted future cash flows. See Section 4.4 for further details.

Abbreviation	**Dimension**	**Description**
Model sets		
$i \in \mathbf{I}, j \in \mathbf{J}$		Nodes, $\mathbf{I}, \mathbf{J} = [1, 2, ...]$
$m, n \in \mathbf{M}$		Zonal markets, $\mathbf{M} = [1, 2, ...]$
$i \in \mathbf{I_m}, j \in \mathbf{J_m}$		Nodes that belong to zonal market m, $\mathbf{I_m} \subset \mathbf{I}$, $\mathbf{J_m} \subset \mathbf{J}$
$i \in \mathbf{I_{m,cb}}, j \in \mathbf{J_{m,cb}}$		Nodes that belong to zonal market m and are connected to a another zone n by a cross-border line, $\mathbf{I_{m,cb}} \subset \mathbf{I_m}$, $\mathbf{J_{m,cb}} \subset \mathbf{J_m}$
Model parameters		
δ_i	EUR/kW	Investment and FOM costs of generation capacity in node i
γ_i	EUR/kWh	Variable costs of generation capacity in node i
$\mu_{i,j}$	EUR/kW	Investment costs of line between node i and node j
d_i	kW	Electricity demand in node i
$PTDF$	$-$	Power Transfer Distribution Factor
$\sigma_{i,j}$	%	Cost share for an interconnector capacity between node i and node j, $i \in \mathbf{I_{m,cb}}$, $j \in \mathbf{J_{m,cb}}$
Model primal variables		
$\overline{G}_i$	kW	Generation capacity in node i, $\overline{G}_i \geq 0$
G_i	kW	Generation dispatch in node i, $G_i \geq 0$
$T_{i,j}, T_{m,n}$	kW	Electricity trade from node i to node j, or market m to market n
X	EUR	Costs of generation
Y	EUR	Costs of TSO
$\overline{P}_{i,j}$	kW	Line capacity between node i and node j, $\overline{P}_{i,j} \geq 0$
$P_{i,j}$	kW	Electricity flow on line between node i and node j
R_i	kW	Redispatch in node i
Model dual variables		
$\kappa_{i,j}, \kappa_{m,n}$	EUR/kW	price for transmission between nodes (i and j) or zones (m and n)
λ_i, λ_m	EUR/kW	nodal or zonal price for electricity

Table 4.2: Model sets, parameters and variables

in this nodal pricing framework, we abstract from economies of scale and lumpiness that would render short-run nodal prices insufficient for long-term investments and jeopardize the existence of an equilibrium as well as its uniqueness (Joskow and Tirole (2005), Rious et al. (2009)). Instead, we assume constant marginal grid costs and continuous generation and transmission expansion.[27] Together with the assumption of inelastic demand, we derive nodal prices which are either based on short-run marginal costs (in off-peak hours) or include marginal capacity costs in peak load hours. Consequently, we

[27]This assumption is certainly more critical for transmission investments which require a certain magnitude to be realized. Generation investment might also be lumpy, but smaller plant sizes are possible.

derive nodal prices that are both efficient and sufficient.[28]

The following optimization problem *P1* corresponds to the formulation of Schweppe et al. (1988) of an integrated problem for operating generation and transmission, extended to include investment decisions. In this formulation, a social planner or an integrated firm minimize total system costs of the operation and investment of generation and transmission.

P1 Integrated Problem

$$\min_{\overline{G}_i, G_i, T_{i,j}, \overline{P}_{i,j}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{4.1a}$$

$$\text{s.t.} \qquad G_i - \sum_j T_{i,j} = d_i \qquad \forall i \qquad | \, \lambda_i \tag{4.1b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.1c}$$

$$|P_{i,j}| = f(T_{i,j}) \leq \overline{P}_{i,j} \qquad \forall i,j \qquad | \, \kappa_{i,j} \tag{4.1d}$$

Indices i, j represent nodes in the system. Generation G_i, generation capacity $\overline{G}_i$ and transmission capacity $\overline{P}_{i,j}$ are explicitly optimized, while trade $T_{i,j}$ is implicitly derived from Equation (4.1b).[29] Additional capacities can be installed at the costs of δ_i for generation and $\mu_{i,j}$ for transmission. Nodal prices are derived from the dual variables λ_i of the equilibrium constraint which states that the demand level d_i at node i can be either satisfied by generation at the same node or trade (Equation (4.1b)). Equations (4.1c) and (4.1d) mirror that generation is restricted by installed generation capacities, and trade by installed transmission capacities. The calculation of flows on transmission lines according to Kirchhoff's law is represented by function f in Equation (4.1d). For instance, function f could represent a PTDF matrix which determines how flows on each line are impacted by trades.[30]

As has been shown, e.g., by Conejo et al. (2006), an integrated optimization problem can be decomposed into subproblems which are solved simultaneously, while still representing the same overall situation and corresponding optimal solution. In our application, we take advantage of this possibility to represent separated generation and transmission levels in problem *P1'*. The generation stage *P1'a* states the market clearing of supply

[28]Note that in the long-term, without additional restrictions, nodal prices would all tend to the costs of the most cost-efficient generation technology. However, in a problem with diversified assets to start with (i.e., an existing generation fleet), and inter-temporal (e.g., ramping) as well as technology-specific (e.g., renewable resource availability) constraints, long-term variations in the nodal prices will persist. These conditions will typically hold true in any applied power system model.

[29]For better readability, we dropped the time index t, which is the only reason why $\overline{G}_i$ and G_i may differ.

[30]We will use the PTDF approach in the numerical implementation, as it enables a linearization of the non-linear grid problem (cf. Hagspiel et al. (2014)).

and demand while respecting generation capacity constraints. As in *P1*, the same nodal prices are obtained by the dual variable λ_i of the equilibrium constraint (4.2b). Instead of including the explicit grid expansion costs in the cost minimization, the objective function of the generation stage now contains transmission costs which assign transmission prices $\kappa_{i,j}$ to trade flows between two nodes i and j. These prices are derived from the dual variable of the equilibrium constraint on the transmission stage (Equation (4.2e)). We assume that the TSO is perfectly regulated to minimize costs of grid extensions accounting for the physical feasibility of the market clearing as determined on the generation stage while considering all grid flows and related costs (problem *P1'b*).

P1'a *Generation*

$$\min_{\overline{G}_i, G_i, T_{i,j}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \sum_{i,j} \kappa_{i,j} f(T_{i,j}) \tag{4.2a}$$

$$\text{s.t.} \qquad G_i - \sum_j T_{i,j} = d_i \qquad \forall i \qquad | \, \lambda_i \tag{4.2b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.2c}$$

P1'b *Transmission*

$$\min_{\overline{P}_{i,j}} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{4.2d}$$

$$\text{s.t.} \qquad |P_{i,j}| = f(T_{i,j}) \leq \overline{P}_{i,j} \qquad \forall i,j \qquad | \, \kappa_{i,j} \tag{4.2e}$$

As can be seen, all terms of *P1* reappear in *P1'*, however, allocated to two separated levels. Both formulations describe the same problem and hence, have the same outcome, namely the first best. In fact, in the optimum of Problem *P1'*, it must hold that the costs of transmission expansion are equal to the marginal of the transmission constraint, i.e. $\mu_{i,j} = \kappa_{i,j}$. Furthermore, the duality property of the problem ensures that in the optimum, $\sum_{i,j} \mu_{i,j} \overline{P}_{i,j} = \sum_{i,j} \kappa_{i,j} f(T_{i,j})$, such that the objectives of *P1* and *P1'* coincide.

4.2.2 Setting II: coupled zonal markets with one TSO and zonal redispatch

In zonal markets, a number of nodes are aggregated to a market with a uniform price. In contrast to nodal pricing, coupled zonal markets only consider aggregated cross-border capacities between market zones during market clearing (instead of all individual grid elements). Thus, the obtained prices for generation do not reflect the true total costs of the entire grid infrastructure. This is due to the fact that zonal prices only reflect those cross-border capacities that limit activities between zonal markets. Cross-border capacities can be allocated in different ways. We consider Net Transfer Capacity (NTC) and flow-based market coupling as cross-border capacity allocation algorithms because they have been used extensively in the European context (see, e.g., Glachant (2010), Commission de Régulation de l'Énergie (2009)). NTCs are a rather simplified version of cross-border trade restrictions, widely neglecting the physical properties of the grid as well as its time-varying characteristics. Flow-based (FB) market coupling shows a much better consideration of the physical grid properties which is crucially important in case of meshed networks. Under flow-based market coupling, cross-border transmission capacities are calculated taking into account the impact of (cross-border) line flows on every line in the system (e.g., Oggioni and Smeers (2013)). As a consequence, more capacity can generally be offered for trading between markets, and a better usage of existing infrastructures is achieved.

Because intra-zonal congestion is neglected in the zonal market-clearing, it needs to be resolved in a subsequent step by the TSO. Besides the expansion of grid capacities, in *Setting II* we provide the TSO with the opportunity of zonal redispatch. The TSO may instruct generators located behind the bottleneck to increase production (positive redispatch), and another generator before the bottleneck to reduce production (negative redispatch). We assume here a cost-based, revenue-neutral redispatch: the TSO pays generators that have to increase their production their variable costs, and in turn receives the avoided variable costs of generators that reduce their supply. As the generator with positive redispatch was not part of the original dispatch, it necessarily has higher variable costs than the generator that reduces supply. Thus, the TSO has to bear additional costs that are caused by the redispatch which amount to the difference between the variable costs of the redispatched entities. Assuming further that the TSO has perfect information about the variable costs of the generating firms, redispatch measures of the TSO have no impact on investment decisions of generating firms as they are revenue-neutral. Hence, additional costs for the economy are induced by inefficient investment decisions of those generators that are not aligned with the overall system optimum due to missing locational price signals.

In the formulation of problem *P2a* zonal pricing is represented by the zonal market indices n, m, each containing one or several nodes i. Market clearing, depicted by the equilibrium Equation (4.3b), now takes place on zonal instead of nodal markets. The corresponding dual variable λ_m represents zonal prices, which do not include any grid costs except for cross-border capacities. This is indicated by the term $\sum_{m,n} \kappa_{m,n} T_{m,n}$ instead of the nodal formulation (with $\kappa_{i,j}$) above. Transmission prices are determined on the transmission stage (Equation (4.3h)). However, contrary to nodal pricing, these prices are calculated based on some regulatory rule (e.g., NTC or FB) and are thus inherently incomplete since they do not represent real grid scarcities. It is noteworthy that nodal trades cannot be used by the TSO anymore to calculate line flows via $f(T_{i,j})$. However, the information on generation (G_i) and demand (d_i), i.e., nodal power balances, is equally sufficient to determine power flows on all lines via function $\tilde{f}(G_i, d_i)$.[31] In addition to grid expansion, the TSO may relieve intra-zonal congestion and optimize the situation by means of redispatch measures R_i at costs of $\gamma_i R_i$.

P2a *Generation*

$$\min_{\overline{G}_i, G_i, T_{m,n}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \sum_{m,n} \kappa_{m,n} T_{m,n} \tag{4.3a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_i - \sum_n T_{m,n} = \sum_{i \in \mathbf{I_m}} d_i \qquad \forall m \qquad \mid \lambda_m \tag{4.3b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.3c}$$

P2b *Transmission*

$$\min_{\overline{P}_{i,j}, R_i} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} + \sum_i \gamma_i R_i \tag{4.3d}$$

$$\text{s.t.} \quad |P_{i,j}| = \tilde{f}(G_i, d_i) \leq \overline{P}_{i,j} \qquad \forall i,j \qquad \mid \kappa_{i,j} \tag{4.3e}$$

$$\sum_{i \in \mathbf{I_m}} R_i = 0 \qquad \forall m \tag{4.3f}$$

$$0 \leq G_i + R_i \leq \overline{G}_i \qquad \forall i \tag{4.3g}$$

$$\kappa_{m,n} = g(\kappa_{i,j}) \tag{4.3h}$$

The following two examples illustrate the fundamental differences between *Setting I* and *II*.

Example for 2 nodes and 2 markets: If the electricity system consists of 2 nodes and 2 markets (Figure 4.1, left hand side), *Setting I* and *II* should be identical. Due to the market regions only consisting of one node, the redispatch Equations (4.3f), (4.3g) as well as the cost term of redispatch in the objective function ($\sum_i \gamma_i R_i$) vanish. In a welfare optimized system, it holds that $\kappa_{m,n} = \kappa_{i,j}$, leading to equivalence of problem *P2* and problem *P1'*.

[31] Note that the duality of the problem would also allow for an alternative formulation of the cross-border transmission constraint by means of quantity constraints instead of prices. Hence, the cost of transmission in the objective function of the generation stage ($\sum_{m,n} \kappa_{m,n} T_{m,n}$) would disappear and an additional constraint for trading would be implemented ($|T_{m,n}| \leq C_{m,n}, \forall m,n$). The restriction of trading volumes $C_{m,n}$ would be calculated on the transmission stage *P2b* via a constraint $C_{m,n} = h(\overline{P}_{i,j})$ instead of the prices $\kappa_{m,n}$. These prices would then be the dual variable of the volume constraint on the generation stage, and necessarily coincide with $\kappa_{m,n}$.

Example for 3 nodes and 2 markets: Figure 4.1, right hand side, shows an electricity system consisting of two markets m and n, where m includes one node (1) and n two nodes $(2, 3)$. Function g for calculating the transmission price $\kappa_{m,n}$ (Equation (4.3h)) between the markets has to be defined, e.g. by averaging the single line prices $\kappa_{m,n} = (\kappa_{1,2} + \kappa_{1,3})/2$. Still, the TSO cannot supply the locational fully differentiated prices $\kappa_{1,2}, \kappa_{1,3}$ and $\kappa_{2,3}$ to the market, and hence, efficient allocation of investments is (partly) achieved *between* the markets, but not *within* the markets. Redispatch does not fully solve this problem, because it is revenue-neutral and does not affect the investment decision.

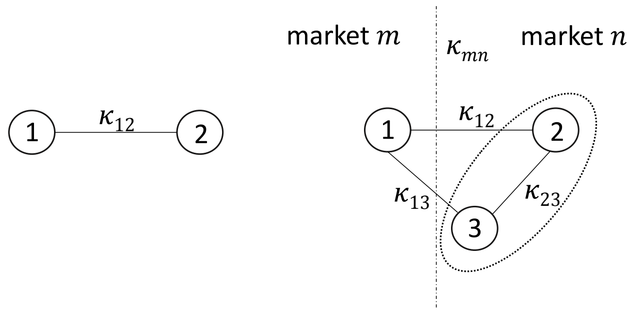


FIGURE 4.1: Two simple examples. Left: 2 nodes, 2 markets. Right: 3 nodes, 2 markets

Overall, *Settings I* and *II* differ in the way grid costs are reflected on the generation stage. Specifically, *Setting II* lacks locational differentiated prices, thus impeding efficient price signals $\kappa_{i,j}$ for the generation stage. Of course, the level of inefficiency depends substantially on the regulatory rule determining the calculation of prices based on a specification of function $g(\kappa_{i,j})$. In general, it is clear that the closer the specification of g reflects real-time conditions and the more it enables the full usage of existing grid infrastructures, the more efficiently the general problem will be solved. While we limit our analysis in this section to this general finding, we will discuss two possible specifications often implemented in practice (NTC and flow-based market coupling) in the empirical example in Section 4.4. Given the inefficiency induced by the specification of function g, the question remains whether and how redispatch measures may help to relieve the problem. We find that the resulting inefficiency cannot be fully resolved by redispatch because the latter remains a zonal measure (Equation (4.3f)). Hence, the TSO cannot induce an efficient usage of generation and transmission across zonal borders. Furthermore, investments into generation capacities are not influenced by redispatch and only zonal prices as well as their costs are considered.[32] Hence, the setting lacks locational signals for efficient generation investments within zonal markets.

4.2.3 Setting III: coupled zonal markets with zonal TSOs and zonal redispatch

In this setting, we consider zonal markets with zonal TSOs being responsible for grid expansion as well as a zonal redispatch. Thus, the problem on the generation stage remains exactly the same as in the previous setting (i.e., *P3a* = *P2a*). However, the transmission problem changes, such that now multiple zonal TSOs are considered. Each TSO solves its own optimization problem according to the national regulatory regime (in our case corresponding to a cost-minimization within the zones). However, cross-border line capacities are also taken into account. Hence, grid capacities, especially cross-border capacities, are extended inefficiently as they do not result from an optimization of the entire grid infrastructure. In addition – just as in the previous setting – inefficient investment incentives for generation and grid capacities are caused by the lack of locational differentiated prices. Hence, overall, system outcomes in *Setting III* must be inferior to those of *Setting II*.

P3a *Generation*

$$\min_{\overline{G}_i, G_i, T_{m,n}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \sum_{m,n} \kappa_{m,n} T_{m,n} \tag{4.4a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_i - \sum_n T_{m,n} = \sum_{i \in \mathbf{I_m}} d_i \qquad \forall m \qquad | \, \lambda_m \tag{4.4b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.4c}$$

P3b *Transmission*

$$\min_{\overline{P}_{i,j \in \mathbf{I_m}}, R_{i \in \mathbf{I_m}}} \quad Y_m = \sum_{i,j \in \mathbf{I_m}} \mu_{i,j} \overline{P}_{i,j} + \sum_{i,j \in \mathbf{I_{m,cb}}} \sigma_{i,j} \mu_{i,j} \overline{P}_{i,j} + \sum_{i \in \mathbf{I_m}} \gamma_i R_i \quad \forall m \tag{4.4d}$$

$$\text{s.t.} \quad |P_{i,j}| = \tilde{f}(G_i, d_i) \leq \overline{P}_{i,j} \qquad \forall i, j \in \mathbf{I_m} \qquad | \, \kappa_{i,j \in \mathbf{I_m}} \tag{4.4e}$$

$$\sum_{i \in \mathbf{I_m}} R_i = 0 \tag{4.4f}$$

$$0 \leq G_i + R_i \leq \overline{G}_i \qquad \forall i \in \mathbf{I_m} \tag{4.4g}$$

$$\kappa_{m,n} = g(\kappa_{i,j}) \tag{4.4h}$$

[32]For obtaining a unique equilibrium we assume that costs differ over all nodes, such that decisions for generation and investments are unambiguously ordered.

In problem *P3*, there are now separate optimization problems for each zonal TSO (indicated by Y_m), with the objective to minimize costs from zonal grid and cross-border capacity extensions as well as from zonal redispatch measures (Equation (4.4d)). For the redispatch, TSOs have to consider the same restrictions as in the previous setting (Equations (4.4f) and (4.4g)). As cross-border capacities are by definition located within the jurisdiction of two adjacent market areas, the two corresponding TSOs have to negotiate about the extension of these cross-border capacities. In fact, cross-border capacities built by two different TSOs may be seen as a Leontief production function, due to the fact that the line capacities built on each side are perfect complements. Corresponding costs from inter-zonal grid extensions are assumed to be shared among the TSOs according to the cost allocation key $\sigma_{i,j}$. According to Equation (4.4h), prices for transmission between zones that are provided to the generation stage ($\kappa_{m,n}$) are determined just as in the previous *Setting II* with only one TSO, depending on the type of market coupling, i.e., the specification of function g. The only difference is that line-specific prices $\kappa_{i,j}$ may now deviate from *Setting II* as they result from the separated activities of each zonal TSO (specifically, from Equation (4.4e), i.e., the restriction of flows on intra-zonal and cross-border lines). Due to the fact that situations may arise where an agreement on specific cross-border lines between neighboring TSOs cannot be reached (which would imply that an equilibrium solution cannot be found), we assume the implementation of a regulatory rule that ensures the acceptance of a unique price for each cross-border line by both of the neighboring TSOs. For instance, the regulatory rule may be specified such that both TSOs are obliged to accept the higher price offer, or, equivalently, the lower of the two capacities offered for the specific cross-border line.

4.2.4 Setting IV: coupled zonal markets with zonal TSOs and g-component

In this last setting, we again consider coupled zonal markets with zonal TSOs. However, instead of having the possibility to perform a zonal redispatch (as in *Setting III*), zonal TSOs may now determine local, time-varying prices for generators, i.e., a g-component, at each node belonging to its zone to cope with intra-zonal congestion. As the g-component reflects the impact of generators on the grid at each node and each instant of time, grid costs are being transferred to the generating firms which consider them in their investment and dispatch decision. In other words, TSOs are able to provide locationally differentiated prices (and hence, generation and investment incentives) for generators within their zone. Noticeably, we do not consider an international g-component here as this would yield the same results as a nodal pricing regime due to generators considering the full set of information concerning grid costs. However, two frictions that may cause an inefficient outcome of this setting remain. When determining nodal g-components, zonal TSOs only consider grid infrastructures within their zone, and not within the entire system. Furthermore, as in *Setting III*, the desired expansion of cross-border lines, which is here assumed to be solved by some regulatory rule ensuring successful negotiation, may deviate between/across neighboring TSOs. However, the negotiation outcome can only be at least as good as an integrated optimization.

P4a *Generation*

$$\min_{\overline{G}_i, G_i, T_{m,n}} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \sum_{i,j} \kappa_{i,j} \tilde{f}(G_i, d_i) \tag{4.5a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_i - \sum_n T_{m,n} = \sum_{i \in \mathbf{I_m}} d_i \qquad \forall m \qquad | \, \lambda_m \tag{4.5b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.5c}$$

P4b *Transmission*

$$\min_{\overline{P}_{i,j \in \mathbf{I_m}, \mathbf{I_{m,cb}}}} \quad Y_m = \sum_{i,j \in \mathbf{I_m}} \mu_{i,j} \overline{P}_{i,j} + \sum_{i,j \in \mathbf{I_{m,cb}}} \sigma_{i,j} \mu_{i,j} \overline{P}_{i,j} \qquad \forall m \tag{4.5d}$$

$$\text{s.t.} \quad |P_{i,j}| = \tilde{f}(G_i, d_i) \leq \overline{P}_{i,j} \qquad \forall i, j \in \mathbf{I_m}, \mathbf{I_{m,cb}} \qquad | \, \kappa_{i,j \in \mathbf{I_m}, \mathbf{I_{m,cb}}} \tag{4.5e}$$

Problem *P4a* is almost identical to *P2a* (and *P3a*), with the exception of one term in the objective function (4.5a). With a g-component, generators pay nodal instead of zonal prices for transmission ($\kappa_{i,j}$ instead of $\kappa_{m,n}$), depending on the impact of their nodal generation level on the grid infrastructure ($\tilde{f}(G_i, d_i)$). These prices are determined by the zonal TSOs via their flow-restriction (4.5e).

4.3 Numerical solution approach

Our approach to numerically solve the problem depicted in the previous section builds on the concept of decomposition. In fact, it follows the approach already depicted in the context of *Setting I* (Section 4.2.1), where we decomposed the integrated problem into two separate levels that are solved simultaneously and showed that they can – in economic terms – be interpreted as generation and transmission levels. According to Conejo et al. (2006), decomposition techniques can be applied to optimization problems with a decomposable structure that can be advantageously exploited. The idea of decomposition generally consists of splitting the optimization problem into a master and one or several subproblems that are solved iteratively. For the problem we are dealing with, namely the simultaneous optimization of generation and grid infrastructures under different congestion management designs and a varying number of TSOs, decomposing the overall problem entails two major advantages: First, the decomposition allows to easily implement variations of the generation and transmission levels including the underlying congestion management design. Hence, the model can be flexibly adjusted to represent the various settings described in the previous section. Second, the iterative nature of the solution process resulting from the decomposition allows to readily update PTDF matrices every time changes have been made in the grid infrastructure. This iterative update of the PTDF matrix, as suggested by Hagspiel et al. (2014), successively linearizes the non-linear optimization problem to ensure a consistent representation of generally non-linear grid properties, and allows for solving a corresponding linear problem. Linear problems can be solved effectively for global optima using standard techniques, such as the Simplex algorithm (e.g., Murty (1983)). Furthermore, the decomposition of linear problems preserves convexity and hence, also guarantees convergence towards the global optimum (e.g., Conejo et al. (2006)).[33] The obtained global optimum corresponds to an intertemporal equilibrium without uncertainty that is – due to the linear (and hence, convex) nature of our problem – unique. Moreover, in economic terms, the iterative algorithm to solve the decomposed problem can be readily interpreted as a price adjustment by a Walrasian auctioneer, also know as tatonnement procedure (e.g., Boyd et al. (2008)).

With some minor modifications, we can directly follow the (economically intuitive) formalization developed in the previous section and implement separate optimization problems representing the different tasks of generation and grid as well as the various settings (*I-IV*). We define the generation stage as the master problem, whereas the subproblem covers the transmission stage.[34] The principle idea of the solution algorithm is to solve the simultaneous generation and transmission stage problem iteratively, i.e., in a loop that runs as long as some convergence criterion is reached. In this process, optimized variables and marginal values are exchanged between the separated generation and grid levels reflecting the configuration of congestion management and TSO organization. For the settings described in the previous section, prices, which are iterated and thus adjusted, differ with respect to the information they contain and hence determine to which

[33]It should be noticed that despite the linearization and iterative solution, the non-linearity of the subproblem constraints still exists. However, even though not formally, Hagspiel et al. (2014) show that the problem converges for both, small test systems as well as large-scale applications. We confirm these findings in several model runs where we vary starting values over a broad range and did not find evidence against convergence. However, an analytical proof is still missing.

[34]Noticeably, the model could be inverted such that the master problem represents the grid sector which would, however, not change any of the results obtained.

degree efficiency can be reached. Compared to nodal pricing (*Setting I*), the other settings provide prices or products that describe the underlying problem only incompletely – and hence, entail an inefficient outcome.

The numerical algorithm to solve the nodal pricing model is sketched below. Parameters that save levels of optimal variables for usage in the respective other stage are indicated by $^{(\cdot)}$. It should be noticed that for the sake of comprehensibility, we still represent a simplified version of a more complete power system model that would need to account for multiple instances in time, multiple generation technologies, etc.. However, the extension is straightforward and does not change the principle approach depicted here.

Information passed from the transmission to the generation stage is captured by α, for which a constraint (lower bound) is added in each iteration u up to the current iteration v. This constraint consists of total grid costs $Y^{(u)}$ as well as the marginal costs each unit of trade $T_{i,j}$ is causing in the grid per node, denoted by $\kappa_{i,j}^{(u)}$. Both pieces of information are provided in the highest possible temporal and spatial resolution. As these components occur in the objective function of the generation stage (via α), the optimization will try to avoid the additional costs it is causing on the transmission stage, e.g., by moving power plant investments to alternative locations. The variable α is needed to correctly account for the impact of the transmission on the generation stage. On the transmission stage, the TSO is coping with the exchange (i.e., trade) of power stemming from the dispatch situation delivered by the master problem, thereby determining the marginal costs the trade is causing on the transmission stage, i.e., κ. Power flows are calculated by linearized load-flow equations represented by PTDF matrices mapping. The TSO then expands the grid such that it supports the emerging line flows at minimal costs.

$v = 1$; convergence=false

While(convergence=false) {

Master problem: generation

$$\min_{\overline{G}_i, G_i, T_{i,j}, \alpha} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \alpha \tag{4.6a}$$

$$\text{s.t.} \qquad G_i - \sum_j T_{i,j} = d_i \qquad \forall i \tag{4.6b}$$

$$G_i \leq \overline{G}_i \qquad \forall i \tag{4.6c}$$

$$Y^{(u)} + \sum_{i,j} \kappa_{i,j}^{(u)} \cdot f(T_{i,j} - T_{i,j}^{(u)}) \leq \alpha \qquad \forall u = 1, ..., v-1 | v > 1 \tag{4.6d}$$

$$- -$$

$$T_{i,j}^{(v)} = \text{Optimal value of } T_{i,j} \qquad \forall i, j \tag{4.6e}$$

Sub-problem: transmission

$$\min_{\overline{P}_{i,j}, T_{i,j}} \quad Y = \sum_{i,j} \mu_{i,j} \overline{P}_{i,j} \tag{4.6f}$$

$$\text{s.t.} \qquad |P_{i,j}| = \left| f(T_{i,j}^{(v)}) \right| = \left| PTDF \cdot T_{i,j}^{(v)} \right| \leq \overline{P}_{i,j} \quad \forall i, j \quad | \, \kappa_{i,j}^{(v)} \tag{4.6g}$$

$$- -$$

$$Y^{(v)} = \text{Optimal value of } Y \tag{4.6h}$$

if(convergence criterion < threshold; convergence=true)

$v = v + 1$

};

As regards the representation of settings *II-IV*, only very few modifications are needed compared to the nodal pricing regime (*Setting I*). The numerical algorithmic implementation of the various settings and modifications directly follows the procedure discussed in Section 4.2 and is thus not discussed again in detail here.[35]

[35] Nevertheless, for the sake of completeness and reproducibility, we have included one more complete model formulation illustrating the main differences of the other settings in Appendix 4.6.

4.4 Large-scale application

In this section, we apply the previously developed methodology to a detailed representation of the power sector in the Central Western European (CWE) region up to the year 2030. The application demonstrates the suitability of the modeling framework for large-scale problems and allows to assess and quantify the welfare losses in the considered region caused by different congestion management designs.

Given its historical, current and foreseen future development, the CWE region appears to be a particularly timely and relevant case study for different congestion management designs. In order to increase the market integration of European electricity markets towards an internal energy market, the European Union (EU) has declared the coupling of European electricity markets, which are organized in uniform price zones, an important stepping stone (see e.g., Glachant (2010), Commission de Régulation de l'Énergie (2009)). As for the cross-border capacity allocation, after a phase of NTC (Net Transfer Capacities) based market coupling, the CWE region is currently implementing a flow-based market coupling which is expected to increase the efficiency of the utilization of transmission capacities as well as overall social welfare (Capacity Allocating Service Company, 2014). Even though nodal pricing regimes have often been discussed for the European power sector (see, e.g., Ehrenmann and Smeers (2005) or Oggioni and Smeers (2012)), it can be expected that uniform price zones that correspond to national borders will remain. In fact, zonal markets coupled via a flow-based algorithm have been declared the target model for the European power sector (ACER, 2014).

In each zonal market, the respective zonal (i.e., national) TSO is responsible for the transmission network. Thereby, TSOs are organized and regulated on a national level, such that they can be assumed to care mainly about grid operation and expansion planning within their own jurisdiction. Although there are an umbrella organization (ENTSO-E) and coordinated actions, such as the (non-binding) European Ten-Year-Network-Development-Plan (TYNDP), the incentives of the national regulatory regime to intensify cross-border action might fall short of effectiveness. At the same time, Europe is heavily engaged in the large-scale deployment of renewable energies, hence causing fundamental changes in the supply structure. Generation is now often built with respect to the availability of primary renewable resources, i.e., wind and solar irradiation, and not necessarily close to load. This implies that the current grid infrastructure is partly no longer suitable and needs to be substantially redesigned, rendering an efficient congestion management even more important than before.

4.4.1 Model configuration and assumptions

The applied model for the generation stage belongs to the class of partial equilibrium models that aim at determining the cost-optimal electricity supply to customers by means of dispatch and investments decisions based on a large number of technological options for generation. As power systems are typically large and complex, these models are commonly set up as a linear optimization problem which can efficiently be solved. Our model is an extended version of the linear long term investment and dispatch model for conventional, renewable, storage and transmission technologies as presented in Richter (2011) and applied in, e.g., Jägemann et al. (2013) or Hagspiel et al. (2014). In contrast to previous versions, the CWE region, i.e., Belgium, France, Germany, Luxembourg and Netherlands, is considered with a high spatial (i.e., nodal) resolution. In order to account for exchanges with neighboring countries, additional regions are defined, but at an aggregated level: Southern Europe (Austria, Italy and Switzerland), South-West Europe (Portugal and Spain), North-West Europe (Ireland and UK), Northern Europe (Denmark, Finland, Norway and Sweden), and Eastern Europe (Czech Republic, Hungary, Poland, Slovakia and Slovenia). Figure 4.2 depicts the regional coverage and aggregation as they are represented in the model. In total, the model represents 70 nodes (or markets) and 174 power lines (AC and DC).

The model determines a possible path of how installed capacities will develop and how they are operated in the future assuming that electricity markets will achieve the cost-minimizing mix of different technologies which is obtained under perfect competition and the absence of market failures and distortions. Among a number of techno-economic constraints, e.g., supply coverage or investment decisions, the model also includes a number of politically implied constraints: nuclear power is phased-out where decided so, and then only allowed in countries already using it; a CO_2-Quota is implemented corresponding to currently discussed targets for the European energy sector, i.e., 20 percent reduction with respect to 1990 levels in 2020, and 40 percent in 2030 (European Commission, 2013b, 2014b); nation-specific 2020 targets for renewable energy sources are assumed to be reached until 2020 whereas from 2020 onwards there are no further specific renewable energy targets. At the same time, endogenous investments into renewable energy technologies are always possible.

The utilized model for the transmission stage is based on PTDF matrices which are calculated using a detailed European power flow model developed by Energynautics (see Ackermann et al. (2013) for a detailed model description). The number of nodes (70) corresponds to the nodal markets implemented in the generation market model and represents generation and load centers within Europe at an aggregated level. Those nodes are connected by 174 high voltage alternating current (AC) lines (220 and 380kV) as well as high voltage direct current (HVDC) lines. Even though the model is generally built for AC load flow calculations, it is here used to determine PTDF matrices for different grid expansion levels. As in Hogan et al. (2010), the law of parallel circuits is applied for the reactances' dependency on line capacity. Hence, each time an existing line with reactance x_1 is expanded by some amount characterized by reactance x_2 (dependent on the added line capacity), the new reactance becomes $1/x = 1/x_1 + 1/x_2$.

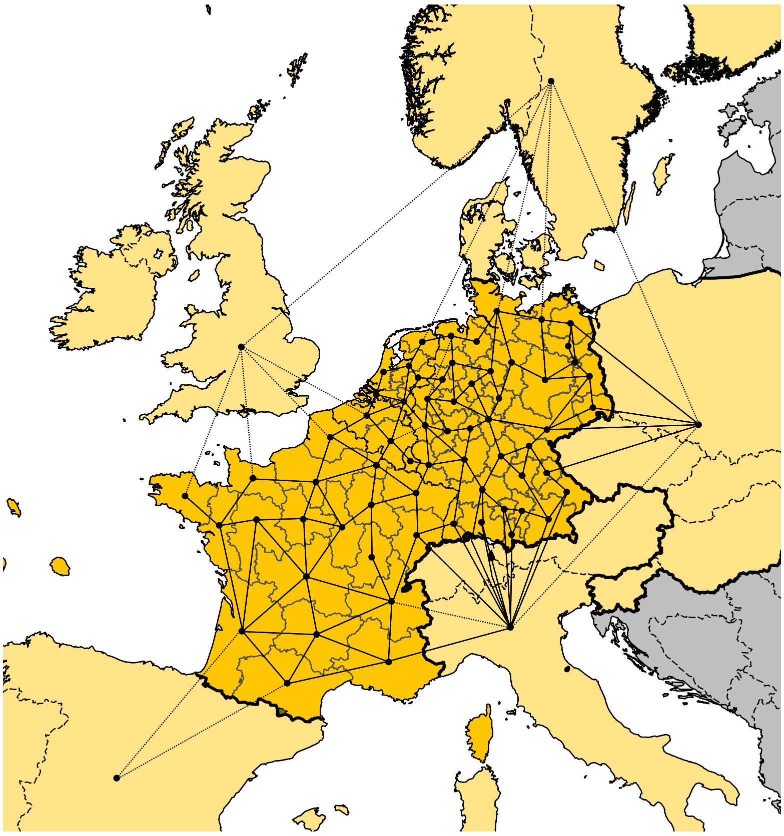


FIGURE 4.2: Representation of the CWE and neighboring regions in the model

As a starting point, the optimization takes the situation of the year 2011, based on a detailed database developed at the Institute of Energy Economics at the University of Cologne which in turn is largely based on the Platts WEPP Database (Platts, 2009). From these starting conditions, the development for the years 2020 and 2030 is optimized.[36] As for the temporal resolution, we represent the operational phase by nine typical days representing weekdays and weekend as well as variations in and interdependencies between demand and power from solar and wind. One of the typical days represents an extreme day during the week with peak demand and low supply from wind and solar. Specific numerical assumptions for the generation and transmission model can be found in Appendix 4.6.

[36]Technically, we implement the optimization routine up to 2050, but only report results until 2030. This is necessary to avoid problematic results at the end of the optimization timeframe.

As in *Settings II-IV* zonal markets are being considered, assumptions about the cross-border price function $g(\kappa_{i,j})$ are necessary. For the NTC-based coupling of market zones, we define function $g(\kappa_{i,j}) = 1.43 \cdot \frac{\kappa_{i,j}\overline{P}_{i,j}}{\sum_{i,j}\overline{P}_{i,j}}$ $\forall i,j \in \mathbf{I_{m,cb}}$ for each market border. The function consists of the weighted average of cross-border line marginals multiplied by a security margin. The security margin is the inverse of the ratio of NTC capacity to technical line capacity and has been derived heuristically by comparing currently installed cross-border grid capacities with NTC values reported by ENTSO-E for the CWE region. For flow-based market coupling, we set this security margin to one, in order to account for enhanced cross-border capacities provided to the power market.[37] In the case of zonal TSOs, we have made the following two assumptions: Differing interest of TSOs regarding cross-border line extensions are aligned by taking the smaller one of the two expansion levels.[38] The costs of cross-border lines are shared half-half by the two TSOs concerned, i.e., $\sigma_{i,j} = 0.5$.

As suggested in Conejo et al. (2006), we define our convergence criterion as the difference between an upper bound (total generation and grid costs) and a lower bound (total generation costs and grid costs that are visible to the generator) of the overall problem and demand it to undershoot a 2.5 percent threshold. This definition is based on the empirical observation that further improvements on the optimality error have little impact on the objective value and optimized capacities.

4.4.2 Results and discussion

We found that all settings converge to the optimal level in a range from around 20 to 60 iterations (corresponding to a solution time of 2 to 7 days). For practical reasons, we let all settings solve for one week and – after having double-checked that our convergence quality is met – take the last iteration for obtaining our final results.

To illustrate the convergent behavior of our problem, the following Figure 4.3, left hand side, shows the development of the optimality error (relative difference between the upper and lower bound of the optimization) along with the (absolute) rate of change of the lower bound obtained during the iterative solution of the nodal pricing setting. The lower bound is observed to change only slightly, reaching change rates smaller than

[37] Of course, this is just a simple representation of the cross-border capacity allocation. However, a more detailed representation is rather complex and would go beyond the scope of this paper. For more sophisticated models of flow-based capacity allocation, the reader is referred to Kurzidem (2010).

[38] Equation (4.7l) in Appendix 4.6. Note that this assumption may influence the equilibrium solution of the coordination between the TSOs. Due to the fact that the minimum of the line capacities is chosen, the solutions for the TSOs are no longer continuous. Hence, some equilibria might be omitted during the iterative solution of the problem. We accept this shortfall in our numerical approach for the sake of the large-scale application. The general approach, however, remains valid, and a process for determining all equilibria could be implemented in the numerical solution method (e.g., through randomized starting values).

0.01 percent after some 40 iterations. Moreover, as can be derived from the numbers presented in Figure 4.3, the relative error decreases at much faster rates with a ratio of approximately 200 for an estimated exponential trend and an iteration count of 60. Taking into account that the decomposition-based algorithm ensures a monotonically changing lower bound, it can be expected that the error further decreases mainly due to changes in the upper bound. Consequently, we argue that the lower bound can be taken as a good guess for the optimal objective value if our convergence criterion is met. A closer analysis of the optimized variable levels that reach stable levels during the same number of iterations is supporting this argument. As an example, the right hand side of Figure 4.3 shows aggregated AC line capacities obtained in the final runs of the nodal pricing setting.

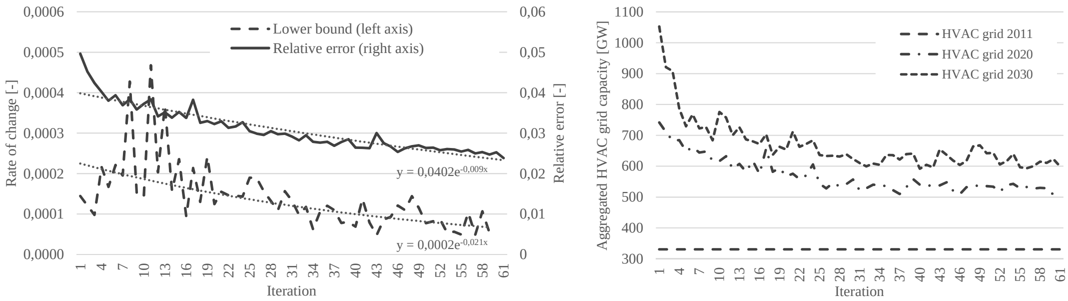


FIGURE 4.3: Development of lower bound, optimality error and aggregated AC line capacities during the iteration in *Setting I*

Costs are reported as accumulated discounted system costs.[39] In the generation sector, costs occur due to investments, operation and maintenance, production as well as ramping, whereas in the grid sector, investment as well as operation and maintenance costs are considered. Overall costs of electricity supply can be considered as a measure of efficiency and are reported in the following Figure 4.4 for the different settings. Besides the absolute costs, which are subdivided into generation and grid costs, the relative cost increase with respect to the overall costs of the nodal pricing setting is also depicted.

As expected, nodal pricing (*Setting I*) is most efficient, with total costs summing up to 899.0 bn. €$_{2011}$ (874.3 bn. for generation and 24.7 bn. for the grid). Overall, costs increase by up to 4.6 percent relative to *Setting I* for the other settings. Thereby, NTC-based market coupling induces highest inefficiencies of 3.8 percent and 4.6 percent for one single TSO or zonal TSOs, respectively, both with the possibility to do redispatch on a national basis (*Setting II-NTC* and *Setting III-NTC*). Hence, offering few amounts of trading capacity to the generation market, as implied by NTC-based market coupling, induces significant inefficiencies. In fact, by increasing trading capacities via flow-based market coupling, system costs can be lowered and inefficiencies amount to 2.5 percent

[39]The discount rate is assumed to be 10 percent throughout all calculations.

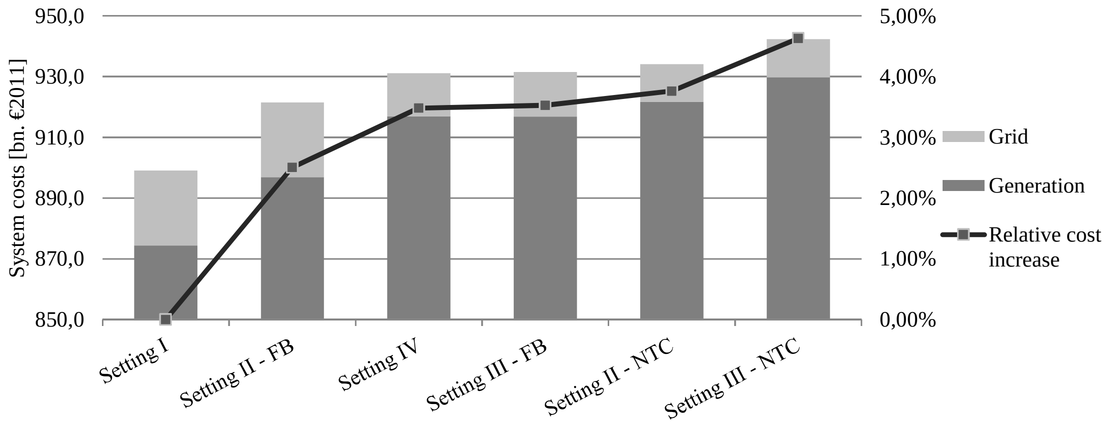


Figure 4.4: Total costs and relative performance of the different settings

for the single TSO, respectively 3.5 percent for zonal TSOs compared to nodal pricing (*Setting II-FB* and *Setting III-FB*). Hence, efficiency gains of 1.1-1.3 percent of total system costs can be achieved by switching from NTC to flow-based market coupling. In turn, enhanced trading activities induced by flow-based market coupling entail greater TSO activity, both in the expansion as well as in the redispatch. For this reason, TSO costs are higher for flow-based than for NTC-based market coupling. However, these additional costs are overcompensated by lower costs in the generation sector. The net effect of a switch from NTC to flow-based market coupling is beneficial for the overall system.

Somewhat surprisingly, the national g-component (*Setting IV*) hardly performs better than the same setting with redispatch (*Setting III-FB*). Hence, the optimal allocation of power generation within market zones is hardly influenced by grid restrictions within that zone. In contrast, the optimal allocation induced by nodal prices throughout the CWE region entails substantial gains in efficiency due to reduced system costs. The setting that comes closest to nodal pricing consists of flow-based coupled zonal markets with a single TSOs and induces an inefficiency of 2.5 percent in comparison to nodal pricing (*Setting II-FB* vs. *Setting I*).

Even though the share of TSO costs on total costs is very small compared to the share of generation costs in all settings (1.3-2.7 percent)[40], the amount of grid capacities varies greatly between the different settings. Figure 4.5 shows the aggregated high voltage (HV) AC and HVDC line capacities.

Grid capacities are generally lower in the case of zonal TSOs where they only agree on the smaller of the two proposed expansion levels for cross-border lines (*Setting III-FB* and *Setting III-NTC*). In these cases, overall AC grid capacities increase from 331 GW

[40] The rather minor role of grid costs compared to costs occurring in the generation sector has already been identified, e.g., in Fürsch et al. (2013).

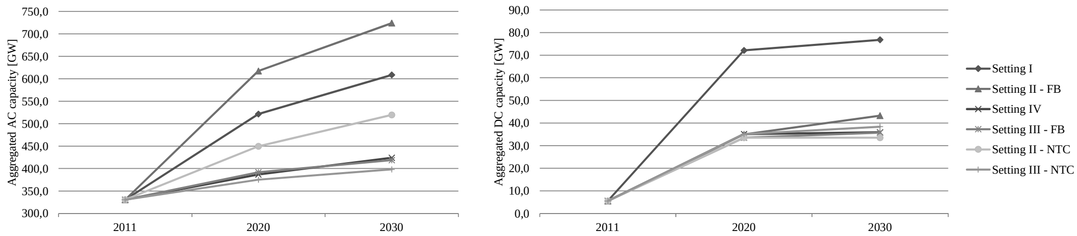


FIGURE 4.5: Aggregated line capacities AC and DC

in 2011 to 398 GW (*Setting III-NTC*) respectively 418 GW (*Setting III-FB*) in 2030, corresponding to an increase of 20-28 percent. In case of a single TSO, cross-border along with overall line expansions are significantly higher compared to zonal TSOs, with 2030 levels reaching 519 GW (*Setting II-NTC*) to 724 GW (*Setting II-FB*). Especially in *Setting II-FB*, the TSO is obliged to cope with inefficiently allocated generation plants by excessively expanding the grid, while not being able to avoid those measures with suitable price signals. DC line expansions appear to be crucial for an efficient system development, especially towards the UK where large wind farms help to reach CO_2-targets and to supply the UK itself as well as the continent with comparatively cheap electricity. Thereby, the high DC expansion level in the nodal pricing regime is remarkable. Whereas in zonal markets prices are "averaged" across the zone, nodal prices reveal the true value of connecting specific nodes via DC-lines and thus enable efficient investments in those projects. In consequence, in the nodal pricing regime, DC line capacities are about double as high as in the other settings. This helps to reduce overall costs to a minimum (*Setting I*).

Besides the overall level of grid and generation capacities, their regional allocation also differs between the various settings, mainly due to differences in the (local) availability of transmission upgrades. As has been seen, higher grid expansion levels result from a single TSO (*Setting I and Setting II*), enabling a better utilization of renewable energies at favorable sites (i.e., sites where the specific costs of electricity generation are lowest). In Figure 4.6, we exemplarily illustrate this effect based on a cross-border line between France and Germany (line 80 in our model). However, the same effect is observable for other interconnections, e.g., between France and Belgium. Higher grid capacities allow the use of high wind speed locations in Northern France and thus foster more expansion of wind capacities in this area. In case of zonal TSOs (*Setting III and Setting IV*) only low amounts of wind capacity are built in France (e.g., in node FR-06) as these areas cannot be connected with the rest of the system. To still meet the European CO_2-target, PV power plants are built in the southern part of Germany (e.g., in node DE-27). Obviously, these locations are non-optimal with respect to other options as they are not used in the setting with one TSO. Thus, implemented market designs significantly influence the amount and location of renewable energies within the system.[41]

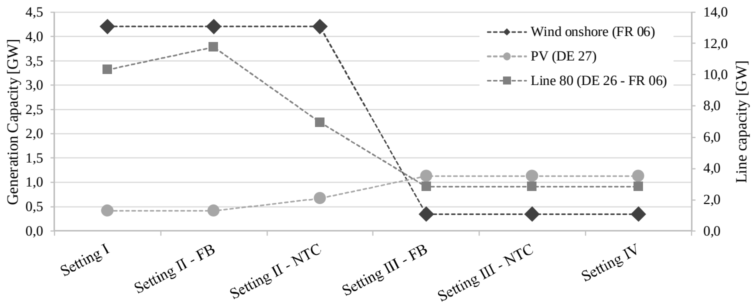


FIGURE 4.6: Exemplary grid expansion and regional allocation of renewable energies

4.5 Conclusions

In the context of liberalized power markets and unbundled generation and transmission services, the purpose of this paper was to develop a modeling framework for different regulatory designs regarding congestion management including both, the operation as well as the investment perspective in the generation and transmission sector. We have presented an analytical formulation that is able to account for different regulatory designs of market areas, a single or zonal TSOs, as well as different forms of measures to relieve congestion, namely grid expansion, redispatch and g-components. We have then proposed an algorithm to numerically solve these problems, based on the concept of decomposition. This technique has shown to entail a number of characteristics that work to our advantage, especially flexible algorithmic implementation as well as consistency of the grid flow representation through PTDF update.

Calibrating our model to the CWE region, we have demonstrated the applicability of our numerical solution algorithm in a large-scale application consisting of 70 nodes and 174 lines along with a detailed bottom-up representation of the generation sector. Compared to nodal pricing as the efficient benchmark, inefficiencies induced by alternative settings reach additional system costs of up to 4.6 percent. Major deteriorative factors are TSOs' activities restricted to zones as well as low trading capacities offered to the market. These findings may serve as a guideline for policymakers when designing international power markets. For instance, our results confirm ongoing efforts to implement flow-based market coupling and to foster a closer cooperation of TSOs in the CWE region. In fact, we find that such a regulatory design could come close to the nodal pricing benchmark, with an efficiency difference of only 2.5 percent. Noticeably, the magnitude

[41] Conventional capacities are also affected. However, the effect is less pronounced as the differences between the site-specific costs of generation are smaller.

of these results should be interpreted as the lower bound of efficiency gains, since we focus on frictions in the congestion management only.

More generally, we find that a single TSO (or enhanced coordination between the zonal TSOs) is key for an efficient development of both, grid and generation infrastructures. Whereas the expansion of grid infrastructure is immediately affected, the generation sector indirectly takes advantage of increased grid capacities and hence, can develop more efficiently. Better allocation of generation units with respect to grid costs through high resolution price signals gains importance for larger geographical areas and larger differences between generation costs and expansion potentials (such as wind or solar power). This has been found for the CWE region, and may prove even more important for the whole of Europe. It should be noted, however, that efficiency gains need to be put into the context of transaction costs occurring from the switch to a different congestion management design. In addition, socio-economic factors such as acceptance for grid expansion are not considered in the analysis, but might also play a role considering the large differences of necessary grid quantities.

Limitations of our approach that leave room for extensions and improvement stem from the fact that we assume linear transmission investments, and do not consider strategic behavior of individual agents, imperfectly regulated TSOs, or uncertainty about future developments (e.g., delays in expansion projects). The assumption of an inelastic demand probably reduces the magnitude of the measured inefficiencies, since demand does not react to any price changes and hence only supply-side effects are captured. Algorithmically, the effectiveness of our solution process could be further improved, e.g., through better usage of numerical properties of the problem (such as gradients, etc.). Nevertheless, in its present form, our framework may serve as a valuable tool to assess a number of further relevant questions, such as the tradeoff between different flexibility options (such as grids, storages or renewable curtailment), the impact of different forms of congestion management in other European regions, or the valuation of grid expansion projects.

4.6 Appendix

Model of NTC-coupled zonal markets, zonal TSOs, and zonal redispatch

In Section 4.3, we have shown the numerical implementation of the nodal pricing regime. For the sake of clarifying the major changes needed to represent the alternative Settings *II-IV*, we here present the model for m zonal (instead of nodal) markets that are coupled via NTC-based capacity restrictions, along with multiple zonal TSOs (instead of only one), all having the possibility to deploy zonal redispatch as an alternative to grid expansion. Hence, the model corresponds to *Setting III* with NTC-based market coupling. Compared to nodal pricing, no more nodal or time-specific information about grid costs is provided. Instead, an aggregated price $\kappa_{m,n}^{(v)}$ for each border is calculated via a function g_{NTC} and passed on to generation level. The model with flow-based market coupling works in the same way, only that the price $\kappa_{m,n}^{(v)}$ is calculated via a different function g_{FB}.

$v = 1$; convergence=false

While(convergence=false) {

Master problem: generation

$$\min_{\overline{G}_i, G_i, T_{m,n}, \alpha} \quad X = \sum_i \delta_i \overline{G}_i + \sum_i \gamma_i G_i + \alpha \tag{4.7a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_i - \sum_n T_{m,n} = \sum_{i \in \mathbf{I_m}} d_i \quad \forall m \tag{4.7b}$$

$$G_i \leq \overline{G}_i \quad \forall i \tag{4.7c}$$

$$\sum_m Y_m^{(u)} + \sum_{m,n} \kappa_{m,n}^{(u)} \cdot (T_{m,n} - T_{m,n}^{(u)}) \leq \alpha \quad \forall u = 1, ..., v-1 | v > 1 \tag{4.7d}$$

$$- -$$

$$G_i^{(v)} = \text{Optimal value of } G_i \quad \forall i \tag{4.7e}$$

Sub-problem: transmission

$$\min_{\overline{P}_{i,j\in\mathbf{I_m},\mathbf{I_{m,cb}}},R_{i\in\mathbf{I_m}},T_{m,n}} \quad Y_m = \sum_{i,j\in\mathbf{I_m}} \mu_{i,j}\overline{P}_{i,j} + \frac{1}{2}\sum_{i,j\in\mathbf{I_{m,cb}}} \mu_{i,j}\overline{P}_{i,j} + \sum_{i\in\mathbf{I_m}} R_i\gamma_i \;\forall m \quad (4.7f)$$

$$\text{s.t.} \quad |P_{i,j}| = \left|\tilde{f}(G_i^{(v)}, d_i)\right| = \left|\sum_{i'} PTDF\cdot(G_i^{(v)} - d_i)\right| \leq \overline{P}_{i,j} \qquad \forall i,j \qquad |\,\kappa_{i,j}^{(v)} \quad (4.7g)$$

$$0 \leq R_i + G_i^{(v)} \leq \overline{G}_i \qquad \forall i \in \mathbf{I_m} \quad (4.7h)$$

$$\sum_{i\in\mathbf{I_m}} R_i = 0 \quad (4.7i)$$

– –

$$Y_m^* = \text{Optimal value of } Y_m \quad (4.7j)$$

$$\kappa_{m,n} = g_{NTC}(\kappa_{i,j}^{(v)}) \quad (4.7k)$$

$$\overline{P}_{i,j\in\mathbf{I_{m,cb}}} = \overline{P}_{i,j\in\mathbf{I_{n,cb}}} = \min\left\{\overline{P}_{i,j\in\mathbf{I_{m,cb}}};\overline{P}_{i,j\in\mathbf{I_{n,cb}}}\right\} \quad (4.7l)$$

if(convergence criterion < threshold; convergence=true)

$v = v + 1$

};

Numerical assumptions for the large-scale application

Country	2011	2020	2030
Belgium	87	98	105
Germany	573	612	629
France	466	524	559
Luxembourg	7	8	8
Netherlands	113	128	137
Eastern	276	328	366
Northern	387	436	465
Southern	450	528	594
Southwest	317	378	433
United Kingdom	400	450	481

Table 4.3: Assumptions for the gross electricity demand [TWh]

To depict the CWE region in a high spatial resolution, we split the gross electricity demand per country among the nodes belonging to this country according to the percentage of population living in that region.

Technology	2020	2030
Wind Onshore	1,253	1,188
Wind Offshore (<20m depth)	2,800	2,350
Wind Offshore (>20m depth)	3,080	2,585
Photovoltaics (roof)	1,260	935
Photovoltaics (ground)	1,110	785
Biomass gas	2,398	2,395
Biomass solid	3,297	3,295
Biomass gas, CHP	2,597	2,595
Biomass solid, CHP	3,497	3,493
Geothermal	10,504	9,500
Compressed Air Storage	1,100	1,100
Pump Storage	1,200	1,200
Lignite	1,500	1,500
Lignite Innovative	1,600	1,600
Coal	1,200	1,200
Coal Innovative	2,025	1,800
IGCC	1,700	1,700
CCGT	711	711
OCGT	400	400
Nuclear	3,157	3,157

Table 4.4: Assumptions for the generation technology investment costs [€/kW]

Fuel type	2011	2020	2030
Nuclear	3.6	3.3	3.3
Lignite	1.4	1.4	2.7
Oil	39.0	47.6	58.0
Coal	9.6	10.1	10.9
Gas	14.0	23.1	25.9

Table 4.5: Assumptions for the gross fuel prices [€/MWh$_{th}$]

Grid Technology	Extension costs	FOM costs
AC overhead line incl. compensation	445 €/(MVA*km)	2.2 €/(MVA*km)
DC overhead line	400 €/(MW*km)	2.0 €/(MW*km)
DC underground	1,250 €/(MW*km)	6.3 €/(MW*km)
DC submarine	1,100 €/(MW*km)	5.5 €/(MW*km)
DC converter pair	150,000 €/MW	750.0 €/MW

Table 4.6: Assumptions for the grid extension and FOM costs

5 The relevance of grid expansion under zonal markets

5.1 Introduction

The market design of the European single market for electricity consists of regional bidding zones, usually aligned to national borders. There is one uniform price per zone, while implicitly neglecting scarce transmission capacities within these zones. In reality, however, this simplification is often inconsistent with physical realities and hence, represents and inherent market incompleteness. In fact, aggregated zonal prices conceal important information regarding scarcities in the transmission grid that would be important to coordinate market participants in an efficient way. In the short term, this incompleteness is addressed by the redispatch of generation facilities: After the market clearing, generation units are requested to modify their scheduled dispatch by increasing or decreasing their production level in order to relieve congestion in the grid. An increase in generation is remunerated with the estimated variable costs, partly covered by the saved variable costs of the decreased generation. If the cost estimations were correct and the redispatch measure succeeded in finding the least cost alternative, the short-term market outcome would be optimal, i.e., statically efficient.[42]

In the long term, functionality of zonal markets shall be ensured by sufficient expansion of the grid infrastructure. In practice, however, grid expansion is often *insufficient* or at least delayed. For Europe, 30 percent of the projects depicted in the Ten Year Network Development Plan (TYNDP) are reported delayed or rescheduled (ENTSO-E, 2015). For Germany, even 50 percent of the projects are reported delayed (Bundesnetzagentur, 2013c). There may be various obstacles causing delays of planned grid expansion. One of the main reasons are long and inefficient approval procedures (e.g., Schneider and Battaglini (2013), Steinbach (2013)). But even if approval procedures are successfully

[42]In practice, this result will probably not be entirely realized due to ramping constraints of redispatched power plants.

completed, missing social acceptance may further delay the realization of infrastructure projects (e.g., ENTSO-E (2010), Schneider and Battaglini (2013)). According to ENTSO-E (2015), these most prominent obstacles account for one third of investment delays.

At the same time, due to the uniform price for all market participants, the resulting intra-zonal scarcity in transmission capacities is not taken into account in the investment decisions of generation. In fact, (zonal) markets should ensure that sufficient capacity is installed to meet demand on a zonal level. However, there might be a misallocation of generation capacities within the zones due to missing locational price signals. These misallocations are exacerbated by missing grid capacities that might not allow to transport electricity to the customers. Thus, missing grid expansion might severly jeopardize the long-term functionality of zonal markets. Especially, although redispatch might induce efficient market outcomes in the short term, it does not suffice to heal the incompleteness of the market design to achieve long-term, i.e., dynamic, efficiency as locational price signals are not considered.[43] As we will show, this might induce severe inefficiencies in the market outcome, which are increasing with the level of grid restriction.

In Europe, the effect of misallocation of generators and missing transmission capacity is particularly relevant due to fundamental changes in the supply and demand structure caused by strong climate protection efforts.[44] A substantial shift from conventional to renewable generation, which is usually far away from current generation and load centers, increases the importance of sufficient grid infrastructure. A blueprint for the described dynamics in a zonal market design with an increasing share of renewables is the case of Germany, where short-term intra-zonal congestion is removed using redispatch measures. In order to avoid situations where redispatch would be necessary but no generation capacities are available at the right location, the German Transmission System Operators (TSO) contract generation capacity in advance at locations which are expected to be relevant for future congestion relieve. This so-called grid reserve ensures locally sufficient generation capacity. Table 5.1 illustrates the development of the renewables share, redispatch measures as well as the grid reserve quantity. As can be seen, redispatch measures broadly increased with an increasing share of renewables, caused by missing transmission grid capacities. Meanwhile, also the grid reserve quantities increase. This development clearly shows the effect and the deficits of the zonal market design.

[43] See Burstedde (2012) for a detailed discussion of the (in-)efficiencies of several redispatch designs.
[44] The European Union (EU) formulated an ambitious 2030 energy strategy, including a EU domestic reduction of greenhouse gases (GHG) by 40 percent compared to 1990, a share of 27 percent renewable energy, and a 27 percent reduction in energy consumption compared to 2005 (European Commission, 2014a).

Year	2010	2011	2012	2013	2014	2015 (Q1/2)
Renewable share of gross electricity demand [%]	16.6	20.2	22.8	23.9	25.8	-
Redispatch volume [GWh]	-	-	4,956	4,604	5,197	5,253
Redispatch costs [Mio. €]	48	129	165	115	139	252.5
Grid reserve [GW]	-	1.6	2.5	2.5	3.1	6.7-7.8

Table 5.1: Development of renewable share, redispatch measures and grid reserve in Germany[45]

In the literature, several papers investigate grid expansions in the short as well as in the long term. Schaber et al. (2011) analyze the importance of transmission grid expansion for the integration of renewables in Europe with a linear dispatch and investment model. They state that grid integration costs amount to 25 percent of renewables investment costs. Schaber et al. (2012) analyze grid expansions for the European system in 2020 given an increasing share of wind and solar power. A cost-minimization model of the European power system is applied under the assumption of a nodal pricing regime. Results indicate lower electricity prices in proximity of renewables and benefits for conventional plants in case of grid expansion. Optimal grid expansion amounts to 20 percent respectively 60 percent of today's capacity - depending on the considered scenario. Fürsch et al. (2013) quantify the benefits of optimal grid expansion up to 2050 by applying a dispatch and investment model coupled with a load flow grid model that determines NTC values for the market coupling. They compare optimal grid expansion with moderate expansion of interconnector capacities, and find that with a large share of renewables, high grid expansion (+76 percent capacity compared to today) proves to be optimal to exploit good renewable sites. The linkage between renewables, grid expansion, and generation backup capacities was investigated by Steinke et al. (2013). They use a stylized model to analyze the effects of grid expansion on the necessity of backup capacities and storage. They find that an ideal grid reduces the need of backup capacities from 40 percent to 20 percent with a share of 100 percent renewables. Hagspiel et al. (2014) analyze the optimal grid expansion until 2050 using a linear dispatch and investment model coupled with an AC grid model via Power Transfer Distribution Factors (PTDF). They find that minimal grid expansion for achieving an ambitious CO_2 reduction target of 90 percent leads to an increase of 21 percent of total system costs. Oggioni and Smeers (2012) as well as Kunz (2013) deal with the impact of zonal markets and redispatch for Germany and Europe in the short run. In doing so, Oggioni and Smeers (2012) use a six node model and Kunz (2013) a European short-term electricity market model. Grimm et al. (2016) build on a trilevel modelling approach investigating the long run impact of different market designs. They apply their theoretical model to a three and six node case study and find that investment decisions of firms and TSOs do not have to lead to the social optimum in a market environment. Bertsch et al. (2016) develop a theoretical

[45]Sources: Bundesnetzagentur (2012), Bundesnetzagentur (2013a), Bundesnetzagentur (2013b), Bundesnetzagentur (2013c), Bundesnetzagentur (2014b), Baake (2014), Bundesnetzagentur (2015a), Bundesnetzagentur (2015b), AGEB (2015)

framework to analyse and compare different market designs. In a large-scale application it is shown that zonal markets with redispatch lead to inefficiencies compared to nodal pricing, representing the first best.

We contribute to the existing literature by investigating the particular relevance of grid expansion under zonal markets. We show that the market design is inherently incomplete due to missing price signals, and that important scarcities in the grid are not properly considered for investment decisions. For this, we build on a long-term fundamental model of the European electricity market developed in Bertsch et al. (2016), allowing the representation of the European flow-based coupled zonal markets with redispatch. The model includes generation dispatch, power flows, as well as generation and grid investments. In contrast to Bertsch et al. (2016), we implement the EU 2030 energy strategy to ensure the results are in line with current European policies. Furthermore, we extend the analysis by designing six scenarios that differ with respect to their level of allowed grid expansion. We are hence able to investigate in great detail the relevance of grid expansion for the market outcome.

Our results show that restricted grid expansion together with the inherent incompleteness of the market design has significant effects. We restrict grid expansion per decade from zero, i.e., no grid expansion at all, to 30 TWkm throughout 6 different scenarios. In case of restrictions ranging from 0-15 TWkm of grid expansion per decade, there are energy imbalances of up to 2 percent (3 percent) for 2020 (2030). Also with less restricted grid expansion, these energy imbalances still amount to more than 0.2 percent for scenarios 15 TWkm in 2020. In 2030, however, significant energy imbalances only occur for the scenarios of restrictions of up to 5 TWkm. The highest energy imbalances are found to be in Southern Germany. Thereby, energy imbalances indicate that generation is missing at some locations, entailing the need to either provide additional generation capacity outside of the market (e.g., by means of a grid reserve as in Germany), or to curtail load. Furthermore, no grid expansion jeopardizes the achievement of the EU 2030 climate targets: the share of renewables is 1.5 percentage points lower than in any other scenario, resulting from a curtailment of up to 7.7 percent of photovoltaic (PV) generation and over 3 percent of wind generation. Missing grid expansion hence results in higher CO_2 emissions in the power sector and implies the need to shift CO_2 emissions from the power sector to other, probably more expensive sectors. DC lines are found to be of particular value for the integration of renewables as they allow point-to-point transfers from renewables generation to load sites.

Overall, the results demonstrate the shortfalls of the zonal market design in the light of restricted grid expansion which is a scenario that appears to be very likely. The more restricted grid expansion is, the more administrative intervention will be needed to avoid energy imbalances possibly causing expensive and politically unwanted load curtailment. One alternative might be to administratively contract generation capacity outside the market. To overcome this problem, a redefinition of zones or introduction of locational price elements may be a suitable way to effectively reduce the amount of administrative intervention. Furthermore, obstacles for grid expansion should be removed in order to ensure sufficient levels of grid to connect generation and load.

The paper is structured as follows: Section 2 introduces the model and numerical assumptions. Results are presented in Section 3. Section 4 concludes.

5.2 Methodology

5.2.1 Model

To simulate the development of the electricity system with a flow-based coupling of zonal markets, we follow the approach described in Bertsch et al. (2016) and combine a cost-minimizing dynamic linear investment and dispatch market model with a model of the AC grid using a linear Power Transfer Distribution Factor (PTDF) representation of the load-flow. To deal with the non-linear dependence of the PTDFs on the grid impedances, the models are solved iteratively by updating PTDF matrices until convergence is achieved as proposed in Hagspiel et al. (2014).

The model represents an intertemporal equilibrium model that simultaneously solves the operation and investment of generation and the transmission infrastructure. The model relies on a set of simplifying assumptions: we assume perfect competition among generators, perfect regulation of TSOs, the absence of transaction costs and uncertainty, inelastic demand, continuous functions and continuous PTDF matrices. For a thorough discussion of the model development and characteristics, the reader is referred to Bertsch et al. (2016).

We make use of a separated representation of generation and transmission in order to represent the status quo of European electricity markets with unbundled generation and transmission firms. The separation of the problem helps us to implement the market incompleteness induced by the zonal market design and the related information deficit. Our iterative solution algorithm is based on two stages that are solved sequentially. First, the generation market equilibrium is determined by minimizing generation and investment costs while meeting an (inelastic) demand and considering inter-zonal transmission capacities. This implies that the zonal market for electricity supply and demand (including both, operation and investment) only considers interconnectors (and no intra-zonal grid congestion). The solution represents the result of perfect competition in the electricity market under flow-based market coupling. Technologies for balancing supply and demand in each zone are conventional and renewable generation technologies as well as storage[46]. In the second stage, the TSO is in charge of investments in transmission

[46]We consider pumped hydro storage, hydro storage dams and the possibility to build new Compressed Air Energy Storages (CEAS) from 2020 onwards.

capacities as well es the operation of redispatch measures across borders given the market results of the first stage. This represents one perfectly incentivized Transmission System Operators (TSO) (or several perfectly coordinated and incentivized TSOs) for all considered markets with the objective of minimizing its (their) costs while keeping the system stable, i.e., matching zonal demand and supply while ensuring that no line is overloaded. At the transmission level, either AC or DC interconnections are available. While the DC interconnections allow direct transfers of electricity between neighboring regions, the utilization of the AC grid is subject to loop flows represented by the PTDF.

Equations (5.1a)-(5.1l) state a simplified yet representative formulation of the problem:[47,48] At the generation stage, total costs X are minimized such that an exogenously defined demand d per zone $m, n \in M$ is met at all points in time t. Zonal demand is determined by aggregating nodal demand levels $d_{i,t}$ for all nodes within a zone $i \in I_m$.[49] Costs for generation technologies consist of the variable costs $\gamma_{i,t}$ for generation $G_{i,t}$ and the yearly fixed and (annualized) investment costs $\delta_{i,y}$ for the generation capacity $\overline{G}_{i,y}$. Both types of costs may change over time (note that y represents instances of investment, e.g., years, while t are dispatch situations, e.g., hours). Generation at a node is restricted by the installed capacity (Equation 5.1c). To balance supply and demand in zone m, generation in that zone may be complemented by trades $T_{m,n,t}$ from other zones n. Thereby, each trade from zone m to zone n equals the negative trade from zone n to zone m and is in turn restricted by inter-zonal transmission capacities $\overline{P}_{m,n,t}$ (Equation 5.1d).

The second stage consists of minimizing costs Y occurring at the transmission level due to grid expansion and redispatch. The grid can be expanded by adding line capacity between two nodes at costs $\mu_{i,j,y}$, while redispatch quantities $R_{i,t}$ have the same variable costs $\gamma_{i,t}$ as in the generation stage. Negative redispatch quantities can be only as high as generation levels obtained at the first stage, while positive redispatch quantities are restricted by generation capacities (Equation 5.1g). The sum of all (positive and negative) redispatch measures has to amount to zero (Equation 5.1h) to keep the system balanced. Generation (including generation and redispatch), demand as well as the existing infrastructure induce power flows on transmission lines that are restricted

[47]The depicted model is a simplified version of the model used for the large-scale application. The large-scale model includes amongst others technical (e.g., minimal load, maximum load, ramping, etc.), political (e.g., nuclear phase-out), as well as environmental (European CO_2 quota) constraints that are for reasons of clarity neglected in the theoretical framework. A more detailed representation of the market model (generation stage) may be found in Richter (2011) or Jägemann et al. (2013), while the AC grid model (transmission stage) is described in Hagspiel et al. (2014).

[48]A detailed overview containing all parameters, variables and sets is depicted in Table 5.5 in the Appendix.

[49]In the numerical simulation, we use interdependent hours and type days and scale the volumes to yearly quantities. Furthermore, costs are discounted to the starting year. Several generation technologies with different characteristics such as peak or base load exist at each node. However, for the sake of simplification we omit these model properties in this formulation.

by transmission capacities $\overline{P}_{m,n,y}$ (Equation 5.1i).[50] The exchange between the generation and the transmission stage takes place via the inter-zonal transmission capacities $\overline{P}_{m,n,t}$.[51] Thereby, function g determines those inter-zonal transmission capacities for each dispatch time t (that are provided to the generation market, i.e., the first stage of the model) based on grid capacities $\overline{P}_{i,j,y}$, generation $G_{i,t}$, demand $d_{i,t}$, redispatch $R_{i,t}$ and a flow-based market coupling regime (see, e.g., Aguado et al. (2012)). The expansion of transmission capacities $\overline{P}_{i,j,y}$ times line length $l_{i,j}$ per decade b is restricted by some value z. The model is re-run with stepwise changes of capacity restriction levels z, thus allowing a fine-grained identification of the effects of limited grid expansion.[52] Due to the functional relationship of trades and transmission capacities, the market clearing condition has to reoccur on the transmission stage (Equation 5.1f). Trades from zone m to n are again equal to the negative trade from zone n to m (Equation 5.1l).

To demonstrate the deficiencies of the zonal market with redispatch, we can compare problem (5.1a)-(5.1l) with a two-stage nodal pricing regime representing the first-best benchmark. The corresponding model can be found in the Appendix. The main difference stems from zonal markets m, n being replaced by nodes i, j. As a consequence, TSOs need no redispatch $R_{i,t}$ as locational price signals are directly incorporated in the dispatch. All relevant information is available to all market participants at all times, making nodal pricing the first-best efficient benchmark. For a thorough comparison of different market designs and their performance including zonal as well as nodal pricing regimes, the reader is referred to Bertsch et al. (2016).

Generation

$$\min_{\overline{G}_{i,y}, G_{i,t}, T_{m,n,t}} \quad X = \sum_{i,y} \delta_{i,y} \overline{G}_{i,y} + \sum_{i,t} \gamma_{i,t} G_{i,t} \tag{5.1a}$$

$$\text{s.t.} \quad \sum_{i \in \mathbf{I_m}} G_{i,t} - \sum_{n} T_{m,n,t} = \sum_{i \in \mathbf{I_m}} d_{i,t} \qquad \forall m, t \tag{5.1b}$$

$$G_{i,t} \leq \overline{G}_{i,y} \qquad \forall i, t \tag{5.1c}$$

$$T_{m,n,t} = -T_{n,m,t} \leq \overline{P}_{m,n,t} \qquad \forall m, n, t \tag{5.1d}$$

[50] In our case power flows are represented by PTDFs that are treated as a parameter while solving the transmission stage, such that Equation (5.1i) becomes a linear constraint. However, we account for non-linearities in the load flow equations by updating PTDFs based on the new transmission capacities when iterating with the AC grid model.

[51] Note that this approach differs from Bertsch et al. (2016), where the exchange worked via transmission capacity marginals.

[52] Note that we use j, k and q as alias for i in order to represent different nodes in the formulation.

Transmission

$$\min_{\overline{P}_{i,j,y},R_{i,t}} \quad Y = \sum_{i,j,y} \mu_{i,j,y}\overline{P}_{i,j,y} + \sum_{i,t} \gamma_{i,t} R_{i,t} \tag{5.1e}$$

$$\text{s.t.} \quad \sum_{i\in\mathbf{I_m}} G_{i,t} - \sum_{n} T_{m,n,t} = \sum_{i\in\mathbf{I_m}} d_{i,t} \qquad \forall m,t \tag{5.1f}$$

$$0 \leq G_{i,t} + R_{i,t} \leq \overline{G}_{i,y} \qquad \forall i,t \tag{5.1g}$$

$$\sum_{i} R_{i,t} = 0 \qquad \forall t \tag{5.1h}$$

$$|P_{i,j,t}(\overline{P}_{k,q,y}, G_{k,t}, d_{k,t}, R_{k,t})| \leq \overline{P}_{i,j,y} \qquad \forall i,j,t \tag{5.1i}$$

$$\overline{P}_{m,n,t} = g(\overline{P}_{i,j,y}, G_{i,t}, d_{i,t}, R_{i,t}) \tag{5.1j}$$

$$\sum_{y\in b} \overline{P}_{i,j,y} l_{i,j} \leq \sum_{y\in b-1} \overline{P}_{i,j,y} l_{i,j} + z \qquad \forall b \tag{5.1k}$$

$$T_{m,n,t} = -T_{n,m,t} \quad \forall m,n,t \tag{5.1l}$$

5.2.2 Numerical assumptions

The geographical scope of the simulation, as shown in Figure 5.1, contains a high-resolution nodal representation of the Central Western European (CWE) region, and an aggregated representation of the neighboring countries.[53] The CWE region consists of 5 zonal markets where nodes within the zones are aggregated and zones correspond to national borders (Belgium, Netherlands, Luxembourg, Germany and France) that are coupled via inter-zonal transmission capacities offered to the market. These transmission capacities are determined based on a flow-based market coupling regime (CWE FBMC, 2014). Due to limitations in publicly available data and models, our approach to model flow-based market coupling only considers cross-border line capacities. Note that in practice, the mechanism applied in the CWE region may in addition include some intra-zonal congestion in the capacity allocation.[54] Physical feasibility of the dispatch on the grid level is ensured by a cross-border redispatch. To account for trades with neighboring countries, 5 satellite regions are included: Southern Europe (Italy, Austria[55] and Switzerland), Eastern Europe (Czech Republic, Poland, Hungary, Slovakia and Slovenia), Northern Europe (Sweden, Norway, Finland and Denmark), South West Europe (Spain and Portugal) and North West Europe (UK and Ireland). The transmission grid of the CWE region is represented by 65 nodes, while the transmission grids of the satellite regions are represented via one node per region. In total, 174 grid connections and 70 nodes are represented in the model.

[53] With this simplification we neglect that a detailed representation of all countries would probably impact congestion in the CWE region.

[54] For further technical details, see CWE FBMC (2014).

[55] Although Austria is currently in the same bidding zone as Germany, we treat it as part of the Southern Europe. This has two main reasons: First, numerical complexity is reduced and second, the effect on the results in case Austria is included with higher granularity is expected to be limited.

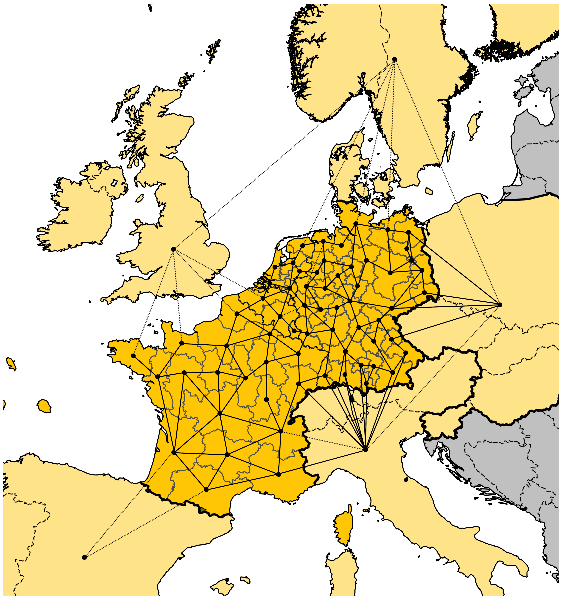


Figure 5.1: Representation of the CWE and neighboring regions in the model

The existing electricity system including power plants[56] and transmission grids[57] as of 2011 was used as the basis for the simulations of the years 2020 and 2030.[58] Existing generation capacities are shut down after reaching the end of their technical lifetime (the model is also allowed to shut down plants earlier if economically beneficial). Investments into new generation capacities (conventional as well as renewables) are subject to political constraints (e.g., no nuclear investments in Germany) or technical restrictions (e.g., areas for renewable sites). The transmission grid topology mainly consists of AC lines, but also includes some DC lines (existing ones plus the projects planned in the 2012 version of the Ten Year Network Development Plan of the European Network of Transmission System Operators for Electricity (ENTSO-E, 2012)). Costs of future years are discounted to 2011-values with a discount rate of 10 percent.[59] Years are represented by nine typical days including different demand levels, wind and solar infeed, distinguished by weekday and weekend.[60] The typical days are coupled to account for

[56]The data for the power plants stems from the power plant database developed at the Institute of Energy Economics at the University of Cologne. This database comprises nearly all European power plants greater than 10 MW and is constantly updated by publicly available sources (e.g., the power plant list of the German regulator) and the Platts WEPP database (Platts, 2009).

[57]The grid model was developed based on the publicly available map and data on the European transmission grid infrastructures from ENTSO-E.

[58]2040 and 2050 are also included in the simulation to control for end time effects. Years in between are accounted for via scaling of the simulated years. Thus, investments into generators as well as the grid infrastructure are possible in 2020, 2030, 2040 and 2050, whereby only 2020 and 2030 are shown in the paper.

[59]This value was chosen to represent typical returns on investment. Note that the costs of capital for investments in the electricity sector are hard to estimate, but considering the rate of return for regulated investments (e.g., around 9 percent for grid expansion projects in Germany) this seems to be a fair assumption.

[60]Typical days are constructed such that they represent statistical features of electricity demand as well as of solar and wind resources along with their multivariate interdependencies found in the original data. Local weather conditions are included through the use of detailed wind speed and solar radiation data (EuroWind, 2011).

seasonal storage, and include one day to cover extreme weather events.

CO_2 emissions are constrained according to the European targets shown in Table 5.2 representing a yearly reduction of 2.2 percent (compared to 2005) up to 2050 (European Commission, 2014a). The maximum amount of CO_2 emissions that can be emitted per year is implemented as a constraint restricting the generation of (conventional) power plants. Thus, a CO_2 quota (in contrast to a CO_2 price) is included in the model. If – due to the restricted grid expansion – the restrictions on CO_2 emissions cannot be fulfilled, the model is allowed to emit additional CO_2. However, these additional CO_2 emissions are costly and amount to 100 €/t CO_2.[61] These additional emissions can be interpreted as shifting CO_2 abatement from the power sector to other sectors of the EU Emissions Trading System (ETS). Although this shifting of CO_2 is not explicitly modeled, this might imply an increase in CO_2 emission costs if more expensive abatement technologies have to be developed.

Year	**2020**	**2030**	**2040**	**2050**
compared to 2005	-21 %	-43 %	-65 %	-87 %

Table 5.2: Assumptions for CO_2 reductions [%]

We assume there is no explicit target for either the share or capacity of renewables in addition to the CO_2 mechanism, meaning that renewables are deployed endogenously due to CO_2 restrictions. However, we will report the deployment of renewables and discuss the implications for the European 27 percent renewables goal in total energy consumption in Section 3. In addition, we assume that the production of PV and wind (onshore and offshore) can be curtailed. This might be necessary if the production of PV and wind capacities exceed the demand in this region and transmission capacities are insufficient to transport the production to other regions. Despite the goals on energy efficiency, the electricity consumption is projected to increase, e.g., due to electrification of heating processes and transportation. Electricity demand is taken from the EU energy road map (European Commission, 2013a).

As the most important trigger in our analysis, we model different scenarios varying the restriction levels for grid expansion, as indicated in Table 5.3. Numbers correspond to the allowed grid expansion z in $TWkm$ per decade. While grid expansion is entirely forbidden in *Scenario 0*, the amount of allowed grid expansion increases throughout the different scenarios. Within *Scenario 30*, where grid expansion is restricted to 30 TWkm per decade, the restriction is not binding any more, hence *Scenario 30* represents an unrestricted scenario. To understand the orders of magnitude, a restriction of 5 TWkm would mean that, e.g., 2 lines, each with 5 GW and 500 km length can be built in a decade. Note that the restriction is imposed as a constraint on the sum of AC and DC lines.

[61] We consider energy efficiency measures as an alternative CO_2 abatement option (see e.g., McKinsey & Company (2009)).

Max. grid expansion	0	5	10	15	20	30

Table 5.3: Scenarios of allowed grid expansion per decade [TWkm/10a]

Due to the imposed constraint on grid expansion, the model may become unable to fully serve demand (except for *Scenario 30* where grid restrictions are not binding). Due to missing local price signals, the market equilibrium might lead to an allocation of generators far away from load centers. Specifically, if transmission capacities are limited, and the congestion in the grid cannot be fully resolved by redispatching generation, energy imbalances occur. Those imbalances can be solved in various ways to ensure technical feasibility. On the demand side, one possibility to overcome an imbalance is load curtailment which is, however, usually avoided as much as possible due to the high value of lost load. In our model, we allow to curtail load with a value of lost load (VOLL) of 7.41 € per kWh (Growitsch et al., 2015). This rather high value forces the model to curtail load as a last resort to ensure feasibility. However, one may also think of other (expansive) measures to relieve imbalances. In fact, measures on the supply side are much more prominent. Additional generation capacities could be contracted in an administrative procurement (i.e., outside the market) to ensure security of supply even in critical situations. This procedure is applied in Germany, for instance.

5.3 Results

5.3.1 Impacts of missing grid expansion

5.3.1.1 Redispatch and energy imbalances

Figure 5.2 shows the yearly energy imbalances in all scenarios, i.e., the mismatch between supply and demand after adjusting the dispatch with a physically feasible redispatch and grid expansion. Energy imbalances might occur due to missing grid infrastructure (if grid expansion is restricted) together with a misallocation of generation capacities. These factors may lead to the fact that not all load in all regions can be served given the installed grid and generation infrastructure. *Scenario 0* shows the highest level of energy imbalances as no grid expansion is allowed and redispatch measures are insufficient. In the CWE region, energy imbalances of around 2 percent of total load in 2020 and nearly 3 percent in 2030 are observed if no grid expansion is allowed. All other scenarios result in energy imbalances of below 0.5 percent in all years. With an increase of the allowed grid expansion, energy imbalances are reduced.

Due to increasing wind capacities built up in the North of Germany without taking into account the ability to transport this generation to load centers in the South, the most severe energy imbalances are in Southern Germany. However, due to the meshed grid

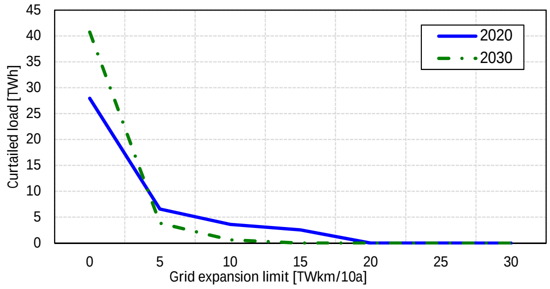


FIGURE 5.2: Energy imbalances in different scenarios

energy imbalances also occur in the BeNeLux-countries and the neighboring regions. Figure 5.3 shows the regional distribution and severity of energy imbalances in *Scenario 0* for 2030.[62] The distribution in the other scenarios is similar, but lower. Noticeably, energy imbalances increase over time in *Scenario 0*, while they decrease in all other scenarios. This is due to the inter-temporal effect of grid expansion (cf. Section 5.3.2.2).

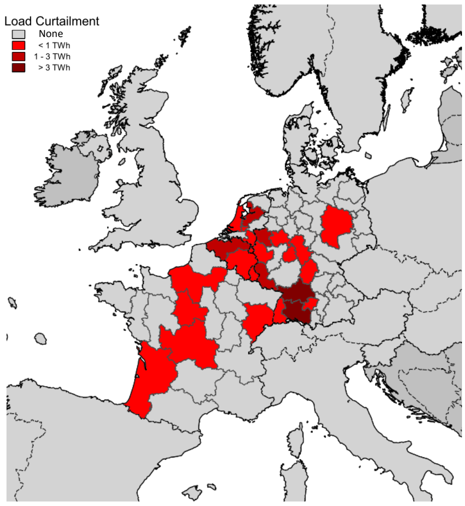


FIGURE 5.3: Geographical distribution of energy imbalances in Scenario 0 for 2030

The overall quantity of redispatch measures shows a similar behavior over the scenarios as the energy imbalances, and are highest for the most restricted case (Figure 5.4). However, the decline in redispatch with a less restricted grid expansion is not as steep as for energy imbalances. This can mainly be explained by the significantly lower overall costs of redispatch, which are only the difference of the variable costs of the redispatched power plants. Even without any restriction posed on grid expansion, a relatively small amount of redispatch measures is still part of the optimal solution when weighed against grid extension costs. The distribution of redispatch, however, shows no distinct pattern.

Figures 5.5 and 5.6 show the number of hours, in which transmission lines are at 100 percent utilization after redispatch indicating the importance of specific transmission lines. As can be seen, the line load decreases with increasing grid expansion. However, in

[62]Note that the map only includes energy imbalances in the CWE region even though there are also imbalances in the satellite regions.

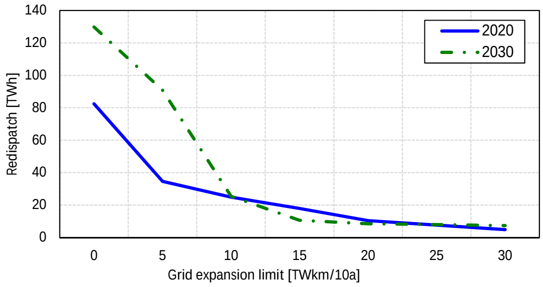


Figure 5.4: Redispatch measures in different scenarios

2030 the pattern for this decrease differs over the scenarios. Different lines are expanded throughout the scenarios and hence, lead to different utilization rates induced by the meshed grid and corresponding loop flows. The line load at the borders of the CWE region shows the importance of the Scandinavian and Iberian countries for the electricity flows in Europe.

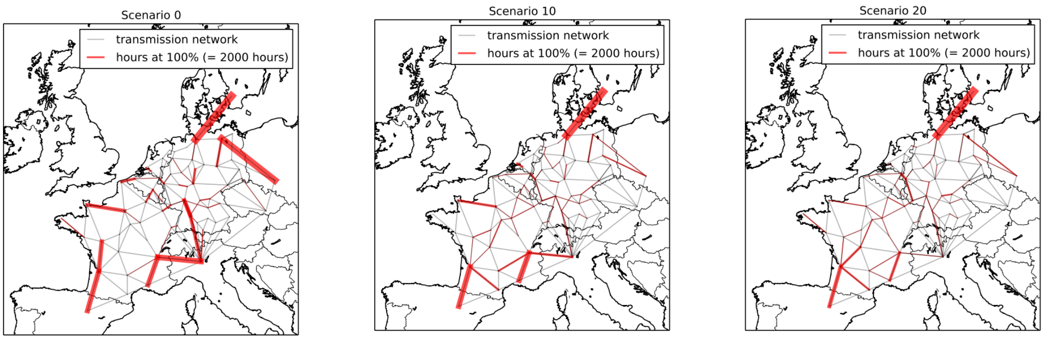


Figure 5.5: Line load after redispatch measures in different scenarios 2020

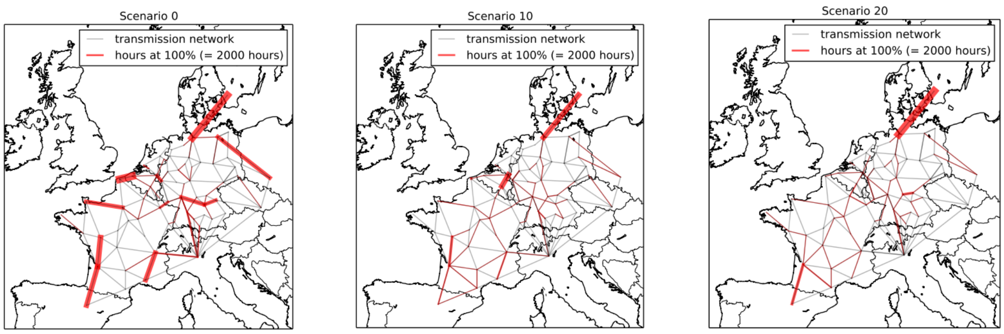


Figure 5.6: Line load after redispatch measures in different scenarios 2030

5.3.1.2 Total system costs

Total system costs are a measure for the overall efficiency of the system. Intuitively, a system with more constrained grid expansion induces higher system costs. For the different scenarios, we find that no grid expansion at all increases total system costs by 138 percent compared to the unrestricted case. Figure 5.7 shows the dependence of total system costs on grid expansion. It can be seen that even a small amount of grid expansion decreases total system costs drastically.

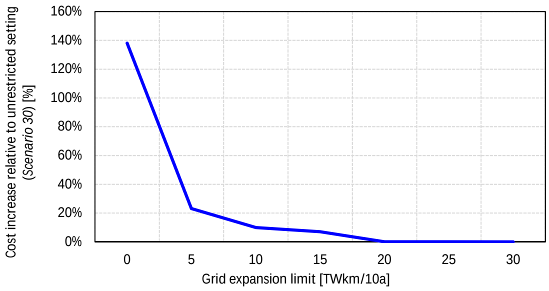


FIGURE 5.7: Total system cost decrease with grid expansion

To further analyze this result, Table 5.4 shows the composition of total systems costs (discounted to $€_{2011}$). The main variation between the scenarios results from differences in the costs to relieve energy imbalances between the scenarios where grid expansion is restricted. The increase of costs aligned to the removal of energy imbalances arises from sub-optimal siting of generators. Due to the market design which is unable to uncover scarcities in the grid within a bidding zone by means of appropriate price signals, investments are made based on supply site characteristics only. As a result, there is not enough generation capacity available at every node and it is furthermore not possible to import sufficient capacity without grid expansion. This in turn leads to situations where redispatch measures trying to overcome internal grid restrictions in each bidding zone are not sufficient any more. Hence, energy imbalances have to be relieved at high costs. In the most extreme scenario with no grid expansion at all, this leads to the additional effect that the implemented CO_2 quota cannot be fulfilled anymore by the electricity sector, meaning that some other sectors have to increase their CO_2 reduction efforts.[63]

Max. grid expansion [TWkm/10a]	0	5	10	15	20	30
Generation [Bn. €]	940.7	938.1	932.5	930.6	930.2	929.5
Grid (including redispatch) [Bn. €]	8.2	8.1	7.9	9.7	10.2	10.7
Clearance of energy imbalances [Bn. €]	1,169.9	211.5	93.5	65.7	0	0
Shifting of CO_2 to other sectors [Bn. €]	120	0.6	0	0	0	0
Total [Bn. €]	2,238.3	1,158.2	1,033.9	1,006.1	940.9	940.2

TABLE 5.4: Total system costs of scenarios (in $€_{2011}$ up to 2030)

Remarkable – while looking at the results on total system costs – is the fact that grid expansion costs are rather low compared to any other cost factor and almost negligible if generation and grid costs are compared. The non-monotonous trend of the grid costs over the scenarios can be explained by the included redispatch costs, which depend on the optimization of the generation and not the transmission level.

[63]The shifting of CO_2 emissions is not explicitly included in the applied modeling framework.

5.3.1.3 Fulfillment of the EU 2030 targets

In case of no grid expansion (*Scenario 0*), the amount of CO_2 emissions that have to be reduced by other sectors (than the power sector) within the EU-ETS amounts to 176 mt CO_2 in 2020 and 391 mt CO_2 per year in 2030. These numbers should be interpreted with care, as feedback loops with other sectors covered by the EU ETS that are induced by an increasing CO_2-price are not considered here. However, the general result that CO_2 emissions are shifted to other sectors should probably hold.

To deal with restricted grid expansion, generation from renewable resources may be curtailed.[64] As can be seen in Figure 5.8, the need to curtail the production of renewables decreases dramatically as the restriction of grid expansion is relaxed. In 2030, 7.7 percent of available PV generation is curtailed in the case of no grid expansion, which drops to just 0.4 percent if 5 TWkm/10a are allowed. The drop for onshore wind from 3.3 percent to 1.4 percent is less dramatic. Furthermore, the regional distribution of curtailment differs. With no grid expansion, curtailment of offshore wind only occurs on the North Coast of France in 2030. For PV and onshore wind, curtailment is concentrated in Southern Germany, and along the North and West Coasts of France, where there are significant grid bottlenecks.

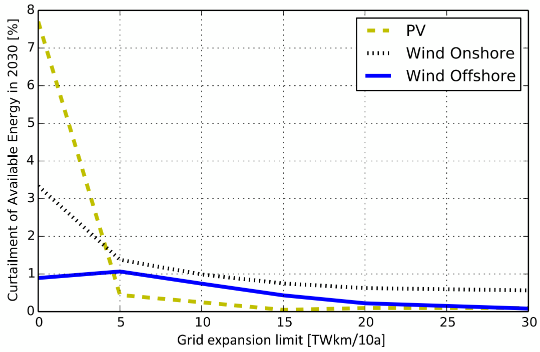


FIGURE 5.8: Curtailment of renewables for the different scenarios

The curtailment of renewables impacts the overall renewables quota only in the most restricted scenario and only in 2030. While the renewables quota for all other scenarios is around 44 percent in 2030, for *Scenario 0* the quota is about 1.5 percent percentage points lower due to the curtailment. Considering the three main targets of European energy policy consisting of a secure, affordable and climate-friendly energy, our results show that missing grid expansion might degrade these targets. Especially with no or only minimal grid expansion, energy imbalances, the fulfillment of the implemented CO_2 quota as well as total system costs increase substantially compared to scenarios with less extreme grid restrictions. Thus, it can be concluded that grid expansions are of high importance in order to meet the European energy targets.

[64] We here focus on weather-driven renewable energy sources which are especially relevant in the context of curtailment.

5.3.2 Development of grid capacities

5.3.2.1 DC and AC capacities

Figure 5.9 shows the grid expansion in the different scenarios for the period from 2011 to 2020 as well as between 2020 and 2030 for AC, DC, as well as the aggregated grid expansion, measured in TWkm. Between 2011 and 2020, the total grid expansion restrictions are binding for the system in *Scenario 0* through *20*, whereas between 2020 and 2030 grid restrictions are binding only for the *Scenarios 0* through *15*. Thus, for *Scenario 30*, grid expansion is not restricted in any decade, which means that the investments made in this scenario are system optimal, such that *Scenario 30* can serve as a benchmark with respect to cost efficiency (see above).

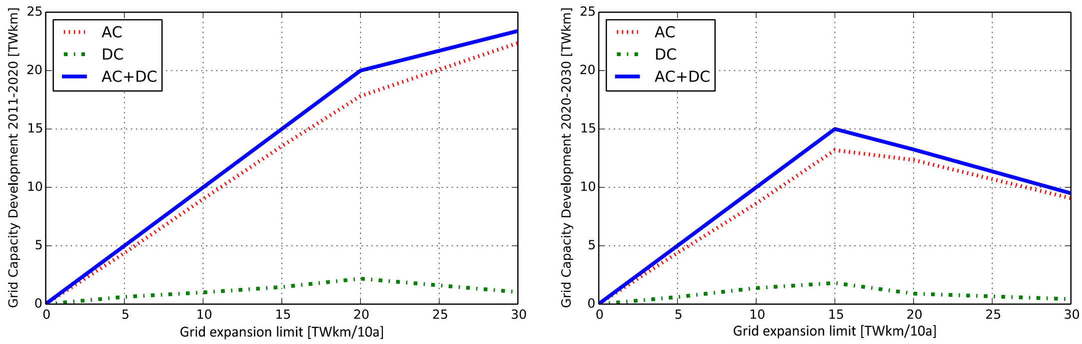


FIGURE 5.9: Capacity development in TWkm for the period 2011-2020 (left) and 2020-2030 (right)

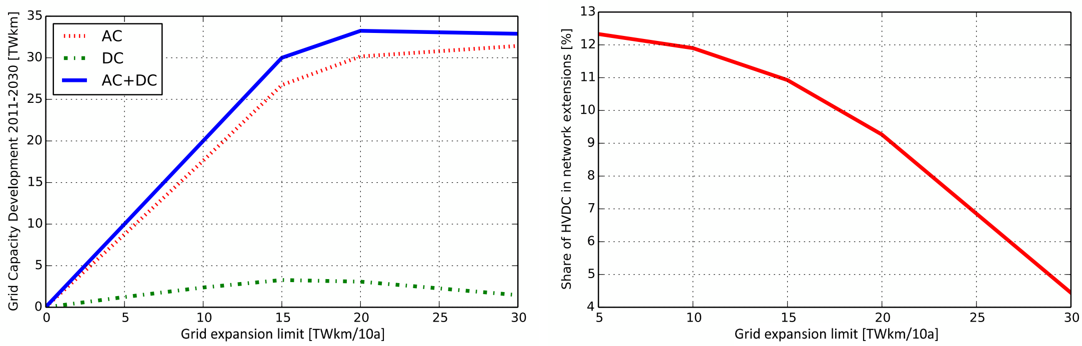


FIGURE 5.10: Total capacity development for 2011-2030 (left) and the share of DC in the total network expansion (right)

To put the total grid expansion into context, the starting grid from 2011 for the CWE region has a capacity of 70.8 TWkm, split between 68.0 TWkm for AC and 2.8 TWkm for DC. In *Scenario 30*, which has a total of 32.9 TWkm of grid expansion between 2011 and 2030, this represents a grid capacity expansion of 46.4 percent.[65]

[65]In the optimal grid scenario considered by Hagspiel et al. (2014), the grid for the entire ENTSO-E area was extended by 48 percent between 2011 and 2030.

In Figure 5.9 and 5.10, it can be seen that the AC network is extended significantly more than the DC network. One reason is that there are simply more AC connections available to the optimizer to extend; only DC connections that already exist and those planned in the TYNDP 2012 are fed into the initial network topology for optimization. Another reason is that DC lines are more expensive because of the costs of the AC-DC converter stations at each end of the line. [66] However, the decrease in DC capacities with a restriction of more than 20 TWkm per decade indicates that DC lines are prioritized when the grid restrictions are enforced. The share of DC in the total network expansion decreases monotonically as the grid restrictions are relaxed (see Figure 5.10). In absolute terms, for the total period 2011-2030, the DC expansion increases, peaks at just over double the existing DC capacity, and then decrease as the grid restrictions disappear. In Figure 5.9 it can be seen that DC capacity increases as the overall capacity limit increases in each decade, but only as long as the overall grid restriction for AC and DC is binding. When grid restrictions are no longer binding for a decade (*Scenario 30* for 2011-2020 and *Scenario 20* for 2020-2030), the DC capacity drops as cheaper AC lines are prioritized over extending DC lines. This shows that DC lines help the system to deal with the grid expansion restrictions and to compensate missing AC lines. A reason for preferring DC to AC is that the power flow is more controllable, so that power transfers can be directed over long distances, rather than spreading out in the AC network in "loop flows", which overload wide areas of the AC network. This underlines the importance of DC lines for example to integrate renewable energies into the system. As a result, whenever grid restrictions are in place, DC lines allow a better system optimum.

5.3.2.2 Inter-temporal effects

In Figure 5.9, an interesting interplay between grid expansion during the two decades 2011-2020 and 2020-2030 can be seen.[67] The less restricted grid expansion are, the more transmission lines are built in the first decade between 2011 and 2020. Grid expansion in the second decade increases first and then decreases, which shows that it is more valuable for the system to have lines installed early, i.e., by 2020. This higher value may be due to the fact that the lines built in the first decade are used for a longer time. The effects also become visible when looking at the imposed grid expansion restrictions: The 2011-2020 restriction is binding longer (up to and including *Scenario 20*) than the 2020-2030 restriction (up to and including *Scenario 15*), which shows the inter-temporal effect of grid expansion and thus, the optimality of building the grid earlier. The inter-temporal effect is strong enough that the total grid expansion from 2011 to 2030 is lower in *Scenario 30* than *Scenario 20* (see Figure 5.10), because of the suboptimal binding grid restriction in *Scenario 20* for the decade 2011-2020.

[66] See Table 5.9 in the Appendix for the transmission cost assumptions.

[67] Recall that 2040 and 2050 are also included in the optimization in order to avoid end-time effects.

5.3.2.3 Geographical distribution

Figure 5.11 shows the geographical distribution of the grid expansion for three scenarios including the optimal grid from *Scenario 30*. Noticeably, many of the grid expansion are concentrated in France and its borders with other countries. This results from the good wind resources in France, that are located particularly along its coastline. The electrical load along the coast is weak, so network extensions are needed to transport the wind power to load centers elsewhere in Europe. These good wind resources are currently under-exploited, but represent the cheapest option to decrease CO_2 emissions in the CWE region.

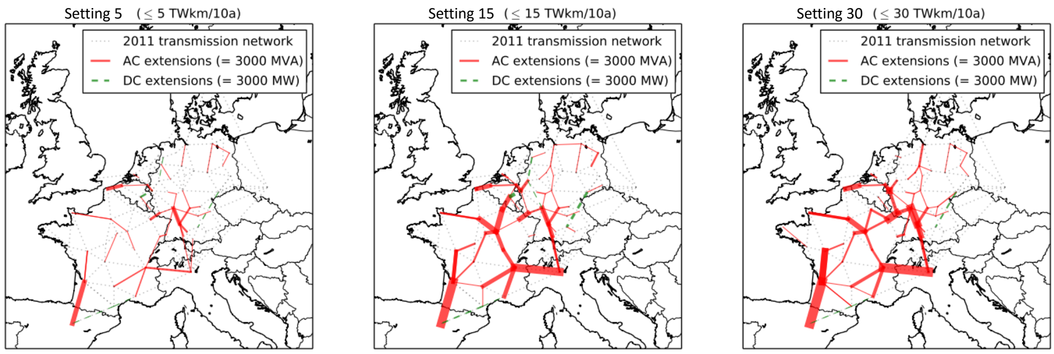


FIGURE 5.11: Maps of grid expansion for *scenarios* 5, 10 *and* 30

There are also grid bottlenecks within Germany, which are overcome with both new AC lines and DC lines along the planned corridors from North to South Germany. The controversial DC line within corridor D, planned by the German TSOs to carry wind and solar power from East to South Germany (Bavaria), is extended in each scenario where grid expansion is allowed; Corridor A ranging from the North Sea to Southern Germany is also expanded in *Scenario 15*.

5.3.2.4 Inter- vs. intra-zonal grid expansion

In the grid model for the CWE region in 2011, 30 percent of the grid capacity measured in TWkm is made up of cross-border lines (this is higher than the actual grid, because of the way countries at the boundary of the CWE region have been aggregated to single nodes, lengthening cross-border lines). However, interconnectors make up 42 percent of all grid expansion in *Scenario 30*, meaning that interconnector capacity is more valuable on average than internal, national grid connections. This is particularly due to the possibility to exploit cheaper generation sites and being then able to transport it to load centers within Europe using interconnector capacities. Between 2011 and 2030, interconnector capacity rises by 65 percent. There is some overlap between the distribution of grid expansion calculated here and the European Commission's Projects of Common Interest[68], particularly for the internal DC lines in Germany and the strengthening of interconnectors between Spain and France and between Germany and Switzerland. However, grid expansions in Figure 5.11 are much more heavily concentrated in France and its interconnectors, due to the significant expansion of wind power in France in the scenarios presented here. Similarly, the dominance of grid expansion in France is not reflected in the 2012 or 2014 TYNDP.

5.3.3 Generation and generation capacities

The total generation capacities and total dispatch in the CWE region in 2030 are shown in Figure 5.12 for each scenario. Overall, there is very little change in installed capacities as grid restrictions are lifted. Comparing *Scenario 0* to *Scenario 30*, there is an increase of wind capacity of 17 GW, which takes place exclusively in France as inland sites with lower capacity factors than the coast are exploited. This raises the wind capacity in France from 55 GW to 72 GW in 2030. This better use of cheap wind resources in France is also reflected in the grid expansion (see Figure 5.11). There is a small drop in solar capacity of 3.8 GW in the British Isles, as grid expansion allow PV to be replaced by cheaper wind generation. In each scenario the offshore wind capacities are identical, amounting to 42.6 GW in 2020 and 42.0 GW in 2030.

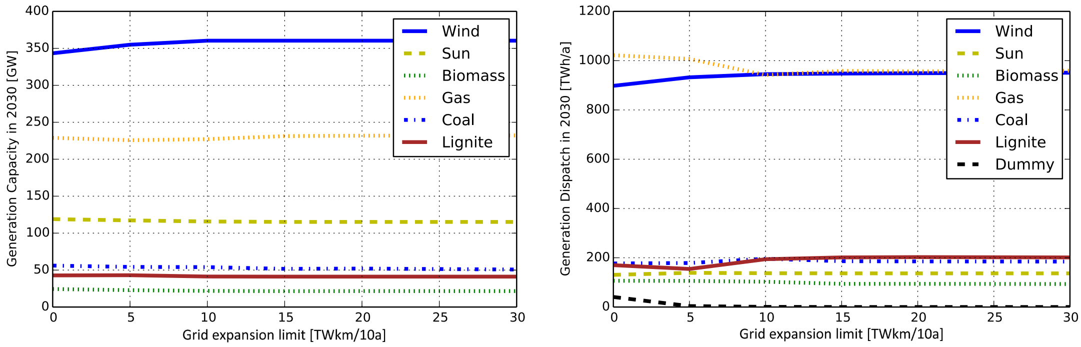


FIGURE 5.12: Total generation capacity (left) and yearly dispatch (right) in 2030 for the different grid restriction scenarios

More change is visible in the yearly dispatch of each technology. Between *Scenario 0* and *Scenario 30* there is a 53 TWh/a increase in wind generation, almost exclusively due to the extra wind capacity in France but also due to reduced curtailment of renewables in France, the Netherlands and Germany, as grid bottlenecks ease. Solar generation increases by 6 TWh/a despite the lower capacity, due to lower renewable curtailment in France and particularly in Germany. Gas generation is reduced by 63 TWh/a and replaced by CO_2-free renewable generation as well as lignite generation, primarily from Central Eastern European countries (increasing by 31 TWh/a). This substitution of gas

[68] `https://ec.europa.eu/energy/en/topics/infrastructure/projects-common-interest`

with lignite as grid capacity increases is induced by lower fuel costs of lignite than gas, which outweigh its higher CO_2 emissions per kWh. There is also 8 GWh/a more coal generation in the Iberian peninsular, enabled by the grid expansion between Spain and France.

The distribution of generation capacity is in general very insensitive to grid expansion because of the way the grid and market are coupled. In the initial dispatch and generation capacity optimization the internal grid constraints of each country are not visible; the internal bottlenecks only become apparent in the next step, as redispatch is performed in each bidding zone. However, the redispatch does not directly affect the optimality of the investment and dispatch decisions in the market. The only impact stems from the indirect effect of altered interconnection capacities, which become visible in the scenarios with little grid expansion.

The capacity and generation of pumped hydro storage and hydro storage dams remains nearly constant throughout all scenarios, which shows that the role of these types of storage in the system is not influenced by restrictions on grid expansion. Thus, storage is no substitute for grid expansion. Pump storage capacity is highest in the Southern region (Switzerland, Austria and Italy) whereas hydro storage capacity is highest in Northern Europe (mainly Norway) followed by the southern region (22 and 14 GW). However, as the potential in these countries is already mostly exhausted, there are no capacity expansion in these regions. Nevertheless, the value of storage is demonstrated when looking at the United Kingdom (UK) where pump storage capacity increases from 3 to 6 GW when grid expansion are highly restricted and to only 5 GW in the less restricted scenarios. At the same time the good and until now not exhausted wind resources in the UK are explored and thus wind capacities increase from 10 GW in 2011 to roughly 72 GW in 2030 throughout all scenarios. As exports to other countries are limited, other sources of demand to absorb this wind generation are needed. Therefore, storage is built. Thus, for the very special case of the UK, storage is a substitute to extending the DC connections (which are limited) to the rest of the CWE region. In addition, grid bottlenecks in France prevent imports of wind power from the UK, which may also drive the expansion of storage capacity in the UK when grid expansions are restricted.

5.4 Conclusions

We investigated the effect of restricted grid expansion for the EU's 2030 energy strategy under the current market design. Specifically, we contributed to the existing literature an in-depth analysis of the long-term effect of grid restrictions in zonal markets with redispatch after market clearing. In case of grid restrictions, this market design reveals its inherent incompleteness due to zonal markets failing to provide efficient locational price signals. Our analysis was based on a large-scale model of the European electricity market with a focus on the Central Western European region. We used a linear model covering the generation and transmission level with endogenous investment and dispatch decisions for both levels. Restrictions for grid expansion were implemented and gradually tightened, reaching from full to non-restricted expansion.

We found that the incompleteness of the market design leads to a misallocation of generation capacities and the inability of the system to transport electricity to where it is needed. Thus, energy imbalances occur. Although they decrease sharply if some grid expansion is allowed, we still see energy imbalances even for an allowed grid expansion of 15 TWkm per decade. Affected regions are mainly those that are characterized by poor conditions for renewables, i.e., comparably low wind speeds and low solar radiation, and far away from (new) generation sites. Most severe energy imbalances appear in Southern Germany. These imbalances indicate the need for additional measures that have to be undertaken in order to ensure system stability, either affecting the demand or the supply side. On the supply side, imbalances can be solved by procuring additional capacity, while on the demand side load curtailment would be necessary. In practice, however, the latter is observed very rarely and usually avoided as much as possible. Therefore, supply side measures are more frequently used, e.g., by contracting additional generation capacity outside the market to ensure security of supply. For example, Germany administratively procures additional generation capacity, especially in Southern Germany where significant imbalances occur.

The restriction on grid expansion has a visible effect on European climate targets only if no or very little grid expansion is allowed. With no grid expansion, the renewables share is 1.5 percentage points lower compared to the other scenarios. As a consequence, conventional generation with higher CO_2 emissions has to jump in, such that the indirect effect of rising CO_2 abatement costs appears. One approach to deal with restricted grid expansion is the utilization of DC instead of AC lines. When overall grid expansion is restricted, DC can bring advantages by directing long-distance power flows, which would otherwise cause loop-flows in the AC network causing wide-spread overloading.

In order to overcome the depicted shortcomings of the zonal market design, several options are available. First, the obstacles of grid expansion could be removed to avoid intra-zonal congestion. As pointed out earlier, the main obstacles are approval procedures as well as social opposition which would need to be addressed by all involved parties. However, from past experience, it seems unlikely that grid expansion could be completely avoided in the future. Thus, an adaptation of the current market design should be considered as a second option. As has been shown, the prevailing market design is inherently incomplete, which may have severe consequences, especially when facing substantial changes in the supply structure. Hence, additional measures are needed, such as administrative intervention to ensure sufficient levels of generation capacity outside the market (as it is currently handled in Germany by means of a grid reserve for redispatch), different shapes of price zones, or via an implementation of locational price elements into the market. Moreover, the issue of the right location should also play a role when designing renewable support schemes, since they are the main driver of the changing infrastructure.

5.5 Appendix

Abbreviation	Dimension	Description
Model sets		
$i, j, k, q \in \mathbf{I}$		Nodes, $\mathbf{I} = [1, 2, ...]$
$m, n \in \mathbf{M}$		Zonal markets, $\mathbf{M} = [1, 2, ...]$
$i \in \mathbf{I_m}$		Nodes that belong to zonal market m, $\mathbf{I_m} \subset \mathbf{I}$
$t \in \mathbf{T}$		Points in time where dispatch decisions are made, e.g. hours , $\mathbf{T} = [1, 2, ...]$
$y \in \mathbf{Y}$		Points in time where investment decisions are made, e.g. years, $\mathbf{Y} = [1, 2, ...]$
$b \in \mathbf{B}$		Decades of grid expansion restriction, $\mathbf{B} = [1, 2, ...]$
Model parameters		
$\delta_{i,y}$	EUR/kW	Investment and FOM costs of generation capacity in node i at time y
$\gamma_{i,t}$	EUR/kWh	Variable costs of generation capacity in node i at time t
$\mu_{i,j,y}$	EUR/kW	Investment costs of line between node i and node j at time y
$d_{i,t}$	kW	Electricity demand in node i at time t
$l_{i,j}$	km	Length of line between node i and node j
z	$TWkm$	Grid Expansion Limit per decade
Model primal variables		
$\overline{G}_{i,y}$	kW	Generation capacity in node i at time y, $\overline{G}_{i,y} \geq 0$
$G_{i,t}$	kW	Generation dispatch in node i at time t, $G_{i,t} \geq 0$
$T_{m,n,t}$	kW	Electricity trade from market m to market n at time t
X	EUR	Costs of generation
Y	EUR	Costs of TSO
$\overline{P}_{i,j,y}$	kW	Line capacity between node i and node j at time y, $\overline{P}_{i,j.y} \geq 0$
$\overline{P}_{m,n,t}$	kW	Capacity between market m and node n at time t determined by function g, $\overline{P}_{m,n,t} \geq 0$
$P_{i,j,t}$	kW	Electricity flow on line between node i and node j at time t
$R_{i,t}$	kW	Redispatch in node i at time t

TABLE 5.5: Model sets, parameters and variables

Representation of a Nodal System

Generation

$$
\begin{aligned}
\min_{\overline{G}_{i,y}, G_{i,t}, T_{i,j,t}} \quad & X = \sum_{i,y} \delta_{i,y} \overline{G}_{i,y} + \sum_{i,t} \gamma_{i,t} G_{i,t} && && \text{(5.2a)} \\
\text{s.t.} \quad & G_{i,t} - \sum_{j} T_{i,j,t} = d_{i,t} && \forall i, t && \text{(5.2b)} \\
& G_{i,t} \leq \overline{G}_{i,y} && \forall i, t && \text{(5.2c)} \\
& T_{i,j,t} = -T_{j,i,t} \leq \overline{P}_{i,j,t} && \forall i, j, t && \text{(5.2d)}
\end{aligned}
$$

Transmission

$$
\begin{aligned}
\min_{\overline{P}_{i,j,y}} \quad & Y = \sum_{i,j,y} \mu_{i,j,y} \overline{P}_{i,j,y} && && \text{(5.2e)} \\
\text{s.t.} \quad & |P_{i,j,t}(\overline{P}_{k,q,y}, G_{k,t}, d_{k,t}, R_{k,t})| \leq \overline{P}_{i,j,y} && \forall i, j, t && \text{(5.2f)} \\
& \sum_{y \in b} \overline{P}_{i,j,y} l_{i,j} \leq \sum_{y \in b-1} \overline{P}_{i,j,y} l_{i,j} + z && \forall b && \text{(5.2g)} \\
& T_{i,j,t} = -T_{j,i,t} && \forall i, j, t && \text{(5.2h)}
\end{aligned}
$$

Country	**2011**	**2020**	**2030**
Belgium	87	98	105
Germany	573	612	629
France	466	524	559
Luxembourg	7	8	8
Netherlands	113	128	137
Eastern	276	328	366
Northern	387	436	465
Southern	450	528	594
Southwest	317	378	433
United Kingdom	400	450	481

TABLE 5.6: Gross electricity demand (without own consumption and pump storage) [TWh]

To depict the CWE region in a high spatial resolution, we split the gross electricity demand per country among the nodes belonging to this country according to the percentage of population living in that region.

Technology	2020	2030
Wind Onshore	1,253	1,188
Wind Offshore (<20m depth)	2,800	2,350
Wind Offshore (>20m depth)	3,080	2,585
Photovoltaics (roof)	1,260	935
Photovoltaics (ground)	1,110	785
Biomass gas	2,398	2,395
Biomass solid	3,297	3,295
Biomass gas, CHP	2,597	2,595
Biomass solid, CHP	3,497	3,493
Geothermal	10,504	9,500
Compressed Air Storage	1,100	1,100
Pump Storage	1,200	1,200
Lignite	1,500	1,500
Lignite Innovative	1,600	1,600
Coal	1,200	1,200
Coal Innovative	2,025	1,800
CCGT	711	711
OCGT	400	400
Nuclear	3,157	3,157

TABLE 5.7: Generation technology investment costs [€/kW]

Fuel type	2011	2020	2030
Nuclear	3.6	3.3	3.3
Lignite	1.4	1.4	2.7
Oil	39.0	47.6	58.0
Coal	9.6	10.1	10.9
Gas	14.0	23.1	25.9

TABLE 5.8: Assumptions for the gross fuel prices [€/MWh$_{th}$]

Grid Technology	Extension costs	FOM costs
AC overhead line incl. compensation	445 €/(MVA*km)	2.2 €/(MVA*km)
DC overhead line	400 €/(MW*km)	2.0 €/(MW*km)
DC underground	1,250 €/(MW*km)	6.3 €/(MW*km)
DC submarine	1,100 €/(MW*km)	5.5 €/(MW*km)
DC converter pair	150,000 €/MW	750.0 €/MW

TABLE 5.9: Assumptions for the grid extension and FOM costs

Bibliography

50 Hertz Transmission GmbH, Amprion GmbH, TenneT TSO GmbH, TransnetBW GmbH, 2018. EEG-Anlagestammdaten. `https://www.netztransparenz.de/EEG/Anlagenstammdaten`, 14.10.2020.

ACER, 2014. Report on the influence of existing bidding zones on electricity markets.

Ackermann, T., Cherevatskiy, S., Brown, T., Eriksson, R., Samadi, A., Ghandhari, M., Söder, L., Lindenberger, D., Jägemann, C., Hagspiel, S., Cuk, V., Ribeiro, P.F., Cobben, S., Bindner, H., Isleifsson, F.R., Mihet-Popa, L., 2013. Smart Modeling of Optimal Integration of High Penetration of PV - Smooth PV. Final Report.

AGEB, 2015. Auswertungstabellen zur Energiebilanz Deutschland 1990-2014.

Agrell, P.J., Brea-Solis, H., 2017. Capturing heterogeneity in electricity distribution operations: A critical review of latent class modelling. Energy Policy 104, 361–372.

Agrell, P.J., Farsi, M., Filippini, M., Koller, M., 2014. The Interrelationship Between Financial and Energy Markets, Lecture Notes in Energy Edition 54. Springer. chapter Unobserved heterogeneous effects in the cost efficiency analysis of electricity distribution systems. pp. 281–302.

Aguado, M., Bourgeois, R., Bourmaud, J., Casteren, J.V., Ceratto, M., Jäkel, M., Malfliet, B., Mestda, C., Noury, P., Pool, M., van den Reek, W., Rohleder, M., Schavemaker, P., Scolari, S., Weis, O., Wolpert, J., 2012. Flow-based market coupling in the Central Western European region - on the eve of implementation.

Baake, R., 2014. Response to Interpellation in the German Parliament, BT 18-5168.

Badunenko, O., Cullmann, A., Kumbhakar, S.C., Nieswand, M., 2021. The Effect of Restructuring Electricity Distribution Systems on Firms' Persistent and Transient Efficiency: The Case of Germany. The Energy Journal 42.

Badunenko, O., Kumbhakar, S.C., 2017. Economies of scale, technical change and persistent and time-varying cost effciency in Indian banking: Do ownership, regulation and heterogeneity matter? European Journal of Operational Research 260, 789–803.

BDEW, 2016. Der aktive Verteilnetzbetreiber in einer dezentralen Energiewelt - Ein Beitrag der Verteilnetzbetreiber im BDEW zum Netzkonzept 2030. https://www.bdew.de/media/documents/Stn_20161130-VNB-Netzkonzept-2030.pdf, 14.10.2020.

Bertsch, J., Brown, T., Hagspiel, S., Just, L., 2017. The relevance of grid expansion under zonal markets. The Energy Journal 38.

Bertsch, J., Hagspiel, S., Just, L., 2016. Congestion management in power systems - Long-term modeling framework and large-scale application. Journal of Regulatory Economics 50, 290–327.

Bollen, K., Jackman, R., 1990. Regression diagnostics: an expository treatment of outliers and influential cases. Modern Methods of Data Analysis, Newbury Park. CA: SAGE. 257-291.

Boyd, S., Xiao, L., Mutapcic, A., Mattingley, J., 2008. Notes on Decomposition Methods.
http://see.stanford.edu/materials/lsocoee364b/08-decomposition_notes.pdf.

Bundesnetzagentur, 2012. Bericht zum Zustand der leitungsgebundenen Energieversorgung im Winter 2011/12.

Bundesnetzagentur, 2013a. Bericht zum Zustand der leitungsgebundenen Energieversorgung im Winter 2012/13.

Bundesnetzagentur, 2013b. Feststellung des Reservekraftwerksbedarfs für den Winter 2013/14.

Bundesnetzagentur, 2013c. Monitoring Report 2012.
https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2014a. Monitoring Report 2013.
https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2014b. Monitoring Report 2014.
https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2015a. Feststellung des Bedarfs an Netzreserve für den Winter 2015/2016 sowie die Jahre 2016/2017 und 2019/2020.

Bundesnetzagentur, 2015b. Quartalsbericht zu Netz- und Systemsicherheitsmaßnahmen: Erstes und zweites Quartal 2015.

Bundesnetzagentur, 2016a. Monitoring Report 2015. https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2016b. Monitoring Report 2016. https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2017. Monitoring Report 2017. https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2019a. EEG in Zahlen 2018. https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/ErneuerbareEnergien/ZahlenDatenInformationen/EEGinZahlen_2018_BF.pdf?__blob=publicationFile&v=2, 14.10.2020.

Bundesnetzagentur, 2019b. Monitoring Report 2018. https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/DatenaustauschundMonitoring/Monitoring/Monitoringberichte/Monitoring_Berichte_node.html, 14.10.2020.

Bundesnetzagentur, 2020a. EEG in Zahlen 2019. https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/ErneuerbareEnergien/ZahlenDatenInformationen/EEGinZahlen_2019_BF.pdf?__blob=publicationFile&v=4, 28.05.2021.

Bundesnetzagentur, 2020b. Monitoring Report 2019. https://www.bundesnetzagentur.de/SharedDocs/Mediathek/Berichte/2019/Monitoringbericht_Energie2019.pdf?__blob=publicationFile&v=6l, 27.01.2021.

Burstedde, B., 2012. Essays on the Economic of Congestion Management - Theory and Model-based Analysis for Central Western Europe. Ph.D. thesis. Universität zu Köln.

Capacity Allocating Service Company, 2014. Documentation of the CWE FB MC solution - As basis for the formal approval-request.

Chao, H.p., Peck, S., Oren, S., Wilson, R., 2000. Flow-based transmission rights and congestion management. The Electricity Journal 13, 38–58.

Colombi, R., Kumbhakar, S., Martini, G., Vittadini, G., 2014. Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. Journal of Productivity Analysis 42, 123–126.

Commission de Régulation de l'Énergie, 2009. Networks - Management and use of electric interconnections in 2008.

Conejo, A.J., Castillo, E., Minguez, R., Garcia-Bertrand, R., 2006. Decomposition Techniques in Mathematical Programming - Engineering and Science Applications. Springer.

Cullmann, A., 2012. Benchmarking and firm heterogeneity: a latent class analysis for German electricity distribution companies. Empirical Economics 42, 147–169.

CWE FBMC, 2014. Documentation of the CWE FB MC solution - As basis for the formal approval-request. Technical Report. Amprion, APX, Belpex, Creos, Elia, EPEXSpot, RTE, Tennet, Transnet BW.

Daxhelet, O., Smeers, Y., 2007. The EU regulation on cross-border trade of electricity: A two-stage equilibrium model. European Journal of Operations Research 181, 1396–1412.

E-Bridge, IAEW, OFFIS, 2014. Moderne Verteilnetze für Deutschland (Verteilnetzstudie) - Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie (BMWi). `https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/verteilernetzstudie.html`, 14.10.2020.

Ehrenmann, A., Smeers, Y., 2005. Inefficiencies in European congestion management proposals. Utilities policy 13, 135–152.

ENTSO-E, 2010. Ten-Year Network Development Plan 2010-2020.

ENTSO-E, 2012. Ten Year Network Development Plan 2012.

ENTSO-E, 2015. 2015 Monitoring update of the TYNDP 2014 Table of projects.

European Commission, 2013a. EU Energy, Transport and GHG Emissions - Trends to 2050: Reference Scenario 2050.

European Commission, 2013b. Green Paper - A 2030 framework for climate and energy policies. COM(2013) 169 final.
`http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52013DC0169`.

European Commission, 2014a. European Council (23 and 24 October 2014) - Conclusions.

European Commission, 2014b. Impact Assessment - A 2030 framework for climate and energy policies. SWD(2014) 15 final. `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52013DC0169`.

EuroWind, 2011. Database for hourly wind speeds and solar radiation from 2006-2010 (not public). Technical Report. EuroWind GmbH.

Farsi, M., Filippini, M., Greene, W., 2006. Application of Panel Data Models in Benchmarking Analysis of the Electricity Distribution Sector. Annals of Public and Cooperative Economics 77, 271–290.

Federal Ministry for Economic Affairs and Energy, 2019. Renewable energy sources in figures – National and International Development, 2018. `https://www.bmwi.de/Redaktion/EN/Publikationen/renewable-energy-sources-in-figures-2018.html`, 14.10.2020.

Fetz, A., Filippini, M., 2010. Economies of vertical integration in the Swiss electricity sector. Energy Economics 32, 1325–1330.

Filippini, M., Greene, W., 2016. Persistent and transient productive inefficiency: a maximum simulated likelihood approach. Journal of Productivity Analysis 45, 187–196.

Filippini, M., Greene, W.H., Masiero, G., 2018. Persistent and transient productive inefficiency in a regulated industry: electricity distribution. Energy Economics 69, 325–334.

Filippini, M., Orea, L., 2014. Applications of the stochastic frontier approach in Energy Economics. Economics and Business Letters 3, 35–42.

Filippini, M., Wetzel, H., 2014. The impact of ownership unbundling on efficiency: Empirical evidence from the New Zealand electricity distribution sector. Energy Economics 45, 412–418.

Fürsch, M., Hagspiel, S., Jägemann, C., Nagl, S., Lindenberger, D., Tröster, E., 2013. The role of grid extensions in a cost-effcient transformation of the European electricity system until 2050. Applied Energy 104, 642–652.

Fürsch, M., Hagspiel, S., Jägemann, C., Nagl, S., Lindenberger, D., Tröster, E., 2013. The role of grid extensions in a cost-efficient transformation of the European electricity system until 2050. Applied Energy 104, 642–652.

Glachant, J.M., 2010. The Achievement of the EU Electricity Internal Market through Market Coupling. EUI Working Papers, RSCAS 2010/87 .

Green, R., 2007. Nodal pricing of electricity: how much does it cost to get it wrong? Journal of Regulatory Economics 31, 125–149.

Greene, W., 2002. Alternative Panel Data Estimators for Stochastic Frontier Models.

Greene, W., 2005a. Fixed and Random Effects in Stochastic Frontier Models. Journal of Productivity Analysis 23, 7–32.

Greene, W., 2005b. Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model. Journal of Econometrics 126, 269–303.

Greene, W.H., 2007. The econometric approach to efficiency measurement. Oxford University Press, Oxford.

Greene, W.H., 2016. LIMDEP. Version 11. Econometric Modeling Guide.

Grimm, V., Martin, A., Schmidt, M., Weibelzahl, M., Zöttl, G., 2016. Transmission and generation investment in electricity markets: The effects of market splitting and network fee regimes. European Journal of Operational Research 254, 493–509.

Growitsch, C., Jamasb, T., Wetzel, H., 2012. Efficiency effects of observed and unobserved heterogeneity: Evidence from Norwegian electricity distribution networks. Energy Economics 34, 542–548.

Growitsch, C., Malischek, R., Nick, S., Wetzel, H., 2015. The Costs of Power Interruptions in Germany: A Regional and Sectoral Analysis. German Economic Review 16, 307–323.

Hagspiel, S., Jägemann, C., Lindenberger, D., Brown, T., Cherevatskiy, S., Tröster, E., 2014. Cost-Optimal Power System Extension under Flow-based Market Coupling. Energy 66, 654–666.

Hess, B., Cullmann, A., 2007. Efficiency Analysis of East and West German Electricity Distribution Companies - Do the “Ossis” really beat the “Wessis”? Utilities Policy 15, 206–214.

von Hirschhausen, C., Cullmann, A., Kappeler, A., 2006. Efficiency analysis of German electricity distribution utilities - non-parametric and parametric tests. Applied Economics 38, 2553–2566.

Höffler, F., Wambach, A., 2013. Investment Coordination in Network Industries: The Case of Electricity Grid and Electricity. Journal of Regulatory Economics 44, 287–307.

Hogan, W., Rosellón, J., Vogelsang, I., 2010. Toward a combined merchant-regulatory mechanism for electricity transmission expansion. Journal of Regulatory Economics 38, 113–143.

Hogan, W.W., 1992. Contract Networks for Electric Power Transmission. Journal of Regulatory Economics 4, 211–242.

Huppmann, D., Egerer, J., 2014. National-Strategic Investment in European Power Transmission Capacity. DIW Discussion Papers, No. 1379 .

IAEW, 2016. Erweiterte Verantwortung der Verteilnetzbetreiber. https://www.bdew.de/media/documents/20161214_Untersuchung-RWTH-\Aachen-DSO-Praesentation.pdf, 15.04.2021.

Incentive Regulation Ordinance (ARegV), n.d. [Original German title: Verordnung über die Anreizregulierung der Energieversorgungsnetze (Anreizregulierungsverordnung - ARegV)]. http://www.gesetze-im-internet.de/aregv/, 14.10.2020.

Jägemann, C., Fürsch, M., Hagspiel, S., Nagl, S., 2013. Decarbonizing Europe's power sector by 2050 - Analyzing the implications of alternative decarbonization pathways. Energy Economics 40, 622–636.

Jondrow, J., Lovell, K., Materov, I., Schmidt, P., 1982. On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model. Journal of Econometrics 19, 233–238.

Joskow, P., Tirole, J., 2005. Merchant Transmission Investment. The Journal of Industrial Economics LIII, 233–264.

Just, L., 2021. Unobserved technological heterogeneity among german electricity distribution network operatiors - a latent class analysis. EWI Working Paper, No. 21/05.

Just, L., Wetzel, H., 2020. Distributed Generation and Cost Efficiency of German Electricity Distribution Network Operators. EWI Working Paper, No. 20/09 .

Kumbhakar, S.C., Lien, G., 2017. Yardstick Regulation of Electricity Distribution - Disentangling Short-run and Long-run Inefficiencies. The Energy Journal 38, 17–37.

Kumbhakar, S.C., Ørjan Mydland, Musau, A., Lien, G., 2020. Disentangling Costs of Persistent and Transient Technical Inefficiency and Input Misallocation: The Case of Norwegian Electricity Distribution Firms. The Energy Journal 41, 143–160.

Kumbhakar, S.C., Wang, H.J., Horncastle, A.P., 2015. A Practitioner's Guide to Stochastic Frontier Analysis Using Stata. Cambridge University Press.

Kunz, F., 2013. Improving Congestion Management: How to Facilitate the Integration of Renewable Generation in Germany. The Energy Journal 34, 55–78.

Kurzidem, M.J., 2010. Analysis of Flow-based Market Coupling in Oligopolistic Power Markets. Ph.D. thesis. ETH Zurich.

Leuthold, F., Weigt, H., von Hirschhausen, C., 2008. Efficient pricing for European electricity networks - The theory of nodal pricing applied to feeding-in wind in Germany. Utilities Policy 16, 284–291.

Llorca, M., Orea, L., Pollitt, M.G., 2014. Using the latent class approach to cluster firms in benchmarking: An application to the US electricity transmission industry. Operations Research Perspectives 1, 6–17.

McKinsey & Company, 2009. Pathways to a Low-Carbon Economy - Version 2 of the Global Greenhouse Gas Abatement Cost Curve.

Murty, K.G., 1983. Linear Programming. John Wiley & Sons.

Neuberg, L.G., 1977. Two Issues in the Municipal Ownership of Electric Power Distribution Systems. The Bell Journal of Economics 8, 303–323.

Neuhoff, K., Boyd, R., Grau, T., Barquin, J., Echabarren, F., Bialek, J., Dent, C., von Hirschhausen, C., Hobbs, B.F., Kunz, F., Weigt, H., Nabe, C., Papaefthymiou, G., Weber, C., 2013. Renewable electric energy integration: Quantifying the value of design of markets for international transmission capacity. Energy Economics 40, 760–772.

Oggioni, G., Allevi, Y.S.E., Schaible, S., 2012. A Generalized Nash Equilibrium Model of Market Coupling in the European Power System. Networks & Spatial Economics 12, 503–560.

Oggioni, G., Smeers, Y., 2012. Degress of Coordination in Market Coupling and Counter-Trading. The Energy Journal 33, 39–90.

Oggioni, G., Smeers, Y., 2013. Market failures of Market Coupling and counter-trading in Europe: An illustrative model based discussion. Energy Economics 35, 74–87.

Orea, L., Jamasb, T., 2017. Regulating Heterogeneous Utilities: A New Latent Class Approach with Application to the Norwegian Electricity Distribution Networks. The Energy Journal 38, 101–127.

Orea, L., Kumbhakar, S.C., 2004. Efficiency measurement using a latent class stochastic frontier model. Empirical Economics 29, 169–183.

Pitt, M., Lee, L.F., 1981. The Measuement and Sources of Technical Inefficiency in the Inndonesian Weaving Industry. Journal of Development Economics 9, 43–64.

Platts, 2009. UDI World Electric Power Plants Data Base (WEPP).

Richter, J., 2011. DIMENSION - a dispatch and investment model for European electricity markets. EWI WP 11/3.

Rious, V., Dessante, P., Perez, Y., 2009. Is combination of nodal pricing and average participation tariff the best solution to coordinate the location of power plants with lumpy transmission investment? EUI Working Papers, RSCAS 2009/14 .

Sauma, E.E., Oren, S.S., 2006. Proactive planning and valuation of transmission investment in restructured electricity markets. Journal of Regulatory Economics 30, 261–290.

Schaber, K., Steinke, F., Hamacher, T., 2012. Transmission grid extensions for the integration of variable renewable energies in Europe: Who benefits where? Energy Policy 43.

Schaber, K., Steinke, F., Mühlich, P., Hamacher, T., 2011. Parametric study of variable renewable energy integration in Europe: Advantages and costs of transmission grid extensions. Energy Policy 42, 498–508.

Schneider, T., Battaglini, A., 2013. Efficiency and Public Acceptance of European Grid Expansion Projects: Lessons Learned across Europe. Renewable Energy Law and Policy Review 42, 42–51.

Schweppe, F.C., Caramanis, M.C., Tabors, R.D., Bohn, R.E., 1988. Spot Pricing of Electricity. Norwell, MA: Kluwer.

Shleifer, A., 1985. A theory of yardstick competition. RAND Journal of Economics 16, 319–327.

Steinbach, A., 2013. Barriers and solutions for expansion of electricity grids-the German experience. Energy Policy 63, 224–229.

Steinke, F., Wolfrum, P., Hoffmann, C., 2013. Grid vs. storage in a 100% renewable Europe. Renewable Energy 50, 826–832.

Sumicsid, EE2, 2008. Projekt Gerner IV - Ergebnisdokumentation: Bestimmung der Effizienzwerte Verteilnetzbetreiber Strom - Endfassung -. `https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/Netzentgelte/Strom/Effizienzvergleich_VNB/1Regulierungsperiode/GA_Sumicsid_EffizienzwerteVNBStrom.pdf?__blob=publicationFile&v=1`, 14.10.2020.

Swiss Economics, Sumicsid, 2014. Effizienzvergleich für Verteilnetzbetreiber Strom 2013. `https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/Netzentgelte/Strom/Effizienzvergleich_VNB/2Regulierungsperiode/Gutachten_EVS2_geschw.pdf?__blob=publicationFile&v=2`, 14.10.2020.

Swiss Economics, Sumicsid, IAEW, 2019. Effizienzvergleich Verteilernetzbetreiber Strom der dritten Regulierungsperiode (EVS3). `https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Sachgebiete/Energie/Unternehmen_Institutionen/Netzentgelte/Strom/Effizienzvergleich_VNB/3RegPer/Gutachten_EVS3_geschw.pdf?__blob=publicationFile&v=3`, 14.10.2020.

Triebs, T.P., Saal, D.S., Arocena, P., Kumbhakar, S.C., 2016. Estimating economies of scale and scope with flexible technology. Journal of Productivity Analysis 45, 173–186.

Tsionas, E., Kumbhakar, S., 2014. Firm Heterogeneity, Persistent and Transient Technical Inefficiency: A Generalized True Random-Effects Model. Journal of Applied Econometrics 29, 110–132.

van der Weijde, A.H., Hobbs, B.F., 2011. Locational-based coupling of electricity markets: benefits from coordinating unit commitment and balancing markets. Journal of Regulatory Economics 39, 223–251.